Quantitative Histology for Assessing Organ Health



Ka Ho Tam

Hertford College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Hilary 2022

Acknowledgements

First and foremost, I would like to thank my DPhil supervisors, Professor Jens Rittscher, Professor Rutger Ploeg, Dr Maria Kaisar, and Dr Korsuk Sirinukunwattana. Over the past four years, they have guided me and given me all kinds of support. I thank Jens for hiring me as a student, bringing different experts together, and giving me advice working on this interdisciplinary project. Rutger and Maria helped provide the resources for me to work on the QUOD biobank. Maria has also spent much time helping me understand the different researches within the QUOD framework and giving me research directions. I also thank Korsuk for giving me advice and ideas on the methodological aspects of this project.

I would like to express my gratitude towards the two pathologists, Dr Maria Soares and Dr Jesper Kers, for their help in assessing so many slides and giving their expert opinions on this work. This project wouldn't have gone anywhere without them. Maria has taught me, as a layperson, some basics of how to interpret histology. Jesper has been very resourceful and provided me with a dataset towards the end of this project.

My colleagues provided a supportive and intellectually stimulating environment for me to work in. It was regrettable that the latter half of the project was shadowed by a pandemic. Otherwise, we would have been able to meet face-to-face more often to cross-pollinate project ideas and socialise. I'd also like to thank my ONBI doctoral training centre colleagues and coordinators for all the fun experiences and for putting me on the right track. Many others have also helped me in the course of this DPhil: Edward Sharples, Clare Verrill, Nasullah Khalid Alham, Thomas Nichols, and it is hard to draw an exhaustive list of people to thank. Next, I would also like to acknowledge the Engineering and Physical Sciences Research Council (EPSRC) and Medical Research Council (MRC) for supporting me with funding throughout my DPhil [grant number EP/L016052/1].

Finally, I would like to thank my family: My loving wife, Qiyan, for her encouragement and accompany; my parents, who have been the pillars who nurtured me and respected my choices. I am indebted to you all, without which this DPhil could not have been achieved.

Abstract

This thesis aims to take the first step in developing a quantitative analysis workflow for characterising histological changes in pre-implantation renal biopsies. Solving this is highly beneficial as there is a significant disparity between organ supply and demand worldwide. It can also help to alleviate the workload caused by a shortage of pathologists. This work would allow quantitative measurement of histological features related to chronic and acute injuries in pre-implantation biopsies. These features can then be linked to scores given by pathologists, post-transplant clinical endpoints, and proteomic biomarkers to help researchers better understand the mechanisms of organ injury and repair.

First, we build the main feature extraction workflow based on convolutional neural networks (CNN) to automatically segment different tissue compartments in renal biopsies. We design handcrafted features and qualitatively show that some of these features can capture visual changes related to chronic and acute injuries.

Then, we find that many tissue compartments in the primary dataset have artefacts not present in the training data. We discuss Bayesian Neural Networks (BNNs) to quantify uncertainties in segmentation and propose a way to improve histological estimates.

Next, we propose using tissue compartment-based features to improve classification and interpretability in a multi-instance learning setting. In addition to learned attention, tissue compartments are weighted by segmentation quality to reduce the impact of instances that bear little resemblance to the training data.

Finally, we combine techniques developed in this project to detect acute injuries from biopsies. We propose an algorithmic score consisting of models trained on pathologists' assessments and donors' clinical states. The score is significantly associated with Delayed Graft Function (DGF) in the test data.

These findings suggest that our workflow can facilitate research in kidney transplant where pathologist time is limited. It also shows histology has great potential in predicting post-transplant complications.

Contents

List of Abbreviations **viii**

1 Introduction **1**
1.1 Research Aims . . . 2
1.2 Contributions . . . 4

2 Background **6**
2.1 Kidney Anatomy . . . 7
2.2 Current Practice for the Assessment of Pre-Implantation Renal Biopsies 7
2.2.1 Time-Zero Biopsy . . . 9
2.2.2 Chronic Kidney Diseases . . . 11
2.2.3 Acute Injuries . . . 14
2.2.4 Tissue Staining . . . 17
2.2.5 Histological Changes in Tissue Compartments . . . 17
2.3 The QUOD Programme . . . 22
2.4 Machine Learning and Computer Vision . . . 24
2.4.1 Classical Machine Learning . . . 24
2.4.2 Deep Learning . . . 26
2.4.3 Digital Pathology in Renal Transplant . . . 31

3 Datasets and Pre-Processing **36**
3.1 Datasets . . . 37
3.1.1 QUOD Dataset . . . 37
3.1.2 NMP Dataset . . . 38
3.1.3 Other Slides . . . 39
3.2 QUOD Cohort Statistics . . . 40
3.3 Slide Annotation . . . 43
3.3.1 Slide-Level Assessment by Pathologist . . . 43
3.3.2 Local Assessment . . . 47
3.3.3 Delineation of Tissues . . . 56
3.4 Biopsy Statistics . . . 59
3.5 Pre-Processing . . . 59
3.5.1 Stain-to-Stain Translation . . . 62
3.5.2 Image Registration and Storage . . . 68

4 Quantitative Analysis Workflow **69**
4.1 Overview . . . 70
4.2 Tissue Segmentation . . . 72
4.2.1 UNet Implementation . . . 72
4.2.2 Improvement of Segmentation . . . 74
4.2.3 Human-in-the-loop Labelling and Ensemble Model . . . 84
4.3 Extraction of Features from Tissue Compartments . . . 86
4.3.1 Handcrafted Features . . . 88
4.3.2 Correlation Between Handcrafted Features and Pathologist Assessment . . . 90
4.4 Extensions of Workflow . . . 97
4.4.1 Clustering and Detection of Anomalous Slides . . . 98
4.4.2 Computational Quantification of Fibrosis in Sirius Red Slides 102
4.4.3 Correlation of Visual Features with Tissue Proteomics Signatures . . . 106
4.5 Limitations . . . 108

5 Improving Histological Measurements with Bayesian Neural Networks **112**
5.1 Why is Uncertainty Quantification Necessary? . . . 113
5.2 Uncertainty Estimation with Bayesian Neural Networks . . . 116
5.2.1 Bayes by Backprop (BBB) . . . 116
5.2.2 Monte Carlo Dropout . . . 117
5.2.3 Quantifying Uncertainty . . . 117
5.3 Method . . . 119
5.3.1 Histopathology Measurements . . . 119
5.3.2 Proposed Method for Improving Histological Measurements . 121
5.3.3 Datasets . . . 124
5.3.4 Tissue Segmentation . . . 129
5.4 Experiment Results . . . 133
5.4.1 Relationship between entropy and histological measurements 139
5.4.2 Hard Thresholding . . . 140
5.5 Conclusion . . . 140
5.6 Synthetic Data . . . 142

6 Predicting Visual Changes and Endpoints with Quality-Weighted Tissue-based Features 143
6.1 Introduction . . . 145
6.2 Method . . . 147
6.2.1 Multi-instance Learning with Soft Attention . . . 148
6.2.2 Datasets . . . 149
6.2.3 Tissue Segmentation . . . 152
6.2.4 Extraction of Histological Features . . . 154
6.2.5 Predicting Slide Level Labels . . . 155
6.3 Results and Discussion . . . 157
6.3.1 Segmentation Results . . . 157
6.3.2 Performance of Different Featuresets . . . 159
6.3.3 Attention Visualisation . . . 163
6.3.4 Confidence-Weighted Predictions . . . 168
6.4 Limitations . . . 172
6.5 Conclusion . . . 174

7 Detection of Acute Kidney Injury and Delayed Graft Function from Pre-Implantation Biopsies 176
7.1 Introduction . . . 178
7.2 Method . . . 183
7.2.1 Dataset . . . 183
7.2.2 Histological Features and Models . . . 186
7.3 Results and Discussion . . . 187
7.3.1 Classification Performance . . . 188
7.3.2 Association between AKI, ATI, and DGF . . . 190
7.3.3 Association between Algorithmic Histology Score and DGF . 193
7.3.4 Utility Value of Algorithmic Histological Scores . . . 199
7.3.5 Visualisation of Relevant Tissue Compartments . . . 200
7.3.6 Limitation of ATI Grades . . . 204
7.4 Limitations and Further Work . . . 207
7.5 Conclusion . . . 208

8 Conclusion 210
8.1 Summary of Thesis . . . 210
8.2 Further Work . . . 211

Appendices

A Appendices **215**
A.1 Impact of Covid19 215
A.2 Donor and Recipient Metadata 216
A.3 Handcrafted Features 218
A.4 DGF Risk Predicted by Different Algorithmic Scores 222
A.5 TCGA Slides 224

References **225**

List of Abbreviations

AH . . . . . Arteriolar Hyalinosis

AKI . . . . . Acute Kidney Injury

AKIN . . . . . Acute Kidney Injury Network

ARF . . . . . Acute Renal Failure

ATI . . . . . Acute Tubular Injury

ATN . . . . . Acute Tubular Necrosis

BBB . . . . . Bayes by Backprop

BNN . . . . . Bayesian Neural Network

CE . . . . . Cross-Entropy

CI . . . . . Confidence Interval

CIT . . . . . Cold Ischemic Time

CKD . . . . . Chronic Kidney Disease

CWSSIM . . . Complex Wave Structural Similarity Index Measure

DBD . . . . . Donation after Brain Death

DCD . . . . . Donation after Circulatory Death

DGF . . . . . Delayed Graft Function

DIA . . . . . Data Independent Acquisition (Mass-Spectrometry)

ECD . . . . . Expanded Criteria Donors

eGFR . . . . . estimated Glomerular Filtration Rate. Measured in $mL/min/1.73m^2$

eGFR12 . . . . 12-month eGFR

eGFR3 . . . . 3-month eGFR

ELBO . . . . . Evidence Lower Bound

ESRD . . . . . End Stage Renal Disease

FCN . . . . . Fully Connected Network

FFPE . . . . . Formalin-Fixed Paraffin-Embedded

FN . . . . . False Negative

FP . . . . . False Positive

FSGS . . . . . Focal Segmental Glomerulosclerosis

GAN . . . . . Generative Adversarial Network

GPU . . . . . Graphics Processing Unit

H&E . . . . . Hematoxylin and Eosin

HOG . . . . . Histogram of Oriented Gradient

ICC . . . . . Intraclass Correlation

IOU . . . . . Intersection-over-Union

JSD . . . . . Jensen-Shannon Distance

KDIGO . . . . Kidney Disease: Improving Global Outcomes

KL . . . . . Kullback–Leibler (distance/divergence)

KW . . . . . Kruskal Wallis Test

LRT . . . . . Local Reparameterisation Trick

MC . . . . . Monte Carlo

MFMC . . . . Max-Flow-Min-Cut

MI . . . . . Mutual Information

MIL . . . . . Multi-Instance Learning

mpp . . . . . Microns-Per-Pixel

MWW . . . . . Mann-Witney-Wilcoxon Test

NHS . . . . . National Health Service

NHSBT . . . . NHS Blood and Transplant

NPV . . . . . Negative Predictive Value

PAS . . . . . Periodic acid–Schiff

PCA . . . . . Principal Component Analysis

PDF . . . . . Probability Distribution Function

PNF . . . . . Primary Non-Function

PR . . . . . Precision-Recall curve

PR-AUC . . . Area-under-the-curve for PR curve

QUOD . . . . QUality in Organ Donation

Remuzzi A . . Remuzzi Vascular/Artery and Arteriole Score

Remuzzi G . . Remuzzi Glomerular Sclerosis Score

Remuzzi IF . . Remuzzi Interstitial Fibrosis Score

Remuzzi IFTA Remuzzi Interstitial Fibrosis and Tubular Atrophy Score

Remuzzi TA . Remuzzi Tubular Atrophy Score

RCNN . . . . . Region-based Convolutional Neural Network

RIFLE . . . . . Risk, Injury, Failure, Loss, and End-Stage Renal Failure

ROC . . . . . Receiver Operating Characteristic curve

ROC-AUC . . Area-under-the-curve for ROC curve

RPN . . . . . Region Proposal Network

RR . . . . . Relative Risk

SIFT . . . . . Scale Invariant Feature Transform

SNR . . . . . Signal-to-Noise Ratio

SR . . . . . Sirius Red

SSIM . . . . . Structural Similarity Index Measure

TN . . . . . True Negative

TP . . . . . True Positive

TPU . . . . . Tensor Processing Unit

TSNE . . . . . t-Distributed Stochastic Neighbour Embedding

VAE . . . . . Variational AutoEncoder

WD . . . . . Wasserstein Distance

WIT . . . . . Warm Ischemic Time

WSI . . . . . Whole Slide Image

1 Introduction

Contents

1.1 Research Aims **2**
1.2 Contributions **4**

Transplantation is often the best and perhaps the only option available for many patients suffering from end stage renal disease (ESRD). A successful transplant would allow patients to be taken off dialysis and improve their quality of life. However, despite advances in donation and transplantation practices in recent years, a significant disparity still exists between kidney supply and demand in the UK and worldwide. With kidneys being the most widely transplanted of all organs [1], in the UK, annual kidney-only transplants had increased steadily from 2495 to 3596 from 2008 to 2018, with a median waiting time of 755 days [2].

To resolve donor shortages, kidneys from Donation after Circulatory Death (DCD) donors are increasingly being utilised. Many transplant centres have also started accepting allografts from "Expanded Criteria Donors" (ECDs) – donors over 60 or younger donors with medical complexities. DCD and ECD allografts are often perceived as inferior to those from Donation after Brain Death (DBD) and younger donors, resulting in higher discard rates. The overall discard rate of all deceased donor kidneys remains high (>40%) worldwide due to concerns with

allograft dysfunction and rejection [3]. However, decisions to discard kidneys are often not justified by reasons other than the donors' age and basic clinical history.

Functional risk assessment of donated organs can help determine the organ's suitability for transplantation. It would also be important to efficiently allocate viable organs, minimise discards, and help with recipient post-transplant management. An ideal assessment should allow the prediction of the organ's long-term function and other post-transplant complications in the recipient.

To date, a number of attempts have been made in renal transplantation to develop assessment criteria to support clinical decision-making and predict the onset of post-transplant complications. However, these criteria are often subjective, and none have been proven very successful in the clinical setting. An improved quantitative workflow that utilises deep learning tools has the potential to provide a more accurate prediction of the organ's functional outcome. Such a tool may allow surgeons to utilise healthy organs that otherwise would have been discarded. Prediction of other potential post-transplantation risks can also lead to better care and long-term management of the recipients.

1.1 Research Aims

In this thesis, we aim to develop a quantitative analysis workflow for characterising histological changes in pre-implantation renal biopsies while overcoming several technical challenges. Some of these hurdles are universal to transplant biopsies, while others are specific to our circumstances. The workflow would allow quantitative measurement of a range of histological features related to chronic and acute injuries in a more objective manner. These features can then be linked to scores given by pathologists, post-transplant clinical endpoints, and proteomic biomarkers to eventually help researchers better understand the mechanisms of organ injury and repair. Specifically, we aim to demonstrate the following objectives in this work:

Analysis of Small Biopsies. As biopsy sampling is an invasive procedure, a larger tissue size which may return more promising diagnostic accuracy may come

at a cost of higher bleeding risk. As such, many biopsies obtained in practice tend to be small in size. Typically, we need a minimum of seven glomeruli and one vessel for adequate assessment of kidneys. Many other publications are conducted on biopsies much larger than this requirement. For instance, in Remuzzi et al.'s study [4], an average of 40-50 glomeruli were obtained per sample. The challenge of working with small biopsies is that they may not be statistically representative of the whole organ. As such, we may expect histology assessment and clinical endpoints to correlate to a lesser extent compared to studies that have artificially selected and analysed only larger biopsies. In addition, small biopsies may also lead to the limiting availability of tissue compartments for training and validating machine learning models. In our primary dataset from the QUOD biobank, biopsies tend to be minimal-sized, and most have fewer than seven glomeruli. While the last chapter (Chapter 7) of this thesis focuses on the prediction of acute tubular injuries, we still need to ensure the biopsies contain at least a certain amount of cortex. To our knowledge, this project is the first significant effort of a histological analysis platform to have been specifically developed and tested on biopsies of this size. The imbalanced distribution of histological grades compounds the challenge as kidneys from high-risk donors are usually far less likely to be chosen for transplantation.

Establish a scalable framework for the detection of visual changes. The severe shortage of renal pathologists is a major problem globally in the UK, with the number of pathologists projected to decline even further [5]. This has constrained our ability to get large quantities of labelled data. We aim to develop an analysis work with limited input from pathologists to avoid too much additional strain on the pathologist workforce. For instance, we rely on mostly assessment at the slide level. The methodological analysis will need to account for tissue compartments that have been wrongly labelled or misclassified. Limited local-level labels are available for evaluation as they are more laborious to obtain. We also attempt to find out which lesions can be assessed by algorithms without explicit local-level

labels given by pathologists.

Histological interpretability of algorithmic assessment. Many retrospective studies find that histological scores given by pathologists are not associated with post-transplant clinical endpoints [6]. There could be multiple reasons why associations are weak. An interpretable quantitative analysis can help us delve into these limitations.

Prediction of post-transplant complications. AKI events have been associated with post-transplantation complications such as Delayed Graft Function (DGF). Donors with severe AKI have also been shown to increase morbidity in recipients [7]. Early stratification of AKI severity and causes may improve the prediction of the post-transplant function of donor kidneys, identify those donor grafts at risk of developing DGF, and potentially lead to better post-transplant patient management.

1.2 Contributions

This thesis is the result of a collaborative effort between biomedical, clinical, and engineering research. Its impact can be divided into three domains: i) development of novel methodological approaches (Chapters 5 and 6); ii) advancements in translational biomedical research (Chapters 7 and Chapter 4, Section 4.4); iii) demonstration of clinical endpoint predictions based on histology (Chapter 7).

The work in Chapter 5 has been published in the *MICCAI 2020 Uncertainty for Safe Utilization of Machine Learning in Medical Imaging Workshop* [8]. In this chapter, I proposed a method to improve histological measurements by weighting tissue instances with confidence measures derived from Bayesian Neural Networks (BNNs).

Chapter 6 has been submitted to the journal Medical Image Analysis, and the manuscript is available on medRxiv [9]. A preliminary version of this work was also presented at The European Society for Organ Transplantation (ESOT) Congress 2021 [10]. In this chapter, I proposed a multi-instance learning scheme whereby features are extracted from individual tissue compartments. This resulted

in significantly improved performance compared to features derived from multi-scale rectangular tiles. The proposal also allows a more intuitive visualisation of attention maps which may help with the interpretability of the model. I have also applied the method developed in Chapter 5 and showed that confidence measures from BNNs can help to improve predictions in data-limited multi-instance learning settings.

I am currently preparing to publish Chapter 7 in a clinical journal. In this chapter, neither donor-clinical nor pathologist-given labels are associated with delayed graft function (DGF) risk. I found that an ensemble of models trained on these labels is significantly associated with DGF.

My collaborator, Dr Philip Charles, has also used the results of Chapter 4, Section 4.4.2, in a manuscript. This will be a co-authored publication. In this section, I found a significant correlation between the fibrotic tissue area computed in Sirius Red slides and 12-month graft function.

I have also produced other co-authored contributions within the duration of this DPhil project. I helped with a) the pre-processing and segmentation of tissue structures in *Tumour irradiation combined with vascular-targeted photodynamic therapy enhances antitumour effects in pre-clinical prostate cancer* [11]; b) scanning and pre-processing of slides in *Research Biopsies Kidneys For Transplant Audit To Compare Clinical Safety With Tissue Quality* [12]; c) voxel normalisation of small patches of mitochondria in *Self-Supervised Voxel-Level Representation Rediscovers Subcellular Structures in Volume Electron Microscopy* [13].

2
Background

Contents

2.1 Kidney Anatomy 7
2.2 Current Practice for the Assessment of Pre-Implantation Renal Biopsies 7
2.2.1 Time-Zero Biopsy 9
2.2.2 Chronic Kidney Diseases 11
2.2.3 Acute Injuries 14
2.2.4 Tissue Staining 17
2.2.5 Histological Changes in Tissue Compartments 17
2.3 The QUOD Programme 22
2.4 Machine Learning and Computer Vision 24
2.4.1 Classical Machine Learning 24
2.4.2 Deep Learning 26
2.4.3 Digital Pathology in Renal Transplant 31

In this chapter, we review several pieces of literature as background to the research work conducted in this thesis. We start by describing the anatomy of the kidney (Section 2.1). In Section 2.2, we outline the current practice and criteria used for renal histopathology assessment with a focus on the reading of pre-implantation biopsies. Section 2.3 is an introduction to the QUOD programme. Finally, in Section 2.4, we review the recent progression in computer vision and deep learning with a focus on digital histopathology.

2.1 Kidney Anatomy

Anatomical illustrations detailing the kidney structure are shown in Figure 2.1 and 2.2.

Cortex. This is the outer portion of the kidney of the renal pyramid located exterior to the medulla. The cortex consists of numerous nephrons which extend down the pyramids. It contains the renal corpuscles and tubules except for sections of the loop of Henle, which descends into the medulla. A varying number of blood vessels are also present in the cortex to supply oxygen to tissues.

Nephron. This is a structural-functional unit of the kidney comprising the renal corpuscle and tubules. The corpuscle consists of a network of capillaries known as a glomerulus or a tuft. The glomerulus is surrounded by a cup-shaped structure called a Bowman's capsule. The gap between the glomerulus and the Bowman's capsule is known as the urinary/capsular/Bowman's space. Blood is filtered as it passes through the capillary walls of the glomerulus. The filtrate is collected at the Bowman's capsule and is channelled through the tubule. The tubule has five different parts: the proximal tubule, loop of Henle, distal tubule, connecting tubules, and collecting duct.

2.2 Current Practice for the Assessment of Pre-Implantation Renal Biopsies

Kidney transplantation is a multi-step process involving organ procurement, preservation, surgery, and recipient blood reperfusion. In many transplant centres worldwide, it has become a common practice to obtain zero-time biopsies in an attempt to predict the potential outcome of the organ. Histology has excellent potential in assisting clinical decisions such as organ allocation, kidney discard, and anticipating the possibility of post-transplant complications for more efficient

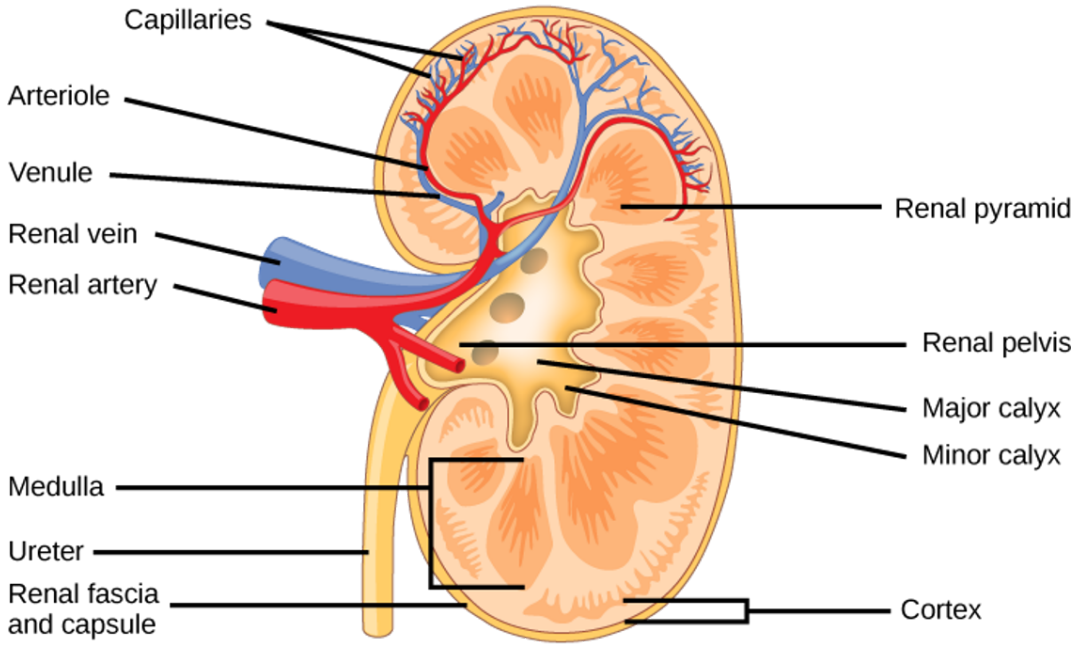


Figure 2.1: Illustration of Kidney Anatomy. *Source: OpenStax [14]; Licensed under CC-BY 4.0*

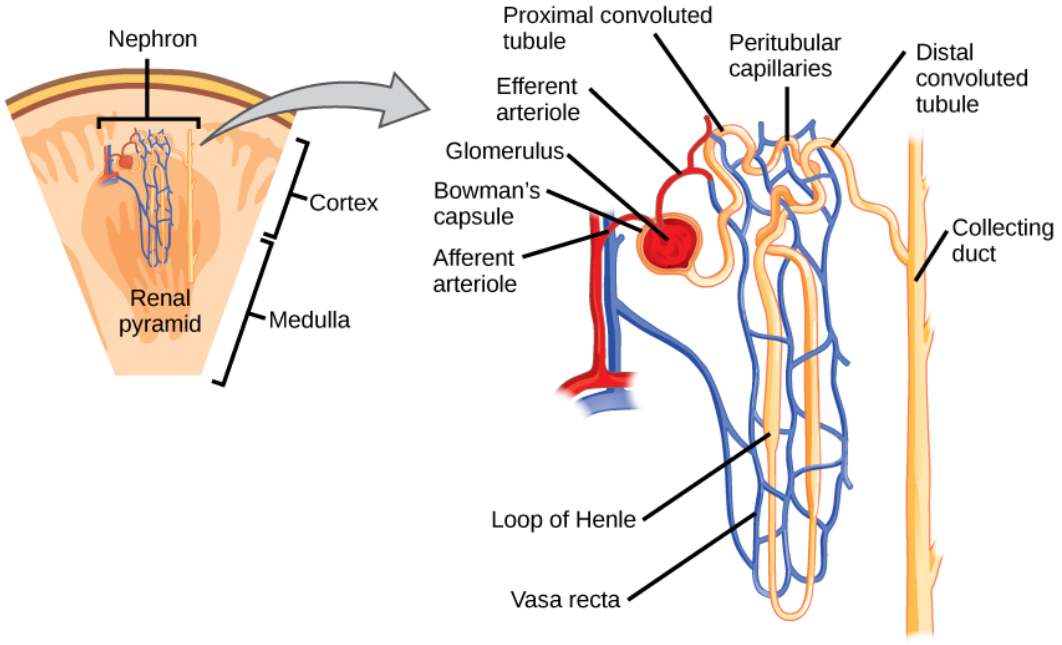


Figure 2.2: Anatomy of the Nephron. *Source: OpenStax [14]; Licensed under CC-BY 4.0*

resource management. Two main types of injuries can often be found in pre-implantation biopsies: chronic kidney disease and acute injuries. Chronic kidney diseases found in histology are usually associated with poorer long term graft function or graft failure, whereas acute injuries are usually associated with longer recovery of function after transplant [15–17]. Donors with severe acute injuries have also been shown to increase morbidity in recipients [7].

2.2.1 Time-Zero Biopsy

While transplantation biopsies could be theoretically obtained at any time between organ procurement (procurement biopsies), on the backtable just before surgery (pre-implantation biopsies), or immediately after surgical completion (post-reperfusion biopsy), most transplant centres in Europe currently only obtain pre-implantation biopsies by routine [18, 19]. Although diagnoses obtained at this point cannot be used for allocation purposes, they would still help decide whether the organ should be discarded at the last minute and make predictions on the extent of after-care that will be required for the recipient.

The exact procedure for sampling and preparing biopsies in different centres may vary. Several works of literature have compared the merits of the different techniques of obtaining biopsies (core needle, skin punch, and wedge biopsies) [19–22]. While wedge biopsies may reliability contain plenty of cortical tissues, they may not reflect vascular intimal thickening reliably [21] and may not contain enough tissues in the subcapsular zone where glomerulosclerosis may be more pronounced. On the other hand, core needle biopsies could provide a more even representation of the cortex at varying depths but may risk bleeding and having inadequate biopsies by sampling the medulla. Protocols of core needle biopsies are also not very well standardised, so the material between different core samples could vary considerably, resulting in more significant variations when graded by different pathologists [3]. Skin punch biopsies are typically wider than core needle biopsies and may be able to provide a balance [22] with a lower risk of bleeding, but further evaluation would be required as to what is best for broader use.

There have also been comparisons between the assessment of frozen samples versus Formalin-Fixed Paraffin-Embedded (FFPE) samples [3, 19]. Screening pre-implantation biopsies are time-sensitive as clinical decisions need to occur before apoptotic processes of kidneys under ischemia significantly affect their long-term function. Therefore, frozen biopsies are often the preferred process for pre-implantation assessment. However, frozen biopsies are prone to suffer from freezing artefacts in epithelial cells. There is typically a higher inter-observer variability intraclass correlations (ICC) for assessing frozen biopsies versus FFPE biopsies of most histological lesions. Regrettably, it is often harder to obtain digitised frozen biopsies for research purposes due to logistic reasons (Section 2.3) - the distribution of liquid nitrogen across different hospitals just to collect frozen samples for digitisation is complex and costly. Thus, this thesis only comprises analysis performed on FFPE biopsies. As most transplant centres use frozen sections in pre-implantation screening, this will limit the applicability of our workflow to preliminary research purposes and possibly the analysis of other kidney diseases. Techniques developed in this thesis will need to be re-trained and tested on frozen sections should such datasets become available.

Despite the current limited clinical utility of pre-implantation biopsies, they are considered histologically similar to procurement biopsies with additional acute histological and molecular changes that reflect the handling and preservation processes. As illustrated in Figure 2.3, pre-implantation biopsies obtained at the time right before surgery may contain information about donor disease and ischemic damage. For organs taken from deceased donors, as is the case of our project, “donor disease” may include chronic diseases from the donor, acute stress to the kidney in the final hours of the donor’s life, and warm ischemic time before procurement (in DCD donors). Procured organs are also often chilled near ice-cold temperatures between procurement and surgery. During this process, kidneys are subjected to anoxia and may manifest mitochondrial swelling along with metabolic reactions of an activated apoptotic pathway [23, 24].

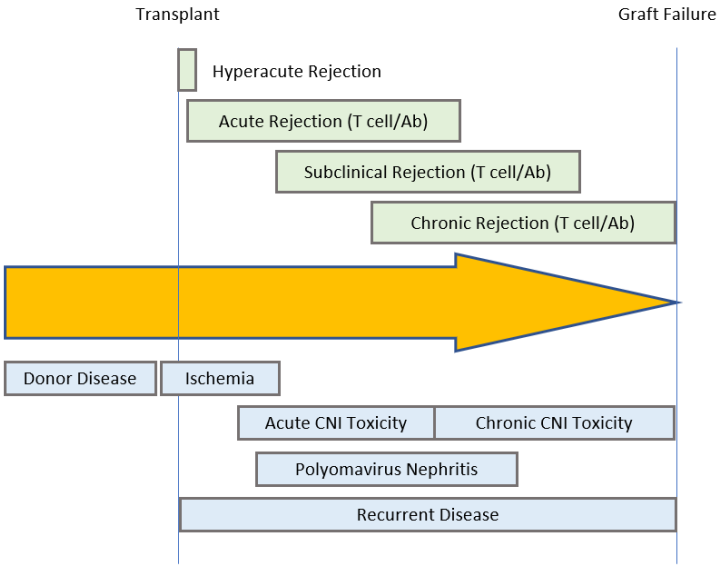


Figure 2.3: Timing of Transplant Complications *Image adapted from Williams et al. [18]*

2.2.2 Chronic Kidney Diseases

At present, several histological features and scoring systems have been proposed. However, the correlation between existing histological scores and clinical endpoints remains poor. In a literature review conducted by Wang et al.[6], it was found that among 47 retrospective studies involving 50 or more donors, approximately half of the authors found no association between chronic histological variables and post-transplant clinical endpoints (either graft failure, delayed graft function, or estimated glomerular filtration rate (eGFR)). While many transplant centres still acquire biopsies before transplant for monitoring purposes, histology is often not used to guide clinical decisions. From the perspective of potential recipients, questions remain open as to whether discarding a kidney with a low (but still substantial) probability of survival in return for a longer waiting time is beneficial for their quality of life.

Several protocols have been proposed to assess donor kidneys in an attempt to predict their function and long-term health after transplantation. Some of the most popular biopsy-based scoring protocols include the Remuzzi score [4], Banff criteria [3], and Maryland Aggregate Pathology Index (MAPI) [25]. Models that combine

histological graft variables and donor clinical parameters are also available [20, 26].

The Remuzzi protocol was initially proposed to help with decision making for the allocation of single vs dual kidneys and is based on a semi-quantitative visual assessment of four different tissue types. The final score is the sum of the four components as detailed in Table 2.1: tubular atrophy (Remuzzi TA), interstitial fibrosis (Remuzzi IF), arterial and arteriolar narrowing (Remzzui A), and global glomerular sclerosis (Remuzzi G). Biopsies are graded as mild if they have a total score ≤ 3 provided all of the components have a score < 3 - these kidneys would be recommended for a single transplant. Kidneys with scores between $4-6$ would be recommended for a double transplant, provided they do not have any components with a grade ≥ 3. Those given a total score ≥ 7 should not be transplanted. To justify the protocol, a study was conducted where only biopsies with ≥ 25 glomeruli were considered for slide evaluation. It was found that the 3-year risk of graft failure for single or dual transplants from donors older than 60 years can be reduced to a level comparable to those of younger donors if biopsy screening was used to allocate organs.

The Banff consensus report is a biennially-updated standard commonly used to assess renal biopsies. Its main aim is to ensure histological reproducibility and to standardise technical issues. The consensus on its own is not a summed point-based system, nor does it give recommendations on how clinical/transplant decisions should be made. However, it does describe criteria for the semi-quantitative grading of various lesions such as tubular atrophy (ct), interstitial fibrosis (ci), glomerulosclerosis (gs), vascular intimal thickening (cv), arteriolar hyalinosis (ah), glomerulopathy (cg), and mesangial matrix increase (mm). One of the main differences between the first four of these grades and those from Remuzzi is that the grade boundaries are defined at different percentage points. Several separate research studies [27, 28] had attempted to combine these lesion scores and found a significant association between their summed scores and eGFR outcome [27, 29] or graft survival [27–29]. Azancot et al. [27] found that correlation was only significant between scores given by one renal pathologist with 12-month eGFR. No association

was found from the assessment by the other 12 on-call pathologists, highlighting the challenges of the evaluation. Snoeijs et al. [28] found a significant association between scores with survival and 12-month eGFR, but the study was only conducted in a small sub-group of older donors after cardiac death.

In addition, the Banff consensus report also provides guidance for processing biopsies, examples of assessment, and schematic visual analogue scales for grading biopsies, which help reduce inter-observer variability. For a renal biopsy to be useful for assessment, there need to be enough cortical tissues present. Since the publication of the 1997 report [30], an adequate biopsy has been defined to be one with ≥ 10 glomeruli and ≥ 2 arteries. Failing this, a "minimal sample" is defined as one with ≥ 7 glomeruli and ≥ 1 artery. It was also recommended that the assessment involves two separate regions of the cortex, either from the same or two separate cores. The guidance for slide preparation is to have seven slides, each with multiple serial sections. Three slides should be stained with hematoxylin and eosin (H&E), three with periodic acid–Schiff (PAS) or silver stains, and one with trichrome stain.

While the aforementioned semi-quantitative approaches allow quick and subjective estimates of compartmental scores, there have also been publications that propose the use of morphometrics [25, 29, 31] to make assessment more objective. This often involves the use of ruler measurements or grid counting on slides. Morphometrics may allow us to extract more information from histopathology as Remuzzi/Banff fails to capture the severity of chronic lesions in individual renal compartments. Comparisons between several semi-quantitative and morphometrics metrics had been studied by Lopes et al. [29]. Although they found a correlation between the severity of compartmental pathological changes and endpoints (3-month eGFR, delayed graft function, and survival), none were more predictive than the "donor chronic damage score" summed from different renal compartments. Thus it was inconclusive whether morphometric features are helpful.

The MAPI is a summed point-based assessment protocol which uses morphometric methods to produce a summed score. The score constitutes the assessment of global glomerulosclerosis, arterial sclerosis, arteriolar sclerosis, cortical scar, and

periglomerular fibrosis. Morphometrics is used to assess arteries and arterioles, the extent of periglomerular fibrosis, and the presence of a scar. However, the metric was developed and tested on the same limited size cohort to maximise the hazard ratio of graft failure; hence, selection bias may be inherent in its development. Derivation of this score tends to require more effort from pathologists as the visual morphological analysis of arterial walls is more complex than the Banff protocol.

Several studies have compared the predictive power of these histological scores and clinical endpoints [20, 32]. De Vusser et al. [20] found that the Remuzzi score gives the highest receiver operating characteristic curve area, whereas the score by Lopes et al. [29] gives the highest hazard ratio for death censored graft failure in a Cox Proportional Hazard model. However, the study had not evaluated MAPI due to the extra complexity of deriving the score. Chen et al. [32] compared Remuzzi [4], Banff [27], MAPI [25], Leuven [20]. They evaluated predictive performance on transplanted kidneys pre-screened according to MAPI. Although they found MAPI had the highest ROC-AUC, they found that none of them was significantly associated with graft survival.

2.2.3 Acute Injuries

Acute Kidney Injuries (AKI) refer to the sudden loss of kidney function in patients. AKI events in donors have been associated with post-transplantation complications such as Delayed Graft Function (DGF) [15–17]. DGF events have been independently linked to higher rates of short and long term graft loss [33]. Donors with severe AKI stage 3, as defined by Acute Kidney Injury Network (AKIN) criteria, have also been shown to increase morbidity in recipients [7].

Since 2004, AKI has been defined under a consensual definition known as RIFLE [34]. Later in 2007, the AKIN criteria [35] were proposed to make the RIFLE criteria more sensitive and reliable. More recently, the Kidney Disease: Improving Global Outcomes (KDIGO) Acute Kidney Injury Work Group proposed further changes [36] to identify more patients with AKI. All these AKI definitions are derived from

Table 2.1: Semi-quantitative scores for evaluating renal biopsies as proposed by Remuzzi et al. For the assessment of vascular lesions, the most severely affected vessel is used for assessment if multiple vessels are available.

Grade	Criteria
Glomerular Global Sclerosis	
0	No global glomerulosclerosis
1+	$< 20\%$ global glomerulosclerosis
2+	$20 - 50\%$ global glomerulosclerosis
3+	$> 50\%$ global glomerulosclerosis
Tubular Atrophy	
0	Absent
1+	Visible in $< 20\%$ of tubules
2+	Visible in $20 - 50\%$ of tubules
3+	Visible in $> 50\%$ of tubules
Interstitial Fibrosis	
0	Absent
1+	$< 20\%$ parenchyma replaced by fibrous connective tissues
2+	$20 - 50\%$ parenchyma replaced by fibrous connective tissue
3+	$> 50\%$ parenchyma replaced by fibrous connective tissue
Arterial and arteriolar narrowing/wall thickening	
0	Absent
1+	Slight increase in thickness, but <lumen diameter
2+	Wall thickness $\gtrsim$ lumen diameter
3+	Wall thickness far exceeds lumen diameter

the serum creatinine and urine output in adult patients. This means the diagnosis of AKI is non-specific and may be caused by different disease and non-disease origins.

Acute Tubular Injury (ATI), formerly also known as Acute Tubular Necrosis (ATN) or Acute Renal Failure (ARF), is the primary pathophysiological change found in patients with Acute Kidney Injury (AKI) [37, 38]. However, the diagnosis of acute lesions has not been broadly standardised; hence inter-observer variability has been reported to be large [39–41]. For these reasons, the assessment of acute lesions is not implemented by many transplant centres as it is possibly regarded as not being worthwhile for pathologists' time.

In the context of transplantation, the most notable consequence of AKI is DGF and primary non-function (PNF). We define DGF as the need for at least one dialysis session the first-week post-transplant with hyperacute rejection, vascular and urinary tract complications, and hyperkalemia excluded [42]. There are several

motivations for predicting DGF. Firstly, DGF may lead to an increased number of recipients needed for post-transplantation dialysis sessions and an increased duration of hospitalisation and costs [43]. Secondly, DGF has been shown to increase morbidity. In a publication by Ojo et al. [33], who studied the outcome of 44,703 DBD transplanted primary renal allografts, DGF was found to be strongly predictive of 5-year graft loss (RR=1.53, P<0.001) independent of other covariates such as cold ischemia time, donor and recipient age.

DGF may originate from the donor, the transplant centre, or the recipient. In a deceased-donor pair study by Louvar et al. [44], it was found that a recipient is at least twice as likely to develop DGF if the recipient in the contralateral kidney also has DGF. The same study found that transplant centre factors contribute to an additional 42% DGF risk after accounting for donor factors. Possible transplant centre factors may include organ procurement or surgical variables such as anastomosis [45, 46].

ATI is often found in histological examination of biopsies with DGF [39, 47]. Based on retrospective histological findings from Rolak et al. [47], 70% of renal biopsies with DGF exhibit signs of acute tubular injury/necrosis (ATI/ATN) and 29% of the biopsies contain findings of acute rejection (with a 13% overlap with both findings). Other DGF causes are T cell-mediated rejection, antibody-mediated rejection, and acute calcineurin toxicity. However, these two studies have only examined post-transplant biopsies, which may contain accumulative stress from the donor, procurement process, and reperfusion injuries in the recipient. While pre and post-transplant ATI may appear morphologically different [48], isolating pre-implantation ATI from post-transplant injuries may be difficult. Therefore these studies do not necessarily conclude how well pre-transplant ATI predicts DGF.

Although this thesis focuses on the effect of AKI on transplantation, it is worth noting that, beyond transplantation, AKI has enormous implications on patients' health and morbidity. The global burden of AKI-related deaths far exceeds breast cancer, heart failure or diabetes [49], but the disease is often under-diagnosed. Repeated episodes of AKI are the precursor of chronic kidney diseases

(CKD) [49–57] through the depletion of immature progenitor cells along nephrons which reduces the capacity of kidneys to replace lost tubular cells. However, studies on how AKI progresses to CKD are limited except for the rare opportunity to study transplant biopsies.

2.2.4 Tissue Staining

A variety of different staining techniques are used for the assessment of renal pathology. The Banff Classification scheme recommends the use of Haematoxylin and Eosin (H&E), Periodic Acid Schiff (PAS), and Silver Methenamine for the assessment of pre-implantation biopsies [58, 59]. H&E is recommended as it is a standard stain – eosin stains basic structures (e.g. cytoplasm) pink, and haematoxylin stains acidic structures (e.g. DNA in the nucleus) purple. On the other hand, PAS stain binds to structures that contain a high proportion of carbohydrates, such as connective tissues, proteoglycans (vessel walls), and basal laminae. Silver Methenamine also stains collagen structures such as the glomerulus, glomerulus basement membrane, and mesangial matrix.

Several other stains can be used for the assessment of renal interstitial fibrosis. Trichrome stain is usually used to assess interstitial collagen content as it binds to tissues' matrix and basement membrane. However, studies [59] have shown that trichrome stain suffers from poor reproducibility either due to the dye's sensitivity to the conditions of the tissue or its low specificity. On the other hand, Sirius Red is a dye that binds to collagen fibres with high specificity and could be examined under polarised light. However, the dye is not widely used due to cost and lack of availability; hence further studies will be needed to justify the dye's effectiveness in assessing renal fibrosis.

2.2.5 Histological Changes in Tissue Compartments

This section shows examples of tissue compartments with different degrees of severity in chronic and acute damage. Examples in this section are extracted from the datasets described in Chapter 3 of this thesis. A mixture of H&E and

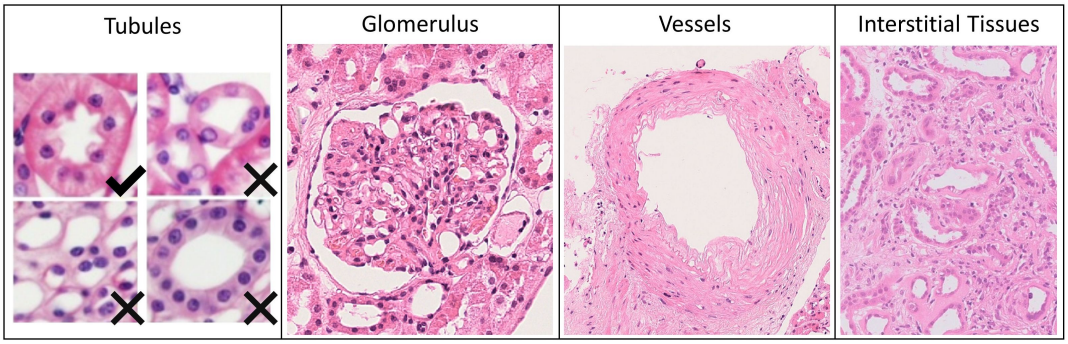
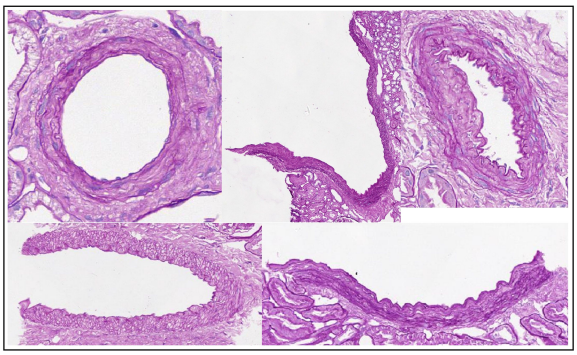


Figure 2.4: Tissues used for Remuzzi Scoring. Note that only proximal tubules (check-marked, top left) are included for assessment. Other parts of the nephron (crossed) are not relevant for scoring.

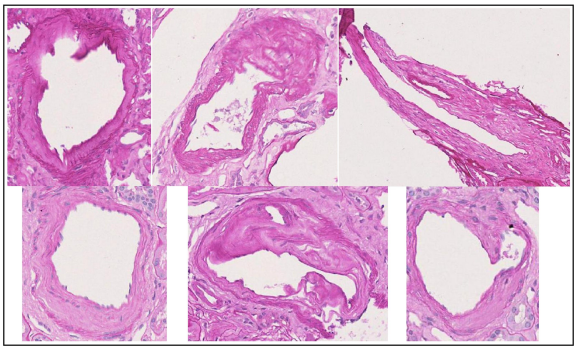
PAS-stained examples are included. For the assessment of chronic lesions, the tissue compartments of interest are proximal tubules (Remuzzi TA), glomeruli (Remuzzi G), vessels (Remuzzi A), and interstitial tissues (Remuzzi IF), as illustrated in Figure 2.4. For assessing acute injuries, predominant changes are in the proximal tubules.

While the assessment of vessels in original Remuzzi A includes both arteries and arterioles, modified versions of the score used by some centres [60] exclude arterioles. For non-pathologists labelling tissue compartments for machine learning, it is also far easier to recognise arteries. For these two reasons, the assessment of vessels was confined to larger muscular arteries with external elastic lamina in this thesis. Examples of arteries from slides with different Remuzzi A grades are shown in Figure 2.5. Note that for Remuzzi A, the pathologist had not marked out exactly which vessel was assessed, so the examples here were chosen based on the artery with the most severe narrowing if multiple arteries were present in a slide. It can be seen that the wall-to-lumen ratio in examples of Remuzzi A grade 2 (Figure 2.5c) is, in general, greater than those given grade 0 (Figure 2.5a). Some slides contain only arteries which have been partially truncated or were elongated in shape but were still scored by pathologists.

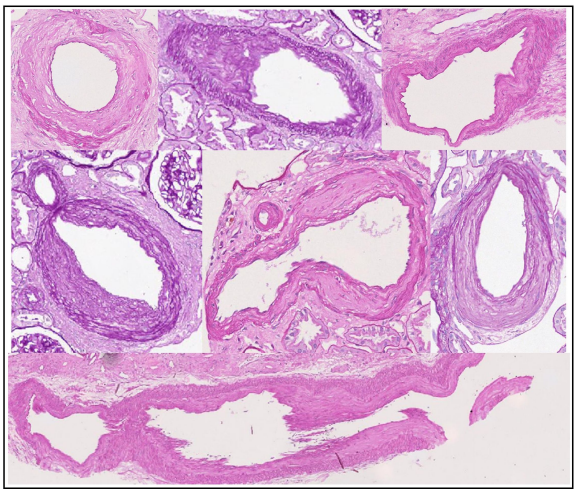
Assessment of Remuzzi G grade is based on the percentage of globally sclerosed glomeruli in each slide. Figure 2.6 shows examples of non-sclerosed (Figure 2.6a) and globally sclerosed (Figure 2.6b) glomeruli. Examples here were handpicked



(a) Remuzzi A 0



(b) Remuzzi A 1

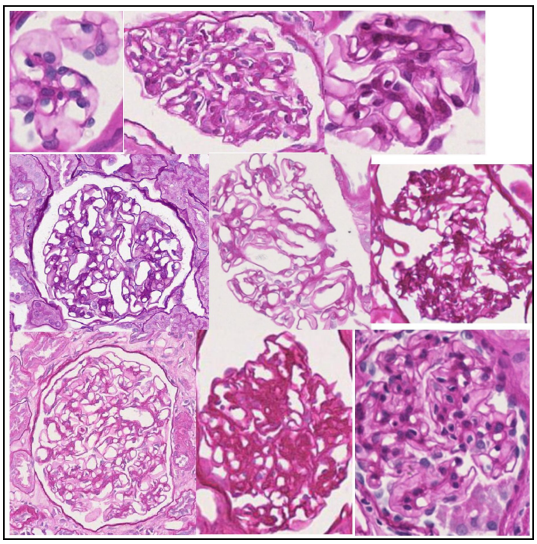


(c) Remuzzi A 2

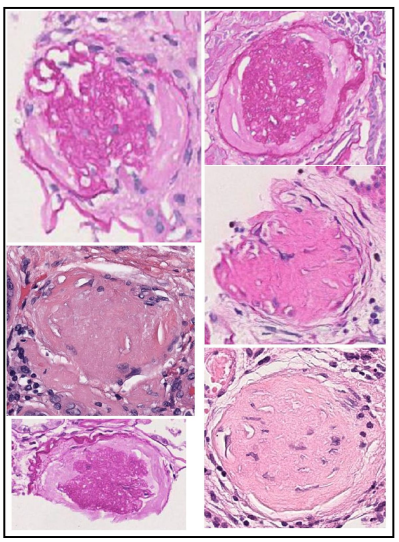
Figure 2.5: Examples of Vessels from slides with different Remuzzi A grades.

and reviewed by a pathologist. Note that urinary casts in tubules (Figure 2.6c) can sometimes be mistaken as sclerosed glomeruli.

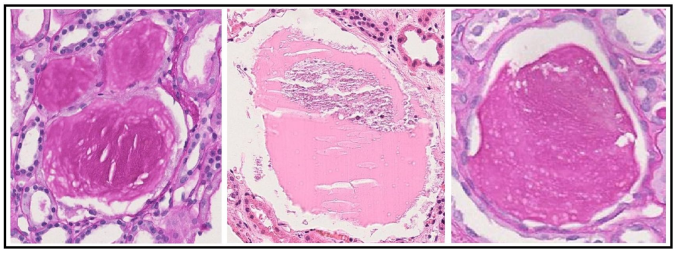
Chronic changes in proximal tubules (Remuzzi TA) are manifested as shrunken tubules with thickened basement membrane [61, 62]. This change is usually accompanied by the presence of fibrotic tissues separating tubules. Many pathologists prefer to score tubular atrophy and interstitial fibrosis as a single grade known as interstitial fibrosis and tubular atrophy (IFTA). Established standards for scoring IFTA lesions are based on the percentage of relevant tissues affected and do not account for the severity of chronic changes in individual compartments. Nonetheless,



(a) Non-Sclerosed Glomeruli



(b) Sclerosed Glomeruli

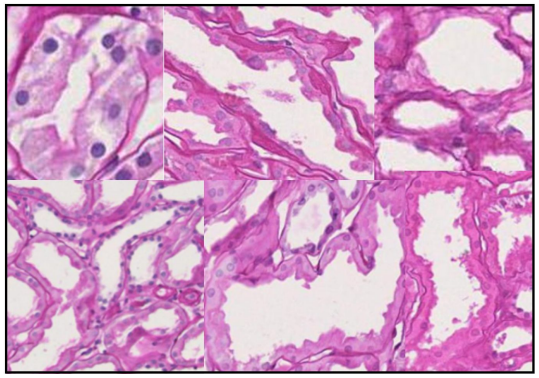


(c) Examples of Casts - these could be confused as sclerosed glomeruli

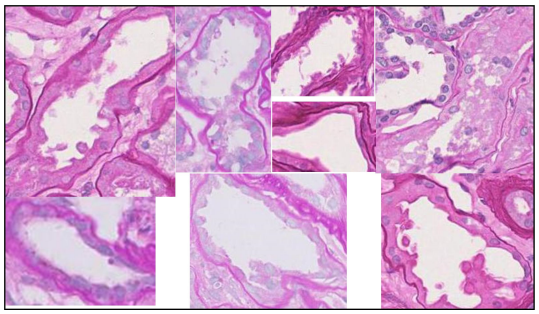
Figure 2.6: Examples of healthy vs sclerosed glomeruli. Note that we have included an example of casts in tubules. These may sometimes confuse algorithms or non-experts.

in our project, we argue that there are no downsides to stratifying individual proximal tubules according to the severity of chronic damage (Chapter 3 Section 3.3.2.2). Here, we show examples of proximal tubules with three different levels of chronic damage in Figure 2.7. It can be seen that tubules with severe chronic damage (Figure 2.7c) exhibit a substantially thicker basement membrane than those with no chronic damage (Figure 2.7a).

ATI's histological changes include loss of brush border, vacuolation of cytoplasm, sloughing of epithelial cells, and flattened epithelium in proximal tubules. In severe cases, visual changes of tubules can progress to cell detachment, cellular casts, and coagulation necrosis (fading of nuclei) [37, 38, 63]. Figure 2.8 shows examples of ATI at different stages, graded according to criteria described in Section 3.3.2.2. Note that these examples only show morphological changes of ATI in pre-transplant biopsies. It has been understood that ATI changes differ in post-transplant biopsies

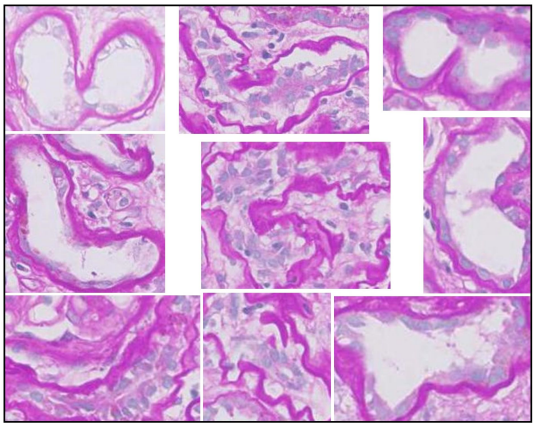


(a) Tubules with no visible chronic changes



(b) Tubules with mild chronic changes

Figure 2.7: Examples of proximal tubules with chronic damage of varying severity.



(c) Tubules with moderate/severe chronic changes

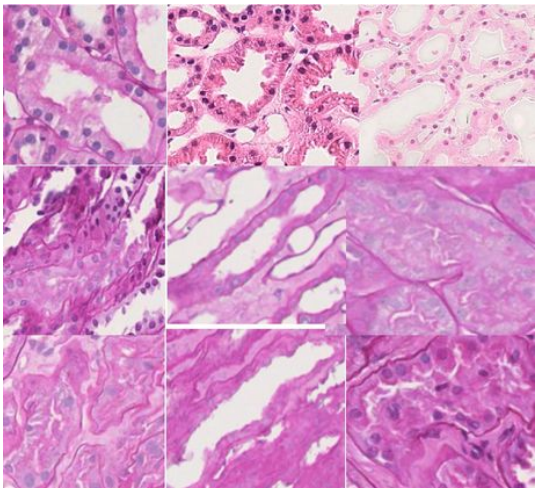
Figure 2.7: (cont.) **Examples of proximal tubules with chronic damage of varying severity**

[48], but this is beyond the scope of this thesis.

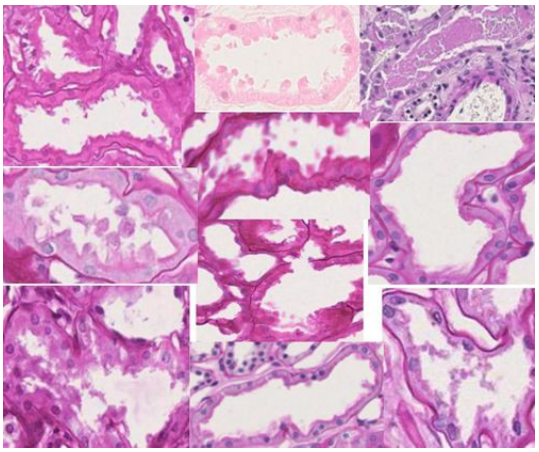
2.3 The QUOD Programme

This project was held in collaboration with the QUality in Organ Donation (QUOD) initiative. The QUOD biobank was established within the research goals of NHSBT for improving donor organ quality, organ utilisation, and transplant outcomes in recipients [64, 65]. The biobank consists of a nationwide collection of clinical samples from deceased donors coordinated by the University of Oxford. Blood, urine, and tissue samples are collected from donors after brain death (DBD) and circulatory death (DCD) from donors over five years old at various time points during donor management and retrieval period. Organs collected include both kidneys, liver, ureter, spleen, and heart. Coordination between the QUOD team and NHSBT Organ Donation and Transplant (ODT) division enables anonymised collection of clinical variables on donors and recipients and long-term follow-up data on the transplant organs.

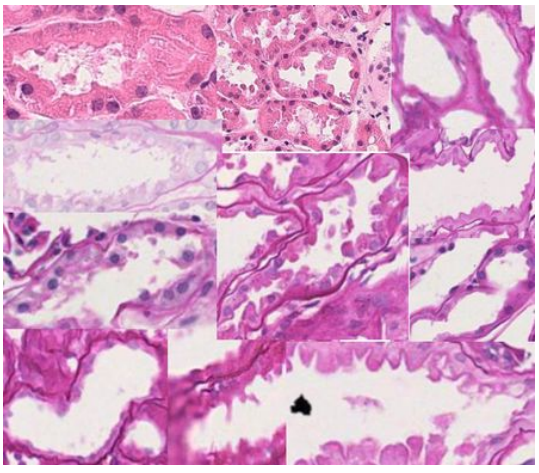
With samples collected from 90% of all deceased donors in the UK, where consent was given for donation and research, the QUOD biobank has grown fast in recent years. As of November 2016, samples have been collected from 1860 deceased



(a) ATI 0

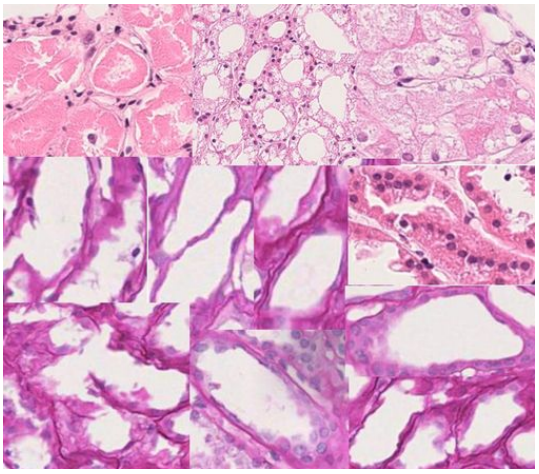


(b) ATI 1



(c) ATI 2

Figure 2.8: Examples of proximal tubules with different ATI grades.



(d) ATI 3

Figure 2.8: (cont.) **Examples of proximal tubules with different ATI grades.**

donors. Towards the end of this DPhil project, in January 2022, samples collected from deceased donors increased to 5,925. A total of 109,561 samples have been collected, of which 15,193 were kidney samples (7,706 biopsies).

Tissue biopsies collected from QUOD are halved at the time of collection and stored in formalin and RNAlater. Formalin samples are processed into formalin-fixed paraffin-embedded blocks, allowing histological processing by researchers at later stages. The RNAlater samples can be used for multi-omics analysis such as proteomics, metabolomics, and transcriptomics. Dividing the biopsies allows for multiple assays and analyses on the QUOD samples, potentially helping researchers link and identify pathways of injury and repair in donor organs.

2.4 Machine Learning and Computer Vision

2.4.1 Classical Machine Learning

Machine learning refers to implementing algorithms that automatically learn features from a data-driven approach instead of entirely based on hard coding. Commonly, machine learning algorithms are used to classify, regress, or cluster data points. Some notable traditional machine-learning algorithms include support vector machines (SVM) and random forests [66–68]. Many of these algorithms

have been proven to be great at solving various problems. For example, boosting trees are commonly used for tabular data and SVMs on lower-dimensional data. On the other hand, typical images used in computer vision consist of at least 10^4 pixels, where each pixel is represented as a separate dimension. The spatial arrangement of the pixels makes them a special form of data. For example, a simple image transformation would result in the reallocation of the value from one dimension to another (dimensional-hopping). While there are still many successful examples of classical machine learning techniques on image analysis, many of these approaches are task-specific and may not generalise well to different datasets without manual adjustments to the models.

Therefore, before the age of deep learning, those working in computer vision had opted for hard-coded methods to reduce the dimensionality of images. For instance, Histogram of Orientated Gradients (HOG) [69], Scale-Invariant Feature Transform (SIFT) [70], and Binary Patterns [71] have been popular handcrafted methods for generating image features. HOG subdivides an image into a grid of arbitrary size and calculates a distribution of pixel gradient orientations in each cell. The histogram of each cell is pooled together to form a single vector, whereas SIFT processes an image by using filters of different sizes to produce key points invariant to scale and orientation. These methods can produce feature vectors that could be subsequently used for object detection and classification.

On the other hand, image segmentation had been classically performed using approaches such as thresholding followed by the watershed method [72], or the clustering of super-pixels [73]. Random forests had also been routinely used to segment images by using handcrafted features as input [68]. However, these methods utilise only low-level saliency features of an image, such as edges and colour. They can be cumbersome to be designed for images with more complex contexts such as foreground objects with highly irregular morphology, variation in scales, overlapping, or poor contrast. As a result, classical algorithms designed for image segmentation tend to be highly task-specific.

The main limitations of classical methods are: 1) pattern-detectors are hand-designed. It is hard to obtain a well-optimised handcrafted pattern detector and they tend to be highly task-specific. 2) A small number of learnable parameters makes it difficult to produce expressive functions to model highly complex features correlations and hierarchical structures. Deep learning overcomes these two limitations by using a large number of data-driven learn-able parameters hierarchically arranged to perform as universal function approximators [74]. As a result, a significant and widening disparity in performance between classical methods and deep learning is also observed in regimes where a large amount of data is available [75].

2.4.2 Deep Learning

Deep learning is a subdomain of machine learning involving algorithms capable of automatically discovering hierarchical layers of feature representations. The models used in deep learning are often known as "neural networks" as they are mainly constructed from layers of simple, biologically inspired, non-linear functions known as "artificial neurons". Each neuron has its associated parameters, known as the weights and biases. These parameters control the relative contribution of the neurons from the previous layer to the next layer. Stacking up layers of these non-linear functions allows neural network models to become universal function approximators [74] that produce the desired output when given different input data. Despite the vast number of parameters in each network, the weights and biases can be automatically "learnt" using the backpropagation algorithm [76] during the training process – backpropagation iteratively updates neurons in each layer by an amount proportional to the difference between the desired output and the actual output, allowing direct learning of pattern detectors without the need to manually engineer individual features.

The basic concepts behind deep learning have been around for a long time. For example, the structure of models in deep learning is biologically inspired by the work of Hubel and Wiesel [66, 77, 78] in the 1960s. LeNet had also been used to recognise hand-written digits on bank notes [79]. However, the adoption of deep

learning was rather limited until recently due to a lack of data, adequate computing infrastructure, and user-friendly software platforms. Their utilisation has only exploded recently when these factors become more accessible.

As neural networks have millions of parameters, they are prone to overfit data. Large amounts of data are required to facilitate their training and validation. This hurdle is gradually being overcome by collaborative efforts to curate large-scale digital databases such as ImageNet [75] for everyday images and TCGA [80] for cancer research. Training large numbers of parameters in neural networks is also enabled by the advances in parallel processing power in modern graphics processing units (GPUs) and tensor processing units (TPUs).

Deep neural networks were incredibly successful at performing various complex computer vision tasks, including not only image classification [75, 81] but also detection [82–85] and segmentation [86, 87]. More recently, there is also increasing interest in using neural networks to develop generative models for synthesising fake images [88, 89] at an unprecedentedly convincing level and representation learning of Bayesian distribution [90, 91]. Examples of the different neural network models for performing these various tasks will be described in later sections.

2.4.2.1 Convolutional Neural Network Architectures

Convolutional Neural Networks (CNNs) are a type of neural network that is particularly useful for extracting features from regularly structured data points with features confined to a locality of the data structure (such as images, videos, and volumes). The weights of the network's local connections are tied, meaning that the output from each layer of a CNN can be effectively regarded as the input convoluted with a kernel filter.

Neocognitron [78] was the original work describing an architecture that closely resembles modern CNNs. The work describes stacked multiple layers of simple and complex cells. While simple cells work as local pattern detectors, complex cells 'pool' together the output of simple cells to generate a shift-invariant response. LeNet-5 is a pioneer CNN reminiscent of the Neocognitron developed by LeCun et

al. in 1998 [79]. It is a network that consists of 7-layers but the pattern detectors are trained using back propagation instead of being hard-wired. It was used for recognising 32x32 pixels' hand-written digits.

However, more recent interest in CNNs only exploded around the year 2012-2013. Research in deep learning was popularised after AlexNet [92] won the ImageNet [75] competition in 2012 for image classification with a top-5 error of 15.3%, achieving state-of-the-art at the time. His network was structurally similar to the LeNet-5, but had more filters per layer and was deeper. Around the same time, in the biomedical community, Ciresan and Schmidhuber also used CNNs to achieve unrivalled performance in the segmentation of neuronal structures in electron microscopy stack [93] and detection of breast cancer cells in mitosis [94]. Since then, various architectural modifications have been made to improve the classification performance of CNNs. The VGGNet [95] led to breakthroughs in classification performance by trading off kernel size with more convolutional layers, and had a record number of parameters for the top-performing networks; ResNet [81] alleviated the vanishing gradient problem in very deep neural networks by introducing the concept of residual learning, where skip-connections between layers are added to prevent outputs of each block from deviating too much from the identity function, allowing a network with 152 layers to be trained; the Inception Network [96] introduced blocks with kernels of different sizes at the same depth to cover different receptive fields, and had 2-dimensional kernels replaced with two 1-dimensional kernels to improve memory efficiency.

Apart from image-level classification, CNN architectures have also been developed to perform more complex computer vision tasks. Semantic segmentation requires localising objects within an image, such that instead of having only one label per image, each pixel is individually classified as either some object class or the background. Apart from Ciresan et al.'s [93] work mentioned earlier, one of the landmark papers on image segmentation is the Fully Convolutional Network (FCN) [86], where up-sampling layers replaced the fully connected layers of a classification network to preserve the spatial relationship between input pixels.

Skip-layer connections have also been added to combine shallow and deep features at the up-sampling stage. However, the segmentation results from the original FCNs were very coarse. Spatial information of the original image is lost as the FCN primarily learns to generate a segmentation map from the lowest resolution feature map. The issue was improved by the U-Net [87], which employed a symmetrical encoder-decoder architecture with multiple feature maps per layer, allowing finer details to be kept at the larger feature maps.

Detection in computer vision involves searching and localising objects or features within an image. In many cases, it can be more challenging than image segmentation - as detection requires the classification of objects and the differentiation of individual instances. At the time of writing, one of the most successful models for object detection is the Faster RCNN [84] in terms of accuracy and YOLO [85] in terms of speed. However, both algorithms claim to be capable of object detection faster than video rate. The Faster RCNN stands out from its predecessors, the RCNN [82] and Fast RCNN [82], by using a single CNN known as a "Region Proposal Network" (RPN) to extract features from objects, outputting a feature map of W×H. A layer is used to classify the presence of objects in a number (k) of anchor boxes with different scales and aspect ratios within the W×H subregions of the feature map. Another layer is used to regress the coordinates of the W×H×k bounding boxes. Regions that score highest in the objectiveness score are pooled, and overlapping bounding boxes are suppressed, giving only bounding boxes with the highest confidence scores. There is also an extension known as the Mask RCNN [83], which adds an FCN branch to the bounding box output of Faster RCNN to produce instance segmentation masks for the object with little overhead computational cost. Contrarily, YOLO works by subdividing the input image into S×S grids. The coordinates of a fixed number of bounding boxes are predicted based on the features from each grid. Each bounding box also predicts an objectiveness score and class probability score conditional to its objective score. As a result of the fixed grid, YOLO performs less well for small or closely spaced objects.

2.4.2.2 Generative Models

In contrast to discriminative CNN models, there were also efforts to develop effective generative models that output synthetic images. Generative models could be enormously useful for several purposes: to serve as a form of data augmentation for the training of discriminative models, translate images between different domains, and help us understand the underlying structure of datasets. Two types of generative approaches related to this thesis are Variational Bayes Methods [90] and Generative Adversarial Networks [97].

The Variational Bayes method hypothesises that data distribution is given by some latent variables z with a probability distribution $p(z)$, which can often be modelled by a multi-variate Gaussian $p(z) \sim N(\mu, \sigma)$ of with means μ and standard deviation σ. Instead of learning deterministic features, a Variational Autoencoder (VAE) [90] learns the representation of this data distribution with a neural network that consists of an encoder and a probabilistic decoder. The encoder deterministically maps input x into the latent distribution parameters μ, σ. A random sample z is then taken from this distribution to produce an output $\tilde{x}$ through the decoder. The model is trained by minimising a loss consisting of two terms. The first term is the Kullback–Leibler divergence between the latent vector z over the data and prior $p(z)$. The second part is the VAE's reconstruction error - usually this would be the $L2$ distance between the input and the output.

Generative Adversarial Networks consist of a generator and a discriminator. The generator aims to produce an image given some random noise as input, whereas the discriminator's purpose is to classify whether a given image is real or fake. During training, these two components work against each other – successful classification from the discriminator would produce a greater loss for the generator, and generated images that successfully fool the discriminator would contribute to a loss to the discriminator. This revolutionary method has successfully generated realistic-looking images that can sometimes even fool humans.

Adversarial methods have also been applied to inter-domain image translation. For instance, Pix2Pix [89] uses a conditional adversarial network to translate images

from one domain to another. The generator of this model is based on a U-Net, whereas the discriminator serves to compare differences between the output fake image to the target real image in the second domain. In addition, the discriminator also compares the original input with the target real image to maintain the ability to distinguish images between the two domains. However, this method requires images to be paired across the domains, making it a form of supervised learning that is expensive to train. Contrarily, CycleGAN [98] could be trained using unpaired images. The CycleGAN consist of two generators and two discriminators. Each generator translates images to the opposite domain, whereas the discriminators attempt to classify the fake image from the real images for each domain. To prevent the generator from collapsing and producing the same output regardless of the input, the authors introduced a cycle consistency loss term – enforcing images translated from domain A to B and back to domain A again have to look like the original image. This loss was added to the adversarial loss terms during training. Because CycleGAN could be trained using unpaired images from two domains, it is regarded as an unsupervised method and is much cheaper to train than Pix2Pix.

2.4.3 Digital Pathology in Renal Transplant

A digitalised histological biopsy typically contains a plethora of details – pathological information can be embedded on multiple scales - from the sub-nuclei level of less than a few microns to the global context of the whole biopsy, approximately 1cm in size. Hence digitalised slides are usually huge – at the order of 100,000 pixels on each side. The data is stored in an image pyramid format for a modern computer to handle such images smoothly. Manual analysis of the slide is a laborious task that is prone to subjective error.

With the rise of deep learning tools and high-resolution slide scanners, digital pathology has become an increasingly helpful tool for the sub-typing of diseases to aid clinical decision-making processes. Implementation of deep learning could enable a more accurate objective assessment of histological samples and the discovery and

quantification of visual biomarkers that could be used to improve our biological understanding.

Aside from renal transplantation, digital pathology has been commonly employed in the analysis of cancer subtypes [99–104]. Some of the techniques may eventually become relevant in renal transplants. For example, Xu et al. [103] investigated the use of k-means, deep, and manual features in a multi-instance learning framework for classifying colon cancer types and found deep features to produce the highest performance; Cruz-Roa et al. [104] trained a CNN on labelled image patches. Manual features are extracted from different scales from nuclei to overall larger tissue structures. This outperformed other state-of-the-art methods at the time; Janowczyk et al. [99] have shown an early approach to deep learning in digital pathology with studies involving segmentation of nuclei, epithelium and glands for the detection of mitosis and grading of patches using an AlexNet architecture; Mobadersany et al. [100] has combined VGGNet's features with genomic data in the final fully-connected layers to predict recipients' survival; Shah et al. [101] constructed a pipeline for the localisation of tumours and correlated both deep and shallow visual features with RNA expressions in breast cancer; Sirinukunwattana et al. [102] trained two CNNs for the respective detection and classification of cells in colorectal cancer biopsies. The classified cells were used to phenotype patches by their cell-cell connection frequencies to assist prediction of the tumour's metastasis-free survival time. Plenty of other early-day publications had also been detailed in surveys [99, 105], but the majority of these focused solely on segmentation, detection, or simple fully-supervised classification.

Comparatively, the number of publications on the use of computer vision and deep learning in the area of organ transplantation has been somewhat limited at the early stages of this project [106–112].

Ledbetter et al. [110] demonstrated using a combination of the output from CNN on kidney tissue whole slide images (WSI) and clinical data to predict the glomerulus filtration rate. The authors manually segmented the renal cortex from the medulla in 80 donors. A VGGNet based CNN model was modified to regress

the eGFR of kidneys using both image patches and initial GFR as input. This "black box" approach attempts to make predictions without utilising visual features already known to pathologists.

Several works of literature have attempted the automated segmentation of glomeruli in renal biopsy slides:

Gadermayr et al. [106] investigated several approaches for segmenting sparse objects (glomeruli). The author tested several networks centred around using one sub-network to detect glomeruli at a low resolution and another sub-network to process the finer details at a higher resolution.

Gallego et al. [107] compared the performance of pre-trained AlexNet and GoogLeNet models for the binary classification of whether an image patch contains any glomerulus. They also attempted to qualitatively compare the difference in activation maps in the hidden layers for images where glomeruli are present/absent.

Simon et al. [108] used local binary patterns from the tissue around the glomeruli to construct a simple feature vector and used these to train an SVM model to classify glomeruli. They have also demonstrated that using these binary pattern features in tandem with the classification results of GoogleLeNet could reduce false positive detection of glomeruli in whole slide images (WSI).

Marsh et al. [109] compared two CNN architectures to assess glomerulosclerosis on WSIs – the first model classifies small patches of the WSI for glomerulosclerosis to yield a probability map in the parent WSI. The second model is an FCN which inputs larger areas of the slide but is down-sampled by a factor of 32. This study showed much better intersection over union (IOU) in the second model, suggesting that larger local areas of the WSI may be desirable for determining glomerulosclerosis pathology.

Kawazoe et al. [111] demonstrated the use of a Faster RCNN on the segmentation of glomeruli in WSI. The network was based on an Inception-ResNet backbone and was trained on one of the largest renal biopsy datasets known (800 WSIs and 33,000 glomeruli). Their network managed to detect glomeruli in each WSI with an average of 64 seconds, significantly faster than the prior art, which took up to two minutes.

There had also been attempts on virtual histology staining using a CycleGAN [112–114] to transfer stain characteristics between two sets of unpaired digital slides. Shaban [113] demonstrated the use of CycleGAN to adapt slides obtained from different scanners. A 12% AUC classification improvement was noted after digitalised slides from a different lab were transferred to match the style of the original training data of a CNN classifier. Gadermayr et al. [112, 114, 115] demonstrated style transfer between different histology stains. One of their studies [114] compared the merits of transferring training data to the test images' domain versus transferring the test images to the domain of the training data. The results suggested that the former method may give better performance. However, they noted that the difference in interpretability and characteristics of the two stains may have influenced the study's conclusions. Two other publications by the group [112, 115] demonstrated the idea of translating images directly to segmentation labels. Under the proposed scheme, the training images in the target domain are simply segmentation mask look-alikes, which could be simply a mask with circles and ellipses of similar sizes/densities as the glomeruli. This scheme may allow the segmentation of the tissue without any manual annotations.

Studies published before the early stages of this DPhil project mainly focused on analysing a single tissue compartment. The segmentation of glomeruli, in particular, has attracted more attention than other compartments. It may be necessary to assess multiple tissues, such as tubules and vessels, for a more holistic assessment of the organ. These tissues may contain features of different underlying diseases that may not be present in the glomerulus. Towards the later stages of this DPhil, several pieces of literature have been published focusing on the segmentation and analysis of multiple tissue types [116–118] or the detection of multiple kidney diseases [119]. These approaches, however, require explicit labelling of a large number of tissue compartments or rectangular tiles as either normal or diseased. To adapt these platforms for a different diagnosis would require additional time from pathologists to go through the entire dataset, adding further to the project's investment. Regrettably, the expert-time cost of fully-supervised learning

is prohibitive and has been a major reason why the number of publications involving deep learning in renal histopathology is still very limited. This work is a unique attempt to develop a renal biopsy analysis platform covering multiple types of tissue compartments where the main cohorts (QUOD and NMP - See Section 3.1) used for development has not been subjected to histological characteristics pre-screening, thus exposing a more realistic challenge of the predictive tasks. This work has made a progressive step to improve slide-level histology predictions by using a Bayesian Neural Network confidence score. The final proposed workflow is a scalable approach that circumvents the bottleneck caused by the lack of renal pathologists. This is also a pioneering work to have successfully found a statistical association of delayed graft function from digital histopathology.

3

Datasets and Pre-Processing

Contents

3.1 Datasets 37
3.1.1 QUOD Dataset 37
3.1.2 NMP Dataset 38
3.1.3 Other Slides 39
3.2 QUOD Cohort Statistics 40
3.3 Slide Annotation 43
3.3.1 Slide-Level Assessment by Pathologist 43
3.3.2 Local Assessment 47
3.3.3 Delineation of Tissues 56
3.4 Biopsy Statistics 59
3.5 Pre-Processing 59
3.5.1 Stain-to-Stain Translation 62
3.5.2 Image Registration and Storage 68

In this chapter, we describe the main datasets used throughout this thesis. A summary of the datasets is detailed in Section 3.1. The primary cohort used in this thesis is the QUOD dataset - we describe the characteristics of this cohort and slides in Section 3.2. Slides in these datasets have also been annotated - both at the local level and the slide level. For local annotations, we design a platform specifically in this thesis for pathologists to annotate various tissue compartments. The annotations generated from these datasets are specified in Section 3.3. Section 3.4 shows some

statistics of the biopsies. Finally, Section 3.5 describes some pre-processing work that is performed on the datasets to homogenise the colour of the slides.

3.1 Datasets

Two main datasets were used to develop our analysis workflow in this thesis: (i) **QUOD**, (ii) **NMP**. Both datasets consist of pre-implantation biopsies. All biopsies used in this thesis are Formalin-Fixed Paraffin-Embedded (FFPE). A summary of the size of the dataset is listed in Table 3.2. Note that multiple biopsies have been taken from some donors, but cross-validation experiments in this thesis had been done such that train/validation/test splits were based on donors, ensuring biopsies from the same donor were all in the same split.

3.1.1 QUOD Dataset

The QUOD Dataset consists of paraffin-embedded 22mm pre-implantation half-core needle biopsies from the *Quality in Organ Donor* Biobank[1], a national multi-centre UK-wide bioresource of deceased donor clinical samples procured during donor management and organ procurement. These biopsies were from a larger cohort of cases where both kidneys from the donor have been transplanted and yielded similar outcomes (12-month eGFR) in both recipients. This cohort selection criteria allow us to reduce the importance of recipient-related factors amongst other variables influencing transplant outcomes. We have received biopsies from n=354 donors from this dataset. Slides from kidneys received by n=373 recipients (348 donors) were stained in PAS, and n=174 recipients (174 donors) were stained in SR. 180 donors had biopsy sections that were only stained with PAS but not SR, and 6 donors were stained vice versa. 90 PAS slides (from 84 donors / 84 recipients) received a full Remuzzi score. Only 2 of these 90 slides (2 donors, 2 recipients) were available in SR so we did not draw any conclusive comparisons between the two stains. According to the Banff criteria, most of these slides do not contain enough

[1]Collection of QUOD samples and the research ethics approval was provided by QUOD (NW/18/0187).

tissues for full assessment [30]. Slides with no arteries or fewer than seven glomeruli are only partially assessed. A large proportion of tissues also suffer from artefacts such as forceps compression, folding, or duplication of serial sections.

3.1.2 NMP Dataset

This dataset originated from an organ normothermic perfusion experiment [120, 121] from 35 donors [2]. The slides were 18 gauge 33mm core needle biopsies obtained from deceased donor kidneys that were discarded as deemed unsuitable as transplants for mechanical reasons. These kidneys were placed into a normothermic machine perfusion (NMP) system. Biopsies were obtained at different time points during NMP; hence most samples suffered substantial ischemic damage. Most of the slides (n=169) were stained with H&E but had been computationally converted to PAS stain using a CycleGAN [98] (Section 3.5.1). The converted images were only used to train the segmentation part of our workflow and may lead to some drop in performance compared to training on slides with the same stains. The conversion was needed as we did not have other PAS slides available at the development stages of the workflow. Features were not extracted from the tissue compartments of the fake NMP slides, so any changes to the content within tissue compartments will be irrelevant to methodological development (Chapters 4-6). A subset of these converted images had been presented to a pathologist MFS[3] and was qualitatively determined to be most indiscernible from real PAS stains at low magnifications. Examples of how these low magnification images were presented are shown in Figures 3.12-3.13 - these are web-page screenshots where the whole page was expanded to regular computer screen size.

We also have a smaller number of slides stained directly with PAS and SR (n=26 from 4 donors for each stain). Although SR is not used by routine assessment, the stain could facilitate a more convenient computational assessment of fibrotic tissue. This was done as a side-project of this thesis detailed in Section 4.4.2.

[2]Research ethics approved by the National Ethics Review Committee of the United Kingdom (12/EE/0273)

[3]Maria F. Soares, specialty doctor in renal pathology

Table 3.1: NMP Donor Characteristics.

Param	DBD (n=19)	DCD (n=17)
Donor age in years (median (min-max))	70 (47-78)	59 (38-74)
Donor BMI (kg/m^2)	27.4±4.7	26.9±6.5
Serum creatinine at admission (µmol/L)	79.5±31.9	87.2±31.6
Serum creatinine at retrieval (µmol/L)	79.3±42.4	69±35
Donor eGFR ($mL/min/1.73m^2$)	82.3±32.9	103.6±60.8
Urine output last hour (mL)	130.5±127.1	92.1±56
Urine output last 24 hours (mL)	2742±1501	2666±2006
CIT (hours)	29.4±28.1	47.4±33.6
WIT (minutes)	n.a.	16.4±7.2

3.1.3 Other Slides

As the original datasets (QUOD and NMP) lack cases with moderate CKD, 12 cases of native biopsies from a National Institute for Health Research (NIHR) study [4] were chosen to include donors with chronic changes. Cases were given a qualitative description by the pathologist: 3 cases had no chronic changes; 4 cases had 'mild' chronic tubular changes; 5 cases had 'moderate' chronic changes. Slides with inflammation, haemorrhage or potential drug effects are not present in these slides. All native biopsy slides were scanned using a Philips IntelliSite scanner at x40 (0.25mpp). Kidney cortical tissues from the Cancer Genome Atlas (TCGA) [80] are also available in the public domain. The TCGA slides were stained in H&E, and we algorithmically converted them to PAS using a CycleGAN as described in Section 3.5.1. TCGA slides were handpicked to include slides that contain plenty of glomeruli so that they can serve as training examples for segmentation. Delineation of tissue compartments was performed only on cortical regions without tumours. A list of the chosen files is given in Table A.4 in Appendices Section A.5.

The native biopsies and TCGA slides were solely used to strengthen the segmentation part of our workflow (Section 4) to improve its performance on test data that were previously not used for model development purposes.

[4]Ethics approved by the Research Ethics Committee of Oxford University Hospital NHS Foundation Trust Research and Governance (19/WM/0215)

Two more sets of slides (AMC and QUOD-AKI) have been obtained towards the end of this DPhil project for validation purposes and were only used in Chapter 7.

3.1.3.1 AMC Dataset

The first set of validation-only slides was from the University of Amsterdam (AMC-PAS) [5]. It contains pre-implantation biopsies from 151 donors and 151 recipients. Slides were stained in PAS and scanned using a Philips scanner at 0.25mpp (40x). The set contains 91 living donors and 60 cadaver donors. These slides are much larger than the QUOD slides, with a median number of 19 glomeruli (standard deviation 9.72).

3.1.3.2 QUOD-AKI Dataset

The QUOD-AKI dataset consists of 39 slides (39 donors / 39 recipients). These were also from the QUOD biobank but were selected based on AKI characteristics for molecular studies [122]. They were selected to have a quarter of cases with and without AKI (as defined by KDIGO [36]) and with good/bad outcomes (as defined by 12-month eGFR (eGFR12) with a cutoff at 45 $mL/min/1.73m^2$). This selection criterion differs from the main QUOD dataset mentioned in Section 3.1.1 which was based entirely on eGFR12. Slides were stained using H&E and scanned using Aperio ScanScope scanner at 0.253mpp - a different scanner from the QUOD-PAS slides. These slides contain a median number of 5 glomeruli (standard deviation 4.88). Note: for clarity, in Chapters 3-6, when we refer to "QUOD Dataset/Slides" we are excluding these QUOD-AKI slides unless specified otherwise.

3.2 QUOD Cohort Statistics

The QUOD dataset is accompanied by a list of metadata describing both the donor and the recipient. Some available metadata examples are shown in Table A.1 in the Appendices Section A.2 of the thesis. Based on all available metadata, we find that donor age correlates more strongly with eGFR12. Figures 3.1 and 3.2

[5] All experiments were done in accordance with the Declaration of Helsinki and were approved by the local ethics and privacy committees at Amsterdam UMC location AMC (19.260).

Table 3.2: Overview of Dataset. This table gives an overview of the size of the datasets used in this thesis. Slides were digitised by at least four different scanners. The total area selected for tissue delineation from each dataset is shown in the right-most column. *Note1*: we have tiles that contain only annotations of cell nuclei. These have not been included in the table. *Note2*: Because glomeruli and vessels are much rarer than tubules, only these two tissue classes are delineated in some tiles. The first number includes all tiles delineated, whereas the number in brackets shows tiles where only glomeruli and vessels are delineated.

Number of Slides	Donors	Kidneys	Biopsies	Slides	Delineated[Note2]	
					Tiles	Area (mm^2)
QUOD → PAS	348	373	394	414	85 (47)	253 (236)
QUOD → SR	173	195	195	215	20	20.1
NMP → HE	35	35	169	169	240 (96)	400 (294)
NMP → PAS	4	4	23	26	12	5.89
NMP → SR	4	4	22	26	20	9.69
Native → PAS	12	12	12	12	26 (14)	235 (222)
TCGA → HE	11	11	11	11	22 (8)	282 (215)
AMC → PAS	151	151	151	151	0	0
QUOD-AKI→HE	39	39	39	39	0	0

show scatter plots of how age affects transplant outcome as measured by eGFR12. Under current allocation practices, kidneys from older donors tend to be matched to older recipients (Figure 3.3). Thus, the effect size of donor age on eGFR12 may be weaker than it appears if stratified by recipient age.

Table 3.3: QUOD Donor Characteristics. n differs for each clinical variable as some donors have missing entries.

Param	DBD		DCD	
	n	**mean±std**	**n**	**mean±std**
Donor age in years	180	47.5±14.4	164	49.7±14.1
Donor BMI (kg/m^2)	180	27.0±5.66	163	27.4±5.27
Serum creatinine at admission (µmol/L)	174	81.2±33.7	159	74.9±27.4
Serum creatinine at retrieval (µmol/L)	172	94.9±74.0	159	70.7±31.8
Estimated GFR ($mL/min/1.73m^2$)	166	101±55.8	149	118±52.6
Urine output last hour (mL)	166	99.9±91.9	152	123±123
Urine output last 24 hours (mL)	113	3260±1870	107	2730±1620
CIT (hours)	180	14.8±4.63	161	13.5±4.58

Apart from eGFR, another metric that we are interested in is whether the recipient has suffered from DGF [42]. DGF is not significantly associated with long

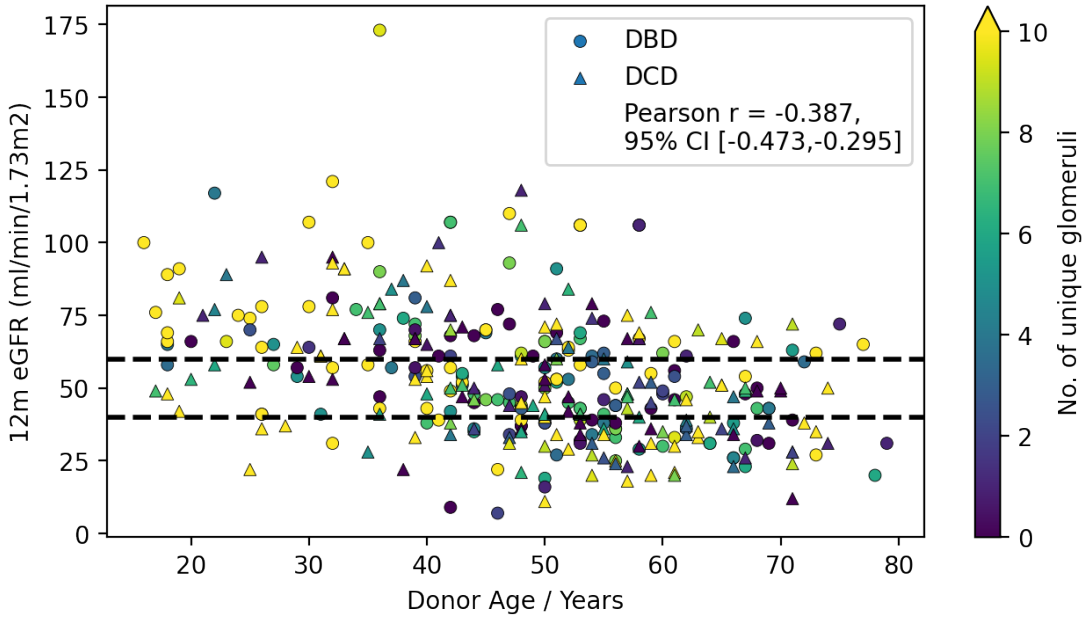


Figure 3.1: Plot showing the correlation between donor age and 12-month eGFR. Data points are coloured by the number of glomeruli in the corresponding slides and shaped according to donor type (DBD/DCD). There is no significant difference between donor type with outcome. Slides from 309 donors contain valid eGFR12 data; Slides from 153 donors (166 recipients) contain valid eGFR12 data with ≥ 7 glomeruli; Slides from 97 donors (104 recipients) contain valid eGFR12 data with ≥ 10 glomeruli. The cohort was selected to have an approximately equal distribution of cases with good/intermediate/sub-optimal eGFR12 outcome. The thresholds for good/sub-optimal outcome is >60 and <40 $ml/min/1.73m^2$ respectively - this is illustrated as horizontal dash lines on the graph.

term outcomes (Figure 3.4: $p = 0.111$ for eGFR3 and $p = 0.0917$ for eGFR12) based on this cohort. Donor age is a significant risk factor for DGF (Figure 3.5, $p = 0.0109$). If donor age is used as the sole predictor for classifying DGF, the ROC-AUC would be 0.584. Unlike prior literature [123], we find no significant association between cold ischemic time and DGF in our cohort ($p = 0.919$, ROC-AUC= 0.496 for 244/77 recipients without/with DGF). There is also no significant association between donor type (DCD/DBD) with any of the available clinical endpoints (DGF, eGFR3, eGFR12).

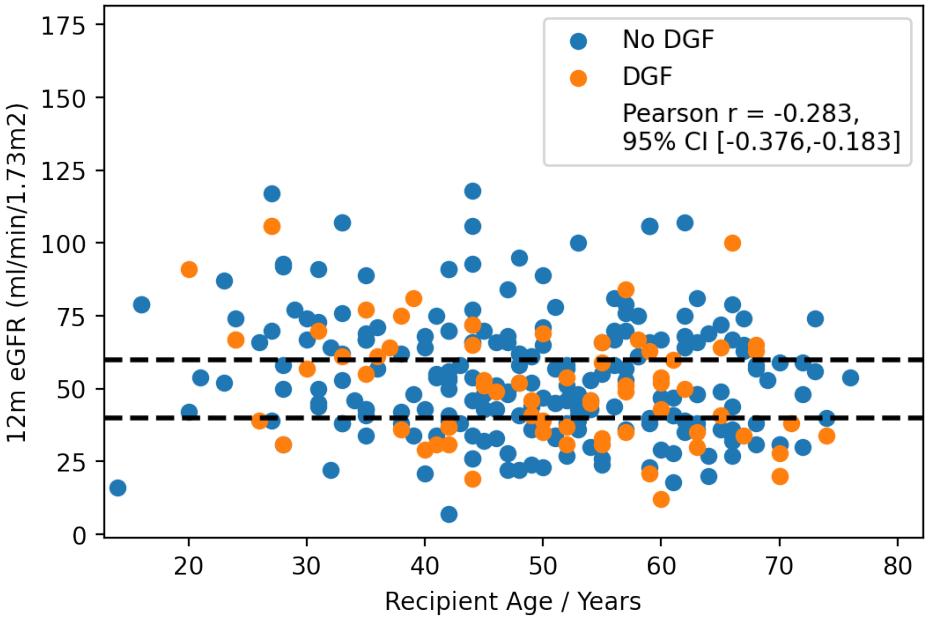


Figure 3.2: Plot showing the correlation between recipient age and 12-month eGFR. Data points are coloured by whether the recipient has suffered from DGF - age is a risk factor for DGF. The horizontal dash lines (40 and 60 $mL/min/1.73m^2$) show eGFR12 cutoff values for selecting the cohort. An approximately equal number of cases from each cutoff were selected.

3.3 Slide Annotation

A number of annotations pertaining to the datasets have been curated throughout this project. Section 3.3.1 shows the statistics of slide-level assessment on the QUOD and NMP datasets as performed by our collaborating pathologist. In addition to these slide-level labels, we have also obtained some local-level labels. These are described in Section 3.3.2. To segment tissue structures, a number of tissue compartments have been delineated by hand using a protocol described in Section 3.3.3.

3.3.1 Slide-Level Assessment by Pathologist

3.3.1.1 Assessment of PAS and H&E Slides

Slide-level assessments were performed by an experienced renal pathologist MFS [6] on the PAS and H&E slides in our dataset. Assessment results were recorded using

[6]Maria F. Soares - specialty doctor in renal pathology

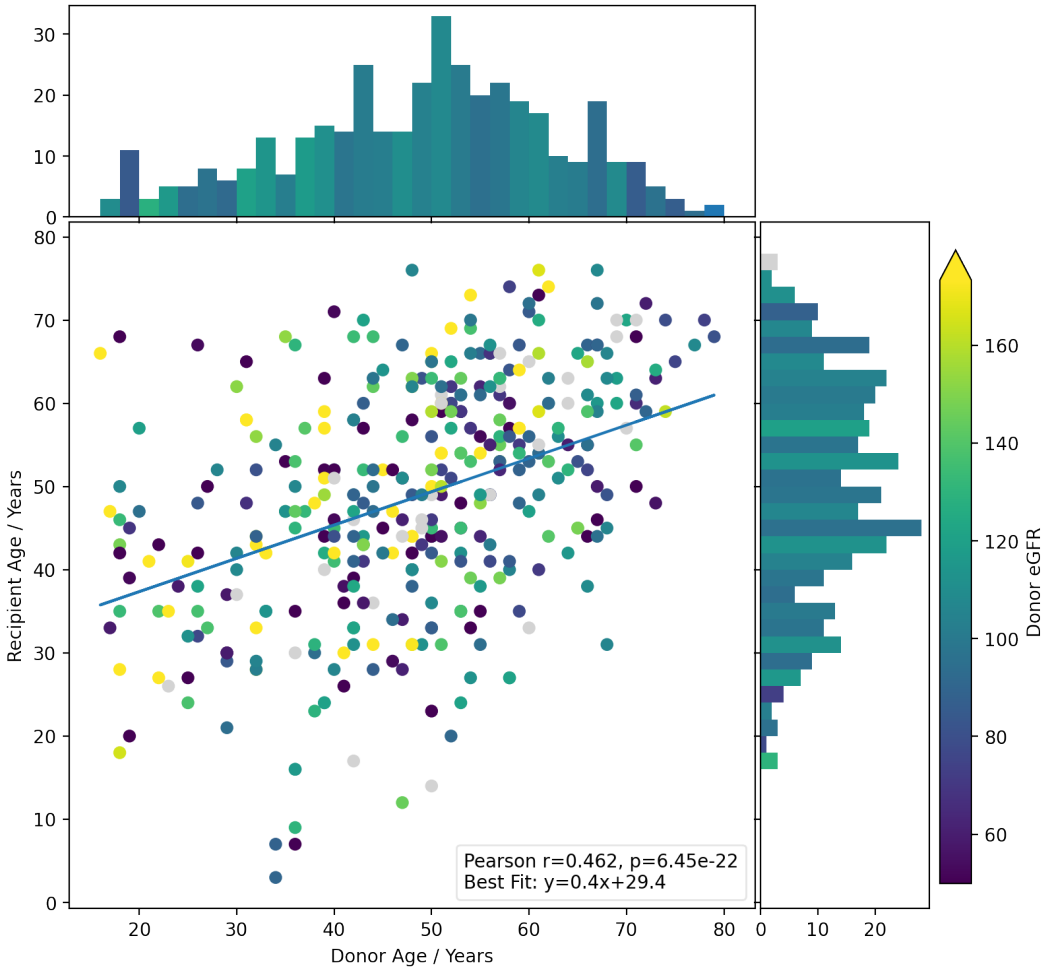


Figure 3.3: Scatter plot showing how kidneys from older donors tend to be matched to older recipients. Donor Age in Years (median (min-max)): 50 (16-79); Recipient Age in Years (median (min-max)): 49.5 (3-76). Markers are coloured by donor eGFR, which is not as good as donor age in predicting eGFR12. Cases with missing donor eGFR values are in grey. Marginal histograms are coloured according to the mean donor eGFR within the age bins.

a spreadsheet template formulated prior to this project for the routine assessment of pre-implantation biopsies. Entries recorded by the template include (i) Remuzzi grade components and the summed score; (ii) Acute Tubular Injury (ATI); (iii) Arteriolar Hyalinosis (Ah). The distributions of these grades are shown in (i) Figures 3.6-3.7, (ii) Figure 3.8, and (iii) Figure 3.9 respectively. Counting only the slides with ≥ 7 glomeruli, we estimate that there are 80 sclerosed glomeruli altogether in the assessed slides (49 sclerosed glomeruli from the NMP biopsies and 31 from the QUOD biopsies). Correlations between the various assessment results relevant

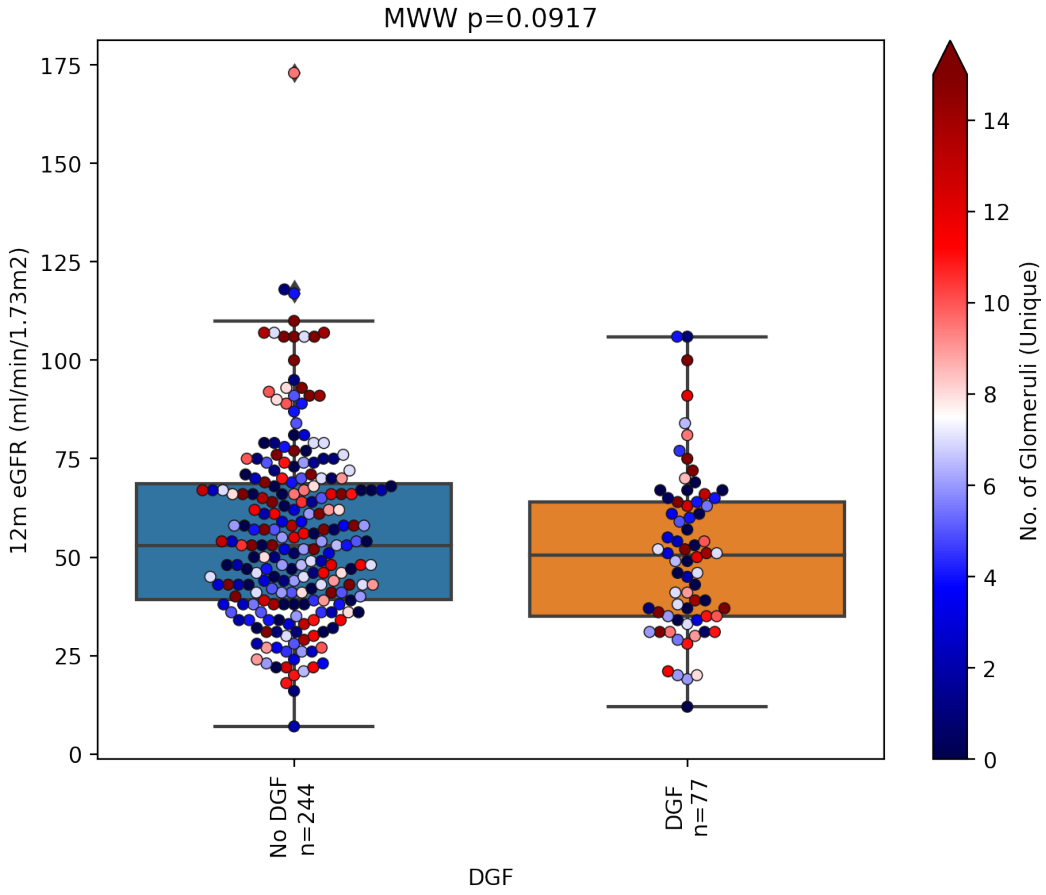


Figure 3.4: Plot showing the relationship between recipients with DGF and eGFR12. No difference is observed when data is split by donor type (DCD/DBD). Colour of the markers reflects the number of glomeruli in the corresponding slides, showing that most slides have < 7 glomeruli.

to this thesis are shown in Figure 3.10. There is a mildly significant correlation between the total score and eGFR12 outcome (Figure 3.11, Spearman $p = 0.014$). When the total scores are broken down into sub-components, we only find Remuzzi A associated with eGFR12 (Spearman $p = 0.046$). Other Remuzzi components are not significantly associated with eGFR12, suggesting that the Remuzzi score itself is made from many weak predictors.

ATI was graded using the following qualitative criteria: *0 – absent; 1 – loss of brush borders/vacuolation of tubular epithelial cells; 2 – cell detachment/cellular casts; 3 – coagulation necrosis.*

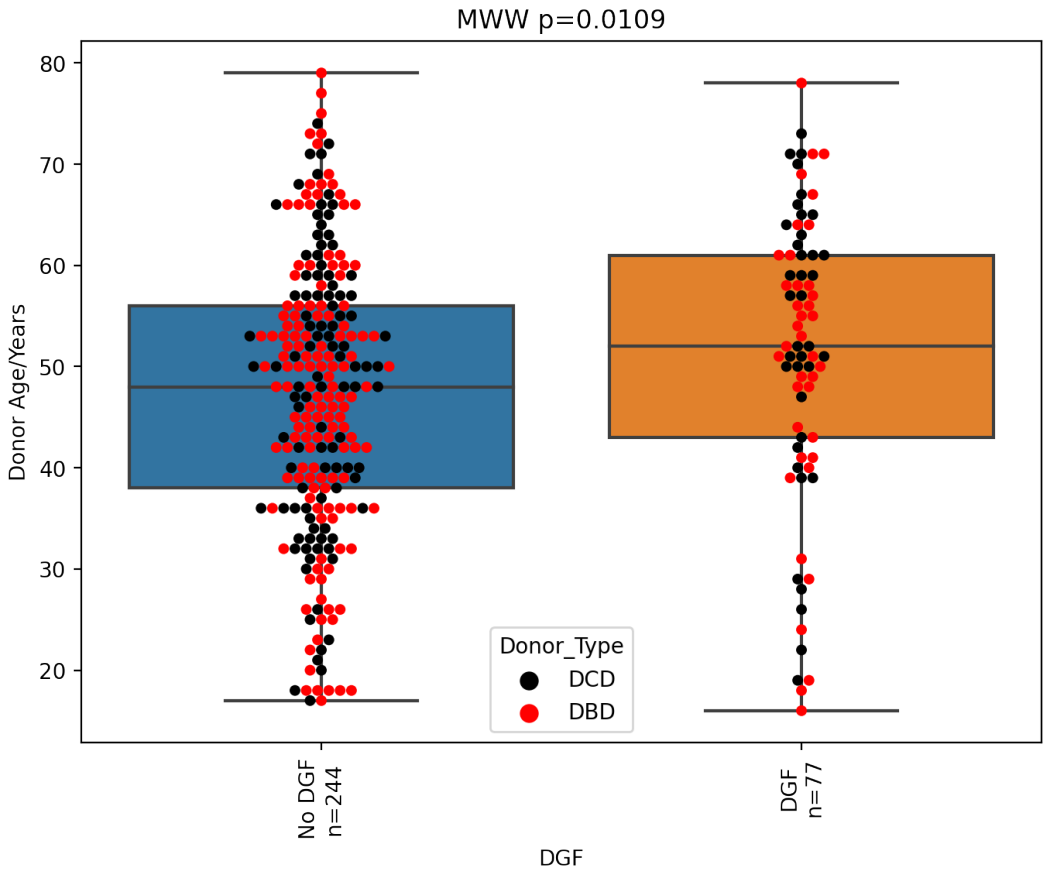


Figure 3.5: Plot showing the relationship between recipients with DGF and Donor Age.

3.3.1.2 Assessment of SR Slides

Apart from PAS and H&E slides, 33 SR slides had also been assessed by our two collaborating pathologists MFS and JK [7] to estimate the percentage of fibrotic tissue in the cortex.

Pathologist 1 assessed 17 slides and rounded the estimated fibrotic content to the nearest 5%. 3 of these 17 slides contain less than 1mm of cortical tissues but were also assessed because these were the only slides available at the early stages of the project. At later stages of the project, Pathologist 2 assessed 20 slides and rounded the estimated fibrotic content to the nearest 10%. Slides were handpicked to contain ≥ 5 glomeruli over a range of eGFR12 outcome. Four slides have been

[7] MFS: Maria F. Soares - specialty doctor in renal pathology; JK: Jesper Kers - renal and transplant specialist

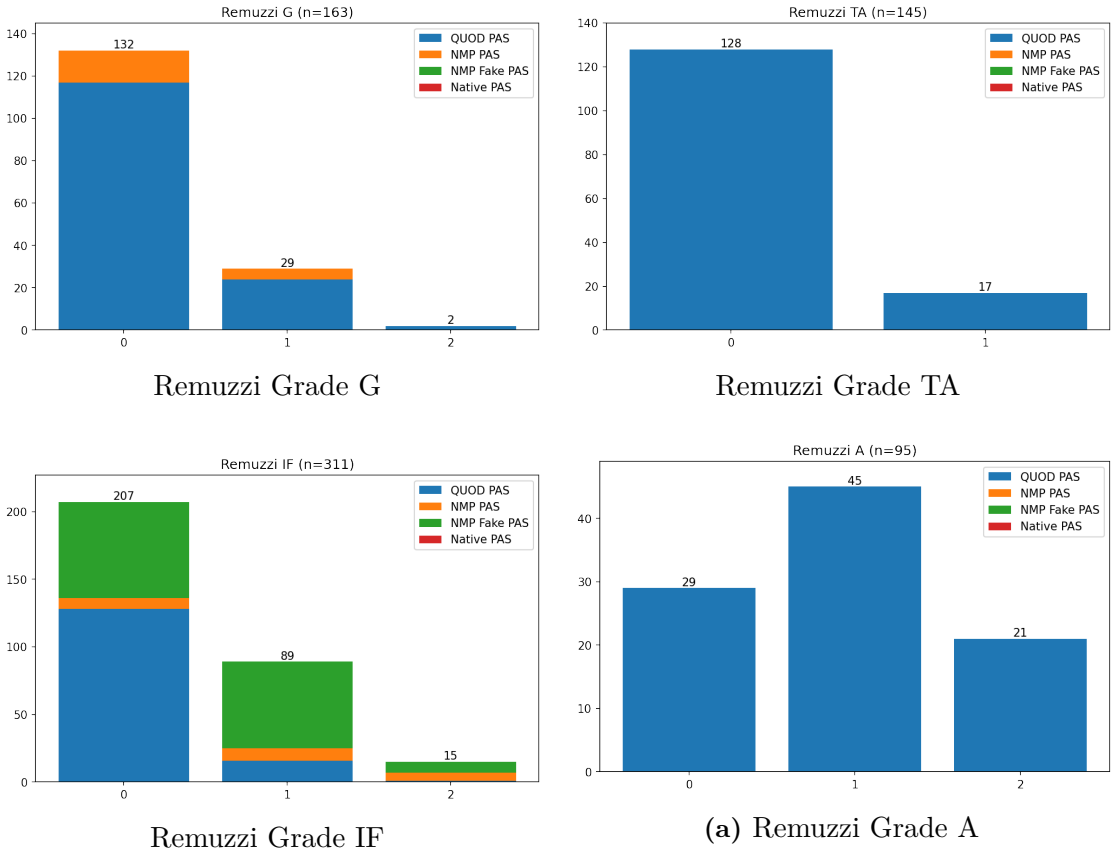


Figure 3.6: Distribution of Remuzzi grade components. Slides that do not contain enough glomeruli/arteries were not scored for that specific category.

assessed by both pathologists. The inter-observer agreement was perfect in these four slides within 5% rounding error Assessment results from the two pathologists were merged (averaged) such that we have a fibrotic percentage for 33 slides.

3.3.2 Local Assessment

3.3.2.1 Patch-based Assessment

At the initial stages of the project, when we only had the H&E NMP slides available, a platform was built [8]. Rectangular patches of physical size $255\mu m \times 255\mu m$ were scored according to two criteria. (i) Informative score - whether it is is relevant for assessment (i.e. contains cortical tissues) based on a score of 0-4; and (ii) Health score - whether it has suffered from any pathological changes based on a score

[8]The interface of the initial assessment in Section 3.3.2.1 was implemented with the help of Nasullah Khalid Alham. The author built the platform in the next stage in Section 3.3.2.2 (Figure 3.12) to assess whether individual tissue compartments were informative.

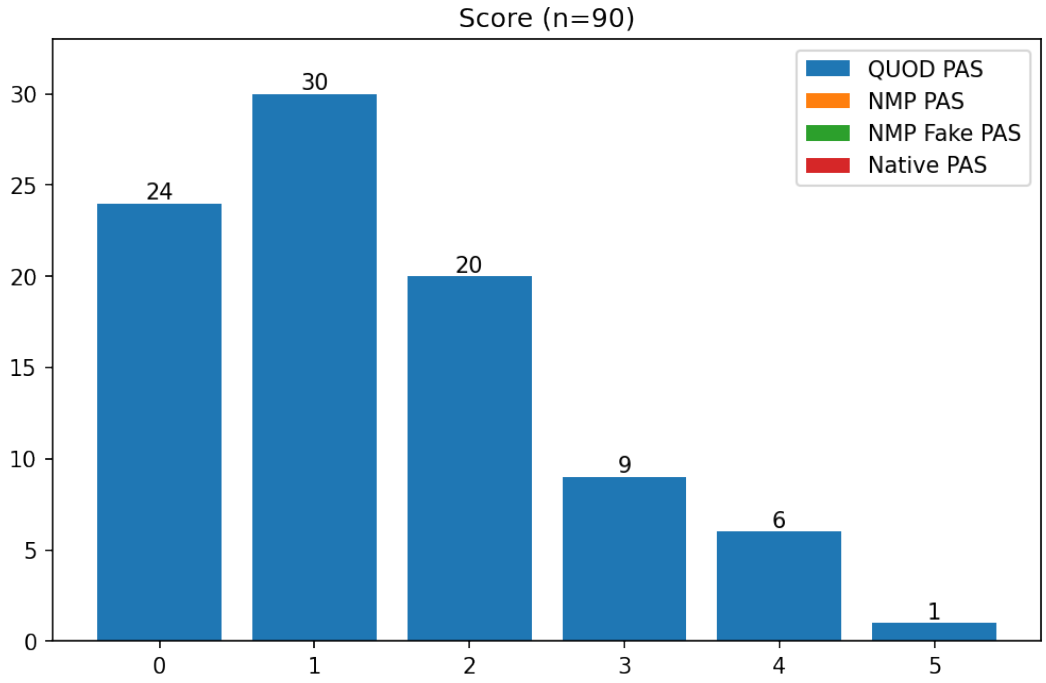


Figure 3.7: Distribution of Remuzzi Score. Only slides containing more than 7 glomeruli and 1 artery are included.

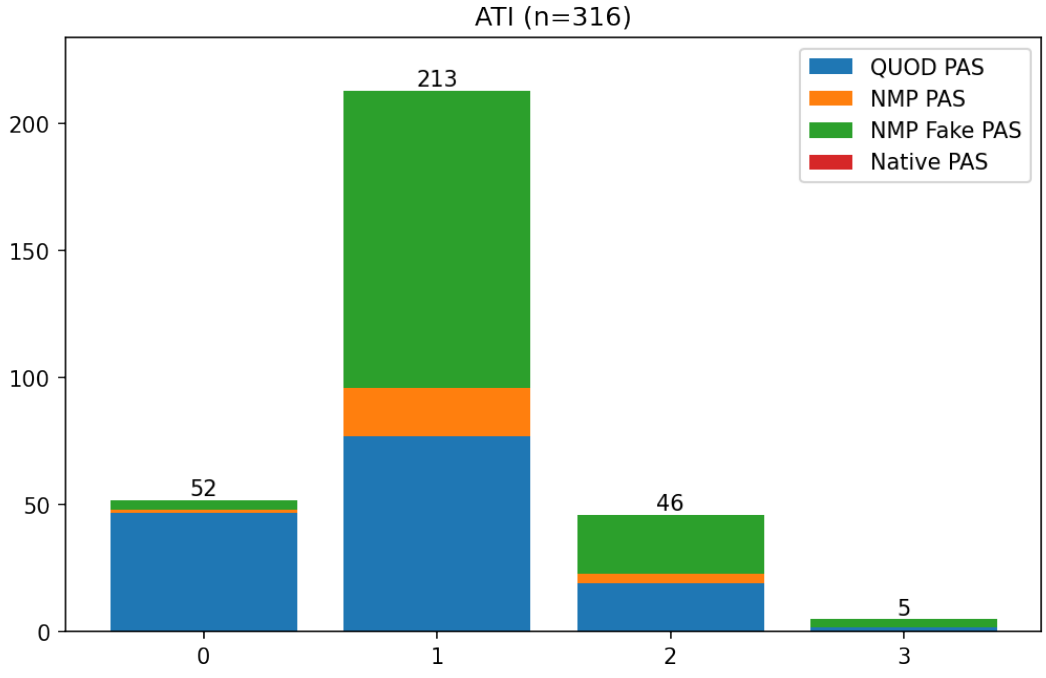


Figure 3.8: Distribution of ATI grades.

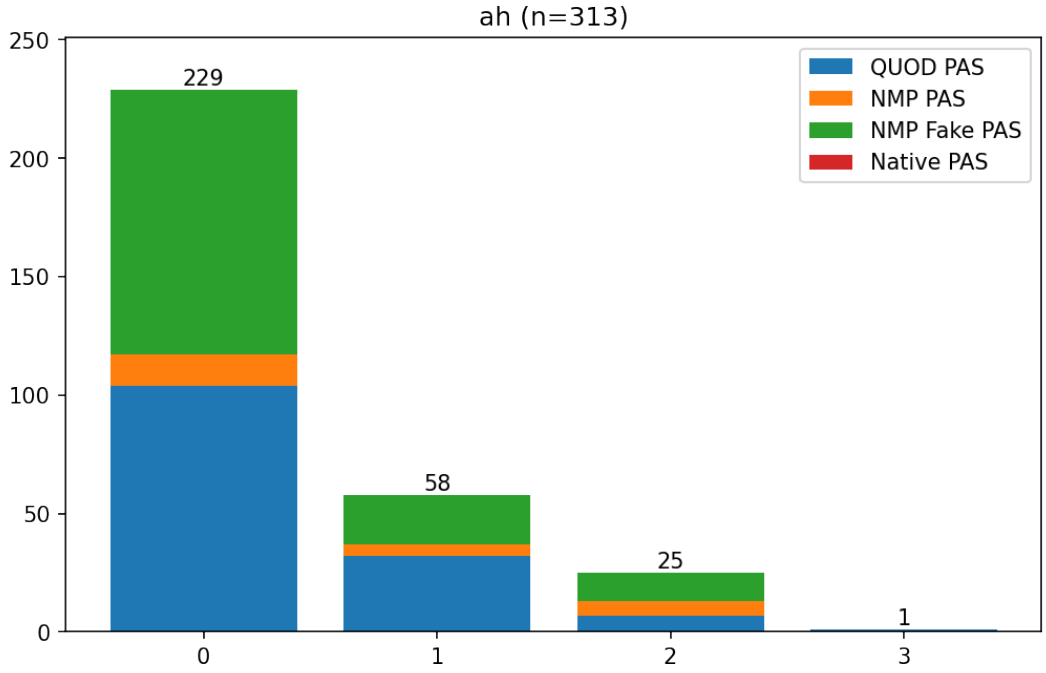


Figure 3.9: Distribution of Arteriolar Hyalinosis grades.

between 1-5 (N/A if tissue is not relevant for assessment). A total of 573 tiles had been scored. The author (KHT) scored 573 slides, our collaborating renal pathologist (MFS) [9] scored 104 slides, and a second pathologist (AS) [10] scored the same 104 slides. For the informative score, the Spearman correlation coefficients between different pairs of scorers are: KHT vs AS: 0.59; KHT vs MFS: 0.744; AS vs MFS: 0.525. For the health score, correlation is weighted by the mean informative score between the pair of scorers and the coefficients are: KHT vs AS: 0.319; KHT vs MFS: 0.387; AS vs MFS: 0.283.

3.3.2.2 Tissue Compartment-Based Assessment

At a later stage of the project, a more detailed assessment platform was designed in-house to assess individual tissue compartments. In contrast to using larger tissue patches as we did earlier, this platform delivers more granularity in disease category to variable size tissue compartments rather than fixed-sized patches. The form consists of some common findings from tissues. For crops containing glomeruli,

[9] Maria F. Soares - specialty doctor in renal pathology
[10] Andrew Smith - consultant histopathologist

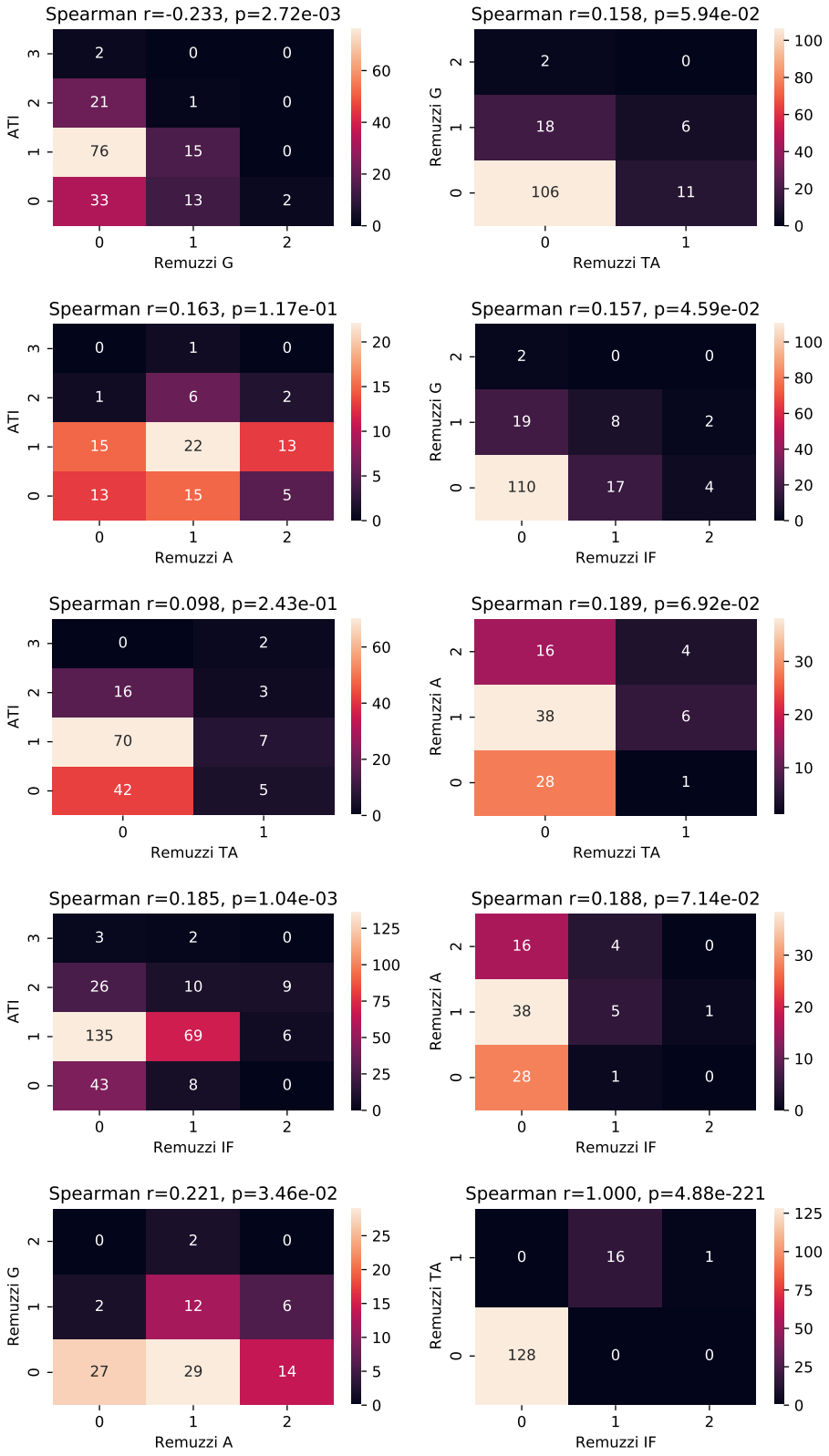


Figure 3.10: Correlation between slide-level grades given by pathologist. Spearman correlation coefficient values are shown. Remuzzi TA and IF are highly correlated, but we have more slides graded for IF but not TA.

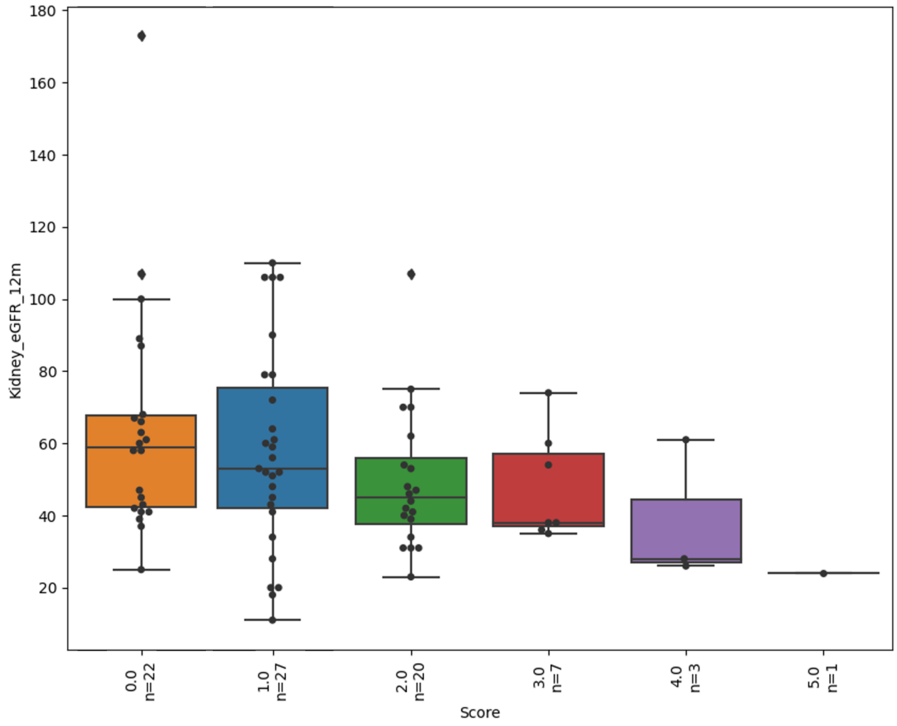


Figure 3.11: Remuzzi Score vs eGFR12, Spearman $r = -0.274, p = 0.014$

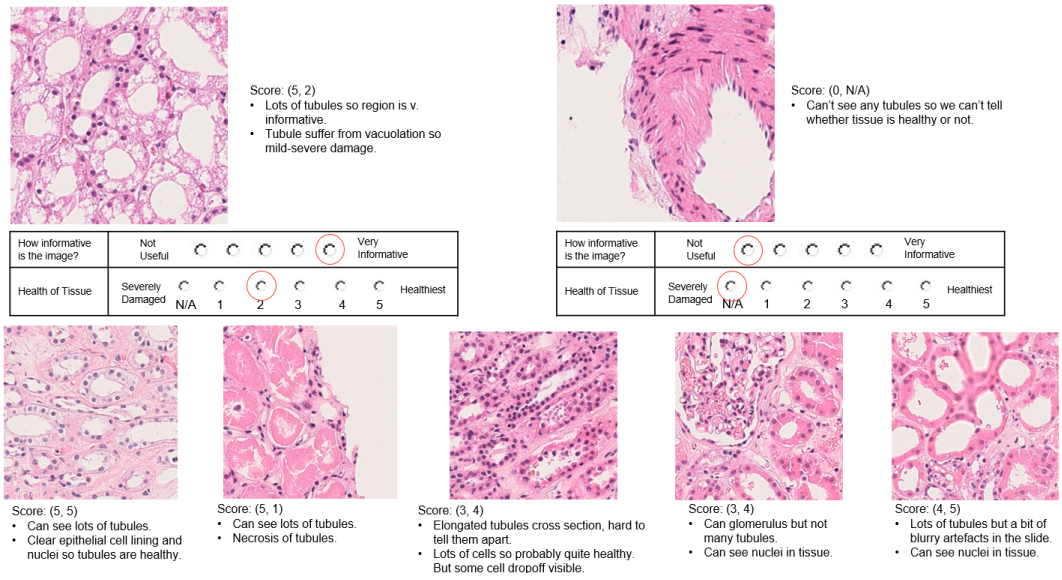


Figure 3.12: Platform used for assessing patches of tissues.

tissues could be labelled as to whether it contains sclerosed tissue, microthrombi, focal segmental glomerulosclerosis (FSGS), or tuft collapse. Crops were graded according to the percentage of fibrosis present. Proximal tubules were graded by the amount of chronic (atrophy) and acute (ATI) changes. An expandable drop-down text box was also available for pathologists to add their list of findings if they were not already available. A snapshot of the interface is shown in Figure 3.13.

In existing scoring criteria, tubular atrophy is graded according to the percentage of tissue area affected and does not directly describe the severity of individual tubules. It would strictly only be necessary to provide binary labels to identify the presence of chronic changes in individual tissues. However, we argue that there are no downsides to grading individual tissues based on severity as these ordinal grades can always be regrouped to binary grades. In the long run, grading individual tissues may help us identify extremities for machine learning. Hence, tubules were graded according to the degree of thickening of the basement membrane. A "0" was given to tubules with no visible thickening; "1" - tubules with barely noticeable thickening, "2" - thickening of basement membrane easily recognisable; "3" - thickness of basement membrane comparable to the epithelial cell lining; "4" - basement membrane thicker than cell lining. Examples given grade 0 are displayed in Chapter 2, Figure 2.7a; examples given grade 1 are shown in Figure 2.7b; examples given grade 3-4 are shown in Figure 2.7c.

We have used this interface to assess individual PAS/H&E tissues delineated by hand by the author and those segmented by our UNets in Chapter 4, Section 4.2.1. Different tissue compartments were mixed together, and a subset was initially picked by hand and fed in random order to the pathologist MFS. However, after assessing several dozen tissue compartments, we narrowed down the subset further due to the pathologist's time constraint. From this point forward, the order of the tissue compartments assessed was chosen to maximise the coverage of their Variational AutoEncoder (VAE) embedding, according to Sener et al. [124]. The VAE was trained with a latent embedding space with dimension $nz = 100$ to minimise a cross-entropy (CE) loss with a weight of 1, structural similarity (SSIM) [125] loss with a

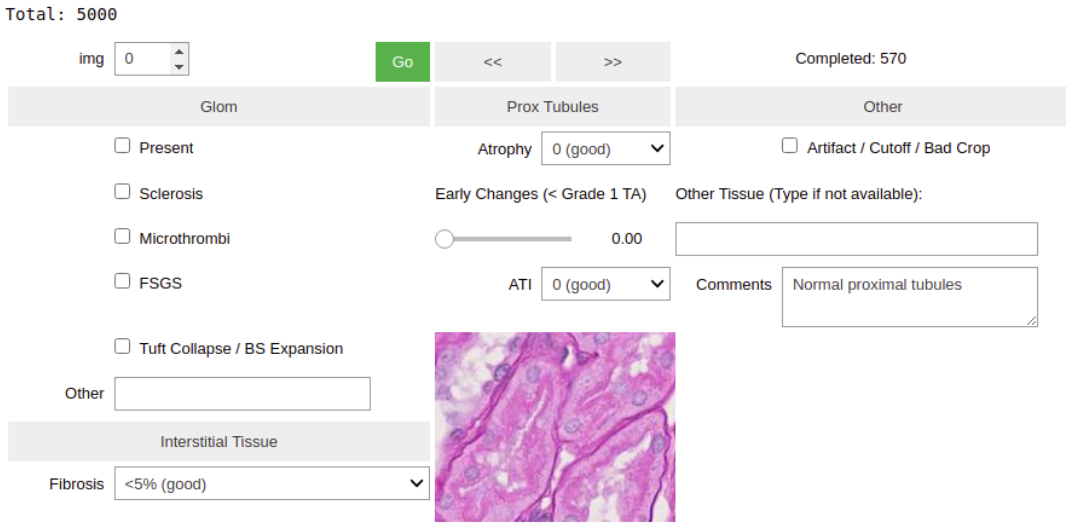


Figure 3.13: Web interface used for assessing individual tissue compartments.

weight of 0.1, maximise the evidence lower bound (ELBO) with a weight of 10^{-2} to restrict the latent vector to a unit Gaussian prior. Hyperparameters of the VAE were optimised using a package HyperOpt's Bayesian optimisation algorithm [126] with Ray Tune platform [127]. After the VAE had learned a meaningful embedding of tissue compartments, the first tissue compartment shown to the pathologist was random. After that, every image was chosen such that the minimal Euclidean distance from all the previous images was maximised. By doing this, we would not need to meticulously plan to have the pathologist assess a fixed number of images.

A total of 1992 tissue compartments had been reviewed by the pathologist using our web-based platform: 1032 of which were output from UNets (Section 4.2.3) to cross-check the predictions of the segmentation algorithm, and 960 were delineated by the author to cross-check the quality of the training set input. It was difficult to foresee how much time the pathologist had available, so we mixed these 1992 tissue compartments and presented them to the pathologist in the same go. This ensured that we would have at least some results available to cross-check the UNet's segmentation and the author's delineation while maximising the overall coverage in VAE embedding.

The pathologist cross-checked objects labelled or predicted as belonging to either tubule or glomeruli class. The statistics of the assessment are shown in Tables

3.4 and 3.5. From Table 3.4, it can be seen that less than half (157/340) of the compartments labelled as "tubules" were actually proximal tubules. However, since this was an ongoing project, the delineation quality might have improved over time. A qualitative estimate suggested that approximately 70% of the delineated tubules are proximal tubules in the up-to-date dataset.

A total of 731 tissue compartments reviewed by the pathologist are proximal tubules. Most of these were graded for chronic (TA 0-5) and acute (ATI 0-5) damages. The distributions of these grades, broken down by dataset, are shown in Figure 3.14. It can be seen that TA grades are heavily imbalanced. Between the two large datasets (QUOD and NMP), only seven and one proximal tubules had been given grade 1, and none have grades above 1. The distribution of ATI grades, on the other hand, is much more evenly spread.

Figure 3.15 shows that the most common glomeruli findings are FSGS and sclerosis based on 759 objects labelled as glomeruli by the pathologist. Note that the pathologist identified six objects as urinary casts based on the tissues that had either been delineated or segmented as glomeruli. Given that there were only twelve globally sclerosed glomeruli, we estimate that the error rate of mislabelling sclerosed glomeruli by non-expert is 1/3.

Table 3.4: Delineated Tissue Compartments Reviewed by the Pathologist. A subset of tissue compartments delineated by the author has been cross-checked by the pathologist. Segmentation of instances is based on coarse labels performed by the author; whereas the pathologist was able to classify these compartments into more fine-grain classes. Hence this table gives a representation of the quality of the input used to train our UNets.

Pathologist Cross-Check \Author Delineation	Tubules	Glomeruli
Proximal Tubules	157	3
Glomeruli	1	600
Vessels	0	0
Other Tissue Compartments	152	1
Artefacts	30	16
Total Relevant	158	603
Total	340	620

Table 3.5: Segmented Tissue Compartments Reviewed by the Pathologist. A subset of tissue compartments segmented by UNets (Section 4.2.3) has been cross-checked by the pathologist. This table gives a partial representation of the quality of the UNets, which could be influenced by the quality of the author's input (Table 3.4) and model design. Note that the numbers in this table include a mix of different datasets as we are mainly interested in the pathologist's input, thus it may not reflect the limitations of the UNets to the full extent. A more realistic assessment of the UNets' performance with respect to author's input labels is discussed in Chapter 4 Section 4.2.

Pathologist Cross-Check \UNet Predictions	Tubules	Glomeruli
Proximal Tubules	553	18
Glomeruli	75	83
Vessels	11	0
Other Tissue Compartments	123	86
Artefacts	76	7
Total Relevant	639	101
Total	838	194

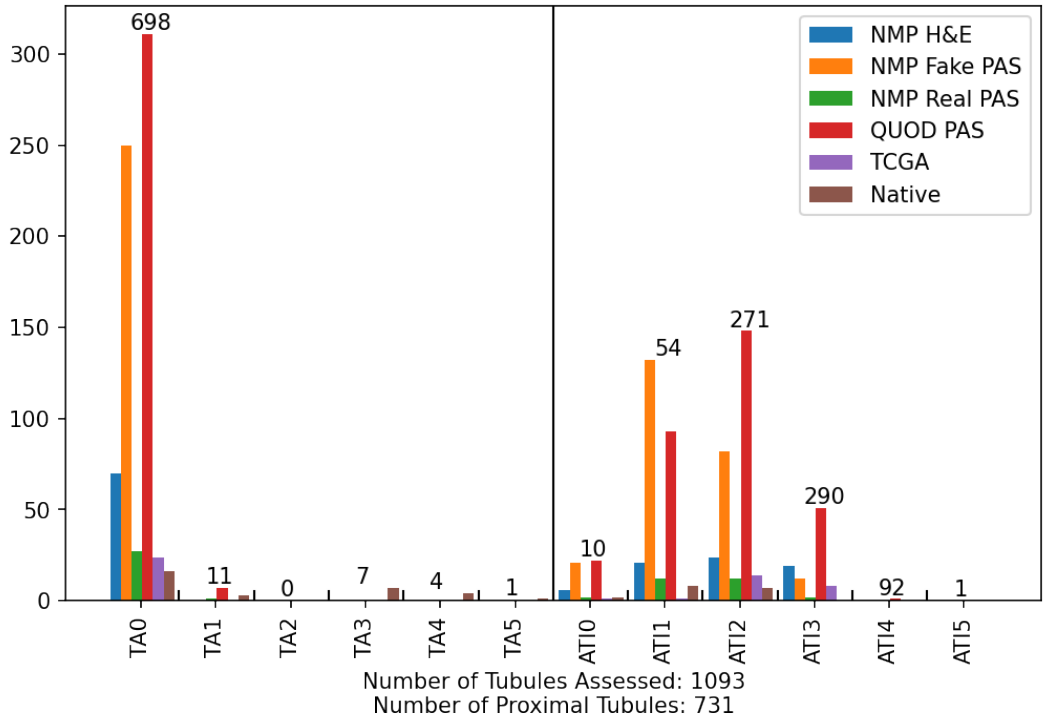


Figure 3.14: Distribution of grades given to individual proximal tubules.

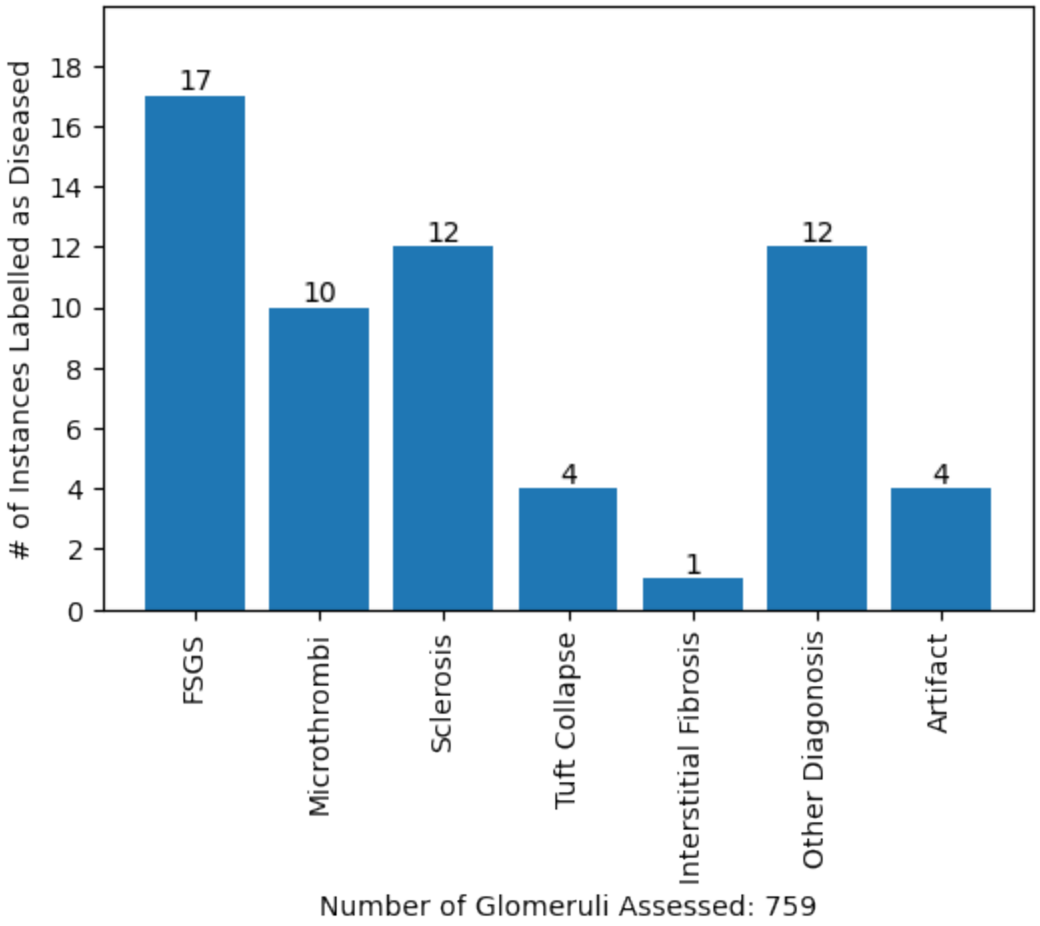


Figure 3.15: Glomeruli disease labels based on a sample of 759 compartments.

3.3.3 Delineation of Tissues

We manually marked out a number of rectangular tiles from each dataset to delineate different tissue compartment types - tubules, glomeruli, vessels, and cell nuclei. The number of tiles selected from each dataset is shown in Table 3.2. Delineating the outline of tissues is a laborious task. It was performed by an engineering student rather than a pathologist to make the task scalable [11]. Thus, we only have the coarse classification of these tissues. For instance, proximal tubules are relevant for assessment, but a notable portion of objects marked as "tubules" were, in fact, distal tubules or the collecting duct. Objects marked as "vessels" consist of a mixture of arteries, arteriole, and veins. However, most arterioles had not been correctly

[11] I performed all of the delineation work but had limited histology knowledge at the beginning of this project.

labelled – many had been labelled as tubules. Some casts had been misidentified as sclerosed glomeruli. The boundaries of glomeruli were inconsistent regarding the inclusion of the urinary space. Due to the small size of many biopsies in the QUOD dataset, tissues truncated near the edge of the biopsies were also delineated as long as they were human-recognisable. The number of tissues delineated from each dataset is summarised in 3.6.

As discussed earlier in Section 3.3.2.2, a subset of these tissue compartments had been cross-checked by our collaborating pathologist in Table 3.4.

Delineation of nuclei is not shown in the tables as it was performed differently from the delineation of larger compartments. Cell nuclei are much smaller and more numerous; hence the ground truth masks were generated using the "Cell Detection" function in QuPath [128], with parameters hand-adjusted depending on the colouration of each tile. A total of 65 tiles containing 7409 cell nuclei masks were directly generated from the NMP H&E slides. Due to limitations of the annotation software, these tiles do not necessarily coincide with the tiles delineating the larger compartments. A UNet was trained to generate an intermediate set of "ground truth" nuclei masks on the 144 NMP H&E tiles from these initial nuclei masks. The UNet was trained using the proposed architecture in Section 4.2.1 on the NMP H&E slides at 0.22mpp with 4:1 train:validation split. Instances were obtained from connected pixels after applying a threshold to the softmax output. The IOU and MI based on the validation data are 0.891 and 0.765, respectively. These 144 tiles altogether contain 276777 nuclei.

Cell nuclei tend to be stained much darker under H&E compared to PAS. Nuclei under PAS stain tend to vary greatly in colour, with many appearing too pale to be detected by hand or using QuPath's tools. When attempting to generate nuclei masks on PAS images, intra-observer variability was found to be very high (IOU <0.3). It may be difficult to interpret whether the UNet is performing well on ground truths of such high variability. Because of this, it may not be particularly meaningful to generate ground truth nuclei masks on PAS slides. Performance of cell nuclei

segmentation in later chapters would only be evaluated on the masks as mentioned earlier generated on H&E on either the original or the converted fake PAS slides.

Table 3.6: Overview of Tissue Instances Delineated by Hand. The total area of tiles delineated is shown in Table 3.2. **Note:* Because glomeruli and vessels are much rarer than tubules, only these two classes were delineated in some tiles. The first number includes all instances delineated, whereas the number in brackets shows instances from tiles where only glomeruli and vessels were delineated.

	Number of Instances*			Area* (mm^2)		
	Tubules	Glomerulus	Vessels	Tubule	Glomerulus	Vessel
QUOD → PAS	2145	323 (253)	119 (52)	4.68	3.87 (3.05)	2.60 (2.10)
QUOD → SR	889	24	22	2.03	0.309	0.254
NMP → HE	13103	581 (330)	175 (64)	31.8	7.39 (3.93)	4.37 (2.86)
NMP → PAS	753	25	40	1.64	0.270	0.236
NMP → SR	1311	38	17	2.70	0.411	0.493
Native → PAS	1156	226 (194)	55 (25)	3.62	4.15 (3.64)	1.11 (0.93)
TCGA → HE	1739	422 (386)	143 (67)	5.84	7.53 (7.07)	9.49 (6.01)

Although the outlines of these tissues mainly were delineated using the same software QuPath [128], the input devices and style of the annotations had changed over time. To grasp the extent of the intra-observer variability, 5 large tiles containing 466 tubules, 12 glomeruli, and 1 vessel were re-annotated by hand. The time elapsed between the two annotation attempts was more than six months. Variability for each class was measured using Intersection over Union (IOU) and normalised Mutual Information (MI) and are shown in Table 3.7. These metrics could be used as a reference for our segmentation workflow's "ceiling" performance in Chapter 4.2.

Table 3.7: Intra-Observer Variability of Tissue Delineation.

Tissue Class	IOU	MI
Tubules	0.764	0.575
Glomeruli	0.876	0.706
Vessels	0.670	0.490

3.4 Biopsy Statistics

There is an inherent trade-off between obtaining biopsies with enough tissue and the risk of complications such as bleeding. As such, in QUOD's protocol, biopsies tend to be very small. A recent audit that replicated the QUOD biobank practice [12] highlighted that other researchers had raised concerns that the amount of tissue might be insufficient for analysis. The audit compared the tissue sizes and amount of proteins recovered from different punch and core biopsy sizes using pig kidneys. It was found that half core needle biopsies contain only approximately 30% of the protein content compared to whole core biopsies. It is possible that a similar challenge also holds for visual analysis of these slides - the perimeter-to-area ratio increases when biopsies are cut, so a larger portion of tissues come close to the edge, leading to more large tissues such as arteries being truncated. The extra processing step in cutting them and the increased number of tissues near the edge may have also made the slides more susceptible to artefacts. Analysing the PAS-stained QUOD slides only, we found that 268 of the 414 half-core biopsies contain duplicated sections on the same slide. Table 3.8 shows the median area and the number of glomeruli in the biopsies (not counting duplicated tissue sections). More detailed distribution of the areas, number of glomeruli, and arteries are shown in Figures 3.16, 3.17, and 3.18, respectively.

Table 3.8: Median values of tissue area, number of glomeruli and vessels by dataset. Tissue areas were computed algorithmically, whereas the number of tissue compartments displayed here were counted by a pathologist.

Dataset	Area (mm^2)	No. of Glomeruli	No. of Arteries
QUOD	3.91	6	0
NMP	7.09	7	2
Native	16.37	27	N/A

3.5 Pre-Processing

During the early stages of the project, only H&E slides from the NMP dataset were available. However, we have subsequently received slides from other datasets

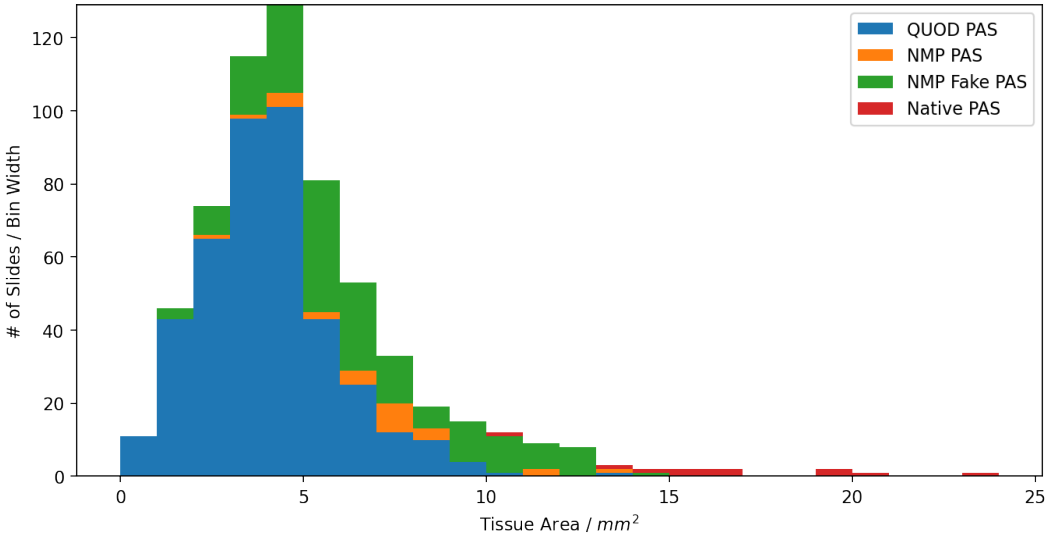


Figure 3.16: Distribution of tissue area in different datasets

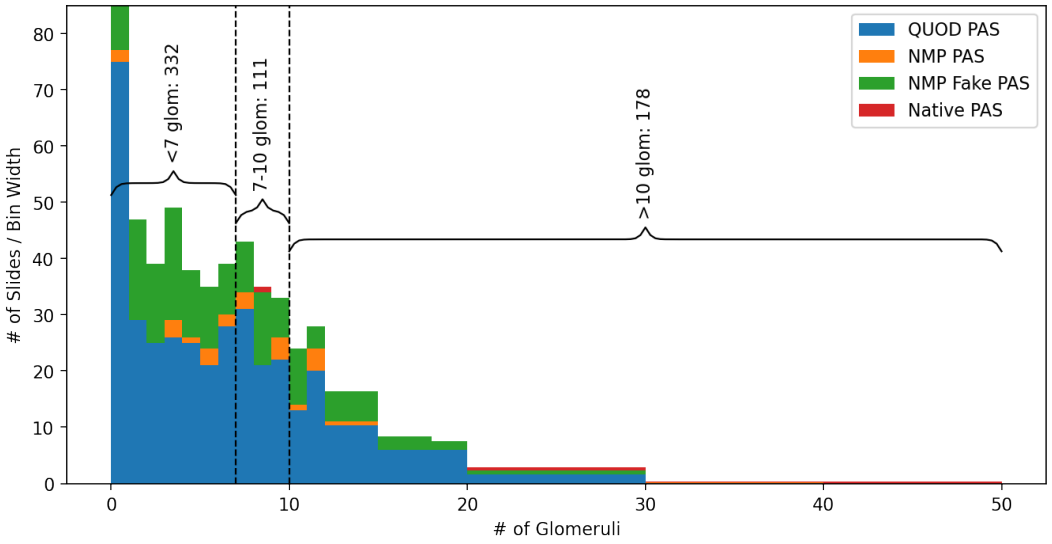


Figure 3.17: Distribution of the number of glomeruli in our datasets. There is an inherent trade-off between obtaining biopsies size and the risk of complications such as bleeding. Some slides contain multiple adjacent sections of the same biopsy - we manually identified these slides and avoided double counting these instances. The majority (332) of slides do not have enough glomeruli for assessment as stipulated by Banff Criteria.

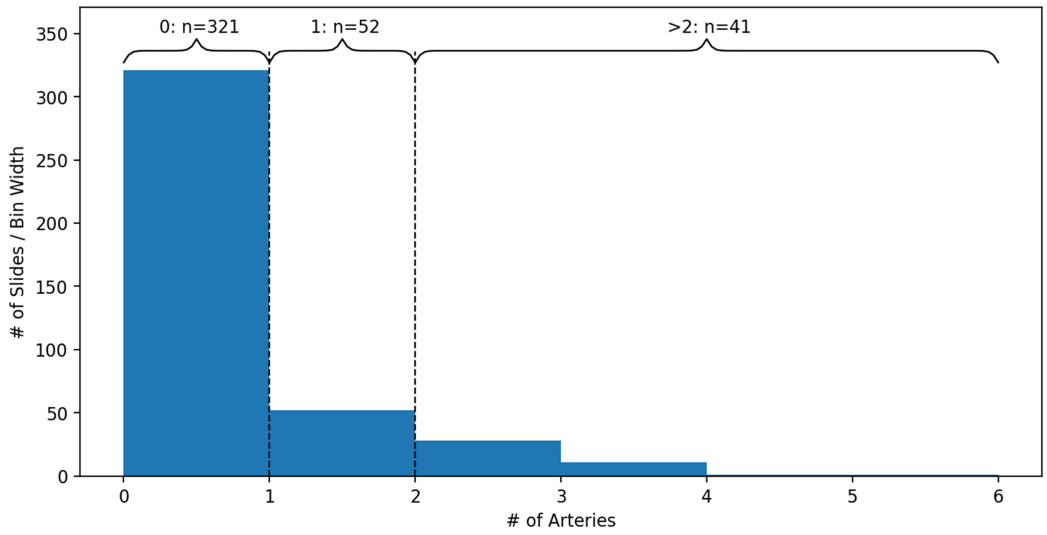


Figure 3.18: Distribution of the number of arteries in the QUOD dataset (PAS-stained slides only). There is some discrepancy between the artery count and those that have received a Remuzzi A grade. This is possibly because some of the slides have partially truncated arteries.

stained with PAS and scanned using different scanners. Thus, we needed to homogenise all datasets to be processed by our workflow without introducing major tweaks to the models.

While the original NMP slides were digitised using a Hamamatsu scanner, over half of the QUOD slides were digitised using a Glissando Desktop scanner. We also have some, albeit fewer, slides digitised using scanners made by Philips, Leica (Aperio), and unknown brands in other training data. While the NMP and QUOD slides were both scanned using a 40x 0.65NA objective, there were differences in the sensor size of these scanners. Thus, this resulted in a pixel resolution of 0.22mpp for the Hamamatsu scanner, 0.25mpp for the Philips scanner, and 0.27mpp for the Glissando scanner. We have chosen to resize all images to 0.44mpp in consideration of a few factors: (i) broader workflow applicability to different scanners; (ii) processing speed and file size; (iii) optical resolution of images in the presence of optical aberrations (e.g. minor defocus is visible in a significant number of slides due to imperfect tissue flatness and technical glitch with the scanner).

3.5.1 Stain-to-Stain Translation

The majority of slides in the NMP dataset were stained in H&E. Only slides from 4 donors in this dataset were stained in PAS. As the original workflow was developed based on this dataset, it would be advantageous to pre-process these slides such that they are in the same staining style as the QUOD slides. Changing the colouration of stains could be done using various style-transfer techniques such as Pix2pix [89], CycleGAN [98] and their derivatives [88, 129]. While Pix2pix is known to produce the best quality results, it is a supervised training method, and paired images are required. This is often not possible in histology since (i) serial sections are not always available; (ii) serial sections are often misaligned due to changes in sizes and shapes of tissues across the thickness of the biopsy. CycleGAN, on the other hand, can be trained using unpaired image patches.

Although we find that the original architectural implementation of CycleGAN [12] was sufficient for generating visually convincing results, there were still a great number of implementation details regarding how domain adaptation could be performed. Firstly, we could either convert the training data to PAS or convert test data to H&E. We chose the former for two reasons: (i) converting WSIs to PAS tends to lead to better segmentation results as tissue boundaries are highlighted under PAS stain. This had been studied extensively by Gadermayr et al. [114]. (ii) The later option requires running CycleGAN on the fly and could result in more computational time.

Furthermore, the training of adversarial networks is known to be unstable. Losses do not necessarily resemble quality output by the generator, and it is often difficult to dictate the translation's quality visually. This raises the question of how to arrive at a quantitative decision to terminate the training of CycleGAN. To this end, we registered a number of serial section WSIs (Section 3.5.2). Based on 23 biopsies with pairs of serial sections, we hand picked 709 coordinates where registration was of better quality. We calculated the Complex Wavelet Structural Similarity Index

[12] We used the CycleGAN architecture from the original author's Github repository out-of-the-box. Link: https://github.com/junyanz/pytorch-CycleGAN-and-pix2pix

(CW-SSIM) [130] between tiles translated with CycleGAN and the corresponding tiles during training. CW-SSIM is the complex domain extension of the Structural Similarity Index (SSIM) and is less sensitive to spatial transformations such as translation and rotation between images than SSIM.

Next, we had to consider the choice of our target dataset. We had the option to either train a CycleGAN to translate between the NMP-H&E slides and (i) the NMP-PAS slides or (ii) the QUOD-PAS slides. While the latter was a viable option to be implemented during test time, there is a possibility that the fake PAS slides generated using (ii) could bias the development of workflow. Therefore we chose to use the former option for the rest of the thesis. We found the difference in CW-SSIM metric between the two options was small when evaluated over the registered PAS tiles - CW-SSIM=0.418 in the former case and CW-SSIM=0.427 in the latter.

Through experimentation, we found that including more tiles in the training of CycleGAN did not always lead to more realistic results. Including too many blank tiles or irrelevant tissues often resulted in divergent losses. In addition, the stain colour of tissues was conditional to the tissue type and morphology. Based on this understanding, we suspected that the distribution of content within the training data might influence the quality of the results. To test whether this was the case, we generated some synthetic histology-stain look-a-like where the stain colour (style) was conditioned to the foreground (content). The foreground was based on the MNIST dataset scaled to 64x64 pixels, allowing the conditional dependencies to be layman-recognisable. Set A of the synthetic data had H&E background (from the NMP Dataset) where the colour of cell nuclei was gradually turned green from 0 to 9. Set B consisted of an SR background where the reds were artificially turned green as the numbers changed from 0 to 9. We trained two cycleGAN models - the first model was trained with a balanced number of 0-9 (approximately 6000 examples for each character); the second model was trained with more “9”s in set A and more “0”s in set B (SR). A montage of the inputs and fake outputs are shown in Figures 3.19 and 3.20. It can be seen that in the case where the distribution of MINST characters was balanced, the CycleGAN managed to learn the conditional

relationship - there are no shades of green visible in the character "0" in either the fake H&E or SR outputs in Figure 3.19. However, the output of the second model demonstrated some crosstalk between content and style: examples in Figure 3.20 show some of the cell nuclei behind the "0" fake images are green when in the real inputs, all nuclei should be dark red.

This mini-experiment highlights a need to regulate the distribution of the training data. If we are training a CycleGAN to translate two datasets, we need to ensure that both datasets have a similar distribution in their tissue content. For instance, the proportion of cortex to medulla may differ in the NMP and QUOD datasets due to the core needle length differences. In this case, it may be optimal to up-sample patches that contain the cortex in one dataset such that the distributions match.

In our thesis, however, training was performed using patches extracted from serial sections from the NMP dataset for each stain (21,464 tiles in each domain for HE $\leftrightarrow$ PAS; 20,906 tiles in each domain for HE $\leftrightarrow$ SR) using tiles of size 400x400 at 0.44mpp with a batch size of 2. Thus, the distribution of tissue types would be similar between the two domains. Despite this, there could still be a benefit to training the CycleGAN based on tiles containing more cortical tissues relevant for predicting clinical complications. Using a network that predicts the relevance score (0-1) of a tile, we up-sampled those that are more informative such that in each epoch the CycleGAN's exposure to tiles with a given score was proportional to the score itself.

Some examples of tiles output from H&E are shown in Figure 3.21. At the tile level, our collaborating pathologist found that the fake results were generally indistinguishable from the real stains for H&E $\leftrightarrow$ PAS at low magnifications. Weaker contrast in basement membrane and nuclei can be seen at higher magnifications for some tissue patches. The majority of fake SR tiles also show decent quality.

To convert WSIs, 2048x2048 tiles containing tissues were processed and stitched together using a sliding window with 50% overlap. Overlapped tiles were combined by weighted averaging based on the squared distance from the centre of the tile. Blank tiles were set to the brightest value of the background and were not processed by the CycleGAN. Examples of stitched fake WSIs are shown in Figures 3.22 and

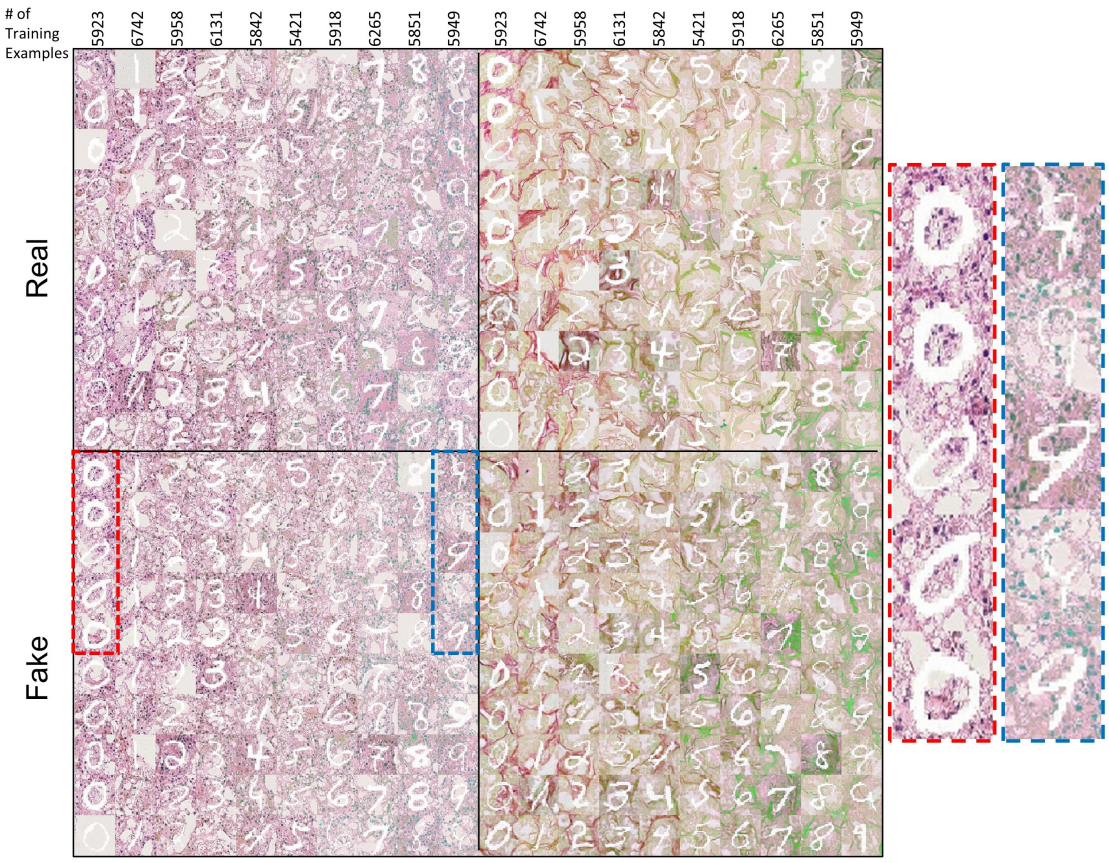


Figure 3.19: Montage of training examples (top) and fake outputs (bottom) for a CycleGAN trained with a balanced distribution of '0's to '9's. The number of training examples for each character is shown on the very top.

3.23. We observe that the stitching of the fake PAS slides appears to be seamless and convincing. However, conversion between H&E and SR often resulted in unrealistic outputs at higher magnifications. Despite the stitching being mostly seamless, SR to H&E often led to the wrong placement of cell nuclei in various interstitial, collagen, and vessel tissues. This could be because no nuclei are visible in our SR stained slides, and there are only a limited number of training examples containing collagen tissues. Vice versa, conversion from H&E to SR often leads to continuous large patches of yellow (picric acid) or reds stain. Although the patches are realistic at the local level, the global distribution of reds and yellows did not resemble the real WSIs.

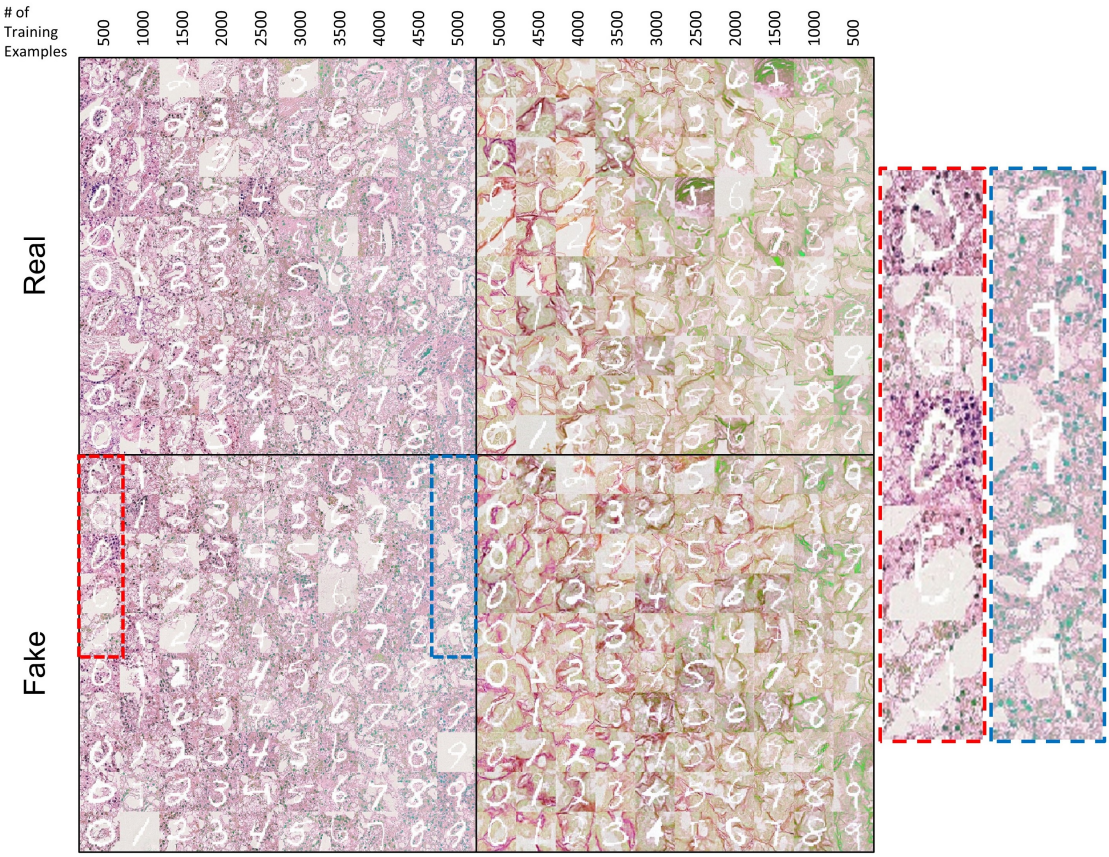


Figure 3.20: Montage of training examples (top) and fake outputs (bottom) for a CycleGAN trained with imbalanced distribution of '0's to '9's. The number of training examples for each character is shown on the very top. Note that some fake '0's were erroneously given green nuclei.

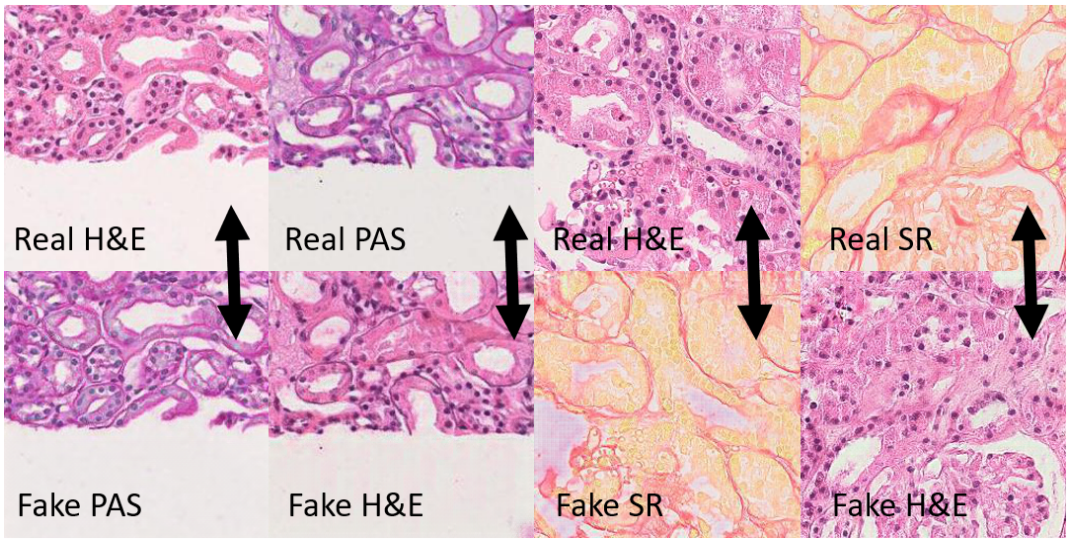


Figure 3.21: Tile examples of stain-to-stain conversion.

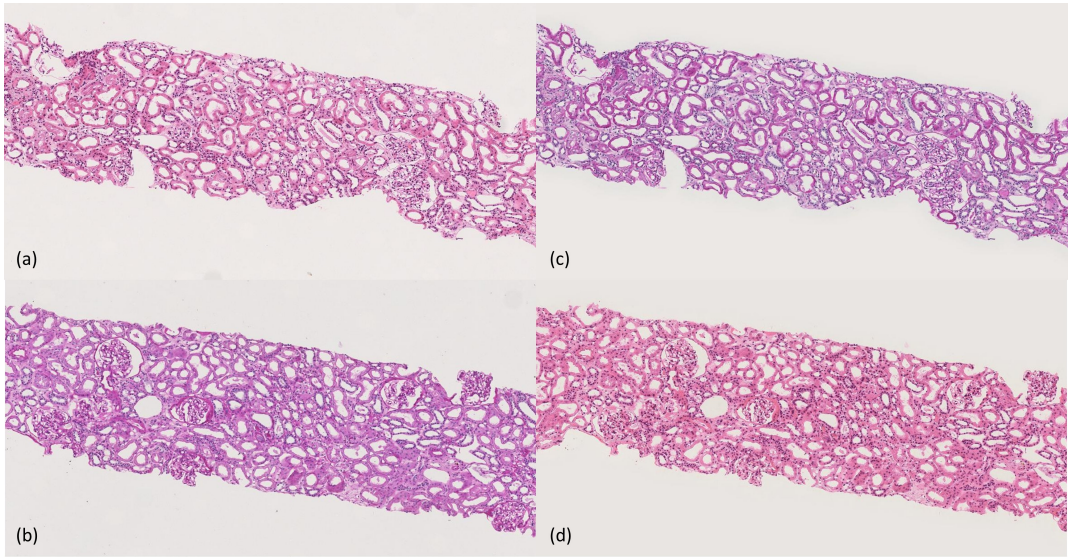


Figure 3.22: Examples of stitched H&E ↔ PAS WSI from a CycleGAN. (a) Real H&E, (b) Real PAS, (c) Fake PAS, (d) Fake H&E. Stitches are generally not visible.

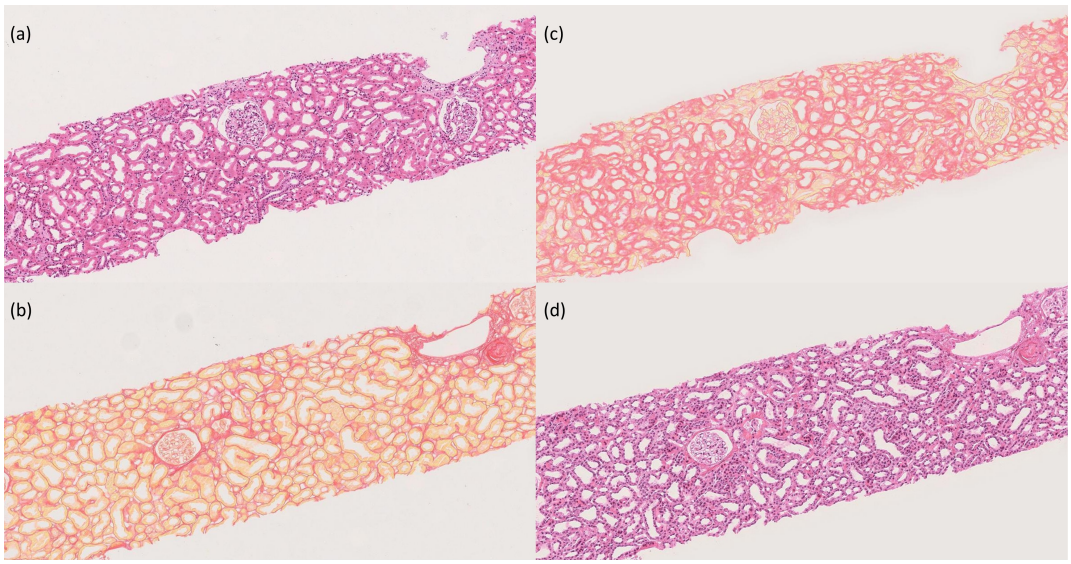


Figure 3.23: Examples of stitched H&E ↔ SR WSI from a CycleGAN. (a) Real H&E, (b) Real SR, (c) Fake SR, (d) Fake H&E. Stitches are generally not visible.

3.5.2 Image Registration and Storage

In WSIs where serial sections were available, images were registered using handpicked landmarks. WSIs were aligned using similarity rigid transformations (translation, rotation, and reflection). Finally, slides were rotated and cropped according to the minimum bounding rectangle of all combined serial sections (Figure 3.24) with padded pixels added to the boundaries. This allowed us to save computational resources as many needle biopsies tend to be long, thin, and smeared diagonally across a slide. Slides without serial sections were also rotated and cropped to reduce the amount of storage required.

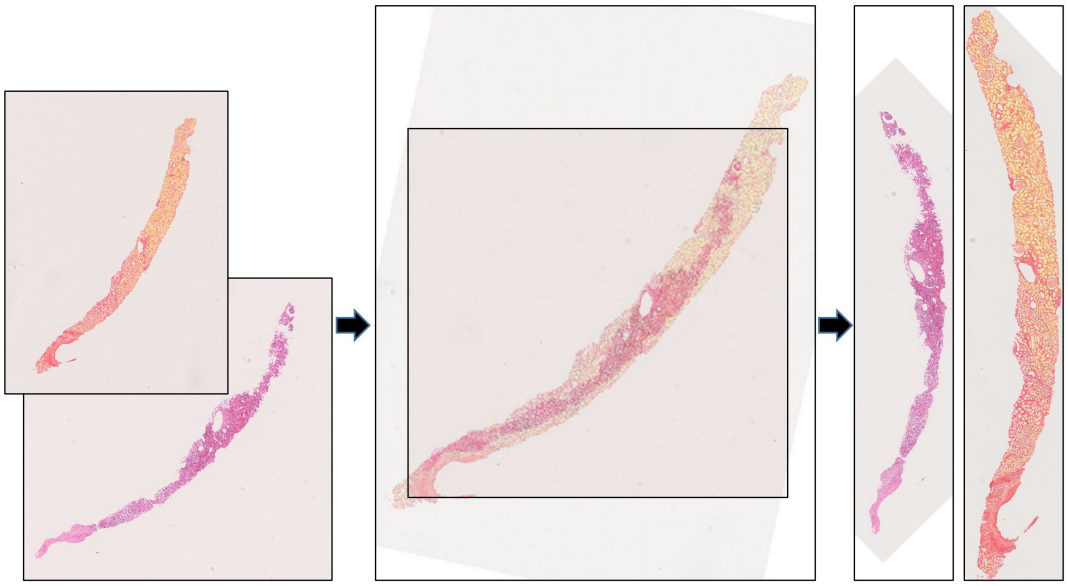


Figure 3.24: Illustration showing the registration and cropping of WSIs

4

Quantitative Analysis Workflow

Contents

4.1 Overview 70
4.2 Tissue Segmentation 72
4.2.1 UNet Implementation 72
4.2.2 Improvement of Segmentation 74
4.2.3 Human-in-the-loop Labelling and Ensemble Model . . . 84
4.3 Extraction of Features from Tissue Compartments . . 86
4.3.1 Handcrafted Features 88
4.3.2 Correlation Between Handcrafted Features and Pathologist Assessment 90
4.4 Extensions of Workflow 97
4.4.1 Clustering and Detection of Anomalous Slides 98
4.4.2 Computational Quantification of Fibrosis in Sirius Red Slides 102
4.4.3 Correlation of Visual Features with Tissue Proteomics Signatures 106
4.5 Limitations 108

The extraction of interpretable histological measurements is an essential first step toward developing a quantitative image-based platform to provide help for renal pathologists and clinical researchers. In this chapter, we describe our main feature extraction workflow. This workflow is the backbone of the subsequent research chapters (Chapters 5-7). Given an input WSI, the workflow segments tissues according to their functional architecture (tubules, glomeruli, vessels, and cell

nuclei) and extracts a number of image-based features from each tissue compartment.

Models in this chapter, unless specified otherwise, were trained primarily on the NMP dataset (see Chapter 3). These NMP slides were available at the early stages of the project, while slides from the QUOD dataset were acquired later. Findings regarding the characteristics from the QUOD dataset and changes to the availability of pathologists have shaped the course of this work. Slides from QUOD tend to be small and contain substantial artefacts caused by forceps. The initial plan for the QUOD slides was to have them assessed by a different pathologist, who preferred to have them stained with PAS as it highlights basement membranes more conspicuously. Thus the primary QUOD slides were stained with a different protocol from the main batch of the NMP slides. Considerable effort had been put into adapting the segmentation workflow to the QUOD-PAS domain.

This chapter can be broken down into four main sections. The overview of the general workflow is described in Section 4.1. Section 4.2 describes the design of our tissue segmentation algorithm used for obtaining individual tissue instances from WSIs. Section 4.3 details the design of handcrafted features inspired by domain knowledge to capture histopathological changes. Finally, Section 4.4 describes several experiments demonstrating potential applications resulting from the immediate histological measurements. We show several applications regarding how these measurements can benefit pathologists and QUOD clinical researchers without the need for further supervisory labels: i) detection of anomalous tissues (4.4.1), ii) quantification of fibrosis in SR slides (4.4.2), iii) correlation with proteomics (4.4.3).

4.1 Overview

An overview of the proposed workflow for extracting tissue features is shown in Figure 4.1. In order to obtain histological measurements from WSIs, we need to segment tissues according to their meaningful functional compartments. Section 4.2 describes the work done to overcome the main challenges in optimising segmentation performance for our datasets using UNets. The next part of the workflow involves extracting visual features from these tissues. We designed a number of domain

knowledge-inspired handcrafted features and explored the use of deep neural network features. Details of feature extraction are described in Section 4.3.

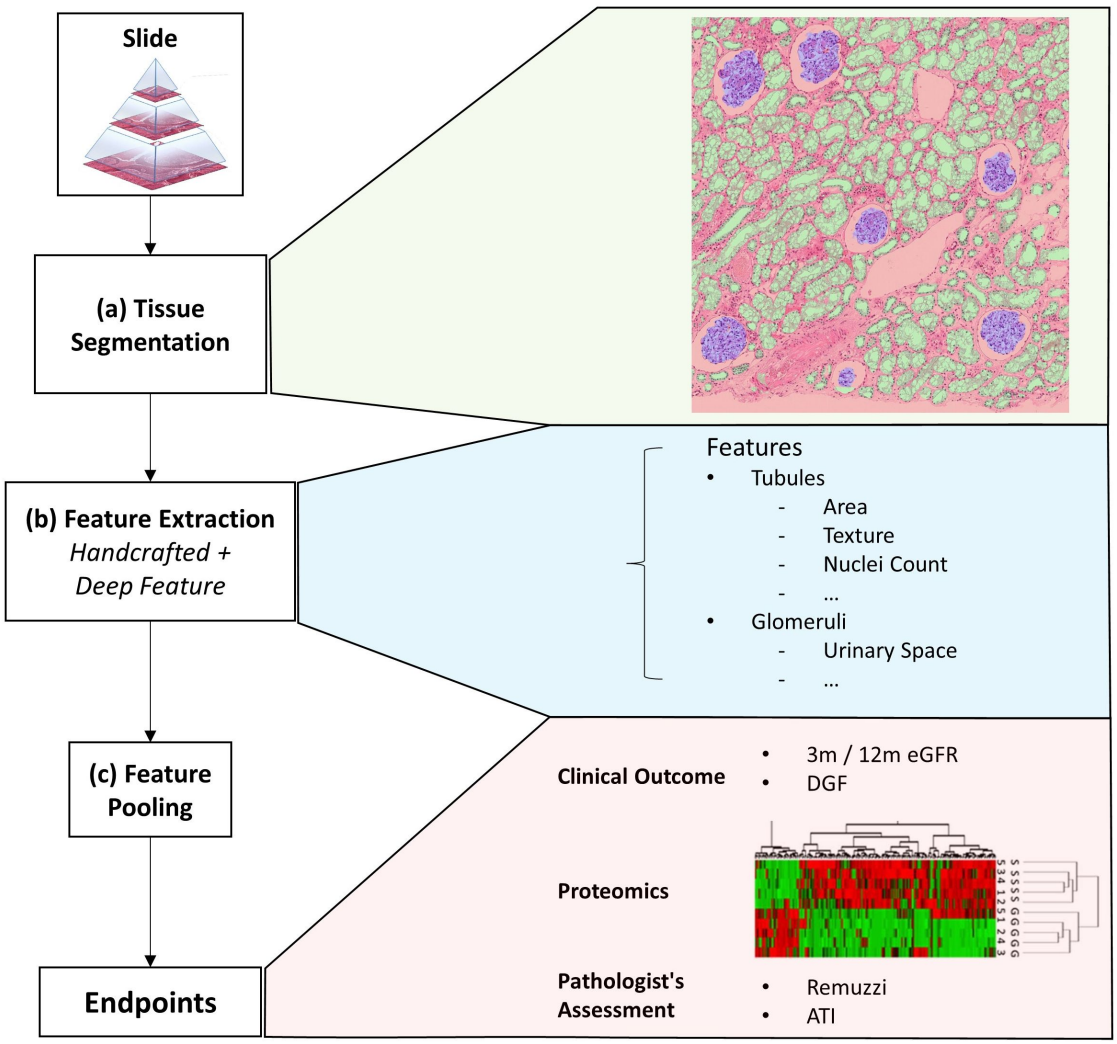


Figure 4.1: Overview of the workflow. This figure shows an overview of the proposed quantitative analysis framework. (a) Tissue segmentation returns the compartments' outline of three different tissue types and cell nuclei. (b) Feature extraction returns a mixture of handcrafted and deep features iterated over each compartment. (c) Finally, features from a variable number of compartments could be pooled together to form a single vectorial description of a slide. The slide label could either be clinical endpoints, assessment results given by pathologists, or other biomarkers. Part (c) of the workflow is explored in more detail in Chapter 6.

4.2 Tissue Segmentation

4.2.1 UNet Implementation

The UNet [87] is the state-of-the-art for semantic segmentation. We base our work on a generic UNet architecture to segment different tissue classes. A standard UNet architecture consists of a symmetrical down-sampling and up-sampling branch with skip connections, as illustrated in Figure 4.2. Filters at the higher levels influence the fine details of the segmentation output as each filter only expands the prior inputs' receptive field by minimal pixels. Vice versa, filters at the lower levels may contribute more to the coarse outline of the final output.

The first step of our experiment was to determine the optimal network architecture. To find the ideal number of filters, we trained UNets with 4:1 train:validation split (based on Donor IDs) on the Fake-PAS NMP (converted from H&E) slides. This consisted of 144 original large tiles used for delineating tissue compartments containing 13103 tubules and 251 glomeruli. Evaluation of segmentation was performed on the QUOD-PAS slides for tubules and glomeruli. This consisted of 38 large tiles with 2145 tubules and 70 glomeruli (More precisely, only the tiles in Table 3.6 with all three tissue classes delineated were used at this project stage). Nuclei segmentation was evaluated on the Fake-PAS NMP tiles (masks discussed in Section 3.3.3 - Nuclei masks were only generated on H&E slides since the intra-observer variability of nuclei masks generated directly on PAS slides was very high). UNets were trained using the architecture illustrated in Table 4.1, outputting four classes: background, nuclei, tubules, and glomeruli. With $m = 16$, the number of filters from Block 1 to 7 are 16, 32, 64, 128, 256, 512, and 1024 respectively, with a symmetric number of filters on the upsampling side. Models were trained with an input size of 400x400 at 0.44mpp with batch size=10, Adam optimiser [131], and an initial learning rate of 10^{-3} decreasing on validation loss plateau with ten epoch patience. Cross entropy losses contributed from each class were weighted inversely proportional to the total area of each class in the training set. Training data was also augmented on-the-fly to include affine transformations, elastic and

piecewise deformations, colour and contrast perturbations, and pixel-wise noise. Note that each of the 144+38 large tiles varies in dimensions. Some of the large tiles could potentially fit in several hundred 400x400 patches. However, saving overlapping patches with extra pixels for cropping rotated/deformed patches would take up too much disk space. Therefore, during training, patches were cropped on-the-fly at a size such that at least 400x400 would remain as valid tissue area after augmentation (e.g. if the rotation is 45 degrees, we would first crop the tile into 565x565, rotate it, and then take a 400x400 crop). The length of one epoch was defined to be the number of 400x400 patches that could fit into all the large tiles. Since the delineated tiles vary in dimensions, tiles were sampled at a frequency proportional to the area of the whole tile within each epoch. Instances were obtained by simply applying a threshold to the final softmax output, clearing small objects (<1600 pixels for tubules and <16,000 pixels for glomeruli and vessels) and detecting connected components. We evaluated segmentation performance using the Intersection-over-Union (IOU) on the semantic segmentation and Normalised Mutual Information (MI) on the instance segmentation.

Table 4.2 illustrates how varying the number of filters from each block influences performance of the UNet. The first three rows show the performance of the proposed UNet on different sets of slides detailed in Tables 3.2 and 3.6. *NMP FakePAS* shows the results on the validation split of the training set. These slides have been algorithmically converted from H&E to PAS (Section 3.5.1); *NMP PAS* shows results on slides obtained from the same biopsies with real PAS stain; *QUOD PAS* is the main QUOD slides to be used for the main experiments in this thesis.

The results from row 4 (Block 1) onward represent an experiment whereby the number of filters was increased from the proposed size. Modest improvement can be seen for nuclei segmentation when Block 1 size is increased (IOU=0.448, MI=0.369), whereas improvement can be seen for tubule (IOU=0.656, MI=0.671) and glomeruli (IOU=0.589, MI=0.569) segmentation when Block 7 size is increased. Increasing the number of filters at Block 3,4 led to decreased performance, possibly resulting from overfitting. However, increasing the number of filters also came at the cost

of higher GPU memory consumption which stifled the potential of implementing other measures. Therefore, we had chosen to implement the proposed architecture for the rest of this chapter unless specified otherwise.

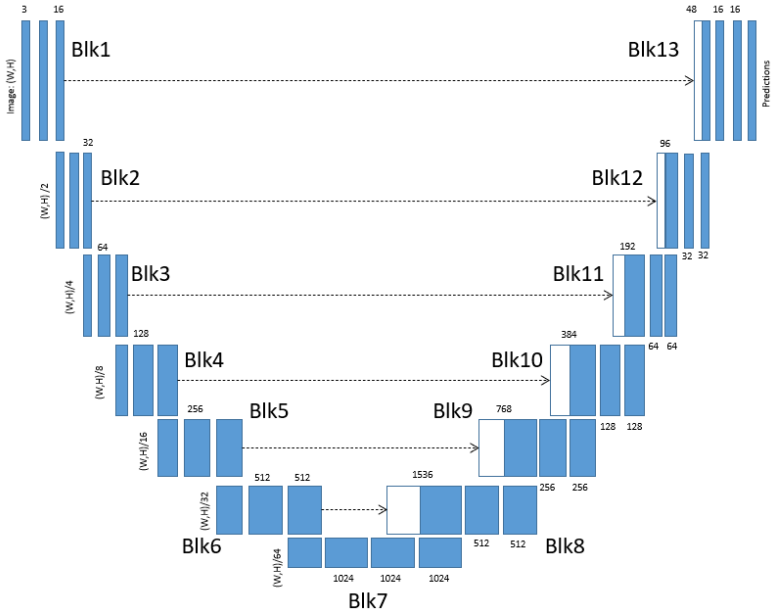


Figure 4.2: Schematic of UNet architecture

4.2.2 Improvement of Segmentation

In Chapter 3, Table 3.7, we delineated different tissues from the same tiles twice. This gave us a grasp of the theoretical segmentation performance "ceiling" of what an ideal UNet should be able to attain. We found a large gap between the vanilla UNet (IOU=0.648 and 0.593 for tubules and glomeruli) and the intra-observer variability (IOU=0.764 and 0.876 for tubules and glomeruli). In particular, owing to the large variation in size and morphology with few training examples, the earlier models fail to converge over the segmentation of vessels. This section describes a number of experiments employing different techniques to improve segmentation performance.

4.2.2.1 Domain Adaptation

Based on Rows 1-3 in Table 4.2, the performance drop appears to be larger between the *NMP PAS* vs *QUOD PAS* slides than between the *NMP FakePAS* vs the *NMP PAS* slides. Therefore, we hypothesised that the performance drop might be due

Table 4.1: UNet architecture The top section of this table denotes the layers that constitute a block. n_1 is the number of channels input for a block and n_2 is the number of out channels.

Up /Down / Center Block (n_1, n_2)	
$3 \cdot 3 \cdot n_1 \cdot n_2$ Conv., Inst. Norm, ReLU, 0.2 (Dropout)	
$3 \cdot 3 \cdot n_2 \cdot n_2$ Conv., Inst. Norm, ReLU, 0.2 (Dropout)	
2×2 MaxPool / Bilinear Interpolation / -	
UNet Architecture	
Block 1	Center $(3, m)$
Block 2	Down $(m, 2m)$
Block 3	Down $(2m, 4m)$
Block 4	Down $(4m, 8m)$
Block 5	Down $(8m, 16m)$
Block 6	Down $(16, 32m)$
Block 7	Down $(32m, 64m)$, Center $(64m, 64m)$
Block 8	Up $(96m, 32m)$
Block 9	Up $(48m, 16m)$
Block 10	Up $(24m, 8m)$
Block 11	Up $(12m, 4m)$
Block 12	Up $(6m, 2m)$
Block 13	Up $(3m, m)$
Block 14	Center $(m, OutputClasses)$

Table 4.2: Filter size vs UNet performance. The first three rows of results show the performance of the UNet on different sets of slides. Nuclei masks were only generated on H&E slides since the intra-observer variability of nuclei masks generated directly on PAS slides was very high (Section 3.3.3)

		Filter Size		IOU			MI		
Dataset	UNet Block	Original	New	Nuc	Tub	Glom	Nuc	Tub	Glom
NMP FakePAS	Proposed	-	-	0.422	0.725	0.678	0.345	0.724	0.731
NMP PAS		-	-	-	0.688	0.506	-	0.702	0.511
QUOD PAS		-	-	-	0.648	0.593	-	0.666	0.570
QUOD PAS	Block 1	16	256	0.448	0.643	0.575	0.369	0.669	0.565
	Block 2	32	512	0.447	0.640	0.584	0.350	0.661	0.557
	Block 3	64	1024	0.382	0.624	0.569	0.302	0.643	0.550
	Block 4	128	1024	0.418	0.610	0.530	0.333	0.633	0.509
	Block 5	256	1024	0.420	0.647	0.578	0.333	0.673	0.562
	Block 6	512	1024	0.428	0.649	0.571	0.350	0.674	0.559
	Block 7	1024	1448	0.432	0.656	0.589	0.346	0.671	0.569

to differences in tissue morphology between datasets rather than simply due to the converted stain colours not being representative of PAS. Here we explored two different techniques of domain adaptation in addition to CycleGAN, which was already being implemented: Domain Adversarial Neural Networks (DANN) [132] and Virtual Adversarial Training (VAT) [133]. Both of these techniques have been shown to be promising in improving the generalisability of neural networks in the target domain without the need to curate additional supervision labels.

DANN encourages learning domain invariant features by adding a domain classifier to the main neural network. This domain classifier is only implemented during training. It takes the intermediate layers of the main network as input and aims to learn differences in features between the training and target dataset. A *reverse gradient layer* is added between these features and the domain classifier. As such, back-propagation would minimise the inter-domain feature differences learned by the main network.

The original implementation of DANN [132] was designed for image classification tasks. It feeds features after the last max-pooling layer to the domain classifier. Kamnitsas et al. [134] first demonstrated the use of DANN on biomedical images with an experiment investigating how using layers of different depths affects performance. Degel et al. [135] adopted the multi-connected configuration to a UNet. DANN has also been applied to histopathology images to adapt models to samples prepared or digitalised under varying conditions [136]. We applied a UNet where features are regularised in the first half of the UNet in a configuration as shown in Figure 4.3. Features at the end of each down block were extracted, max-pooled, concatenated, and fed to the domain classifier. The aim of this architecture is to allow UNets to unlearn domain-specific features, preserving only generalised features present in the target dataset.

Likewise, VAT is a technique appended only during the training of neural networks. In many neural networks trained using limited data, the curse of dimensionality means there is little guarantee for the output distribution to be smooth with respect to the input. This could be particularly problematic in the

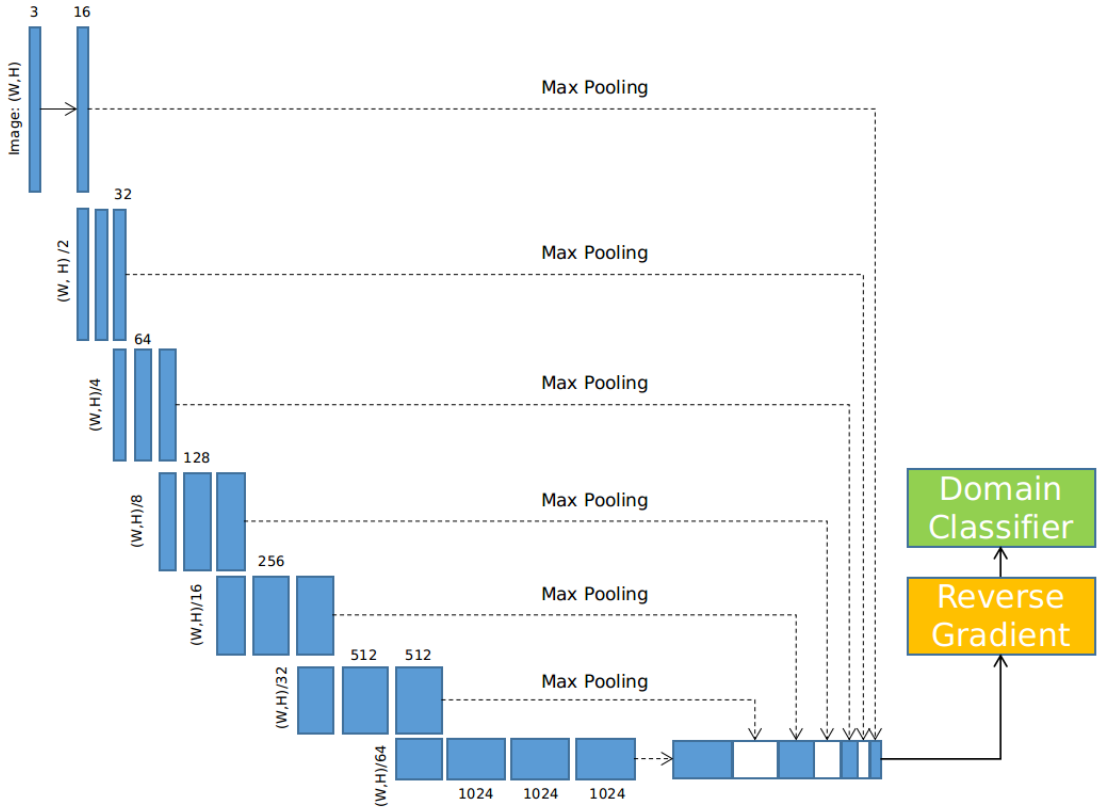


Figure 4.3: Domain adversarial classifier in UNet. Only the left half of the UNet is shown. Output from each downward block is pooled and fed to the domain classifier.

presence of domain shift where the input does not resemble any of the training examples. VAT applies adversarial training to promote local smoothness of the output distribution by determining the adversarial direction from the input during the training of neural networks. It ensures that small changes to the input in this direction do not lead to significant changes in predictions by enforcing a Kullback–Leibler loss term between the original and the perturbed output. In our experiment, VAT loss was applied to both the training and target dataset with augmentation and hyperparameter $\epsilon = 2.5$, $\lambda = 0.001$ (Variables defined in Miyato et al. [133]). We find that a large VAT loss term tends to produce segmentation results favourable to predicting the majority class, so a small value of λ was chosen to balance the contribution of VAT loss against the performance of rarer classes. Since no supervisory label is needed, the application of VAT to all data points would ensure smooth input-output mapping of both domains. We use a network as specified

by [132] trained with negative log likelihood loss for the domain classifier itself.

In contrast to Section 4.2.1, several changes have been made to the training of the UNet to improve performance: 1) the learning rate scheduler was changed to a cyclic learning rate [137] with an initial maximum learning rate of 0.1; 2) focal loss [138] was used instead of cross entropy loss. Table 4.3 shows the results of our mini experiment. 3) A larger tile size of 512x512 pixels was used. 4) A "boundary" class was added to the segmentation of tissues to help separate touching instances. This was achieved by generating a separate set of masks using Sobel or Canny filters from the original masks. We generate boundaries for all foreground classes (nuclei, tubules, glomeruli, and vessels). The UNet predicts 7 classes (background, nuclei boundaries, tissue boundaries, nuclei, tubules, glomeruli, and vessels). The output layer of this UNet consists of two softmax units. The first softmax unit takes logits from 3 classes (background, nuclei boundaries, and nuclei). The second softmax unit takes logits from 5 classes (background, tissue boundaries, tubules, glomeruli, vessels). The combined focal loss from the two branches was back-propagated during training.

In the segmentation masks used for training, the "boundary class" consisted of only thin lines. However, we find that "boundary class" output by the UNet tends to be thicker lines of lower values as the UNet's performance was limited by boundary ambiguities from the training examples. If we simply predict segmentation classes using argmax, we would end up with fragmented boundaries. Hence, we obtained the foreground and boundaries by applying a threshold to convert semantic predictions to instance segmentation. Each connected component of each foreground class was treated as an instance. We iterated through each instance multiple times and dilated its mask until the instance touched another instance or exceeded the mask of the boundary class. The dilation does not extend into pixels of the background class. Instances smaller than 400 pixels were discarded as we could not recognise tissues smaller than this size.

We compare performance from the default UNet against the same architecture trained with DANN and VAT, with the best performing cases highlighted in bold. Based on the MI values from instance segmentation, DANN's performance was

slightly higher than VAT. Higher MI was obtained for 3 out of 4 classes in DANN compared to VAT. Nevertheless, the difference was not so clear-cut when measured in terms of IOU. It is also important to note an opportunity cost trade-off: columns BS5 and BS10 show the amount of GPU memory used during training for batch sizes 5 and 10 when implementing the different measures. While VAT increases the memory footprint by a little, DANN consumes much more memory as it takes features from all levels of the UNet. It may also be possible for future work to combine both DANN with VAT for better segmentation performance as they are mutually compatible. However, we had limited GPU memory, so we had to choose between the two. After considering the trade-offs with GPU memory, we have decided to apply only VAT in our subsequent experiments.

Table 4.3: Comparison of domain adaptation methods implemented in the training of our UNet. GPU memory cost (gigabyte) for different batch sizes (BS) is also shown in the table. The IOU/MI results were generated with a batch size of 10.

	GPU RAM		IOU				MI			
	BS5	BS10	Nuc	Tub	Glom	Ves	Nuc	Tub	Glom	Ves
Default	2.6	4.4	.583	.662	.763	.259	.335	.301	.471	.550
DANN	5.4	10	**.623**	.680	**.779**	.294	**.369**	.303	**.531**	**.597**
VAT	2.6	4.5	.588	**.697**	.770	**.303**	.354	**.310**	.485	.592

4.2.2.2 Obtaining Instances using Graph Cuts

To obtain instances of different tissue compartments from soft semantic segmentation predictions, we needed to implement several extra image processing steps. Previously, instances were obtained by applying a threshold to the final softmax output, clearing small objects and detecting connected components. However, this simplistic approach has led to a poor reconstruction of instances as the threshold method does not account for predicted values of neighbouring pixels.

Graph algorithms have been successfully applied to a number of image processing problems and can be used to divide an image into background and foreground while accounting for adjacent pixels. Boykov et al. [139] described how to define graphs based on greyscale images. Pixels are defined as vertices, and there are two

types of edges in this graph. The first edge type connects the neighbouring pixel vertices to a given pixel. The second edge type connects a "source" or "sink" to each pixel vertices. In our case, we define the source and sink as fictitious pixels with values 0 and 1. The max-flow-min-cut (MFMC) algorithm would serve to minimise the following energy function:

$$E(\mathbf{x}) = \sum_i D_i(x_i) + \sum_{(i,j)\in C} K|x_i - x_j| \tag{4.1}$$

Where $x_i \in \{0,1\}^N$ is the value of the restored class predictions in an image with N pixels and pixel index i. $D_i(x_i)$ is the penalty term for assigning the value $1 - x_i$ to the ith pixel. Each D_i depends on the softmax values of the UNet predictions p_i. In our case, $D_i = p_i$ if $x_i = 0$ and $D_i = 1 - p_i$ if $x_i = 1$. K is the regularisation term that preserves spatial smoothness in the resulting cut.

Ultimately, the optimal smoothness factor K depends on the size of the tissue compartment. We obtained optimal values of K for different tissue classes, as shown in Table 4.4. Here, a UNet was retrained to predict only 5 classes: background, boundaries, tubules, glomeruli, and vessels. We decided to use a UNet that predicts only the larger tissue classes as the addition of nuclei class in the UNet negatively impacted the performance of other classes. Segmentation of nuclei was hardly affected by MFMC as instances rarely touch each other. Other settings of this UNet remained the same as that of the VAT model in Table 4.3.

We found that tissue compartments obtained from MFMC tend to be smoother, less likely to be an artefact, and instances were less likely to be merged. Figure 4.4 shows an example comparison between tubule instances obtained using (b) threshold vs (c) MFMC. Slide (a) is a slightly challenging example where tubules are compressed (the magnified crop). We can see from this region that some instances obtained by threshold are wrongly merged while instances obtained by MFMC maintain a reasonable compact tubule shape. The slide also contains bits of folded tissue to the right, which do not contain viable tubules. The UNet failed to output a high value for this region for the background class as not many of our training examples contain folded tissues. This led to patches of folded tissue

being detected as tubules in (b) but not (c). It is worth noting that because the test images contain only delineation of tissue compartments that the annotating person can recognise, there may be some selection bias in the test labels. As a result, some of these differences are only available as qualitative examples and were not always reflected by changes in IOU/MI.

Table 4.4: Results from MFMC. For each segmentation class, the setting leading to the highest MI is in bold. We find that MFMC generally delivers better performance

	IOU			MI		
Smoothness K	Tub	Glom	Ves	Tub	Glom	Ves
0	0.643	0.708	0.245	0.528	0.687	0.550
0.02	**0.645**	0.711	0.207	**0.528**	0.698	0.500
0.05	0.644	0.711	0.232	0.526	0.702	0.558
0.1	0.644	0.707	0.236	0.521	0.695	0.518
0.2	0.642	0.714	0.241	0.514	0.689	0.541
0.4	0.646	0.723	0.237	0.502	0.701	0.539
0.8	0.650	0.720	0.252	0.434	0.706	0.585
1.6	0.653	**0.711**	0.243	0.363	**0.722**	0.604
3.2	0.646	0.715	0.254	0.304	0.712	0.624
6.4	0.623	0.704	**0.251**	0.252	0.718	**0.685**

4.2.2.3 Nuclei Segmentation

Nuclei segmentation is a task that has attracted much literature due to its broad applicability. It is most commonly performed on microscopy images or histological slides stained in H&E [87, 140]. In our context, segmentation needed to be performed in the PAS domain. Spatial and colour distribution derived from epithelial cell nuclei in tubules may reveal the physiological or clinical state of kidneys. We expect the optimal models to differ slightly from prior literature due to exact acquisition parameters and greater nuclei colour variation under PAS stain. Here, we briefly compare the different options of nuclei segmentation.

As cell nuclei are significantly smaller than other tissue compartments, the performance of UNets trained only using nuclei (and nuclei boundaries) under different settings was evaluated. Table 4.5 shows the segmentation performance of cell nuclei trained and tested under different stains (Other settings remained the

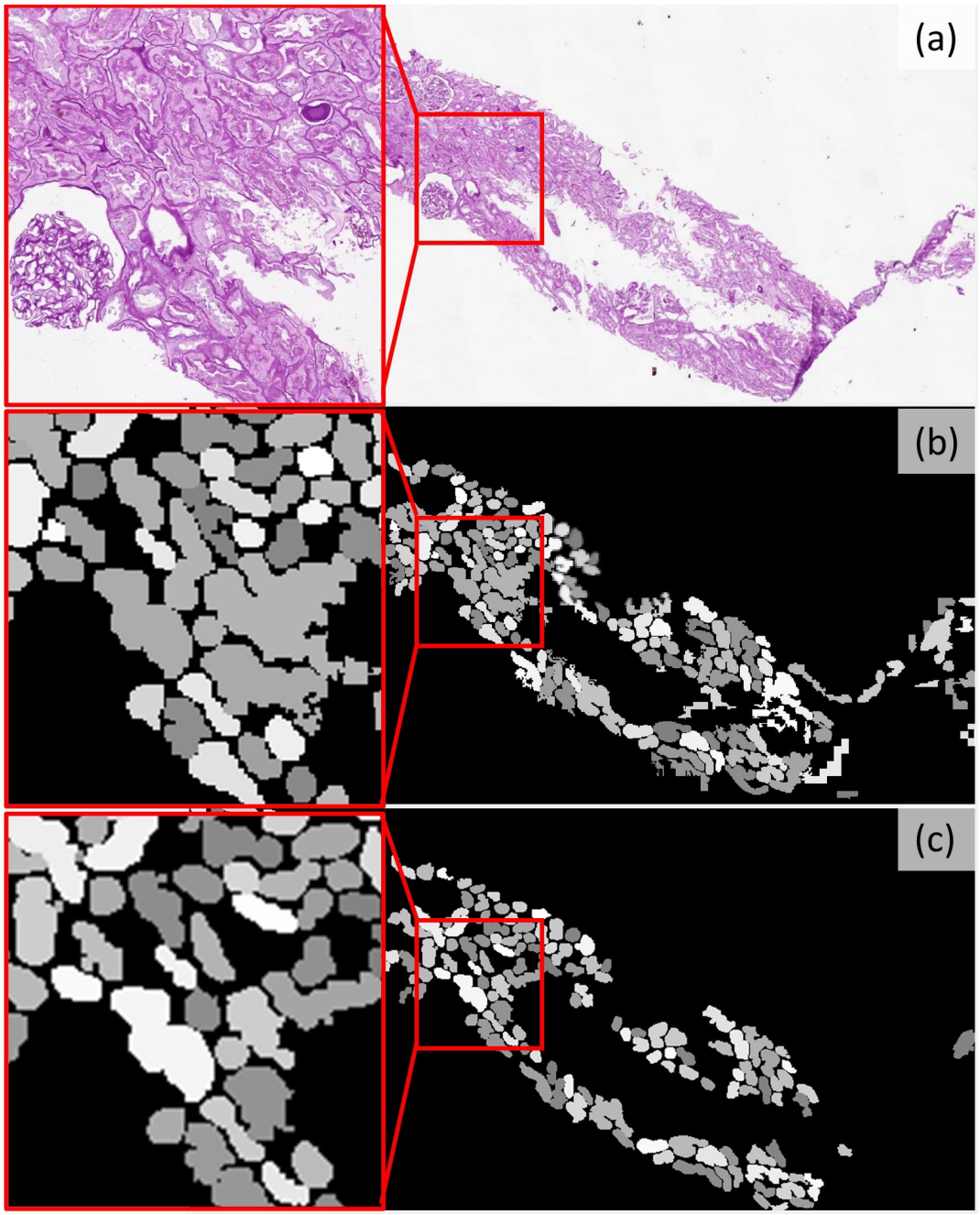


Figure 4.4: Visual comparison of instance segmentation. a) PAS stained slide; b) Tubule instances obtained by applying a simple threshold; c) Tubule instances obtained by applying max-flow-min-cut.

same as the VAT model in Table 4.3). It can be seen that performance is notably higher when networks were trained on images of 0.22mpp resolution (rows 1 and 6). The performance also drops when UNets were trained/validated on FakePAS images for two reasons: 1) CycleGAN might have failed to convert some nuclei from H&E to PAS; 2) Nuclei stained in PAS exhibit a continuous colour variation, whereas nuclei stained in H&E tend to have much less variation and appear darker. Despite having very few touching instances, we also see segmentation performance improves with the inclusion of boundary class (rows 4 and 5) compared to segmentation without boundary class (rows 2 and 3).

Nuclei segmentation on different network architectures had also been tested. For example, Mask-RCNNs [83] can be used to obtain instances of individual nuclei directly without the need for post-processing semantic masks. The architecture gave reasonable performance (Table 4.5, row 6) with minimal tuning. However, we found that Mask-RCNN did not perform well when input images were scaled to 0.44mpp. Besides this, we also found that our Mask-RCNN did not converge when trained to segment tubules. We also tried implementing an adversarial network [89] to the UNet and found that while this leads to more realistic, sharper segmentation masks, there was no significant improvement to IOUs.

We decided to adhere to the settings of a 3-class UNet for images at 0.44mpp (row 5) for three reasons: 1) In later stages of this thesis, we have slides scanned at lower resolution as the author could not access the original high-resolution scanner due to the covid pandemic; 2) We wanted to keep the workflow homogeneous and use a similar network architecture for segmenting different tissue compartments; 3) We argue that the pursuit of maximising IOU/MI is likely to produce only marginal returns for improving what this thesis ultimately aims to predict. IOU/MI tends to be lower in PAS slides because some paler nuclei are hard to detect. In the next part of our workflow (Section 4.3), we describe a number of handcrafted histological measurements to detect nuclei colour and count the number of nuclei per tissue compartment. A model that fails to segment pale nuclei will result

in lower nuclei count. In this case, histological changes will be captured either way by one feature or another.

Table 4.5: Segmentation of cell nuclei under different stains and magnifications

	Stain	Model	Resolution (mpp)	IOU	MI
1	H&E	UNet (2 Classes)	0.22	0.862	0.841
2	H&E	UNet (2 Classes)	0.44	0.526	0.516
3	FakePAS	UNet (2 Classes)	0.44	0.431	0.311
4	H&E	UNet (3 Classes)	0.44	0.704	0.677
5	FakePAS	UNet (3 Classes)	0.44	0.514	0.355
6	H&E	Mask-RCNN (2 Classes)	0.22	0.831	0.813

4.2.3 Human-in-the-loop Labelling and Ensemble Model

By inspecting the segmentation results in the QUOD dataset, we found that performance, as benchmarked by IOU/MI, did not always translate to better segmentation quality over the whole slide. The reason was that tiles delineated by hand were chosen manually. There is an inherent human bias towards picking areas that are more legible. To reduce this bias, we implemented a human-in-the-loop process to identify and hand-delineate new regions where the algorithm performs worst: 34 slides were selected (9 slides from NMP-HE and 25 slides from QUOD-PAS) as the most difficult cases by visual inspection of the segmentation output generated by UNets trained at various stages during the development of the segmentation workflow. We delineated the glomeruli and vascular structures of these 34 whole slides by hand using photo editing software on the predicted soft segmentation maps. Tubule classes were not re-delineated because the performance was already comparable to human intra-observer variability. By choosing to delineate whole worse-performing slides instead of choosing to delineate specific regions within the slides, the person is 'forced' to make a decision on regions that look ambiguous. Evaluating models on these examples exposes the true difficulty of the segmentation task.

Table 4.6 shows the total number of tissue compartments delineated through this human-in-the-loop process. We are primarily delineating new data on whole slides here, but there are small regions that inevitably overlap with those given earlier in Chapter 3, Table 3.6. Many of the QUOD slides chosen here contain large portions of tissues with artefacts caused by forceps handling. Many of these artefacts had been falsely detected as glomeruli or vessels by the UNets trained solely on the original labels. The newly delineated slides also contain areas with artefacts such as loose objects - some of these "blank" regions have been added as part of the new delineation set.

Table 4.6: Additional tissues delineated through a human-in-the-loop process

	Instances		Area (mm^2)	
	Glomeruli	Vessels	Glomeruli	Vessels
QUOD→PAS	342	239	3.56	1.89
NMP→HE/FakePAS	380	89	5.20	1.54

To see how the inclusion of these 'hard examples' affects performance, we retrained the earlier UNets+VAT setup using the original ground truths under the same settings as Section 4.2.2.2. Except that, in this section, the original labels were divided to train an ensemble of five models using 5-fold cross-validation. The numerical output of these models was averaged before further processing using graph cuts. The ensemble model's output was evaluated on the original (Table 3.6) and new (Table 4.6) QUOD-PAS tiles combined. We see a dramatic drop in IOU and MI when the models were evaluated on these combined labels (first row in Table 4.7) compared to earlier results where the evaluation was done on only the original labels (i.e. the maximum values of each class in Table 4.4). We found that the ensemble had failed to segment local regions with artefacts and large vessels. Many tubules were falsely detected as glomeruli or vessels. The problem remained regardless of the number of ensembles used and aggressive data augmentation as artefacts were detected at the same locations. These results reveal a more realistic perspective of the challenges in trying to segment these WSIs.

To improve segmentation performance on these hard examples, we trained additional models by combining the original training data with the 9 newly delineated NMP slides. The presence of tissue compartments at varying scales motivated us to extend the ensemble to multiple magnifications. To show the performance achievable with a multi-magnification ensemble, we trained an ensemble of 11 models: 2 models to segment tubules, glomeruli, and vessels at 0.44mpp; 3 models to segment tubules and glomeruli at 0.44mpp; 3 models to segment glomeruli and vessels at 0.88mpp; 3 models to segment glomeruli and vessels at 1.76mpp. Models that include prediction of tubule class were only trained on the original labels; while models without tubule class were trained using both the original labels and the new labels from the NMP-FakePAS slides. This ensemble was evaluated on the old and new ground truths combined and the results are shown in the bottom row of Table 4.7. We found that using multi-scale input helps to suppress artefacts. Segmentation performance of vessels and glomeruli is improved due to larger receptive fields and the cancelling of artefacts detected with models at different magnifications. Thus we propose to use a multi-scale ensemble of models for the later chapters of this thesis (Chapters 6 and 7).

We also compare the number of glomeruli counted by our algorithm versus a pathologist on the QUOD-PAS slides in Figure 4.5. We find that the mean L1 difference between the two is 1.02. This is a reasonable value considering some glomeruli are detached from the main biopsy, and some may be too small due to the plane it was sliced. Note that some slides contain two sections of the same biopsy. In these cases, the pathologist counted glomeruli from only one random section, but we have not recorded which section it was. The algorithm counted both sections, and the count was manually divided by two to produce this plot. This may also have contributed to counting differences between the pathologist and the algorithm.

4.3 Extraction of Features from Tissue Compartments

After obtaining individual tissue compartments, we extract visual features from the cropped regions. Ideally, we want to find visual features that are representative

Table 4.7: Segmentation performance including "hard examples". "Ensemble (x5) Single Scale": five models trained on the original (NMP+TCGA+Native) labels at 0.44mpp and evaluated on the original and new QUOD-PAS labels combined. "Ensemble (x11) Multi-Scale": eleven models trained on multiple scales and evaluated on the original and new QUOD-PAS labels combined. Five out of these eleven models were trained on only the original (NMP+TCGA+Native) labels as the new labels do not contain tubules. The six other models trained on lower magnifications were trained on both the original and new training labels combined. A dramatic drop in performance can be seen in the single-scale ensemble when compared to earlier results (e.g. maximum value in each column in Table 4.4) evaluated on only the original labels.

	IOU			MI		
Models	Tub	Glom	Ves	Tub	Glom	Ves
Ensemble (x5) Single Scale	.456	.335	7.66e-3	.230	.400	8.14e-2
Ensemble (x11) Multi-Scale	.490	.625	.588	.290	.610	.510

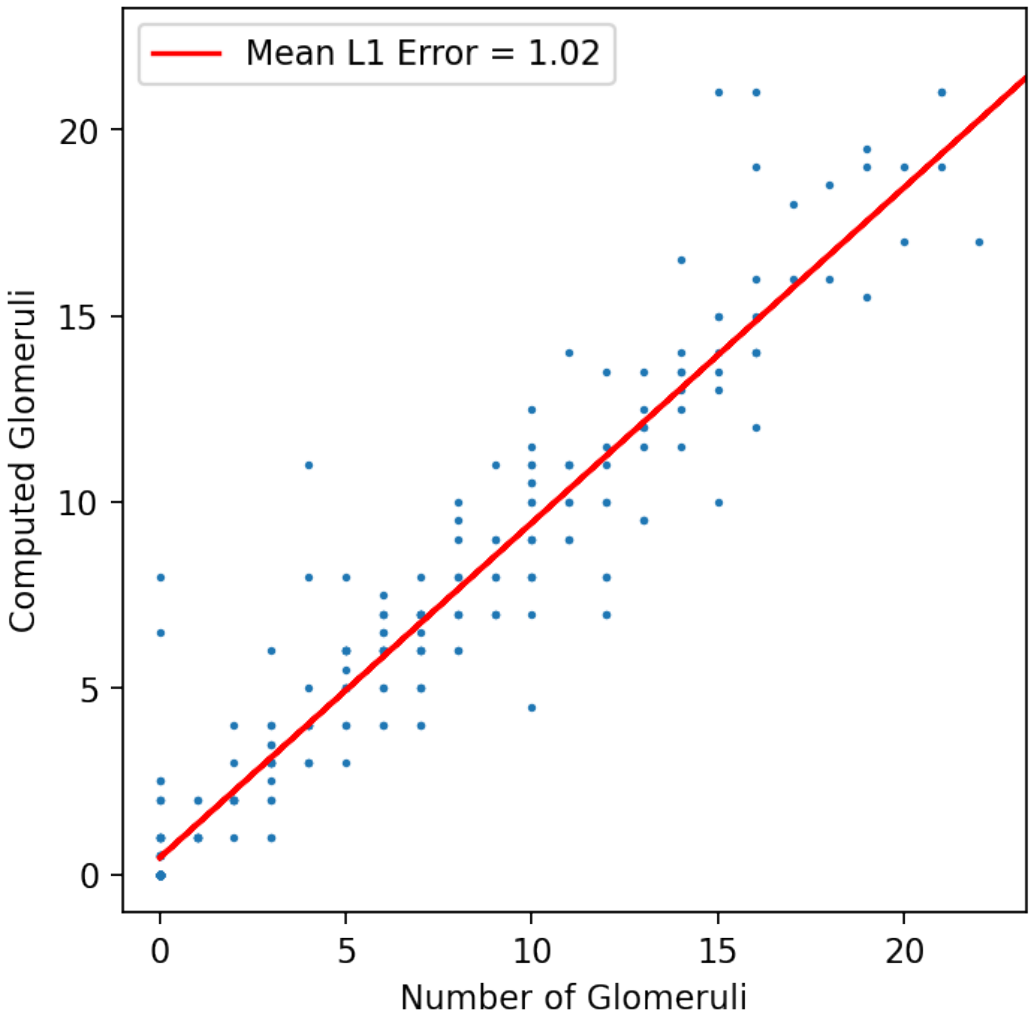


Figure 4.5: Comparison of glomeruli counted by pathologist vs algorithm.

of various physiological changes in the kidney. There are two main types of visual features - handcrafted features and deep neural network features. While deep features have gained popularity in recent years due to their self-forming properties and superior classification performance on large everyday-image datasets, handcrafted features are generally more easily interpretable by clinicians. They may be valuable in medical datasets that typically have orders of magnitudes fewer images. In this section, we discuss the design of domain-inspired handcrafted features for the assessment of kidney histology.

4.3.1 Handcrafted Features

There is extensive evidence that a number of histological features can be used for assessing the health of kidney tissues [4, 31, 63, 141]. Based on works of literature and experience gained through discussion with our collaborating pathologist, we design a number of domain-knowledge inspired handcrafted features. These features are extracted from the tissue compartments segmented in Section 4.2.

Broadly speaking, three main types of handcrafted features have been explored in this thesis, as illustrated in Figure 4.6:

a) Textural/colour descriptors are designed to capture changes in spatial and colour distribution of histological stains from the original image. We give two examples of such handcrafted features. The first example measures the variance of blue pixel values of the entire tissue compartment - we name this b^{σ}. Some illustrations of tubules with varying values of b^{σ} are shown in Figure 4.7.

The second example represents a calculation of the spatially-weighted variance of the epithelial cells' pixel values when evaluated on tubules - we name this y^{σ^2}, and it is defined as follows:

$$y^{\sigma^2} = \frac{\sum_{m \in M}(d_m x_m)^2 - (\sum_{m \in M} d_m x_m)^2}{\sum_{m \in M} d_m \sum_{m \in M} x_m} \ , \tag{4.2}$$

where d_m is the distance transform of an instance's mask measured from the edge; x_m is the pixel value of the image bounded by mask M of the instance. Some illustrations of tubules with varying values of y^{σ^2} are shown in Figure 4.8.

b) Morphological descriptors operate on the tissue compartments' segmentation masks. We give two examples of such handcrafted features. The first morphological feature is the urinary space area in glomeruli. This is obtained through classical image operations by finding white space surrounding each glomerulus. Some illustrations of glomeruli with varying values of urinary space area ratio are shown in Figure 4.9. The second morphological feature is the lumen area ratio of vessels. Some illustrations are algorithmically selected from the QUOD dataset and are shown in Figure 4.10.

c) Nuclei descriptors are features of cell nuclei nested within larger compartments segmented by the workflow. Each parent compartment (tubules, glomeruli, and vessels) may contain a varying number of nuclei. This many-to-one relationship between cell nuclei and tissue compartments makes nuclei features a form of hierarchical feature. Nuclei features include colour, area, and spatial distribution with respect to the parent tissue compartment. A more detailed description of these features is shown in the Appendices, Table A.2, features 8-29. One way to describe spatial distribution is to calculate how close they are to the boundary of the compartment using the distance transform D of the mask of image I. We call one of the features *Nuclei moments* as the calculation resembles how we calculate 'moments of inertia' in physics and is defined as $\sum_{i \in N} d_i/n$. Where N is the set of nuclei, d are values of the distance transform D at the centre of each nuclei, and n is the number of nuclei within the tissue compartment segmentation mask. An illustration of these features is shown in Figure 4.6. Note that nuclei outside the tissue compartment are not counted.

We only get a single value per tissue compartment for the example feature above. However, we get a variable number of nuclei per compartment for other nuclei features, such as those describing nuclei colour. To get a fixed length descriptor per compartment, these are represented by quantile values. Compartments with no nuclei detected are filled with nans.

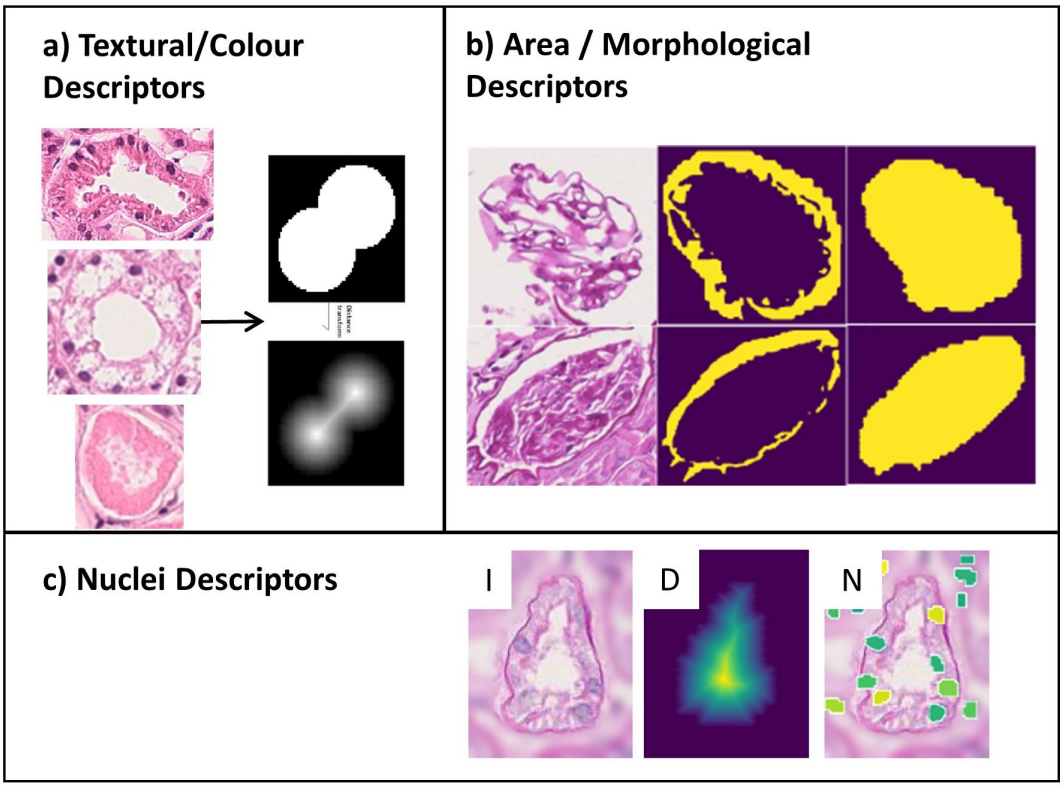


Figure 4.6: Illustration showing an example for each of the three types of handcrafted features (a) Texture/Colour descriptors are used for capturing global and spatial variation of colours within the tissue compartment. One example feature describes how stain colour is distributed with respect to the tissue compartment. This can be derived from the distance transform-weighted pixel values within the segmentation mask. (b) Morphological descriptors are used for capturing the area or shape of the compartment. An example would be the urinary space area of the glomeruli. (c) Nuclei descriptors capture the colour and distribution of nuclei within each compartment. An example is to take all the nuclei N from compartment I and calculate their mean distance transform D value. For this example feature, tubules with compact epithelial cells would result in a low value.

4.3.2 Correlation Between Handcrafted Features and Pathologist Assessment

The example handcrafted features mentioned above (b^{σ}, y^{σ^2}, vessel lumen area, glomeruli urinary space area, tubule nuclei moment) are all inspired by known histological changes in diseased kidneys. To see how well they correlate with pathologists' assessments, we find the median values of each feature across the whole slide from the relevant segmentation class. These median values are converted into a rank across the whole QUOD dataset and compared with the ordinal assessment

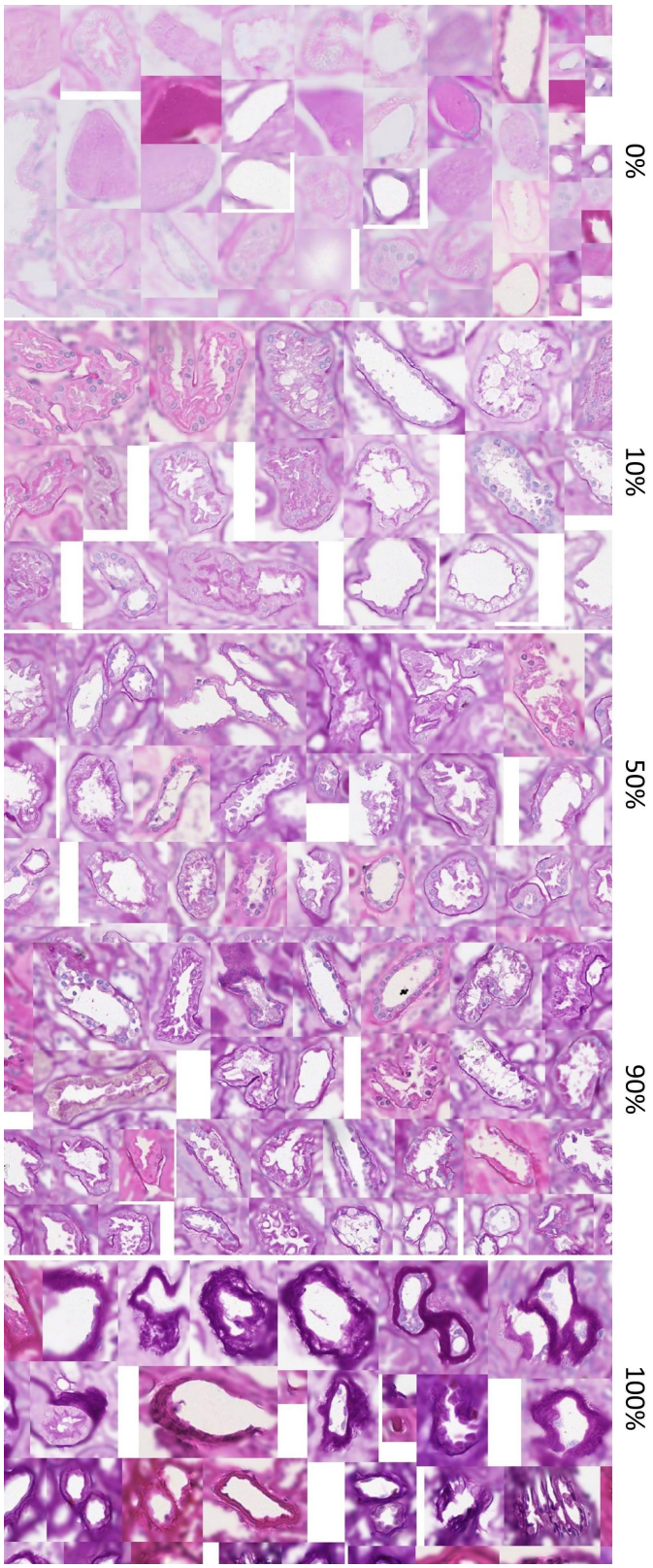


Figure 4.7: Tissue stain colour variation in the blue channel (b^{σ}). Numbers next to each block denote the quantile from which the example tissue compartments are algorithmically selected from the QUOD dataset.

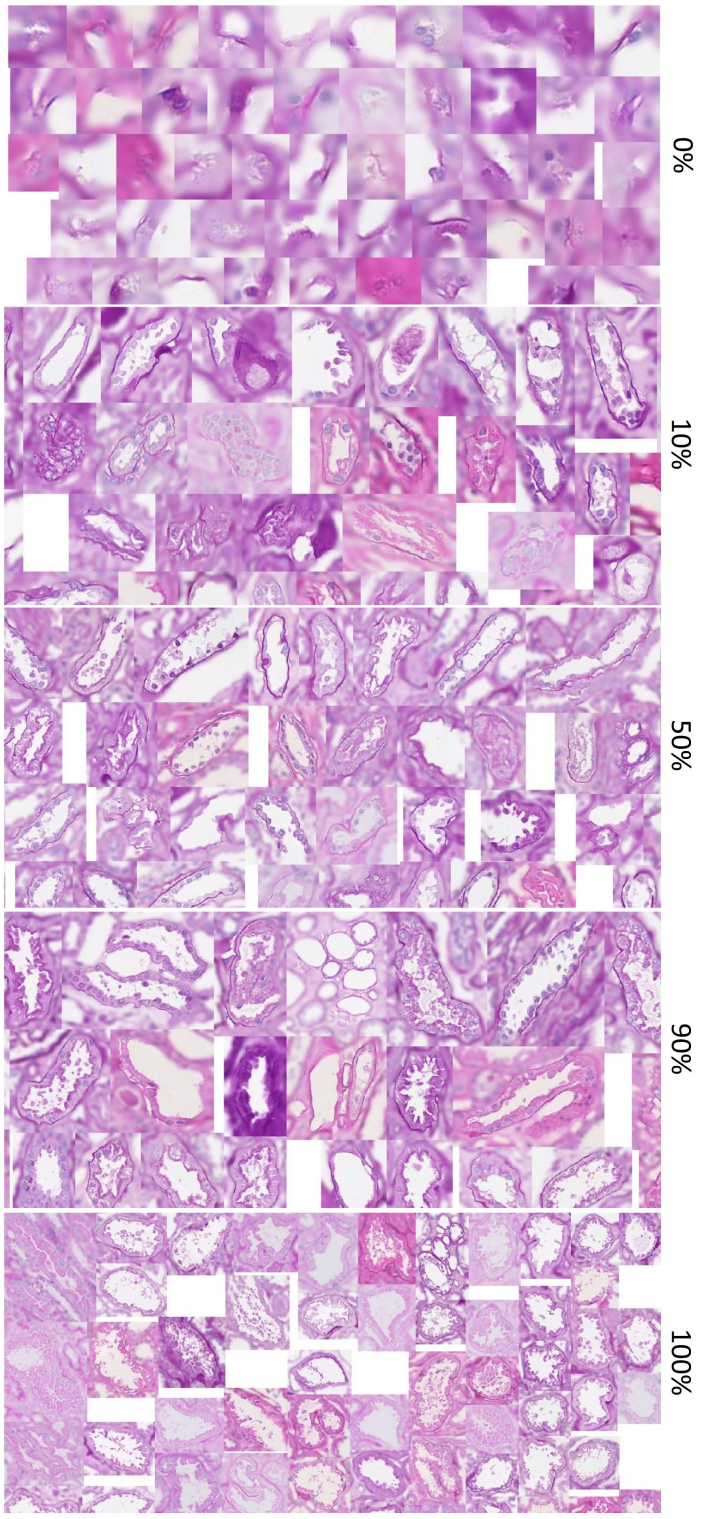


Figure 4.8: Spatially-weighted variance of the epithelial cells' stain (y^{σ^2}). Numbers next to each block denote the quantile from which the example tissue compartments are algorithmically selected from the QUOD dataset.

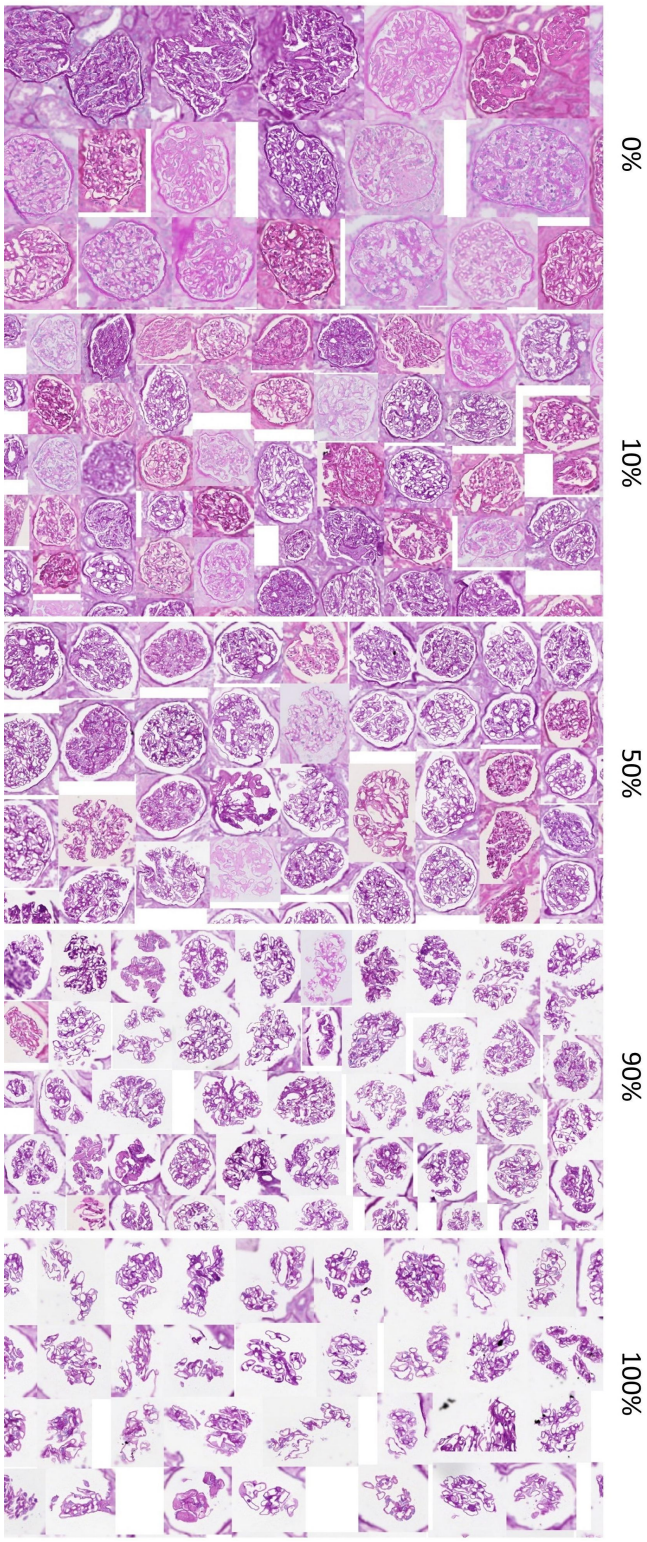
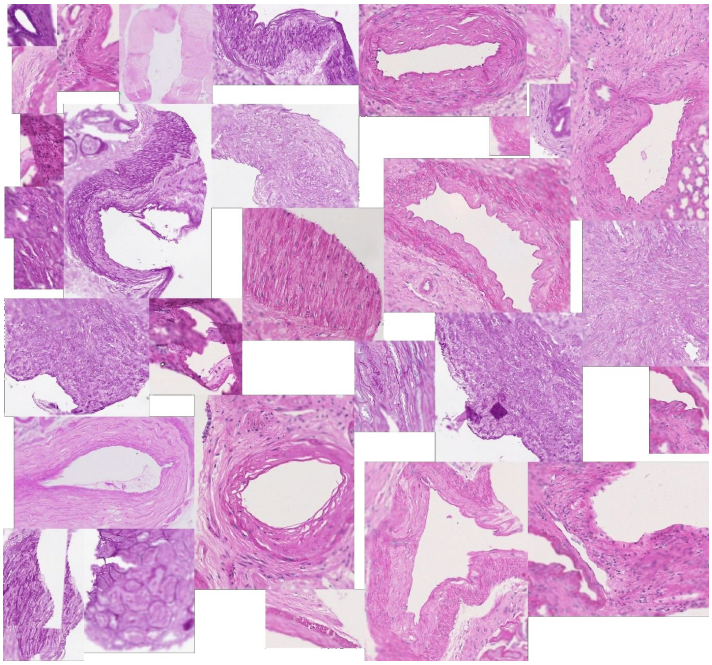
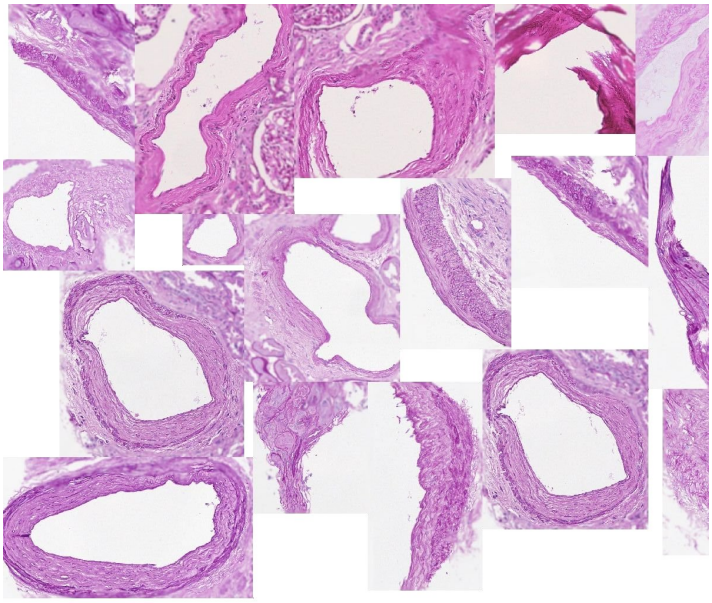


Figure 4.9: Urinary space to glomeruli area ratio. Numbers next to each block denote the quantile from which the example tissue compartments are algorithmically selected from the QUOD dataset. Visual differences of individual handcrafted features can be easily interpreted by inspecting the collection of tissues with low vs high values. Note that regions outside the tissue are blurred.



(a) Segmented vessels with minimum lumen to total tissue area ratio



(b) Segmented vessels with maximum lumen to total tissue area ratio

Figure 4.10: Vessels with (a) minimum / (b) maximum ratio between the lumen area to total vessel area. This feature corresponds to one of the Remuzzi Score criteria. Examples are algorithmically selected from the entire QUOD dataset. We choose to calculate the ratio of areas instead of diameters to avoid geometric template fitting, as most vessel sections are not round.

grades using Spearman correlation, as shown in Table 4.8.

Nuclei Moments, y^{σ^2}, and b^{σ} are intended to be tubular features describing acute (ATI) changes. When we compare the median values with the slides' ATI scores (Section 3.3.1), correlations are positive and significant ($p=1.88 \times 10^{-3}$, $p=9.60 \times 10^{-8}$, and $p=6.37 \times 10^{-7}$, respectively). From examples of the latter two features, as shown in Figures 4.7 and 4.9, it can be seen that tubules at the 90th percentile contain many more loose pieces compared to the $10th$ percentile. These changes reflect the local assessment scores provided by the pathologists (Figures 2.8a-2.8d).

One assessment criteria of the Remuzzi score [4] is the vessels' lumen to wall cross section ratio. Computationally, it is more straightforward to compute the ratio of the lumen to vessel area. Area and cross section ratio measurements are closely related as the two scales monotonically for vessels of fixed aspect ratio. The former has the advantage as it can be evaluated on non-circular vessels. We find that the lumen to vessel area correlates negatively with Remuzzi A grade (r=-0.276). This correlation is confirmed by a visual inspection of the examples (Figure 4.10). However, the correlation is not found to be significant. There are three possible reasons why statistical significance is not met: 1) The number of arteries in our QUOD dataset is rather small (Distribution of artery numbers in Figure 3.18); 2) Segmentation of vessels in our dataset is confined only to larger arteries (Many arterioles have been misclassified as tubules in both the ground truth and our UNets); 3) Some Remuzzi A grades were given to slides with only a partial cross-section of an artery (For example, the bottom right of Figure 2.5a).

In transplant cases with poor outcomes, it has been found that glomeruli sometimes expand in volume to compensate for the loss of function in other regions of the kidney [31]. In our QUOD-slides, we find that glomeruli areas neither correlate with eGFR12 (Spearman r=-0.0443, p=0.456) nor Remuzzi G grade (Spearman r=0.0352, p=0.656). One limitation of our dataset is that the average cross section areas do not reflect glomeruli volume as most QUOD slides have a sub-optimal number of glomeruli. In this case, the ratio of the urinary space to the glomeruli

area may be a more suitable feature as it is less dependent on where the microtome cuts through the glomeruli. Figure 4.9 shows that our feature does capture this change successfully. However, the correlation between the urinary area to glomeruli ratio is still weak when it is compared against eGFR (r=-0.0148, p=0.800) and Remuzzi G grade (r=-0.0813, p=0.302).

Table 4.8: Correlation between individual features and grades given by a pathologist. A few selected handcrafted features are chosen. We take the median values from the relevant tissue type from each slide and correlate them with grades given by our collaborating pathologist.

Handcrafted Feature	Tissue Type	Assessment	Spearman r	Spearman p
Nuclei Moments	Tubules	ATI	0.174	1.88e-3
y^{σ^2}	Tubules	ATI	0.295	9.60e-08
b^{σ}	Tubules	ATI	0.278	6.37e-07
Lumen Area Ratio	Vessels	Remuzzi A	-0.276	0.301
Urinary Space Area Ratio	Glomeruli	Remuzzi G	-0.0813	0.302

We recognise that correlation between individual handcrafted features may not necessarily correlate strongly with clinical variables or visual assessment grades. Hence, a longer list of features has been designed. A brief description of $m = 98$ handcrafted features is shown in Table A.2 of Section A.3.

There are two ways to incorporate these handcrafted features for machine learning. The first way is a "tissue compartment agnostic" representation - the same features are evaluated on different tissue compartments regardless of the c tissue classes (tubules, vessels, glomeruli). In this approach, each slide with n tissue compartments would be described by a matrix of size $\mathbb{R}^{n \times m}$. The second approach is to evaluate features for each of the c tissue classes separately. We call this "tissue compartment specific" approach. This would result in a matrix of size $\mathbb{R}^{n \times cm}$ for each slide (with missing values being imputed). The choice between the two approaches would depend on the quality of the ground truth segmentation labels. For methodological development in Chapter 6, we want predictions to be robust to mislabelling of the training labels (since the delineation of tissue compartments was performed by the author and the pathologist did not have time to cross-check all of them) as only a small portion of delineated compartments

had been checked by a pathologist (Section 3.3.2.2), so slides are represented by the tissue compartment agnostic approach.

Some of the features are strongly correlated with each other but differ in how they are normalised. A correlation matrix of these features in the QUOD slides, as plotted using the tissue compartment agnostic representation, is shown in Figure 4.11.

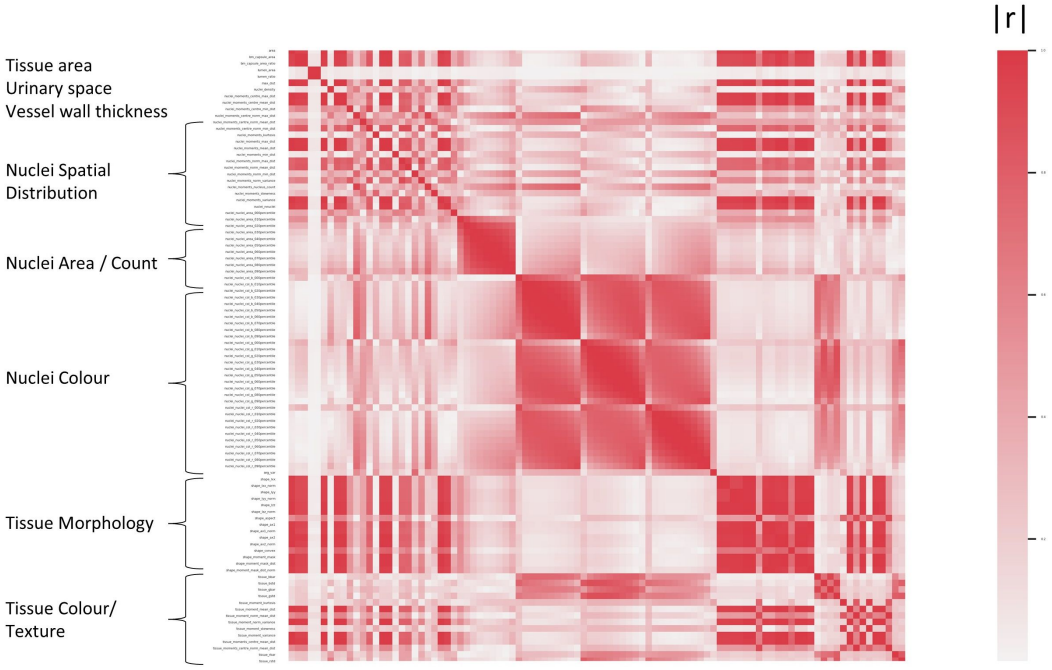


Figure 4.11: Illustration correlation between handcrafted features

4.4 Extensions of Workflow

Here we detailed three short extensions of this workflow that were found to be promising but may require further validation and interpretations from clinical experts.

Section 4.4.1 discusses a workflow to cluster tissue compartments based on visual features. It was found that when tissue compartments are restricted to cortical tiles containing only proximal tubules, visual features can successfully capture histological changes. However, feature variations within slides dominate the principal components forming the clusters when irrelevant tiles are included. In this case, clustering can be used to detect tissue compartments with artefacts without the need for explicit local labels.

Section 4.4.2 describes a feasibility study which aims to quantify the amount of fibrotic tissue in Sirius Red slides. We find that the amount of fibrotic tissue computed by our workflow significantly correlates with the recipient's 12-month eGFR (r=-0.259, p=0.00533, n=114). This highlights the potential of the workflow as a tool to assess donor kidney quality by predicting kidney function by generating a donor kidney fibrotic score.

Section 4.4.3 describes a preliminary application of our workflow on a cohort of histological slides of kidneys that we have also obtained proteomic data by data independent acquisition (DIA) spectrometry analysis. The experiment describes an attempt to explain two PCA-generated proteomics clusters of patients with no significant clinical differences using computational histology. It demonstrates how visual features can be used to capture subtle histological changes to provide potential interpretations of two clusters with what are possibly sub-clinical differences.

4.4.1 Clustering and Detection of Anomalous Slides

This section describes a short experiment in an attempt to cluster histological features to detect abnormal tissue compartments within slides. Clustering tissue compartments according to their visual features is a common approach for obtaining slide-level descriptions of biopsies. This method works when biopsies have a reasonably homogeneous distribution of compartment classes. However, in renal needle biopsies, the morphology of nephrons changes progressively with depth from the cortical capsule. As such, variation of tubules is often greater within each slide than between slides. A shift in the distribution of tissue may often be a sign of the slight differences in biopsy procurement (e.g. depth of the needle) rather than actual pathological changes.

A schematic principal component analysis (PCA) example of tissues from different QUOD-PAS slides is shown in Figure 4.12. Here, each marker is a tubule and is coloured by the eGFR12 outcome from their corresponding slide. The most important principal components include compartment area and aspect ratio. Tubules from each slide are scattered across the whole PCA space. Small shifts in distribution

from one slide to another do not tell us the quality of the slide. Similar results can be seen if deep features are included or if compartments are clustered using alternative techniques such as Kernel PCA [142], TSNE [143], or k-means clustering.

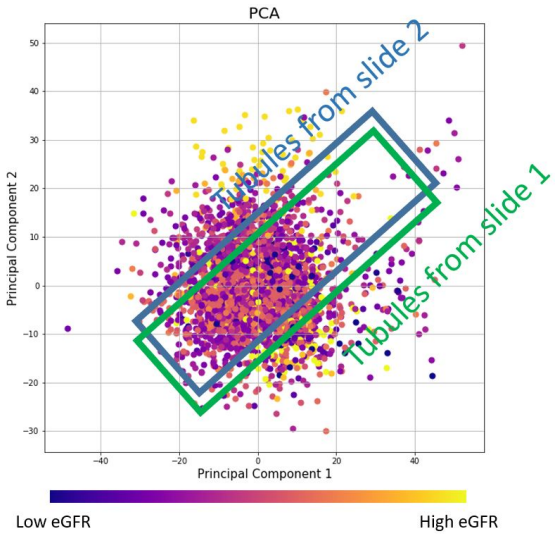


Figure 4.12: Schematic example of PCA of handcrafted features from tissue compartments of multiple slides. Each marker is a tubule and is coloured by eGFR12 of the slide.

While these techniques were designed to visualise the dominant trends in one dataset, features important for the identification of abnormal tissue are subtle in many settings. Contrast PCA (cPCA) [144] can be used to identify low-dimensional structures enriched in a dataset relative to some background comparison data. We find that by manually picking a set of slides with normal (healthy and free of artefact) tissue appearance, tubules with abnormal morphology conspicuously stand out. Figure 4.13 illustrates an example of cPCA: Markers in blue denote tubules from the QUOD slides with no artefacts. The green markers denote tubules from the remaining QUOD slides. Outlier tubules denoted by green markers tend to be wrinkled or were from thick sections. The original implementation of cPCA requires a hyperparameter α. We propose to determine α automatically by minimising the convex hull IOU between the background and test data markers. Overall, this approach allows us to identify slides containing abnormal tissues without explicitly labelling them locally.

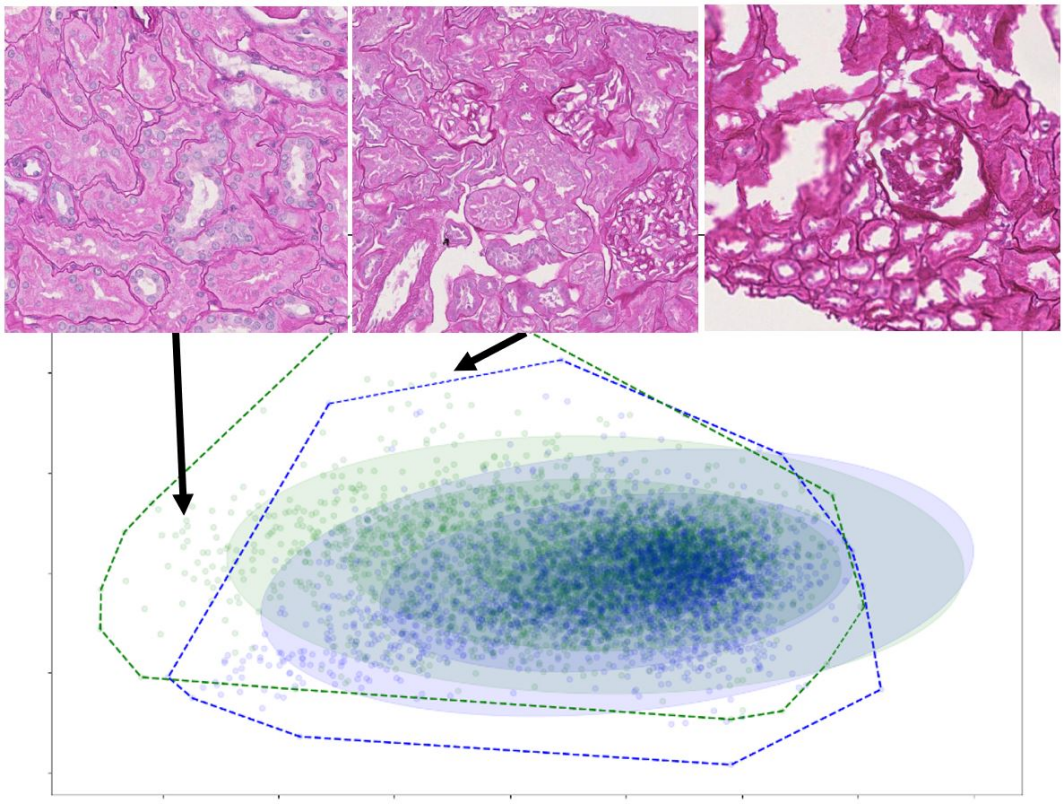


Figure 4.13: cPCA plot of tubules from the QUOD-PAS slides

While clustering could help identify slides with artefacts, it did not identify tissue compartments with pathological changes in the QUOD slides. We suspect there were two reasons: i) too few proximal tubules with chronic changes in the whole dataset (see Figure 3.14 in Section 3.3.2); ii) tubules from the medulla had been included in the analysis.

We have some evidence suggesting that excluding tissues irrelevant for assessment may help identify tubules with pathological changes: At the beginning of this project, we had 104 tissue patches scored by a pathologist (MFS) at a scale between 0-4 indicating how relevant a patch is for assessment (i.e. whether it contains cortical tissue) - we called this the 'informative score'. Details of these labels are described in Section 3.3.2.1. The informative scores were normalised to a numeric scale between 0-1. These scores were used to train a ResNet18 [81] regressor with the last layer's softmax predicting the informative score. A model was trained on 4:1 train:validation split with patches scaled to 224x224. Adam optimiser was used

to optimise the model based on L2 loss between the output and the normalised informative score. This model was iterated over all NMP-HE slides. Patches that scored above 0.5 were considered cortical tissues. The trained model was found to be only able to regress the informative score on the NMP dataset (r=0.82). It does not appear to generalise well to other datasets at this stage due to the small number of training examples available. However, the biased performance of this ResNet18 does not matter as we are only trying to explore the feature extraction part of the workflow by excluding irrelevant tissues. By combining this ResNet18 with a UNet, we were able to select tissue compartments that were exclusively within the cortex. Figure 4.14 shows that plotting the handcrafted features of these relevant tissues using TSNE [143] reveals a clear distinction between tubules with different acute damage stages.

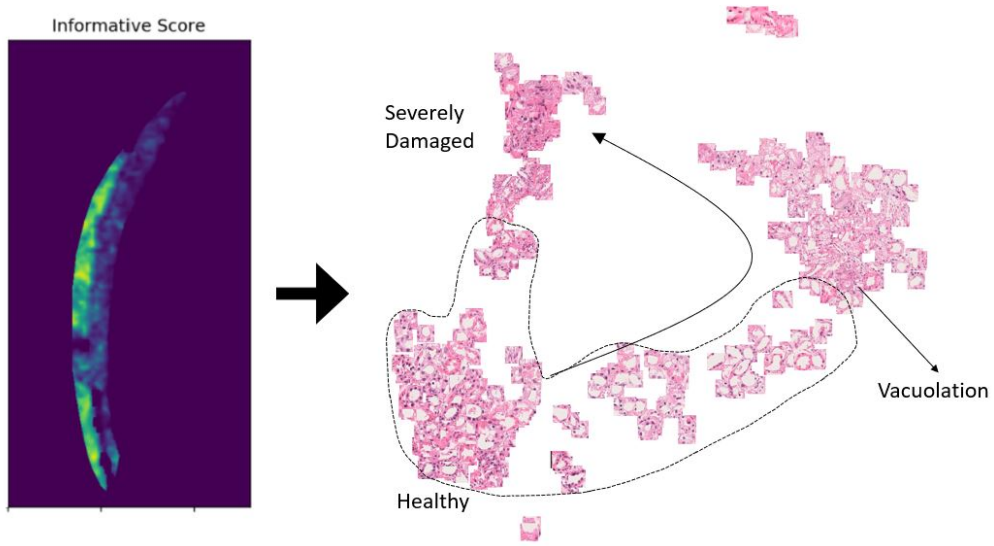


Figure 4.14: TSNE visualisation of handcrafted features in the NMP dataset. Left: A ResNet18 was trained to predict an informative score indicating whether a local region is relevant for assessment. Tissue compartments with informative score above 0.5 were considered 'relevant' (i.e. part of the cortex). Right: The handcrafted features of these tissue compartments are plotted using TSNE visualisation showing a very intuitive clustering of healthy and damaged tubules.

4.4.2 Computational Quantification of Fibrosis in Sirius Red Slides

In this section, we applied our workflow to analyse the fibrotic content of kidney histological slides from tissue obtained from 141 deceased donors (141 kidneys, 141 donors). Slides were stained by Sirius Red (SR), and n=33 slides were assessed by pathologists as described in Section 3.3.1.2. The analysis showed that the amount of fibrotic content in donor slides with $> 1mm$ of cortex tissues (n=114) as computed by our workflow correlates significantly with post-transplant 12-month function (eGFR12). The computed values also correlate with pathologists' assessment (n=33, r=0.391, p=0.0243). However, for this cohort of n=33 slides the pathologists' assessment is not associated with eGFR12 (n=33, r=-0.0977, p=0.589). This work shows our workflow has the potential to predict outcomes from histology.

Chronic kidney disease is manifested by several known histological changes in renal biopsies. One of these changes includes the accumulation of extracellular matrix, the precursor of interstitial fibrosis. While several different assays [3, 59, 145, 146] can be used to reveal the extent of fibrosis, Sirius Red (SR) stain binds specifically to collagen types I and III when combined with enhancement of birefringency [147]. It offers a high signal-to-noise ratio (SNR) for detecting collagen fibres when slides are examined under crossed polarisers. This property makes SR a convenient assay for the computational analysis of fibrosis. While SR slides are usually inspected under a polarising microscope for optimal SNR, the efficacy of examining SR slides under unpolarised light is also possible [145, 146]. In fact, fibrosis quantification in unpolarised SR slides was shown to have higher inter-observer reproducibility [145] and correlates with recipient eGFR more strongly than SR under polarised light and several other stains [145, 146]. Under SR stain, a variety of tissues such as basement membrane are also stained red but are not birefringent. The red stain is also picked by vessel walls which need to be excluded from the analysis. This makes it harder to analyse SR slides by simple computer algorithmic techniques as it would no longer be possible to just measure the total red areas. Despite the added challenge, it may possibly still be easier to

computationally measure fibrosis from unpolarised SR compared to other stains. To our knowledge, computational quantification of fibrosis has only been performed on polarised SR slides [148–151] and not on unpolarised SR.

Here we performed a feasibility study to quantify the amount of fibrosis in unpolarised SR slides. To achieve this, we used our segmentation workflow to mimic pathologists. The fibrosis percentage was evaluated as the percentage of all cortical tissues (excluding tubules and glomeruli) occupied by fibrous tissue [145]. A schematic of our method is shown in Figure 4.15. From each SR slide (a), we manually excluded the medulla. The mask of the cortex (b) was segmented using saturation and value thresholding. Regions with tubules and glomeruli were evaluated using an ensemble of UNets (c) which predicted tubules, glomeruli, vessels, and boundaries. The UNet ensemble architecture details are the same as what will be used later in Chapter 6, Section 6.2.3. As the UNets predict both tissue compartments (c1, green) and boundaries (c2, red), we argue that some fibrotic content might be present at the boundaries between different tissue compartments. Hence, we calculate the amount of fibrosis as follows:

$$Fibrosis = (b - c_1 - 0.5c_2)/b, \tag{4.3}$$

where b, c_1, c_2 are the cortex, tissue compartment, and boundary areas as illustrated in Figure 4.15.

The amount of fibrosis obtained through computation is compared against both eGFR12 and the pathologist's assessment (details of slide assessment: Section 3.3.1). A summary is shown in Table 4.9. We find that computed fibrosis is significantly associated with a lower eGFR12 when all 114 slides with $1mm^2$ of cortical area are considered (Pearson r=0.259, p=0.00533). We also find that computed fibrosis significantly correlates with pathologists' assessment (r=0.391, p=0.0243) based on the 33 assessed slides. However, pathologists' assessment does not associate with eGFR12.

Because we evaluated computed fibrosis on a larger number of slides than the pathologists had, for a fairer comparison, we compare correlations on the same

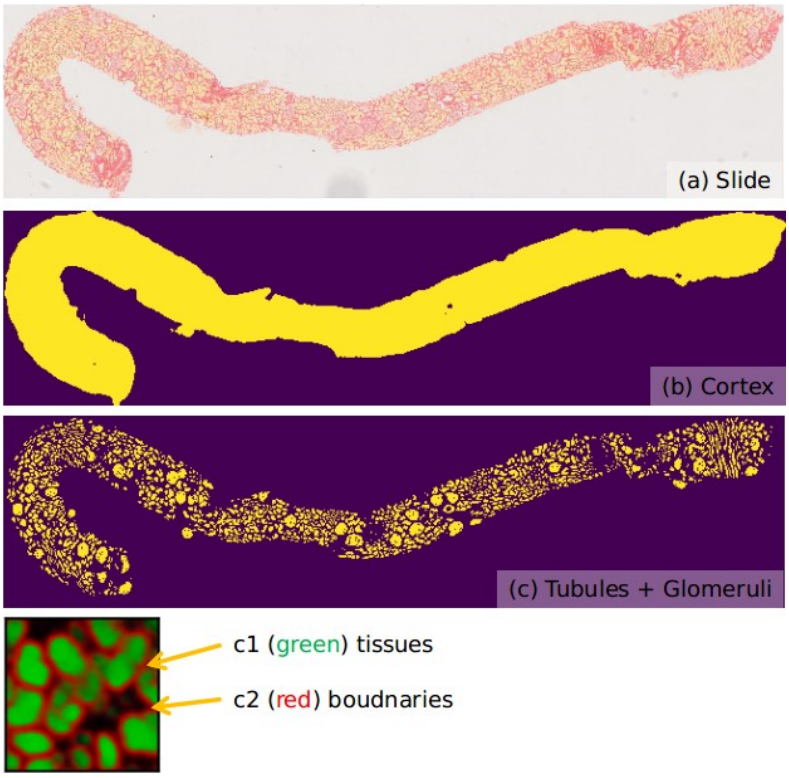


Figure 4.15: Computational estimate of Fibrosis in SR Slides

30 slides that contain $> 1mm^2$ cortex. It was found that computed fibrosis still correlates with eGFR12 (r=-0.417, p=0.0218), but the pathologists' assessments do not correlate with eGFR12 (Table 4.9 (a)).

Figures 4.16 and 4.17 show scatter plots of the 3-way comparison between pathologists, computed fibrosis, and eGFR12. These figures confirm our earlier findings that computed fibrosis may be useful for predicting recipient eGFR.

Apart from subjectivity, one possible explanation for why our algorithm performed better is that our collaborating pathologists assessed fibrotic content differently from ours. When pathologists assess SR slides, they tend to look for large patches of fibrotic red areas, whereas our algorithm partially counts red areas between tissue compartments.

It is worth noting that, in theory, fibrotic content can be assessed in both PAS (a routine stain) and SR slides. In PAS slides, interstitial fibrosis is a criterion in the Remuzzi score (Remuzzi IF). Hence, we also attempted to compare fibrosis assessed on the SR slides against the Remuzzi IF sub-component on adjacent PAS

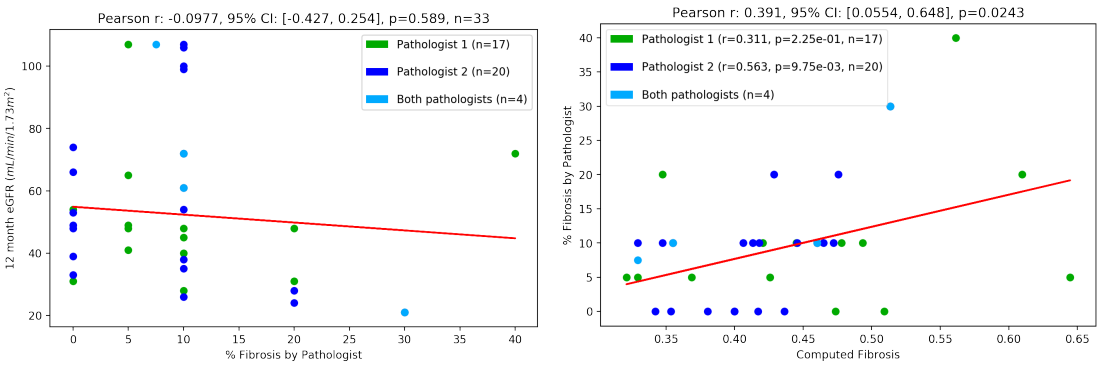


Figure 4.16: Pathologist Assessment of Fibrosis in SR Slides (n=33) (a) vs eGFR12; (b) vs Computed Fibrosis

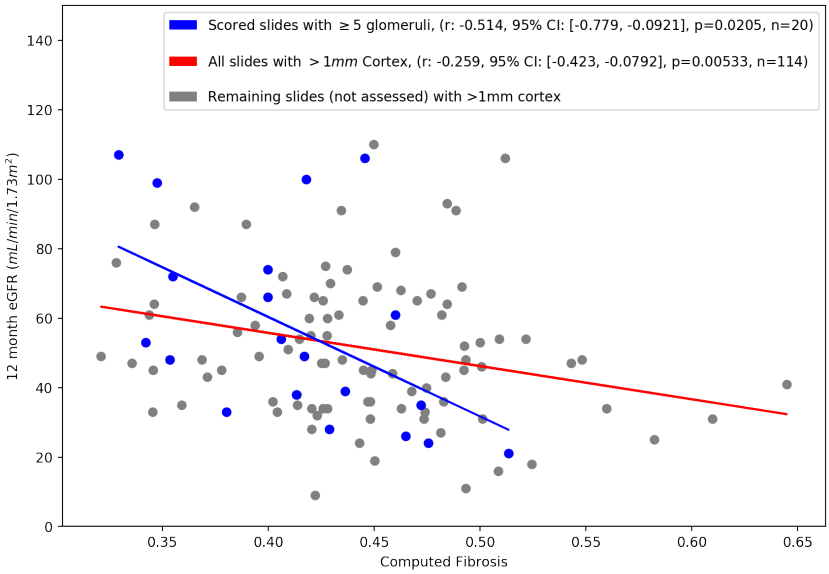


Figure 4.17: eGFR12 vs Computed Fibrosis % from SR Slides. The plot shows a marginally significant correlation between computed fibrosis and outcome. The trend is stronger when only the larger slides (n=20 with ≥ 5 glomeruli, handpicked to cover a range of eGFR12 and scored by pathologists.) were analysed. Correlation is still mildly significant when we include all slides with $> 1mm^2$ of cortex.

Table 4.9: Correlation (Pearson) between Computed Fibrosis, Pathologists' Assessment, and eGFR12. We find the correlation between computed fibrosis and eGFR12 significant when only slides with >1mm of cortex are considered. Pathologists' assessments do not correlate with eGFR12. As only a subset of slides has been scored, this table includes comparisons of where only the score slides are considered (a, b), only slides with >1mm cortex are considered (a, c), and all slides (d). Significant correlations are in bold.

(a)	Scored Slides with >1mm Cortex (n=30)	Computed	eGFR12
	Pathologist	r=0.357, p=0.0527	r=-0.233, p=0.215
	Computed	-	**r=-0.417, p=0.0218**
(b)	Scored Slides (n=33)	Computed	eGFR12
	Pathologist	**r=0.391, p=0.0243**	r=-0.0977, p=0.589
	Computed	-	r=-0.341, p=0.0519
(c)	Slides with >1mm Cortex (max n=114)	Computed	eGFR12
	Pathologist	r=0.357, p=0.0527, n=30	r=-0.233, p-0.215, n=30
	Computed	-	**r=-0.259, p=0.00533, n=114**
(d)	All Slides (max n=141)	Computed	eGFR12
	Pathologist	**r=0.391, p=0.0243, n=33**	r=-0.0977, p=0.589, n=33
	Computed	-	r=-0.148, p=0.0805, n=141

sections. No significant association was found between either the computed (n=70) or pathologist (n=20) SR scores against Remuzzi IF. This could be because only a few slides were available as adjacent sections.

Given that Remuzzi IF grades on the PAS slides from the same cohort are not significantly associated with eGFR12, this work may indicate the value of using SR as an assay for computational histopathology - either 1) as a means for assessment or 2) to just curate more training data to bootstrap the training of a numerical version of Remuzzi IF grade in adjacent PAS slides.

4.4.3 Correlation of Visual Features with Tissue Proteomics Signatures

Advances in mass spectrometry technologies and computational biology have opened new opportunities in disease stratification and the development of diagnostics. The profiling of the human proteome has increased in sensitivity and depth with data independent acquisition (DIA) mass spectrometry [152]. Here we aimed to identify associations of handcrafted features in histological kidney slides with molecular

profiles in biopsies of the same kidneys. In short, deceased donor kidney tissues were selected based on paired 12-month post-transplant outcomes. Donor kidneys were selected to include the continuum of transplant outcome (full range of eGFR from primary non-function to eGFR $> 80ml/min/1.73m^2$) and, where possible, to exclude extreme demographic or clinical factors. All clinical samples were linked to corresponding donor and recipient demographic and clinical metadata (Table 3.3), provided by NHS Blood and Transplant National registry (NHSBT). Analysis of proteomic data resulted in the identification and quantification of 2970 proteins that cluster the included donors in two main principal component analysis generated clusters without significantly different clinical characteristics. Here in this analysis, we aimed to identify whether donor kidney histological characteristics could explain the clustering of donors. We present some of our findings based on the 162 donor kidneys available to us as PAS-stained histology slides.

To analyse whether there are visual differences between the two donor clusters, we converted tissue compartment specific visual features into a slide representation. Each feature is represented by the median value across all tissue compartments for a given slide. Figure 4.18 shows a screenshot of an interactive volcano plot representation of the visual features. The x-axis shows the ratio in mean values between the two clusters (Cluster 2 / Cluster 1), whereas the y-axis shows the significance (Kruskal Wallis p-values). To reduce the number of spurious correlations, we also calculated mutual information between the cluster label and binarised visual features - this is represented by the marker size in the plot. By inspecting the plots, we find three groups of visual features to be significantly different between the clusters:

- Cell nuclei in Cluster 2 are significantly paler ($p < 0.01$)
- Lumen areas of arteries in Cluster 2 are smaller ($p < 0.05$)
- Epithelial cell cytoplasm is less compact in Cluster 2 ($p=0.0107$)

All three groups of visual features paint a consistent picture suggesting donors in Cluster 1 might be in a better quality state. For instance, paler nuclei and

loose objects in the cytoplasm are signs of acute changes (examples in Figure 2.8 of Section 2.2.5). A smaller lumen in Cluster 2 is also a sign of chronic damage (Figure 2.5 Section 2.2.5). There is a possibility that our workflow might be able to detect some sub-clinical changes within the cohort.

The major limitations here are: 1) Some findings may require further validation from pathologists. Regrettably, validation of subtle features such as nuclei colour will require extensive labour input. 2) Many slides in the analysis contain no arteries; hence the significance level is only mild ($0.01 < p < 0.05$).

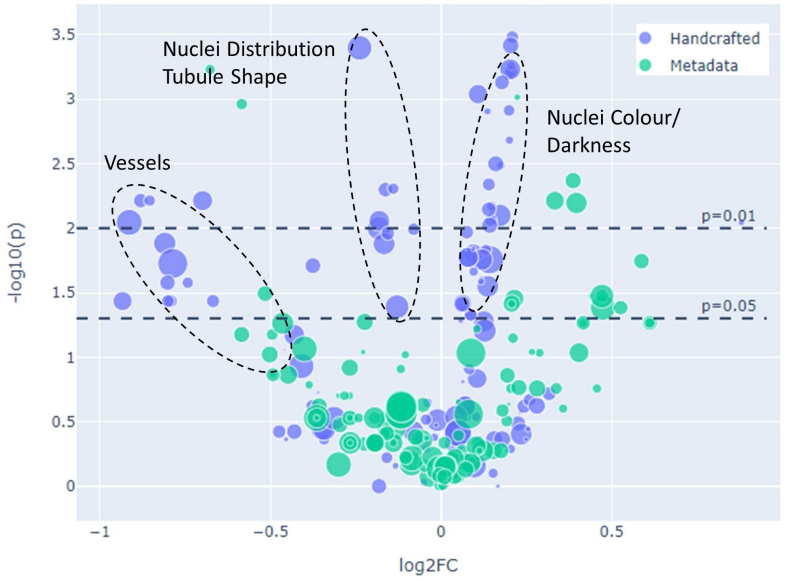


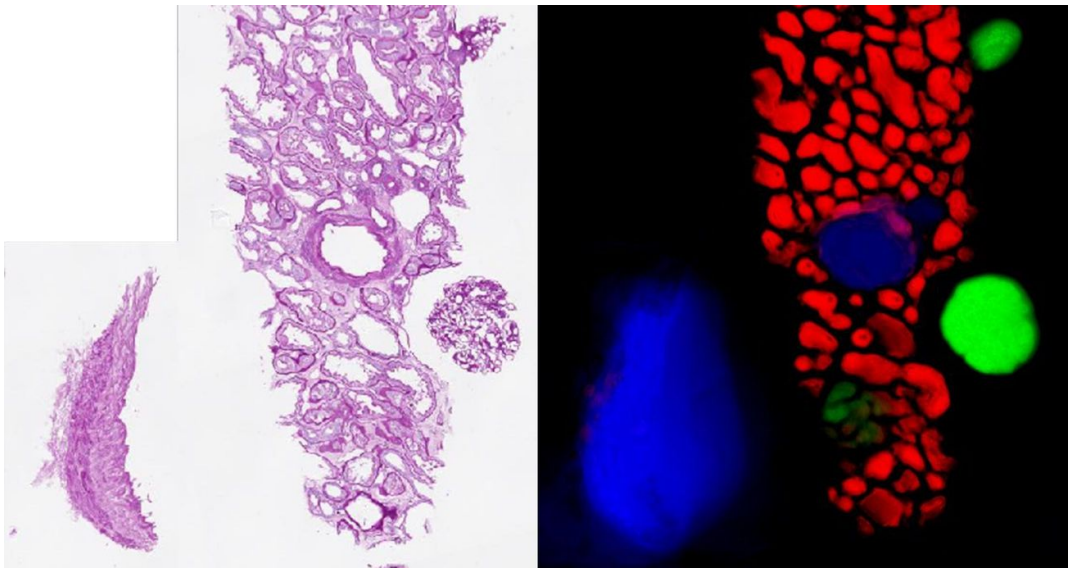
Figure 4.18: Volcano plot showing differences in visual features between two different proteomic clusters

4.5 Limitations

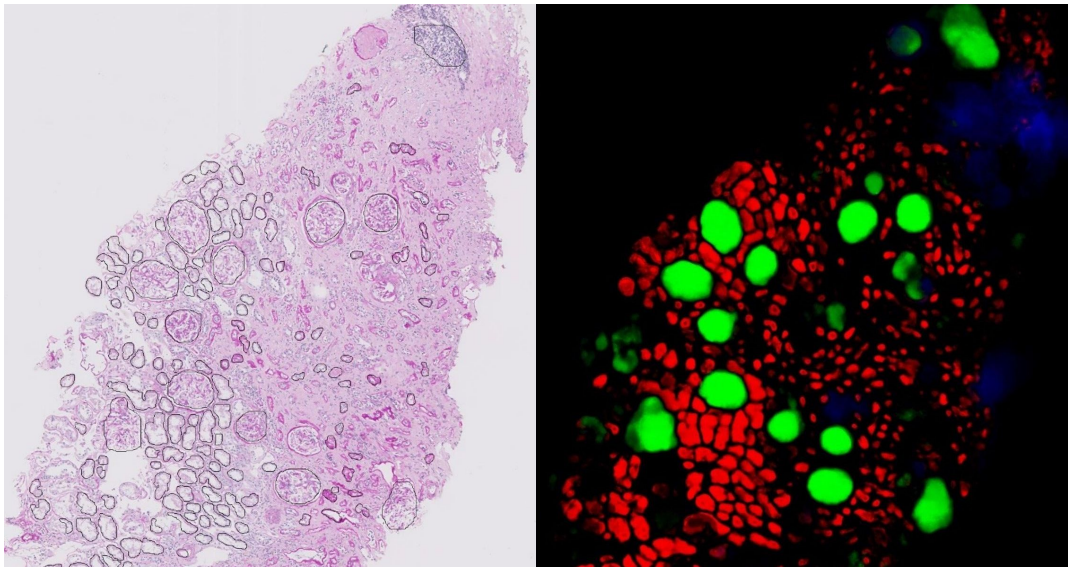
So far, we have developed a workflow that segments and extracts visual features from four tissue compartment classes. The primary constraints to this workflow are i) small tissue area in the QUOD slides and presence of artefacts not available in the training data; ii) Lack of segmentation examples of tissue compartments with chronic damage; iii) Lack of time for the pathologist to cross-check annotations. These factors have limited our ability to segment chronic lesions. To illustrate the impact of these limitations, we present qualitative examples of how our workflow

performs on several challenging biopsies in Figure 4.19. Figure 4.19a is from the QUOD-PAS dataset, whereas Figures 4.19 (b)-(c) are from the AMC-PAS dataset – these were external data obtained towards the end of this DPhil project.

Limitation 1: Due to the small size of needle biopsies, many compartments were truncated. Figure 4.19a shows examples of a truncated vessel and two glomeruli on the edge of the biopsy. Many tubules in the QUOD slides also contain wrinkles

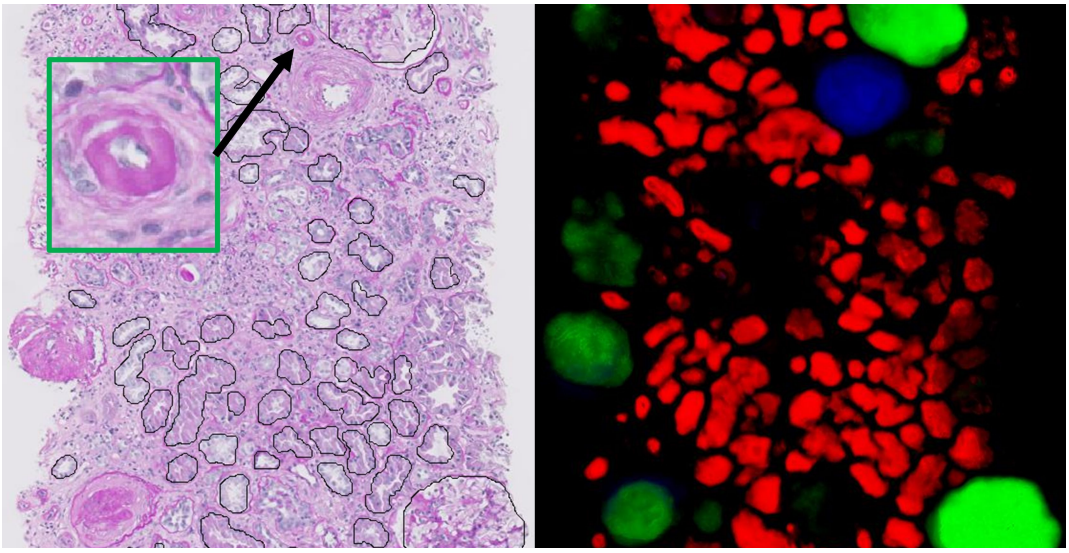


(a) Example from QUOD-PAS showing vessels and glomeruli detached or near the edge of the biopsy



(b) Example from AMC-PAS showing patches with IFTA

Figure 4.19: Challenging Segmentation Examples



(c) Example from AMC-PAS showing sclerosed glomeruli and arteriole (arrow)
Figure 4.19: Challenging Segmentation Examples (cont.)

which tend to be poorly segmented (e.g. Figure 4.13). Some handcrafted features (e.g. vessel lumen area) may not serve their intended purpose in compartments near the edge of the biopsy or are poorly segmented.

Limitation 2: The lack of chronic lesions is apparent in the NMP and QUOD-PAS datasets. We only have 80 sclerosed glomeruli among the scored slides in these two datasets. Among all proximal tubules reviewed by the pathologist, approximately 1% exhibit barely-noticeable signs of chronic damage (Section 3.3.2.2). The lack of cases has impacted segmentation performance on these chronically injured instances. Examples of segmentation performance on atrophic tubules and sclerosed glomeruli can be seen in Figures 4.19 (b)-(c), respectively. We do not report the IOU of sclerosed glomeruli because a significant proportion of the delineated examples might be urinary casts (Section 2.2.5 Figure 2.6c), but most have not been reviewed by the pathologist (Section 3.3.2.2).

Limitation 3: Quality of the segmentation ground truth. Due to the limited availability of pathologists, tissue instances were delineated by the author, an engineering student with several months of experience in renal pathology. Many

tissues have been mislabelled and might have led the UNets to misclassify arterioles as tubules (Figure 4.19c) or urinary cast as sclerosed glomeruli. It is also hard to distinguish proximal tubules from other parts of the nephron for kidneys with ATI. Furthermore, precise details on how compartments were delineated had changed over time.

Limitation 4: The workflow was validated on only one cohort (QUOD) with limited diversity where some slides were digitised using the same scanner as the training data. This may produce overly optimistic results. Further work to address this limitation would be to include slides from different centres and scanners for a more realistic benchmark for the evaluation of segmentation quality and the reliability of various visual features.

5

Improving Histological Measurements with Bayesian Neural Networks

Contents

5.1 Why is Uncertainty Quantification Necessary? 113
5.2 Uncertainty Estimation with Bayesian Neural Networks 116
5.2.1 Bayes by Backprop (BBB) 116
5.2.2 Monte Carlo Dropout 117
5.2.3 Quantifying Uncertainty 117
5.3 Method 119
5.3.1 Histopathology Measurements 119
5.3.2 Proposed Method for Improving Histological Measurements 121
5.3.3 Datasets 124
5.3.4 Tissue Segmentation 129
5.4 Experiment Results 133
5.4.1 Relationship between entropy and histological measurements 139
5.4.2 Hard Thresholding 140
5.5 Conclusion 140
5.6 Synthetic Data 142

In Chapter 4, we trained CNNs to perform segmentation of various renal tissue followed by feature extraction from these tissue structures. Cluster analysis (Section 4.4.1) showed that some histological features were highly dependent on the quality of the image segmentation mask. In this chapter, we discuss the use of Bayesian Neural

Networks (BNNs) to quantify uncertainties in image segmentation maps and propose a way to improve histological estimates using confidence measures derived from these BNNs. Our results indicate that the estimated distributions are consistently improved when instances are weighted by the entropy-derived confidence.

This chapter is structured into three main sections. Firstly we explain in Section 5.1 why uncertainty quantification is important in digital pathology, and in particular, for the assessment of renal transplant biopsies. The second section (Section 5.2) describes related work on the quantification of uncertainty in BNNs. The third section (Section 5.3) details our proposed approach to improve histological measurements using uncertainty measures derived from BNNs. This is followed by an experiment performed on both a test dataset and a synthetic dataset generated based on our renal biopsy images. We demonstrate an application showing how uncertain estimates can be used to improve the results of various histological measurements.

5.1 Why is Uncertainty Quantification Necessary?

The core goal of computational pathology is to extract clinically relevant information from digitised slides to support the work of human experts. Examples of such measurements include the size and shape of certain tissue architecture components (e.g. glands, tubules, and cells) as well as their colour and texture. Accurate evaluation of these measurements is crucial but could also be very challenging for a number of reasons. Firstly, histological changes are often subtle and could easily be overwhelmed by the biopsies' preparation artefacts. Secondly, colour and morphological appearances could vary greatly between different datasets, leading to overfitting errors. Thirdly, like many medical image processing tasks, delineation of histopathological structures could be subjected to errors caused by underlying ambiguities in the dataset and annotators.

Segmentation errors could lead to the miscounting of tissue instances and measured areas and inaccuracies in other downstream statistics that depend on the segmentation results. As segmentation uncertainties are often heteroscedastic, erroneously segmented instances could lead to systematic errors in the estimated

distributions. Representing classification results from neural networks as point estimates is risky as it has been argued that deep neural networks often produce over or under-confident results. Quantification of classification uncertainty is, therefore, necessary to help us manage the risks involved in these black-box models.

Bayesian neural networks (BNNs) provide a framework that captures model uncertainties in the form of a probability distribution. They are less prone to overfitting and often perform better in small datasets. To date, a number of methods have been developed to approximate Bayesian processes in neural networks. Examples include Monte Carlo (MC) dropout during test time [153], the use of variational parameters in the classification network trained by Bayes by Backprop [154, 155], injection of variational embedding from an auxiliary network [156], and re-calibration of models [157] using the aforementioned models as the teacher model.

These probabilistic models could be applied to evaluate uncertainties in semantic segmentation tasks to obtain pixel-wise uncertainty maps. For example, this had been demonstrated by Kwon [158] and Kendall [159] in an open image dataset; Mukhoti [160] had also developed methods to evaluate Bayesian uncertainties for semantic segmentation. Uncertainty maps generated from Bayesian networks have also been used to quantify or visualise semantic segmentation errors in medical images in several modalities, including CT images [156], retinal OCT Scans [161], skin lesion images [162], Cardiac [163] and stroke lesion MRI images [158].

On the other hand, object detection is more challenging than semantic segmentation as it involves identifying separate objects in an image. Only a limited number of publications had focused on extending BNNs to object detection tasks. Miller [164, 165] employed MC dropout in the fully connected layers of a Single Shot Detector. The network performs multiple forward passes at test time and produces a different set of bounding boxes in each pass. The boxes are clustered together based on their spatial and semantic similarity. The similarity between the boxes is also used to derive uncertainty measures of the task.

Instance segmentation is yet more difficult than object detection as it requires us to delineate the segmentation masks of individual objects rather than just bounding

boxes. In this regard, Morrison [166] extended Miller's implementation to work with MaskRCNN [83]. The resulted model produced not only bounding boxes but also masks for instance segmentation, allowing uncertainties to be evaluated based on similarity measures between masks. However, these modalities involve the use of architectures that require a large number of training data, making it hard to be applied to most biomedical image datasets.

In digital histopathology, we are often interested in measuring histological features from densely populated touching instances of the same class. Some of these measurements will depend on how well we obtain the masks of individual tissues through instance segmentation. If the tissue sample is large, the easiest solution to circumvent the problem is to ignore the regions with high ambiguity. However, in renal transplant, biopsy procurement is a highly invasive procedure where larger bore needles could lead to higher complication rates [167]. Thus the slides typically contain only a minimal amount of tissue. Our test dataset of 144 biopsies is no longer than 11mm (as each of the biopsies was cut in half and donated to another institute), smaller than the standard 22mm biopsies used for transplant assessment. If we discard all tissue compartments with high uncertainty, there may not be much tissue remaining for quantitative analysis. In addition, test data may also contain preparation artefacts not present in the training set. Since artefacts are associated with a slide, we find it difficult to determine a threshold to reject instances with high uncertainty. Fortunately, in biopsy assessment, we are sometimes not concerned about measuring an individual instance. For example, in our case, we are interested in measuring the distribution of histological measurements over the whole slide. Consequently, we propose a method to include all segmented tissue into our workflow and weight them by some confidence measures derived from BNNs. We test several variations of the proposed method in subsequent sections.

5.2 Uncertainty Estimation with Bayesian Neural Networks

Although deep neural networks have achieved state-of-the-art results in many classification tasks, in traditional feed-forward neural networks, the fixed weights and bias parameters only allow the output of point estimate predictions. It has been known [153] that the numerical value from the softmax output in the $[0, 1]$ range from these deterministic networks cannot be directly interpreted as the classification confidence. On the other hand, Bayesian approaches allow inferences with reduced overfitting. Sample-averaged estimates from the probability distribution of Bayesian models may also produce a better-calibrated estimation of model errors. However, the exact probability distribution of Bayesian neural networks is often intractable. Therefore, in practice, approximations are usually used in place of true posteriors.

5.2.1 Bayes by Backprop (BBB)

Blundell [154] proposed parameterising the weights ω of neural networks as a variable drawn from a stochastic distribution $\omega \sim P(w|D)$ to approximate a Bayesian network. In their proposed approach, the parameters governing the weights' distribution are learnt from the data D. For example, if we choose a Gaussian as the prior, then the learnable parameters for each weight will be the mean μ and standard deviation σ. Training the model would then involve an objective function that aims to minimise the distance between the prior and the sampled weights in addition to the classification loss. During test time, we obtain multiple predictions from different sets of sampled weights. Inference can be made by averaging the ensemble predictions, which ideally should give us a more realistic value for model confidence.

Instead of sampling the weights of the model, an alternative approach is to sample the activation layers using the local reparameterisation trick (LRT). Following Kingma [168], for a convolutional network with Gaussian prior [155, 169], we can write the sampled activation from layer i to the next layer j as the following:

$$b_j = \sum_{k}^{k \in A_{ij}} y_{ik}(\mu_{ijk} + \epsilon_{ijk}\sqrt{\alpha_{ijk}\mu_{ijk}^2}) \ , \tag{5.1}$$

where $\epsilon \sim N(0, 1)$; y_{ik} is the output from hidden layer i; μ_{ijk} and α_{ijk} are the learnt parameters of the convolutional filter such that the contribution y_{ik} from each pixel k is scaled by a stochastically sampled value generated from a Gaussian with mean μ_{ijk} and variance $\mu_{ijk}^2\alpha_{ijk}$; A_{ij} is the receptive field of the filter that connects the pixels from layer i to j.

5.2.2 Monte Carlo Dropout

Dropout [170] involves setting a random subset of network weights to zero each time data is fed through a model. Although dropout was originally added to neural networks during training to prevent overfitting, Gal [153] later proposed that it could also be implemented during test time to simulate a Gaussian Process. This method can be seen as equivalent to a specialised version of BBB where the weight follows a Bernoulli distribution. MC dropout is easier to implement than BBB as the former only requires half the number of trainable parameters per convolutional filter and needs no additional coding if the dropout is already applied during training.

5.2.3 Quantifying Uncertainty

Sample averaged BNNs may be able to produce confidence values that scale more linearly with the probability that the model makes a correct prediction. However, interpretation of the statistical distribution of BNN's output is important as it would allow us better understand to what extent the model can be further improved. Given that a BNN has been sampled N iterations during test time, one possible measure to quantify uncertainty is the predictive entropy H [171]:

$$H(\hat{\boldsymbol{\theta}}) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{C}\hat{y}_{ij}log(\hat{y}_{ij}) \ , \tag{5.2}$$

where C are the classes output by the model; $\hat{y}_{ij}$ is the probability for each class j and sample i predicted by a network stochastically instantiated based on some variational parameters $\hat{\theta}$.

Kendall [172] proposed another measure to quantify uncertainty by decomposing the predictive variance from neural networks into aleatoric and epistemic terms. Aleatoric variance measures uncertainty as a result of inherent ambiguity or noise of the data, whereas epistemic variance measures uncertainty due to limitations of the model. An approximate variational predictive distribution $q_{\hat{\theta}}(\hat{y}|x)$ defined by optimised parameters $\hat{\theta}$ can be evaluated as an integral over point predictions $\hat{y}$ from models instantiated by some randomly drawn parameter $\omega \sim P(\hat{\theta})$ as follows:

$$q_{\hat{\theta}}(\hat{y}|x) = \int p(\hat{y}|x,\omega) q_{\hat{\theta}}(\omega) d\omega \ , \tag{5.3}$$

where x is the data input of the model and $\hat{y}$ is the predicted output. It can be shown that this predictive distribution has the following variance [172]:

$$\begin{aligned} Var_q(\hat{y}) =& E_q[\hat{y}^{\otimes 2}] - E_q[\hat{y}]^{\otimes 2} \\ =& \underbrace{\int [diag(E_p[\hat{y}]) - E_p[\hat{y}]^{\otimes 2}] q_\theta(w|D) dw}_{\text{aleatoric}} \\ &+ \underbrace{\int (E_p[\hat{y}] - E_q[\hat{y}])^{\otimes 2} q_\theta(w|D) dw}_{\text{epistemic}} \end{aligned} \tag{5.4}$$

where $v^{\otimes 2} = vv^T$ and $E_q[g]$ is the expected value of function g over the predictive distribution $q_{\hat{\theta}}(\hat{y}|x)$ formulated in Equation 5.3, and D is the data. The original method to calculate the different uncertainties requires an extra output channel in the model. An alternative method proposed by Kwon [158] allows the evaluation of these two uncertainty terms using an unmodified BNN. With Kwon's method, the aleatoric and epistemic variances can be obtained in summation form from the stochastic models' ensemble predictions:

$$\sigma^2_{aleatoric} = \frac{1}{N} \sum_{i=1}^{N} diag(\hat{y}_i) - \hat{y}_i^{\otimes 2} \ , \tag{5.5}$$

$$\sigma^2_{epistemic} = \frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - \bar{y})^{\otimes 2} \ , \tag{5.6}$$

where $\bar{y} = \frac{1}{N}\sum_{i=1}^{N}\hat{y}_i$, and $\hat{y}_i$ is the soft prediction output by model instantiated by sampled parameters ω_i.

5.3 Method

In this section, we summarise the theoretical underpinnings of our methodology and design an experiment to validate the proposed approach. Firstly, We elaborate on several exemplary histopathological measurements derived from proximal renal tubules. After that, we describe methods of obtaining confidence scores from BNNs. Next, we would like to test whether confidence-weighted instances can improve measurements on slides that contain ambiguous tissue instances. To achieve this, we generate a synthetic dataset where subsets of images differ gradually from the original training data. Finally, we measure the distance between various confidence-weighted and non-weighted distributions against the ground truth.

5.3.1 Histopathology Measurements

A quantitative histology pipeline typically involves the segmentation of relevant tissue components, followed by the extraction of features developed based on domain knowledge. However, many of these features and downstream measurements are sensitive to uncertainties in the segmentation masks.

The simplest features would be tissue area measurements, obtained by summing up the pixels that belong to an instance. If the predictions are over-confident, we may end up with fewer, larger instances (under-segmented). A network that produces under-confident predictions may produce more small instances than an adequately calibrated network (over-segmented).

Another useful clinical feature that could be extracted from the renal biopsies is the thickness of the proximal tubules' epithelial cells that outlines the tubules. If the cells are damaged, they could vacuolate and expand, causing the epithelial lining to appear thickened. However, it is often difficult to determine the boundaries

of the epithelial cell lining in damaged tissue. Hence, a workaround is to measure the H&E / PAS stain distribution with respect to the mask's distance transform. We name these measurements y^{γ} and $y^{-\gamma}$, which are calculated as follows:

$$y^{\gamma} = \frac{\sum_{m \in M} d_m x_m}{\sum_{m \in M} d_m \sum_{m \in M} x_m} \ , \tag{5.7}$$

$$y^{-\gamma} = \frac{\sum_{m \in M} (d_M^{max} - d_m) x_m}{\sum_{m \in M} (d_M^{max} - d_m) \sum_{m \in M} x_m} \ , \tag{5.8}$$

where d_m is the distance transform of an instance's mask measured from the edge; x_m is the pixel value of the image bounded by mask M of the instance, and d_M^{max} is the maximum value of the distance transform d for each instance. The terms in the denominator of both equations are normalisation terms that reduce the dependency of y^{γ} and $y^{-\gamma}$ on the tissue's size and mean pixel values. An intuitive illustration of y^{γ} is shown in Figure 5.1 - instances with a clear lumen would result in higher values of y^{γ} as the white pixels would coincide with pixels with large d. Instances with pink pixels near the centre would tend to have smaller y^{γ} - this could be indicative of loose cells, a feature of damaged tissues. As an alternative to the feature y^{γ}, $y^{-\gamma}$ describes stain distribution with respect to the centre of the instance. While y^{γ} allows comparison between instances of different areas, the measurement $y^{-\gamma}$ may be more robust to over/under confident predictions of the tissue boundaries.

As we are measuring the 2D cross section of 3D objects, area measurements of the tissue structures on their own may not be a very informative histological feature. To augment such limitations, we can also measure the tissue's aspect ratio. This could help us estimate the plane of how the tissue was cut. This measurement could be determined from the major and minor axis of the minimum-area rectangle fitted to the segmentation mask. We present instance area, y^{γ}, and major axis measurements in the following experiment.

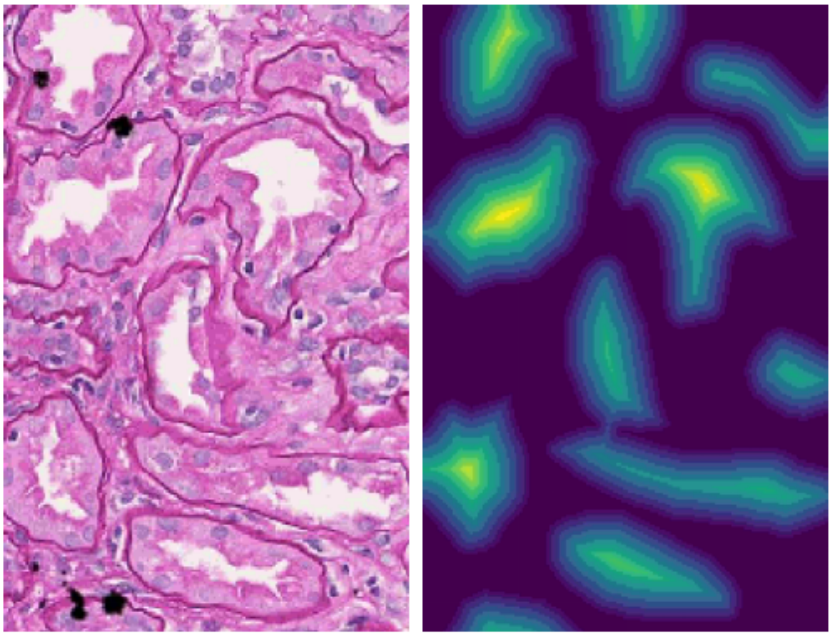


Figure 5.1: Illustration of the calculation of y^{γ}. Pixel values x of each tubule is multiplied by their distance transform d, summed and normalised as described in Equation 5.7 . Instances with thin epithelial cells and a large lumen would have higher values of y^{γ} as the value of x would be large (white) where d is also large.

5.3.2 Proposed Method for Improving Histological Measurements

Previous studies that explored the quantification of segmentation uncertainty on medical images [156, 158, 161–163] were tested on modalities where the images contain a fixed or known number of object instances. While this is often the case for MRI/CT images, the number of objects could vary greatly between different slides in histopathology, making an instance-to-instance comparison of tissues impossible. Hence, the most sensible approach to analysing these slides is to compare the measurement distribution between different slides.

It had been previously demonstrated that such a distribution estimate could be improved by weighting each instance by some confidence measure [173]. However, from our understanding, this approach has only been applied to hand-crafted segmentation algorithms, and the choice for confidence weights was arbitrary.

This chapter proposed and explored several approaches to weighting instance-wise histological measurements by confidence obtained from Bayesian Neural Networks.

Consider an object instance x in a dataset characterised by a distribution $x \sim P(X)$. Suppose we want to derive some metric estimate y from x by applying some function f, we would have the following:

$$y = \tilde{y} + \epsilon = f(x) + \epsilon(x, f) \ , \tag{5.9}$$

where $\tilde{y}$ is the estimator of y, ϵ is the error of f, which could depend on both f and x (heteroscedastic). If we have a dataset where we know the ground truth of y for all x, the easiest way to demonstrate that f is a good estimator would be to simply calculate ϵ directly. However, in many cases, the number of objects N in the dataset is also subjected to other sources of errors. Due to over/under-segmentation errors, we cannot be certain that a particular segmented object $\tilde{x}$ corresponds to the ground truth instance x. In this case, we may only be able to quantify the performance of f by using some similarity measures $M(P(Y), \tilde{P}(Y))$ on the distribution estimate $\tilde{P}(Y)$ against the ground truth's distribution $P(Y)$ where

$$P(Y) = \int_{x \in S} [f(x) + \epsilon(x, f)] P(x|X) dx \ , \tag{5.10}$$

And in summation form

$$\int_{y}^{y+\Delta y} P(y) dy \approx \frac{1}{N} \sum_{k}^{k \in S} [f(x_k) + \epsilon(x_k, f)] \ , \tag{5.11}$$

In this case, the corresponding distribution estimate would be

$$\tilde{P}(Y) = \int_{x \in \tilde{S}} f(x) P(x|X) dx \ , \tag{5.12}$$

And in summation form:

$$\int_{\tilde{y}}^{\tilde{y}+\Delta \tilde{y}} \tilde{P}(y) d\tilde{y} \approx \frac{1}{\tilde{N}} \sum_{k}^{k \in \tilde{S}} f(x_k) \ , \tag{5.13}$$

$S = \{k \in \mathbb{N} | (x_k \in D) \wedge (y < f(x_k) < y + \Delta y)\}$ is the set of instances in our dataset with measurable parameters between y and $y + \Delta y$; $\tilde{S} = \{k \in \mathbb{N} | (x_k \in D) \wedge |(\tilde{y} < f(x_k) < \tilde{y} + \Delta \tilde{y})\}$ is the corresponding set of detected instances.

In general, ϵ may depend on the input data x (heteroscedasticity) and the function f. However, if we assume that ϵ does not strongly correlate with y, then it may be possible to obtain an improved distribution estimator $\tilde{P}_g$ by weighting the instances by some confidence measure $g(x, f)$ as follows:

$$\tilde{P}_g(Y) = \frac{\int_{x \in \tilde{S}} f(x) g(x, f) P(x|X) dx}{\int_{x \in \tilde{S}} g(x, f) P(x|X) dx} \ , \tag{5.14}$$

$$\int_y^{y+\Delta y} \tilde{P}_g(y) dx \approx \frac{\sum_k^{k \in \tilde{S}} f(x_k) g(x_k, f)}{\sum_k^{\tilde{N}} g(x_k, f)} \ , \tag{5.15}$$

In renal histopathology, x could be a tissue compartment such as a tubule or glomerulus. In this case, y would be a clinical feature such as area or texture. The number N depends on the size of the whole slide images, and its estimate $\tilde{N}$ could depend on the exact details of how the instances are counted. Fortunately, in digital pathology, we are often only interested in the distribution of features within each slide, so getting an accurate estimate of $\tilde{P}(Y)$ may already be sufficient.

In our practical example, the weight g could simply be a normalised function derived from the entropy H (Equation 5.2). Such that, for an instance k we have the following:

$$g_k^H = 1 - \frac{1}{|M_k| log(C)} \sum_{m \in M_k} H_m \ . \tag{5.16}$$

where M_k is the predicted segmentation mask for instance k with pixels m.

Alternative, we could also derive g from the aleatoric or epistemic standard errors (Equations 5.5) as follows:

$$g_k^\sigma = max \left(1 - \frac{1}{|M_k|} \sum_{m \in M_k} \sigma_m(w), 0 \right) \ , \tag{5.17}$$

where σ_m is either the aleatoric or epistemic standard error at a given pixel m. We use Kwon *et al.*'s [158] implementation to evaluate σ_m in our experiment. The

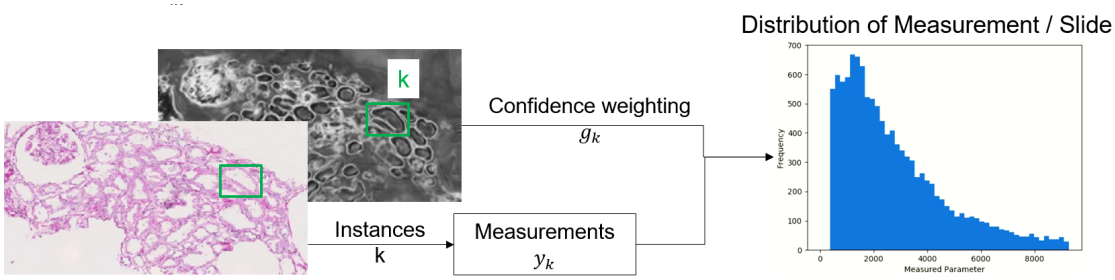


Figure 5.2: Schematic of obtaining confidence-weighted histological measurements from tissue instances k. The confidence weighting g_k could be obtained from BNNs' uncertainties as proposed in Equations 5.16 and 5.17.

confidence weighting in Equations 5.16 and 5.17 can be applied to each instance to allow us to obtain a corrected distribution, as shown in Figure 5.2's schematic.

5.3.3 Datasets

The dataset we use for training originated from an organ normothermic machine reperfusion experiment [120, 121]. The slides were obtained from kidneys deemed unsuitable for transplantation, and most samples suffered from significant cold ischemic damage. The tissues were stained with H&E but had been computationally converted to PAS stain using a CycleGAN [98], with image quality largely indiscernible by our collaborating pathologist (described in Section 3.5.1). The training dataset consists of 1642 patches from 169 slides containing 13,103 manually annotated tubules.

Our test dataset consists of 144 PAS-stained slides obtained from the QUOD Oxford biobank from kidneys that had been transplanted. We annotated 30 patches with a total of 1741 tubules in this dataset.

5.3.3.1 Generation of Synthetic Dataset

For the purpose of our experiment, we also synthesise a dataset that deviates progressively from the training data. We would like the data to fulfil three requirements. Firstly, the images need to contain tissue structures that exhibit diverse morphological, textural, and colour variations. Secondly, the images need to be accompanied by a corresponding segmentation mask to be used as the ground truth. Thirdly, to simplify how these images are generated, we would like to control

these visual changes using a single parameter. While the simplest way to generate controlled morphological changes is to deform the images and masks simultaneously, this would provide only limited morphological variations.

Consequently, we build a pipeline to generate synthetic images that fulfil the above requirements using a Variational AutoEncoder (VAE) [90, 174] and a BicycleGAN [88]. A simple schematic and some examples of the synthetic images are shown in Figure 5.3. The VAE learns the underlying morphological variations of the relevant renal tissues, whereas the BicycleGAN is used to convert masks from the VAE to a diverse set of realistic-looking images. To obtain images that progressively differ from the real data, we sample the VAE 10 times with a latent vector $\sigma_{VAE} \in [0.0, 1.8]$. This produces masks with varying morphological changes. Next, the BicycleGAN is used to generate 6 images from each mask, one of which is generated from the encoded vector from the real image (BGAN0), and the other 5 are generated from a sampled latent vector (BGAN1).

In order to produce tubules with morphological diversity, we train a VAE to learn the underlying morphological variations of the relevant renal tissues. The inputs to our VAE are the hand-annotated masks of size $256 \times 256 \times 5$ ($0.88\mu m/pixel$), with the five segmentation classes being the background, tissue boundaries, tubules, glomeruli, and vessels. We only use the first 3 classes for the remaining experiment (as labelled in Figure 5.3). The tissue boundary class is generated by edge filtering other foreground classes (tubules, glomeruli, and vessels) as our earlier experiments suggested that a distinct boundary class could help us to separate touching instances. The encoder uses convolutional layers with a kernel size of 4 (5 in the last layer) with leaky ReLU [175], giving a latent representation z of size $8 \times 8 \times 500$. The decoder construction is symmetrical to the encoder's, where up-sampling is achieved using fractionally-strided convolution layers. The schematic of our VAE is shown in Figure 5.4.

As with the training of typical VAEs, we need to balance its reconstruction loss versus the regularisation of the latent vector. To produce realistic reconstruction, cross-entropy (CE) and structural similarity (SSIM) [125] losses are implemented

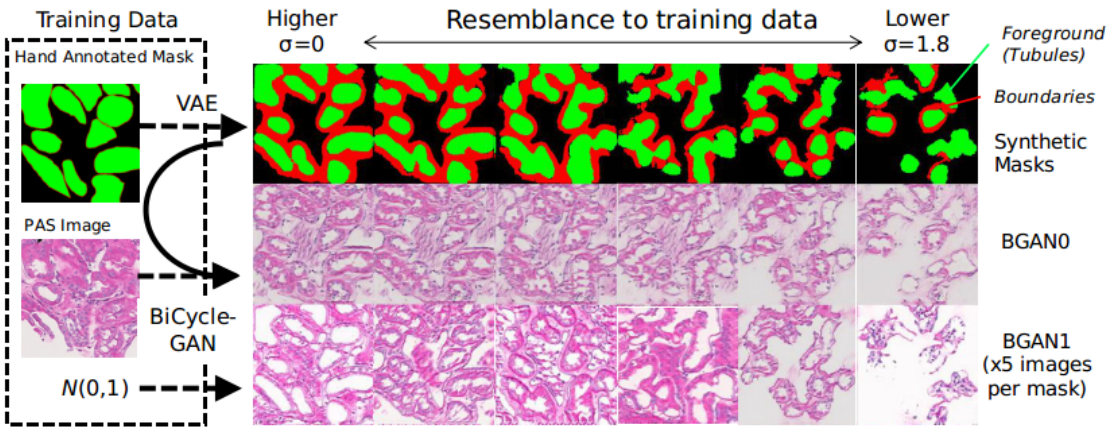


Figure 5.3: Schematic for the generation of synthetic images. We train a VAE to generate synthetic masks at varying σ_{VAE} based on the encoded embedding (1st row). By sampling the latent vector radially at different σ from the original mask, we obtain masks that are increasingly dissimilar to the original. A BicycleGAN is used to convert these masks to sets of images with colour and textural diversity without significant changes in morphology.

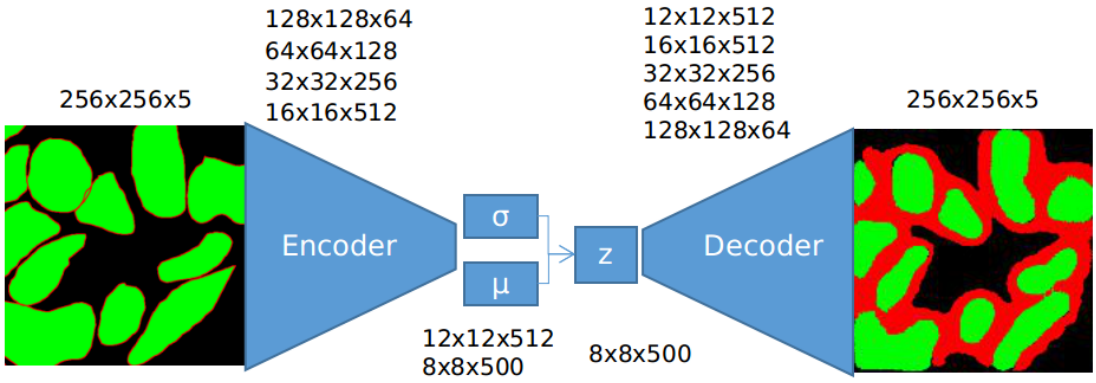


Figure 5.4: VAE used for generating a synthetic segmentation masks for the histology images. The size of the latent layers is shown above the figure.

during training. We use a CE loss weight of 10:50:10:20:20 for the five aforementioned segmentation classes respectively, a SSIM weight of 1.0, and evidence lower bound (ELBO) weight of 10^{-2} to restrict the latent vector to a unit Gaussian prior. We notice that larger weighting for the boundary class would generate masks with thicker boundaries, but the output would also vary more smoothly with σ_{VAE} as we resample the latent vector.

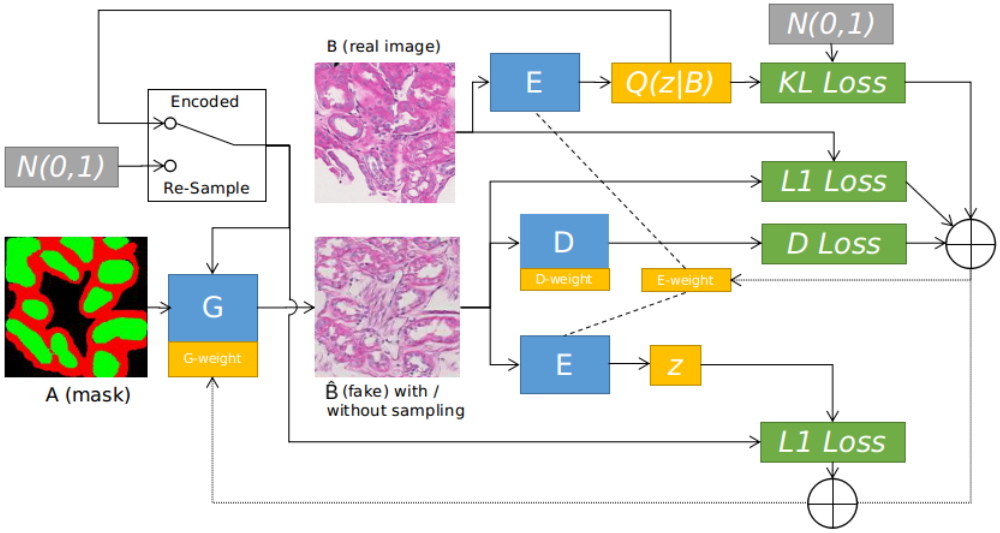
Next, we would like to produce realistic looking images based on the mask output from the VAE. One way to convert segmentation masks into images is to use a pix2pix network [89, 176], but the output from this model would be deterministic.

However, in our segmentation masks, several non-relevant tissues, such as interstitial tissue, are also represented as the background class. A deterministic output may collapse the reconstruction of the background. In order to generate a diverse set of images, we train a BicycleGAN [88] using the VAE's reconstructed masks ($\sigma_{VAE} = 0$, i.e. without latent vector resampling) paired with the images. A schematic of its training is shown in Figures 5.5a and 5.5b. The generator G of the BicycleGAN is trained to convert masks A to images $\hat{B}$ while being conditioned to either a normal latent vector $N(0, 1)$ or an encoded embedding $Q(z|B)$ generated by the encoder E from the real image B. The purpose of E is to learn a smooth embedded representing the style of the images, while the discriminator D learns a discriminative loss to discern the fake images $\hat{B}$ from real ones B. We find that the default implementation from the original paper [88] works well with our dataset.

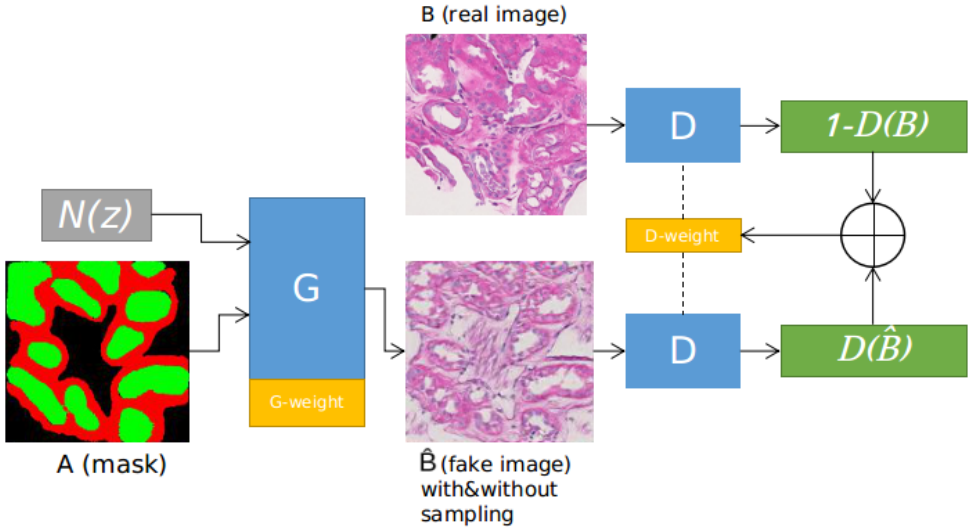
The combination of VAE and BicycleGAN allows all synthetic images and masks with $\sigma_{VAE} \geq 0$ to be generated from a single training. While our VAE-BicycleGAN pipeline appears to be capable of generating images with decent visual quality, a significant number of reconstructions still fail to resemble the original input images. In order to ensure the quality of the output, we select only a subset of masks generated from the VAE as follows:

We apply a pre-trained UNet ($UNet$) on both the original image B and the reconstructed image $\hat{B}$ (no latent vector sampling in either the VAE and BicycleGAN). The L1 difference between the soft predictions is calculated between each pair of soft predictions and is plotted in Figure 5.6, where each count in the histogram is derived from $UNet(B_i) - UNet(\hat{B}_i)$ for an image tile i. Ideally, the segmentation mask generated from the original image should be similar to that generated from a reconstructed image. Discarding image patches above a certain L1 threshold allows us to remove many cases where the UNet predicted the existence of tubules in patches that did not originally contain the tissue.

We obtain the tubule instances of the entire synthetic dataset from the UNet and calculate the area distribution to determine this threshold. We evaluate the Kullback–Leibler (KL) divergence of the area distribution from the ground truth's



(a) Schematic of the training of the BicycleGAN's Encoder (E) and Generator (G)



(b) Schematic of how the BicycleGAN's Discriminator D is trained.

Figure 5.5: Schematic of how the BicycleGAN's Encoder (E), Generator (G), and Discriminator (D) are trained. G takes a latent vector and a segmentation mask as its input. At the same time, E is trained to output a latent vector from an image. The latent vector input to G could either be a random sample or an encoded vector from $E(B)$. In the latter case, we calculate the L1 loss between the encoded vectors $(Q(z|B), z)$ from the real and fake images to encourage G to embed this latent information into $\hat{B}$.

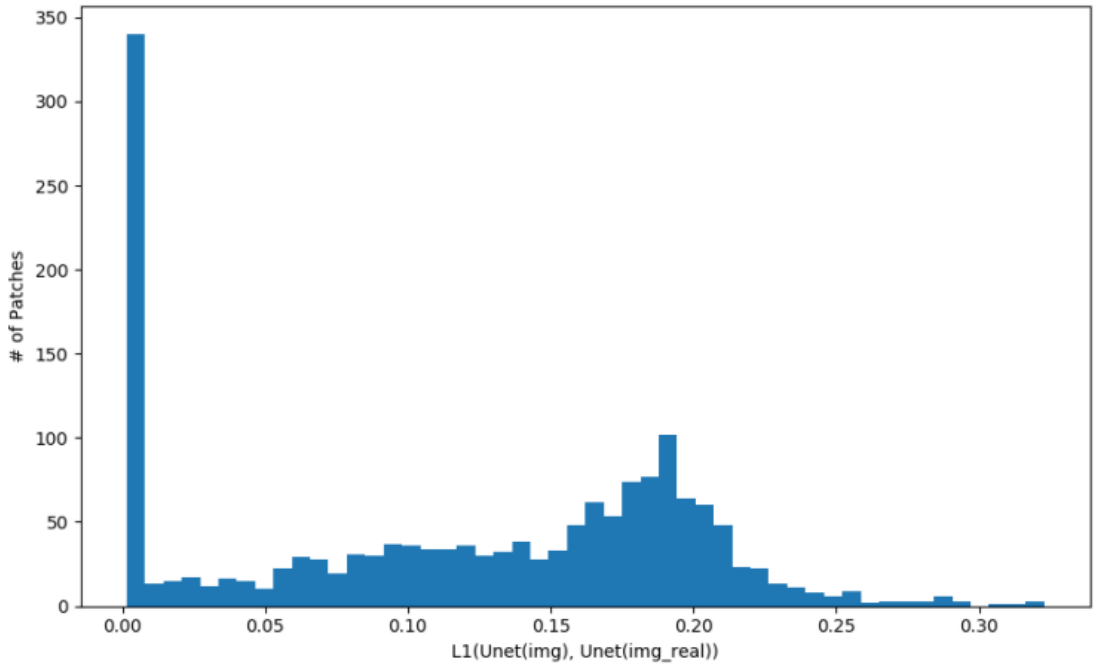


Figure 5.6: Distribution of L1 between semantic predictions generated by a UNet on the original image vs on the encoded synthetic image output by BicycleGAN. Each unit on the y-axis represents a pair of images. A total of 50 bins are used in this histogram. Note: the bin on the left is caused by real images that contain no relevant tissue compartments (i.e. the segmentation mask is blank). In many of these cases, the BicycleGAN also generates tissue patches containing no objects of interest under reconstruction mode hence L1=0.

distribution. Then we plot how the KL changes with respect to the L1 cutoff. The results are shown in Figure 5.7. We choose a threshold that would discard 60% of the synthetic dataset based on this plot. This allows us to select 39,400 better quality images from 98,400 synthetic tiles for our subsequent experiment. An illustration of some of the chosen images is shown in Figure 5.8.

5.3.4 Tissue Segmentation

To demonstrate that our method is robust to the choice of hyperparameters, we experiment on two different networks. We train a UNet [87] with Monte Carlo Dropout [153] of 0.2 in all but the first and last layer, and another UNet using Bayes by Backprop with local reparameterisation trick (BBB_LRT) [154, 168, 169] using a unit Gaussian prior to replacing the dropout layers. We train the networks with 3-fold cross validation while ensuring that tiles from the same slide

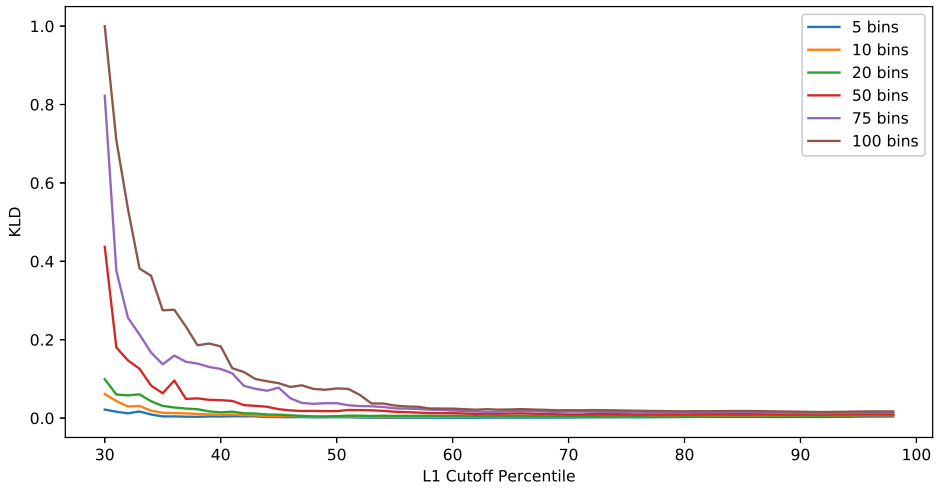


Figure 5.7: Kullback Leibler divergence between the distribution of instance areas of the hand-annotated masks versus the predictions. The x-axis of the figure represents the percentage of synthetic images discarded based on L1 measures such as that in Figure 5.6. The current figure also shows how KL varies with bin size (For instance, $n = 50$ bins are used in Figure 5.6). For the rest of the experiment, we have decided to discard 60% of masks generated by the VAE-GAN as they appear too different from the original masks even without latent vector resampling.

are in the same split. After that, we obtain 5 Bayesian samples for each network over each of the 3 cross validation models and take the mean softmax over the 15 predictions. We obtain instances using the max-flow-min-cut graph algorithm [139] and classical morphological operations.

We base our experiment on a standard U-Net [87], as detailed in Table 4.1 in Chapter 4. The architecture is modified from the original [87] based on our earlier experiments which used larger image patches to contain approximately 50.3M trainable parameters. In the context of this experiment, however, the networks are trained and evaluated on small cropped patches of 256×256 pixels ($0.88\mu m/pixel$). In the MC Dropout implementation, we use dropout in every block except Block 1 and 14 with $m = 16$. In the BBB_LRT implementation, we replace normal convolutional layers with BBB_LRT convolutional layers on every block except Block 1 and 14 with $m = 8$ and remove all dropout layers. Halving the number of filters in the BBB_LRT model allows us to keep the number of parameters similar to the MC Dropout implementation.

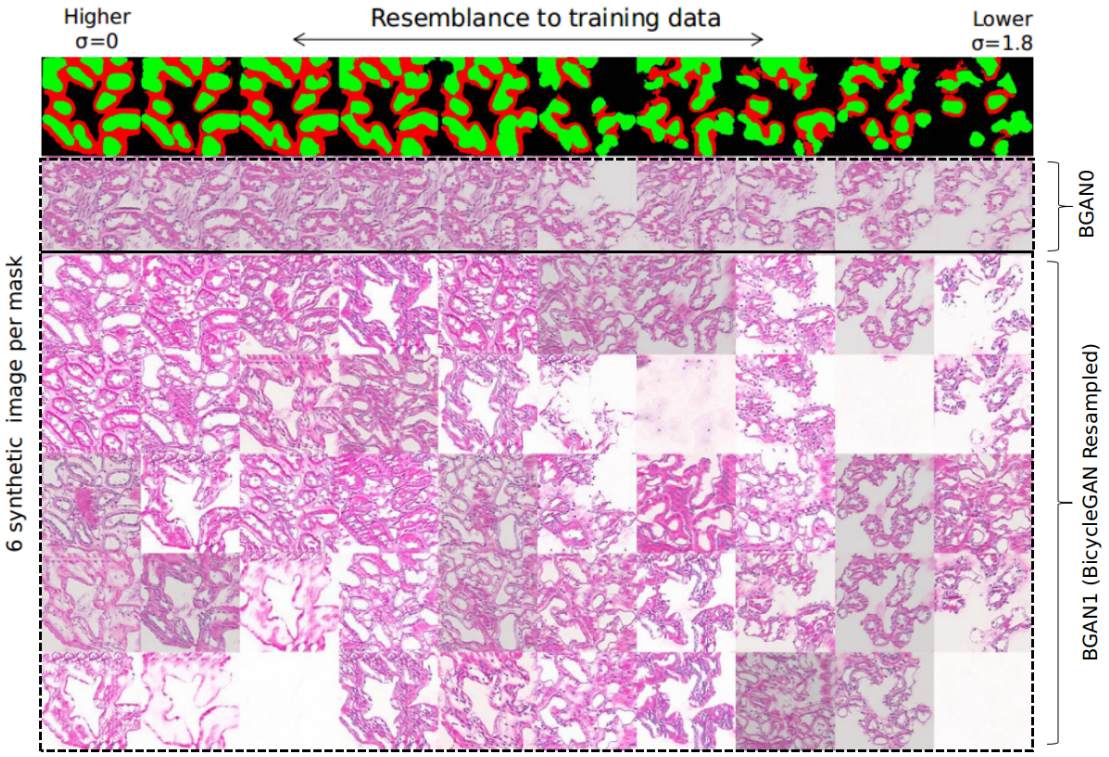


Figure 5.8: Examples of Synthetic Images. This figure shows further examples of synthetic images generated using the proposed VAE-BicycleGAN schematic. The $2^{nd} - 7^{th}$ rows show examples generated by the BicycleGAN based on the mask directly above. We generate 6 images per mask, including 1 generated using the encoded vector from the real image (2nd row, BGAN0); the other 5 are generated from a sampled vector ($3^{rd} - 7^{th}$ rows, BGAN1) to produce images with diverse tissue texture and background.

For the training of the BBB_LRT UNet, instead of sampling the weights of the model, we sample the activation layers using the local reparameterisation trick (LRT) [168] with a Gaussian prior. We add a KL loss term with a weight of 0.005 to regularise the variational parameters.

We train and validate our model 3-folded over the training patches for each mini-experiment. Since a different number of tiles are extracted from each whole slide image, the cross validation splits are randomised such that no patches from the same parent slide are divided into different subsets. We use focal loss [138] with hyperparameter $\gamma = 1$ and a class weighting of 5:100:5:10:50 for the five segmentation classes. The dataset is augmented with flipping, arbitrary rotation, scaling, shear, elastic deformation, and colour/contrast enhancement. Training is performed over a number of epochs until the validation intersect-over-union (IOU)

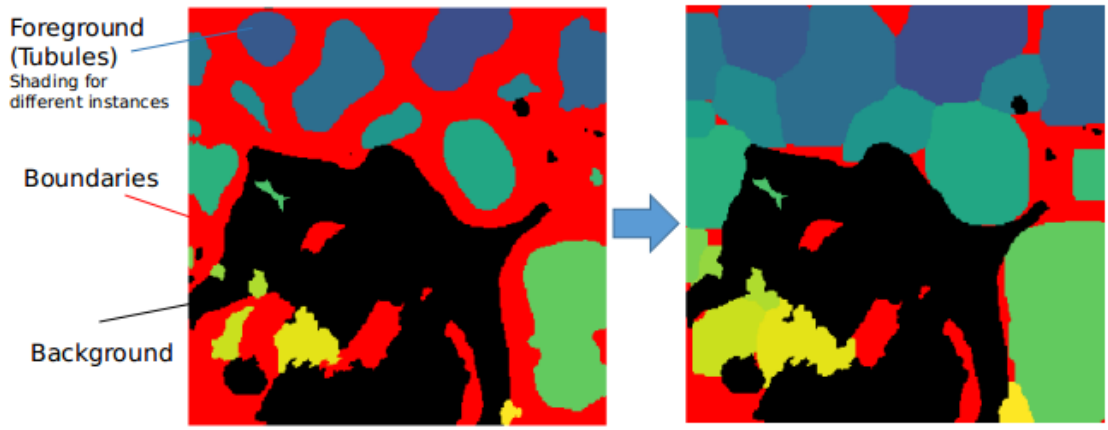


Figure 5.9: Schematic showing how tubule instances are obtained. The figure on the left shows the pre-processed binarised predictions where different tubule instances are labelled in shades of blue/green/yellow. We loop through the instances iteratively and morphologically dilate each mask with a 3x3 kernel until the instance exceeds the boundary or touches another instance. The figure on the right shows an intermediate result of the processed instances.

for all foreground (Tubules, Glomeruli, Vessels) classes start to fall off. As per an earlier experiment, validation accuracy for glomeruli and vessels tend to peak much later than for tubules. We select the model based on the latest epoch in which the validation accuracy is maximum for at least one of the foreground classes. However, we only performed the experiment on the tubules in this experiment as it is the most abundant tissue in our dataset.

5.3.4.1 Instance Segmentation

After obtaining the semantic predictions, each foreground class is processed using the max-flow-min-cut graph algorithm [139], resulting in a binary mask for each class. As tissue boundaries (labelled in Figure 5.9) are classified as a distinct class, masks for each tubule are generally well separated. Each connected component of the tubule class is treated as an instance. We iterate through each instance multiple times and dilate its mask until the instance touches another instance or exceeds the mask of the boundary class. The dilation does not extend into pixels of the background class. Instances smaller than 400 pixels are discarded as we cannot recognise any tissues smaller than this size. If we include these small objects in the analysis, the results shift further in favour of our proposed method.

5.4 Experiment Results

The tubule instances' area, y^{γ}, $y^{-\gamma}$, minor axis, and major axis measurements are obtained as described in Section 5.3.1 on the synthetic and the real test dataset. We evaluate the performance of different versions of our proposed method by comparing the various weighted and unweighted estimated measurements against the ground truth. Here, we present most comparisons of probability distributions (PDF) using the first Wasserstein Distance (WD) [177]. Other metrics such as Kullback–Leibler (KL) and Jensen-Shannon (JSD) distance could also be used and would result in a similar experiment conclusion. However, WD is most suitable as it considers the metric space and does not approach infinity where the PDFs are disjoint.

We initially expect WD to be higher for the examples with higher σ_{VAE}, where the tissue instances increasingly differ from the training images. However, we observe the opposite trend in most cases. An example of this is plotted in Figure 5.10. To understand this counter-intuitive trend, we inspect the images and observe that instances tend to shrink in size as σ_{VAE} increases. Although the instances become morphologically less plausible with increasing σ_{VAE}, having fewer touching instances might have reduced ambiguity in the segmentation. This observation is illustrated on the top row masks in Figure 5.8. Because no drop in performance is observed in the examples with large σ_{VAE}, we conclude that there is no evidence of overfitting and our UNets appear to generalise well to unseen morphology. The hyperparameters of the UNets are empirically tuned for the MC Dropout implementation. Despite this, we find that the IOU of the BBB_LRT implementation is marginally better. Further tuning of the latter will be required if we are to present meaningful comparisons of the two BNNs.

Next, in order to evaluate the improvement of our proposed method, we define an evaluation metric ΔWD as follows:

$$\Delta WD = WD(\tilde{P}(Y), P(Y)) - WD(\tilde{P}_{Baseline}(Y), P(Y)) \ , \tag{5.18}$$

where WD is the function for calculating the Wasserstein metric; $\tilde{P}(Y)$ is the estimated PDF of a histological measurement; $P(Y)$ is the ground truth PDF;

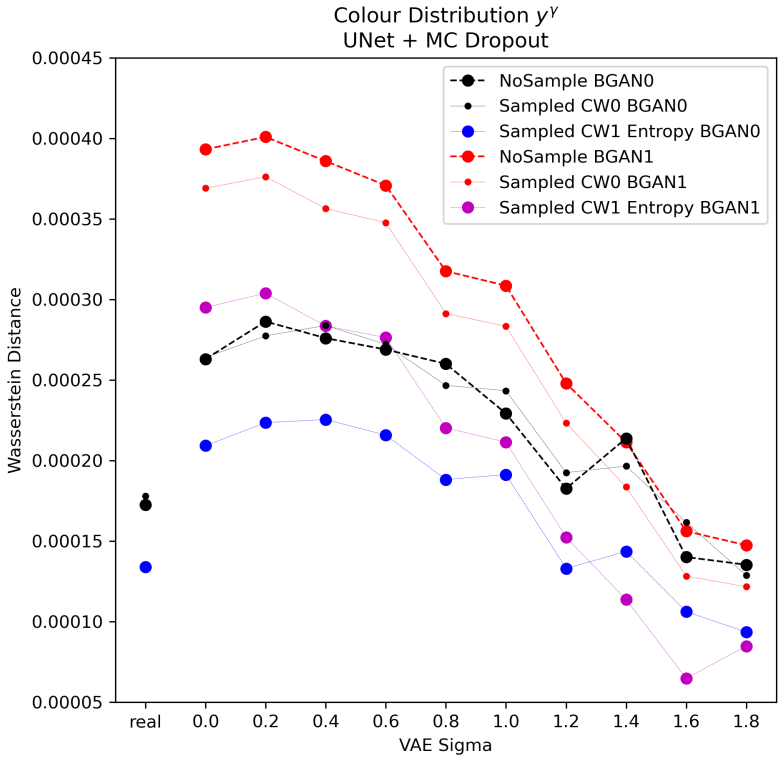


Figure 5.10: Example of Wasserstein Distance (WD) plots comparing distributions between measurements derived from the predicted instances vs measurements derived from the ground truth masks. This plot is generated from UNets trained with MC Dropout for the histological measurement y^γ. The x-axis shows how WD changes as the instances differ from the training data at varying σ_{VAE}. For comparison, the real test data is also added to the left-most tick on the x-axis. Legends with *NoSample* denotes measurements derived from UNet without MC Dropout averaging at test time; *CW0/1* stands for confidence unweighted/weighted respectively; *BGAN0/1* stands for images with/without BicycleGAN resampling respectively. The *CW1* lines are weighted by the entropy-derived confidence (Equation 5.16). Note that WD tends to decrease with σ_{VAE}.

$\tilde{P}_{Baseline}(Y)$ is the baseline estimated PDF derived from instances segmented by the UNets without MC Dropout. We plot ΔWD values in Figure 5.11. Each plot shows the results of a different measurement by the UNet with either MC Dropout (*a.i, b.i, c.i, d.i, and e.i*) or BBB_LRT (*a.ii, b.ii, c.ii, d.ii, and e.ii*). The x-axes show the different subsets of the synthetic data ($\sigma_{VAE} \in [0, 1.8]$) and the real test data (*real*). A more negative ΔWD represents a greater improvement.

For the confidence-weighted (*CW1*) measurements, we weight each instance by confidence derived from entropy (Equation 5.16), aleatoric, and epistemic uncertainty (Equation 5.17) to obtain the estimated PDF. We also split our synthetic dataset by

quality - *BGAN0* denotes synthetic encoded images and is better quality than those generated using a sampled vector with the BicycleGAN (*BGAN1*). The weighted measurements always appear better than the unweighted (*CW0*) for both networks. Instances weighted by entropy generally perform the best, followed by those weighted by aleatoric uncertainty. Instances weighted by epistemic uncertainty receive the least improvement because the dominant uncertainty in our dataset is aleatoric. Unsurprisingly, we also find that improvements in correlated measurements y^{γ}, $y^{-\gamma}$ appear to be similar (Figures 5.11 b and c).

In order to compare improvement across different measurements, we normalise the WD metric as follows:

$$WD_{norm} = (WD_{CW1} - WD_{CW0})/WD_{CW0} \ , \tag{5.19}$$

where WD_{CW1} is the confidence-weighted WD; WD_{CW0} is the unweighted WD. Table 5.1 shows the entropy-weighted normalised WD averaged over different σ_{VAE}. It can be seen that improvement is greater on the image subset with poorer quality (BGAN1) than better quality (BGAN0). A similar trend can be seen in Figure 5.11 where greater improvement can be seen on the synthetic dataset than on the real data. These results suggest our approach becomes more valuable the more different the test data deviates from the training data.

Furthermore, if we take the mean improvement over different models and datasets, we can see that improvement appears greatest for $y^{-\gamma}$ (-0.312) and y^{γ} (-0.310). Improvement is smallest on major axis measurements (-0.155). We reason that the improvement is dependent on the measurement's sensitivity to segmentation error. Intuitively, $y^{\pm\gamma}$ is the most sensitive to segmentation errors. If we assume the tissue is a circular disc with radius r and uniform pixel values, then the incremental change in y^{γ} would contain a term that depends on r^2; whereas the incremental error in tissue area would scale linearly with r (i.e. $dA = 2\pi r dr$). Major axis measurements are generally more robust to segmentation errors as it requires a lot of misclassified pixels to change the size of an instance's bounding box. Hence $y^{\pm\gamma}$ would benefit the most from confidence weighting.

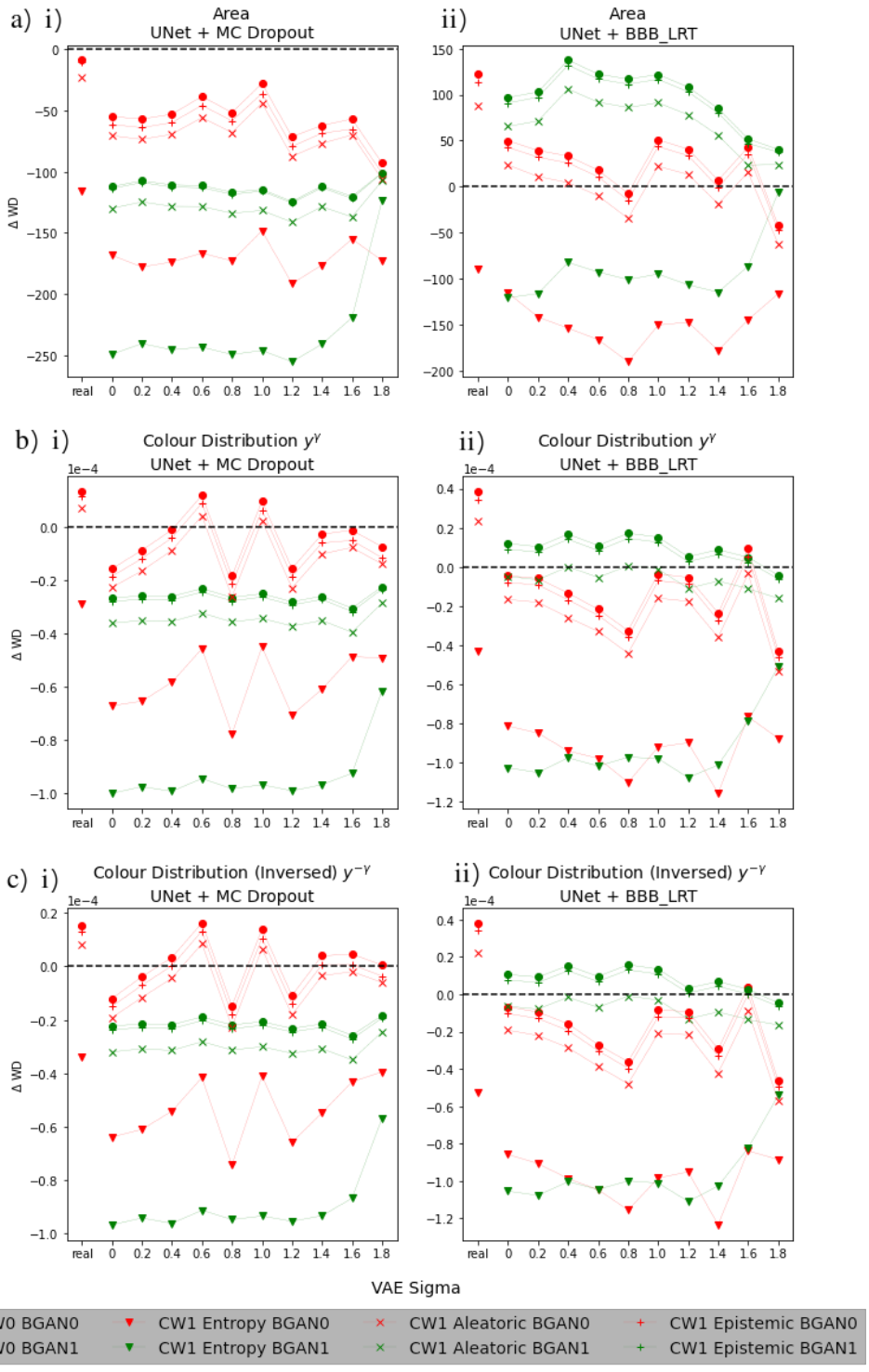


Figure 5.11: Comparison of Measurements. Plots show ΔWD comparing distributions between measurements derived from predicted instances vs derived from the ground truth masks. The x-axes show how the improvement changes as instances differ from the training data at varying σ_{VAE}. The test data (labelled "real" in the graph) is also added to the plot for comparison. Values are relative to the baseline. Lower ΔWD means greater improvement. *CW1/0* stands for confidence weighted/unweighted respectively. *BGAN1/0* stands for images with/without BicycleGAN resampling respectively.

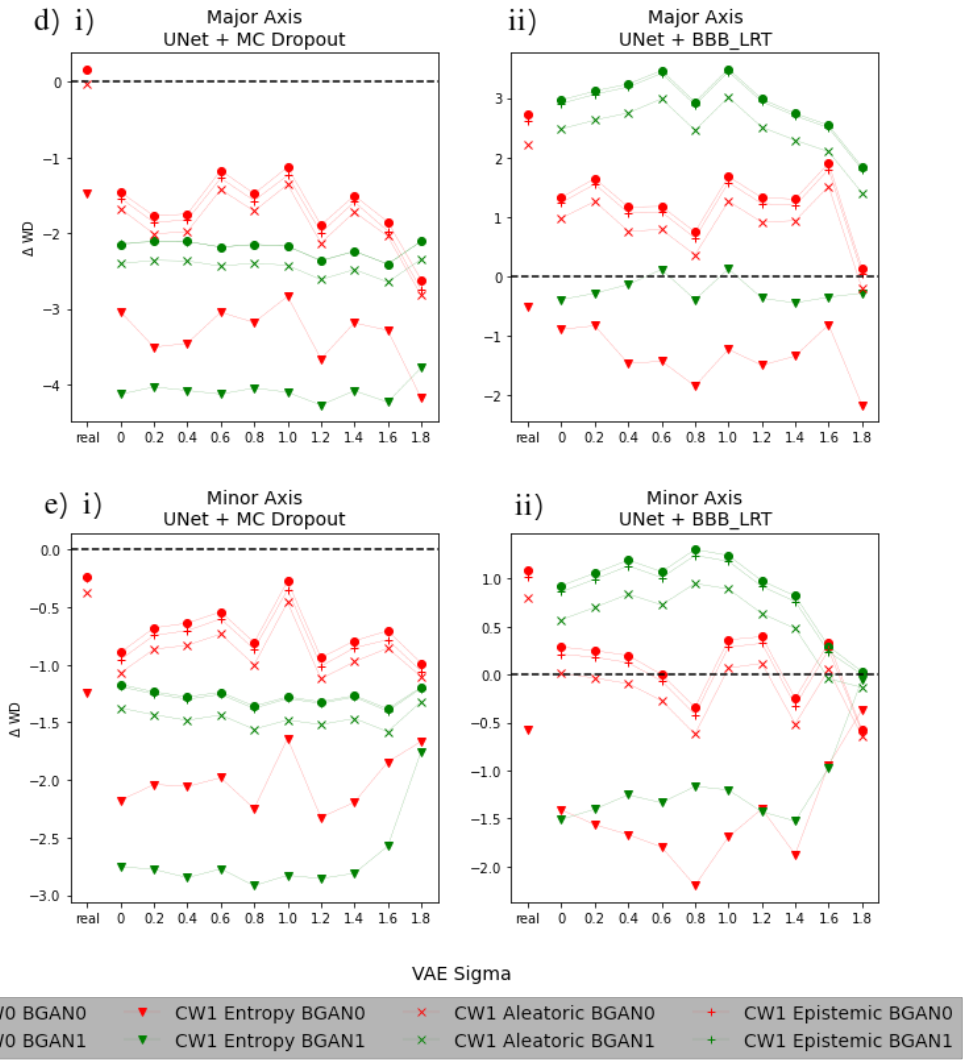


Figure 5.11: Comparison of Measurements. Plots show ΔWD comparing distributions between measurements derived from predicted instances vs derived from the ground truth masks. The x-axes show how the improvement changes as instances differ from the training data at varying σ_{VAE}. The test data (labelled "real" in the graph) is also added to the plot for comparison. Values are relative to the baseline. Lower ΔWD means greater improvement. *CW1/0* stands for confidence weighted/unweighted respectively. *BGAN1/0* stands for images with/without BicycleGAN resampling respectively.

To demonstrate that our proposal could deliver reliable improvement, we also present our results in JSD. We use an equation analogous to Equation 5.19 (substituting WD with JSD) to obtain a normalised metric. The results are shown in Table 5.2. Again, we can see consistent and largest improvements in $y^{\pm\gamma}$ (-0.0977 and -0.100). Entries in the table are mostly negative, with several positive values suggesting that there may be little or no improvement for the major/minor axis measurements. However, as we have noted earlier, JSD tends to behave poorly when PDFs are disjoint, so, overall WD is the preferred metric.

Table 5.1: WD improvement in confidence-weighted distribution estimates. We compare how our approach performs on different BNNs, image quality, and measurements. (Lower is better)

	MC Dropout		BBB_LRT		
Measurement	BGAN0	BGAN1	BGAN0	BGAN1	Mean
Area	-0.136	-0.158	-0.214	-0.229	-0.184
y^{γ}	-0.246	-0.288	-0.337	-0.369	-0.310
$y^{-\gamma}$	-0.251	-0.291	-0.340	-0.367	-0.312
Major axis	-0.106	-0.124	-0.188	-0.202	-0.155
Minor axis	-0.167	-0.199	-0.232	-0.267	-0.216
Mean	-0.181	-0.212	-0.262	-0.287	

Table 5.2: JSD improvement in confidence-weighted distribution estimates. We compare how our approach performs on different BNNs, image quality, and measurements. (Lower is better)

	MC Dropout		BBB_LRT		
Measurement	BGAN0	BGAN1	BGAN0	BGAN1	Mean
Area	-0.0742	-0.0782	-0.0962	-0.1000	-0.0872
y^{γ}	-0.0833	-0.0786	-0.121	-0.108	-0.0977
$y^{-\gamma}$	-0.0856	-0.0861	-0.124	-0.106	-0.100
Major Axis	-0.0519	-0.0123	-0.0693	0.0014	-0.0330
Minor Axis	-0.0206	0.0172	-0.0192	0.0273	0.0012
Mean	-0.0631	-0.0476	-0.0859	-0.0571	

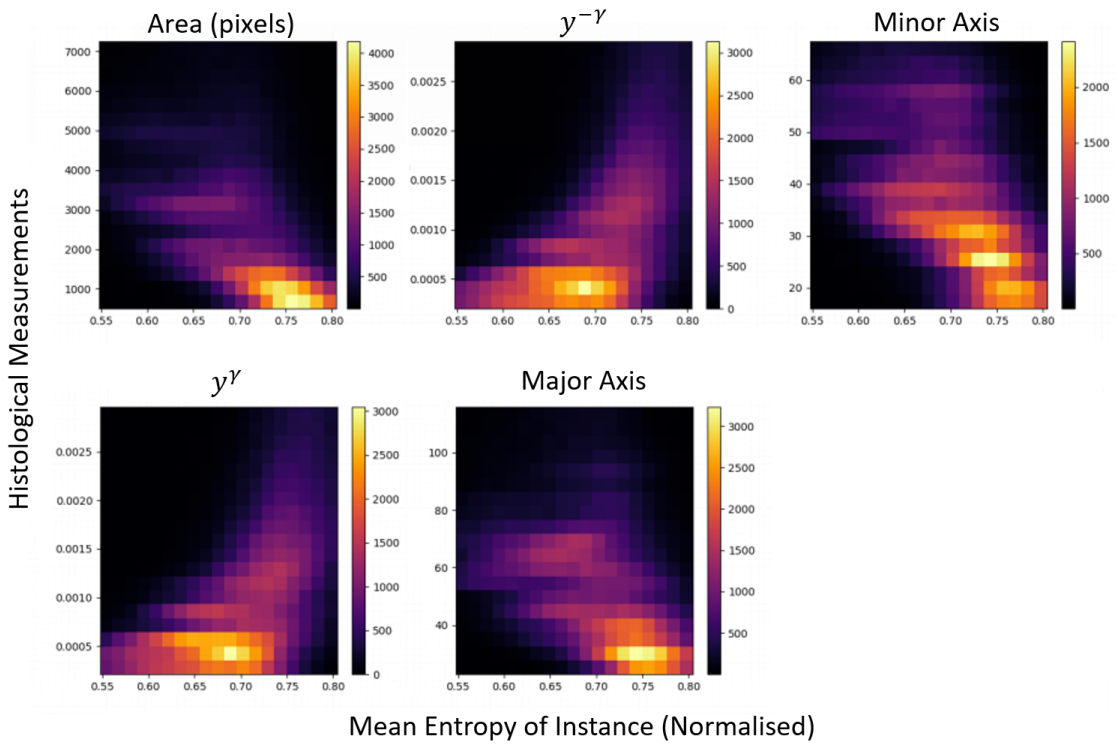


Figure 5.12: Distribution of tissue instances as a function of entropy and various histological measurements. The plots are generated from the UNet + MC Dropout implementation. The three histograms show the distribution of tissues as a function of area, y^{γ}, and major axis, respectively, against the mean entropy of the instance.

5.4.1 Relationship between entropy and histological measurements

Figure 5.12 shows histograms of how the measured parameters of individual instances correlate with their mean predictive entropy. From the plot on the top left, it can be seen that smaller instances tend to have higher entropy. This is because uncertainty is the highest near the edge of the instances. A similar relationship can be seen in the major axis plot. A slightly weaker but still significant correlation can be seen in the $y^{\pm\gamma}$ plots as y^{γ} is normalised by the area of the instances. We attribute our approach to work despite the correlation because the segmentation algorithm might have picked up lots of false positive small instances, so weighting these instances less would counter this correlation.

5.4.2 Hard Thresholding

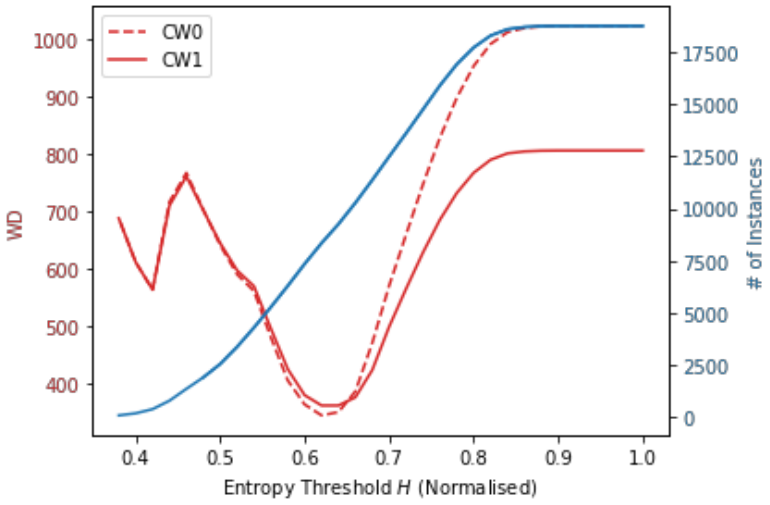
Some may wonder, would it be possible to simply discard the instances with the highest uncertainty in order to improve histological measurements? We experiment to find out. We compare the statistics on our resampled (BGAN1) synthetic dataset with a portion of instances higher than a certain entropy H removed. Figure 5.13 shows how WD changes with this threshold for tissue area (Figure 5.13a) and y^{γ} (Figure 5.13b). Starting from the far right of the plot, where all instances are included, it can be seen that measurements are significantly better when weighted by confidence (CW1) compared to non-weighted (CW0). As we move toward the left, instances with the highest uncertainties are removed from the analysis. As such, CW0 and CW1 also gradually improved until reaching a minimum at around $H \approx 0.63$. At $H < 0.63$, too few instances remain in the analysis, so WD increases again.

The plots show that our confidence weighting scheme has several advantages. Firstly, it can be seen that CW1 consistently outperforms CW0 when only a small number of instances are removed. Secondly, our proposal does not require removing any instances. This is vital for our project because, in renal needle biopsies, many slides barely contain enough tissues for assessment. Thirdly, it can be seen that the trough in the plots is rather narrow, meaning that it could be challenging to find a hyperparameter to reject poorly segmented instances. Weighting instances by BNN confidence, on the other hand, could lead to improvements without the need to find such hyperparameters.

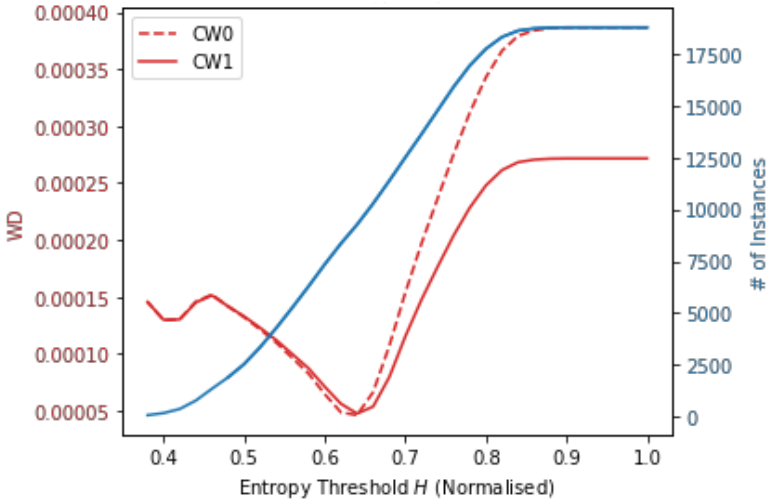
5.5 Conclusion

We propose a method to improve histological measurements by weighting tissue instances with confidence measures derived from BNNs.

To demonstrate the performance of our proposal, we design a method to generate synthetic data from an embedding that progressively differs from the original images. The method allows independent control of morphological parameters in the ground



(a) Area



(b) Colour Distribution y^{γ}

Figure 5.13: Comparison of Histological Measurements vs Ground Truth with the Least Confident Instances Removed. The figures show that instances weighted by entropy-derived confidence measure (CW1) consistently outperform the unweighted measurements (CW0) when only a small number of instances are removed. At $H \approx 0.63$, the weighted and unweighted estimates perform almost equally well. The plots above are generated using UNet + MC Dropout with *BGAN1*, but a similar trend can be seen for the results generated with BBB_LRT and other histological measurements.

truth masks and synthetic images' textural/colour variations. Confidence-weighted measurements are obtained on a real test dataset and the synthetic dataset using uncertainty derived from UNets with either MC Dropout or Bayes by Backprop. The weighted distribution is improved for all histological measurements. This demonstrates our proposal is robust to the choice of hyperparameters, such as the architecture of the BNN and the choice of histological measurement. We also notice that the improvement is greater on data with lesser resemblance to the training data and on measurements that are more sensitive to segmentation errors. This suggests that our approach could be beneficial for cross-domain datasets. We also demonstrate that there is no need to find a hyperparameter to discard instances with high segmentation uncertainties.

For future work, we will investigate whether decorrelating histological parameters with confidence measures could help to improve the histological metrics even further.

5.6 Synthetic Data

Our synthetic dataset containing 98,000 synthetic images has been made publicly available under the Creative Commons license (CC BY-SA) at
`https://tinyurl.com/vaebicycleganpas` .

6

Predicting Visual Changes and Endpoints with Quality-Weighted Tissue-based Features

Contents

6.1 Introduction 145
6.2 Method 147
6.2.1 Multi-instance Learning with Soft Attention 148
6.2.2 Datasets 149
6.2.3 Tissue Segmentation 152
6.2.4 Extraction of Histological Features 154
6.2.5 Predicting Slide Level Labels 155
6.3 Results and Discussion 157
6.3.1 Segmentation Results 157
6.3.2 Performance of Different Featuresets 159
6.3.3 Attention Visualisation 163
6.3.4 Confidence-Weighted Predictions 168
6.4 Limitations 172
6.5 Conclusion 174

Two common obstacles limiting the performance of data-driven algorithms in digital histopathology classification tasks are the lack of expert annotations and the narrow diversity of datasets. Multi-instance learning (MIL) can address the former challenge for the analysis of whole slide images (WSI), but performance is often inferior to full supervision. We show that the inclusion of weak annotations can

significantly enhance the effectiveness of MIL while keeping the approach scalable.

In Chapter 4, an analysis framework was developed to process periodic acid-Schiff (PAS) and Sirius Red (SR) slides of renal biopsies. The workflow segments tissues into coarse tissue classes. Handcrafted and deep features were extracted from these tissues and combined using a soft attention model to predict several slide-level labels: delayed graft function (DGF), acute tubular injury (ATI), and Remuzzi grade components. A tissue segmentation quality metric was also developed to reduce the adverse impact of poorly segmented instances. The soft attention model was trained using 5-fold cross-validation on a mixed dataset (NMP, TCGA, Native - discussed in Section 6.2.3) and tested on the QUOD dataset containing n=373 PAS and n=195 SR biopsies.

The average ROC-AUC over different prediction tasks was found to be 0.598 ± 0.011, significantly higher than using only ResNet50 (0.545 ± 0.012), only handcrafted features (0.542 ± 0.011), and the baseline (0.532 ± 0.012) of state-of-the-art performance. In conjunction with soft attention, weighting tissues by segmentation quality has led to further improvement ($AUC = 0.618 \pm 0.010$). Using an intuitive visualisation scheme, we show that our approach may have the potential to support clinical research work by highlighting individual tissue compartments the model uses for making predictions.

This Chapter is divided into 5 sections. Section 6.1 introduces the resource bottlenecks of our projects and the limitations of current state-of-the-art multi-instance learning methods. Section 6.2 details an experiment to compare the use of the proposed tissue-compartment based featuresets with a baseline. We also reiterate the work done in Chapter 5 and how it can be applied in this context. Section 6.3 shows the results supporting the proposed method as well as an intuitive tissue-compartment attention visualisation scheme. Section 6.4 discusses the limitations and Section 6.5 concludes the chapter.

6.1 Introduction

Computational pathology can assist pathologists by providing an automated second opinion on their assessment. Moreover, it may help us to better understand the mechanisms of organ injury by detecting and quantifying subtle histological changes in biopsies. While these models can help us improve the discriminative power of assessment tasks with performance unrivalled by classical image processing algorithms, several challenges are limiting their applicability. Firstly, training neural networks often requires large amounts of labelled data. However, most datasets contain no more than several hundred slides. In our setting, samples with known outcomes are biased due to pre-transplantation screening (either based on the donor's clinical information or histology). The only available data are "hard examples" that have either been missed by pathologists or are plagued by factors not guaranteed to be visible in biopsies (Kidneys that were transplanted have already passed some form of screening. Screening could be based on biopsies or donor clinical data.). There is also a "bootstrap" problem - while a severe shortage of pathologists is a primary motivation to expedite the development of an automated tool, this shortage limits the speed and scale in which labelled data can be procured.

To date, most deep-learning-based computational pathology platforms are designed to predict or assess only a bespoke set of narrowly defined clinical outcomes or visual changes (collectively known as slide-level labels). Most existing work [109, 116, 178–180] is limited to fully-supervised learning, which requires labelling a large number of tissue compartments or rectangular tiles as either normal or diseased. To adapt these platforms for a different diagnosis would require additional time from pathologists to go through the entire dataset, adding further to the project's investment. Regrettably, the expert-time cost of fully-supervised learning is prohibitive and has been a major reason for the limited number of publications applied to renal histology.

To address the lack of local expert annotations, a number of multi-instance learning approaches [181–184] have been developed to train classification tasks using only slide-level labels. However, available multi-instance learning models typically

need to be trained on large datasets or slides with plenty of tissue area with good diagnostic quality. In our setting, where many slides contain sub-optimal tissue areas, existing approaches fail to deliver acceptable classification performance.

Furthermore, models trained under multi-instance learning tend to have limited diagnostic transparency - features are often extracted from rectangular tiles, which are inconsistent with the anatomical, irregular tissue compartments within the biopsies. It is common for different functional tissue structures to vary in size by several orders of magnitude (e.g. cell nuclei vs arteries). This may be partially addressed by using tiles from several magnifications [184], but this solution could make visualisation considerably more challenging.

Any histology analysis framework also needs to be robust to artefacts. This is particularly true in our application for two reasons. Firstly, in needle biopsies, a large proportion of tissue resides close to the edge and is often distorted or truncated. Exclusion of these tissues is not always possible as biopsies are often narrow and have limited material available. Secondly, decisions on transplantation are time-sensitive. Hence the long-term aim is to eventually read histology from frozen biopsies which are often plagued with artefacts. There is a lack of definitive attempts to reduce the impact of artefacts. Treatment of artefacts is often either not mentioned or is excluded manually in most experiments. The most common approach to tackle artefacts includes explicitly labelling artefacts and aggressive data augmentation [109, 179]. However, these approaches would only work on objects that resemble the training data.

This chapter aims to conduct a pilot study to show the feasibility of a flexible yet scalable platform for providing quantitative insight into visual associations to transplant dysfunction using renal biopsies stained with periodic acid-Schiff (PAS) and Sirius Red (SR) while addressing the aforementioned issues. Our workflow extracted a number of histologically relevant visual features from tissues and was developed with minimal laborious labelling by expert pathologists. We used these visual features in combination with convolutional neural network (CNN) features to predict several slide-level labels such as Delayed Graft Function (DGF) - defined

as recipients who need dialysis within the first week after transplantation [185], Acute Tubular Injury (ATI), and Remuzzi Scores [4]. We compared the predictive performance of our framework with features based on tissue compartments with a standard workflow that relied only on CNN-derived features and on rectangular tiles and showed that the proposed workflow produces consistently higher area-under-the-curve (AUC) in the models' receiver operator characteristics (ROC) and Precision-Recall (PR) curves.

Furthermore, we developed a visualisation scheme that works specifically for our proposed workflow. Compared to prior work [181–183], using tissue-derived features enabled us to pinpoint a diagnosis of specific tissues irrespective of their size and shape, enabling the potential for transparent diagnosis and visualisation.

Finally, through our experiments, we also find that tissue quality and quantity may play a significant role in the predictability of a slide. By incorporating a metric derived from Bayesian Neural Networks (BNN) describing tissue quality similar to that described in Chapter 5 derived from an ensemble of CNN models, we were able to improve the quality of predictions measured by AUC consistently.

6.2 Method

In Chapter 4, we propose a computational framework to extract visual histopathological features from different functional components of the biopsy. A schematic of the workflow is shown in Figure 4.1 in Section 4.1. In the first step, we identify specific tissue compartments (Figure 4.1(a)). Details of the segmentation algorithm are discussed in Section 6.2.3. Subsequently, we extract a set of tissue compartment specific features (Figure 4.1(b)), which are described in Section 6.2.4).

Finally, we combine instance level features into a fixed-length description of the whole slide (Figure 4.1(c)). We evaluate different approaches to combining these features. The most trivial method is to simply perform an average/max pooling from the feature values of all the tissues. However, if the segmented tissues were only coarsely categorised, a large portion may be irrelevant for diagnosis. Pooling

features from different tissues irrespective of their histopathological importance could lead to erroneous predictions that lack transparency.

6.2.1 Multi-instance Learning with Soft Attention

Multi-instance learning (MIL) [183, 186, 187] is another approach commonly applied to histology analysis for making slide level (bag) predictions from a variable number of instances. A slide is classified as positive if at least one positive instance is detected. Implementation of the original MIL algorithm involves predicting a probability value for each instance and then converting these instance probabilities into a bag-level value using max-pooling. However, this approach has several limitations that make it unsuitable for predicting kidney function. Firstly, the use of max-pooling means that the method is particularly sensitive to noise. Our slides often have artefacts or tissues with morphology not previously seen during training. There is an inherent risk that the presence of these artefacts could sway the predictions as the classifiers have not been trained on embeddings beyond the original data. Secondly, histological changes can progress in several different ways - changes can affect a small number of patches severely or a large number of patches mildly. Standard MIL classification where predictions are based on the presence of a single positive instance would be less successful at capturing changes due to the latter. It would likely rank the former is more severe even if kidney function is more heavily dependent on the percentage of healthy tissues.

To partially mitigate the aforementioned limitations, we implement a soft attention mechanism. As a result, we can use attention-weighted averages of the instances to make predictions, as shown in the schematic in Figure 6.1. The model consists of multiple stages: Firstly, it converts each instance's features into a permutation-invariant embedding. From these embeddings, a gated attention mechanism [187] assigns weights to each instance depending on their relative importance for the bag-level predictions. A gated mechanism is used to enhance the non-linearity of *tanh* when the function's input values are small. Because soft attention is learned, theoretically, it should be capable of rejecting instances not

relevant for the assigned bag-level prediction task. The fractional contribution a_k of an instance k to the final prediction is:

$$a_k = \frac{exp(\mathbf{w}^\top(tanh(\mathbf{V}\mathbf{h}_k^\top) \odot sigmoid(\mathbf{U}\mathbf{h}_k^\top)))}{\sum_{j=1}^{K} exp(\mathbf{w}^\top(tanh(\mathbf{V}\mathbf{h}_j^\top) \odot sigmoid(\mathbf{U}\mathbf{h}_j^\top)))} \tag{6.1}$$

Where $\mathbf{h}$ is an embedding derived from the feature vector of one instance, $\mathbf{w}, \mathbf{U}, \mathbf{V}$ are learnable weights in our neural network.

In practice, however, in many biomedical datasets, the number of instances in each bag could be greater than the number of bags, making it very difficult to train a reliable attention mechanism. The attention network may also fail to assign a meaningful score for instances that do not resemble any of those from the training examples. To address this challenge, we propose to include an additional factor $\mathbf{g}$ into the weighted average. The weight g_k over an instance k can be described as a confidence score of the neural network on an instance indicative of its resemblance to the training data. Such a score can be derived using probabilistic predictions from BNNs [153, 154, 172]. As we have a plethora of delineated tissues, we decided to obtain $\mathbf{g}$ using our UNet ensemble (described in Section 6.2.3). To account for the Bayesian uncertainty of different instances, we simply incorporate $\mathbf{g}$ into the attention as follows:

$$a_k = \frac{exp(\mathbf{w}^\top(tanh(\mathbf{V}\mathbf{h}_k^\top) \odot sigmoid(\mathbf{U}\mathbf{h}_k^\top))g_k)}{\sum_{j=1}^{K} exp(\mathbf{w}^\top(tanh(\mathbf{V}\mathbf{h}_j^\top) \odot sigmoid(\mathbf{U}\mathbf{h}_j^\top))g_k)} \tag{6.2}$$

6.2.2 Datasets

Two main datasets are used in this study. A breakdown of these datasets was summarised earlier in Chapter 3, Table 3.2. The datasets contain slides stained using PAS and SR. While PAS is a routine stain for renal biopsy assessment, SR could be promising for the computational quantification of fibrotic tissues. SR is normally viewed under polarised light [149] for maximum signal-to-noise ratio

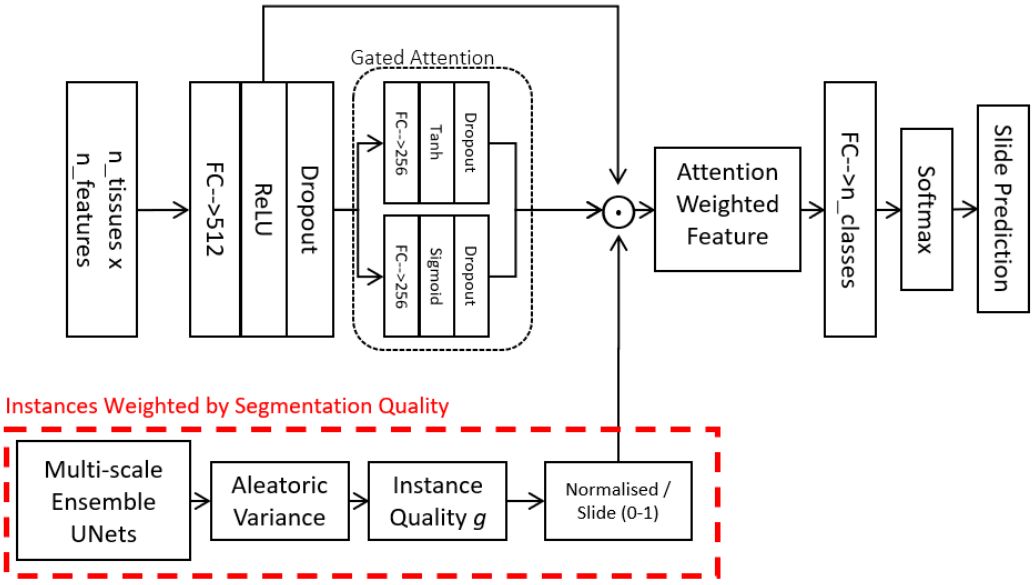


Figure 6.1: Attention Model - this model is used for combining feature vectors from a variable number of tissues into a single vector describing a slide. The gate attention module learns which instances are relevant for the bag-level prediction tasks. Optionally, in addition to the learned attention, we also weight instances by their segmentation quality and how much the instances resemble our locally-delineated training examples.

but attempts to quantify the extent of fibrosis under unpolarised light have also been shown to be highly reproducible [145].

Background and some statistics of the QUOD dataset were discussed earlier in Section 3.1.1. A characteristic of these QUOD biopsies is that they are very small for two reasons. Firstly, a small needle is used to minimise bleeding complications after the transplant. Secondly, biopsies obtained at the backtable were cut into two halves. Half was stored in formalin for histology analysis and half in RNAlater for biochemical analysis. The majority of slides do not contain enough tissues for full assessment according to the Banff criteria [30], which states that ≥ 7 glomeruli and ≥ 1 artery is necessary for assessment. (Distributions of glomeruli and arteries in Figures 3.17 and 3.18.) Slides that contain no arteries or < 7 glomeruli are only partially assessed. Hence only 90 PAS slides had received a full Remuzzi score. Several slides containing only fractions of an artery also received a Remuzzi artery score. A large proportion of tissues also suffer from artefacts such as forceps compression, folding, or duplication of serial sections.

The NMP dataset originated from an organ normothermic perfusion experiment [120, 121] from 35 donors. Most of the slides (n=169) were stained with Haemotoxylin and Eosin (H&E) but had been computationally converted to PAS stain using a CycleGAN [98] with image quality largely indiscernible by our collaborating pathologist at low magnifications. Details of this dataset were discussed in Section 3.1.2.

Apart from the two main datasets, we have also included additional slides from native biopsies in PAS and slides from TCGA stained in H&E. These slides were solely used for strengthening the segmentation algorithms to ensure they would generalise well to unseen data.

We manually marked out a number of rectangular tiles from each dataset to delineate different tissue compartments - tubules, glomeruli, vessels, and cell nuclei. The number of delineated tissues for each dataset is shown in Table 3.6. Delineating the outline of tissues is a laborious task. To make the task scalable, initial annotations were performed by an engineer with limited training in pathology. Thus, we only have the coarse classification of these tissues. For instance, proximal tubules are relevant for assessment, but a notable portion of objects marked as "tubules" were actually distal tubules or the collecting duct. Objects marked as "vessels" consist of a mixture of arteries, arteriole, and veins; some casts might be misidentified as sclerosed glomeruli. The boundaries of glomeruli were inconsistent regarding the inclusion of the urinary space. Owing to the small size of many biopsies, tissues that were truncated near the edge of the biopsies were also delineated as long as they were human-recognisable. A subset of these tissues was cross-checked by our pathologist (Details in Section 3.3.2.2 Table 3.4).

The tiles selected for tissue delineation may contain a mixture of tissue-containing and blank areas and are of different sizes and aspect ratios such that it covers a diverse range of tissue morphology. As the number of tubules is far greater than the number of glomeruli and vessels, in 165 out of 425 tiles (967 out of $1210mm^2$ in terms of area) listed in Table 3.2, we only delineated the glomeruli and vessels but not tubules.

6.2.3 Tissue Segmentation

Segmenting tissues according to functional compartments allows us to incorporate known visual features into histology analysis and maximises the interpretability of predictions made by algorithms. Tissue segmentation was performed using an ensemble of UNets [87]. As we have a varying number of annotations available, we chose to use a different number of UNets for PAS and SR-stained slides.

For segmenting tissues in PAS-stained slides, we used a total of 13 models as follows: 2 models to segment cell nuclei at 0.44 microns-per-pixel (mpp); 2 models to segment tubules, glomeruli, and vessels at 0.44mpp; 3 models to segment tubules and glomeruli at 0.44mpp; 3 models to segment glomeruli and vessels at 0.88mpp; 3 models to segment glomeruli and vessels at 1.76mpp. Models that process identical tissue classes at the same magnification were trained on a different train:validation (4:1) split. These UNets were trained and validated using tissues delineated from the NMP, TCGA, and native biopsy slides.

For SR-stained tissues, we used 4 models as follows: 2 models to segment tubules, glomeruli, and vessels at 0.44mpp; 2 models to segment the same tissues at 0.88mpp. Cell nuclei are not segmented as they are not visible under SR.

We implemented the ensemble of UNets with dropout to simulate Bayesian Neural Networks [153, 159] with the same hyperparameters as Tam et al. [8] to process the test data. The motivation for using BNNs is that they generally output predictions where the soft values are more representative of the probability of a correct prediction. However, in cases where uncertainties are data-limited (aleatoric uncertainties) rather than model-limited (epistemic uncertainties), there could still be notable discrepancies between the predictions and the actual probabilities. In particular, if the relevant class is rare or looks very different in the test data, it could lead to under/over-confident predictions. Thus, we propose to correct the predictions using a data-driven approach, as shown in Equation 6.3:

$$\tilde{p} = max(\bar{p} - A\sigma_p, 0) \tag{6.3}$$

Where $\bar{p}$ is the mean output from the neural network ensemble; σ_p is the standard deviations from the ensemble over a single pixel p; A is a constant to be empirically determined from the training data; $\tilde{p}$ is the corrected probabilistic output from the network ensembles, which is clamped to a value above zero. As we shall see in Section 6.3.1, this serves to remove overconfident pixels and would help suppress false-positive pixels caused by artefacts.

From the ensemble-averaged (Equation 6.3) segmentation maps, we obtain tissue instances using the max-flow-min-cut [139] algorithm. Individual tissues are cropped from the original slide with $1.32\mu m$ padding on each side. In order to perform localised diagnostics based on individual tissues, areas outside of the tissues are blurred. This helps prevent extra-tissue regions from contributing to visual features at later steps.

In our case, we have chosen to derive $\mathbf{g}$ from the UNet ensemble. For a slide with K tissue instances, the weight assigned to the instance k is given as:

$$g_k = \frac{\max\limits_{j \in K}(s_j^2) - s_k^2}{\max\limits_{j \in K}(s_j^2) - \min\limits_{j \in K}(s_j^2)} \tag{6.4}$$

$s_j = \sum\limits_{p \in M_j} \sigma_p / |M_j|$ is the mean value of σ_p over the segmentation mask M_j for instance j; σ_p is the standard deviations from the ensemble over a single pixel p. If all UNets from the ensemble predict similar values for the pixels within instance k, g_k would have a value close to 1. Otherwise, high discordance between different UNets would result in g_k close to 0.

Details of the UNet architecture used in this study are shown in Table 4.1 in Chapter 4. A value of $m = 16$ is used for our default UNets. For the segmentation of cell nuclei, we use $m = 1$ for Block 6-8 and $m = 8$ for other blocks. For the segmentation of tissues at $1.76mpp$, we use $m = 8$ for Block 6, 8 and $m = 4$ for Block 7. A smaller number of filters is used to save memory resources as a large receptive field is less relevant for the segmentation of cell nuclei and in the segmentation at low magnification. In addition to the foreground tissue classes, each UNet also outputs a background class and a boundary class. After obtaining instances using

max-flow-min-cut, we expand the area of each instance by looping through each instance iteratively and dilating each mask with a 3x3 kernel until the instance exceeds the boundary or touches another instance. We find that the inclusion of boundary class helps separate tissues with ambiguous boundaries.

6.2.4 Extraction of Histological Features

We aim to demonstrate (i) the benefits of using handcrafted features to augment deep features and (ii) how extracting features from functional tissue structures can boost performance and interpretability in multi-instance learning settings. As there are currently very few studies that quantitatively assess how individual histological features correlate with physiologically relevant measurements, we tested a wide range of features in our study, including both handcrafted and deep features. Using the aforementioned workflow, we extracted a number of histological features from our slides from tissues. We designed handcrafted features that comprise tissue morphological descriptors, colour, texture, and second-order features such as how colour/texture are distributed with respect to the tissue compartment. The majority of handcrafted features were designed with one tissue type in mind but implemented across all tissue classes such that the feature vector is the same length for all tissue types.

Some of these features are designed to reflect visual changes in tubules that have undergone chronic or acute injuries. For PAS-stained slides, cell nuclei are typically visible within each tissue, so their colour and distribution may also shed light on the state of the biopsy. In proximal tubules, darker nuclei in epithelial cells may be a feature of mitosis and cellular repair, whereas cell nuclei located far away from the boundary of proximal tubules may signify cell dropout or cytoplasm expansion which is a feature of acute tubular injury [39, 188]. As the number of nuclei in each tissue is variable, the values are pooled together at every tenth percentile: nuclei features within each compartment are converted into a distribution and the values of every 10 percentile ($10^{th}...90^{th}, 100^{th}$ percentile) are used as a feature in

the parent compartment. This has an advantage over max-pooling of being less sensitive to artefacts/falsely detected cell nuclei.

The handcrafted features used in this study are discussed in 4.3 in Chapter 4. Several features are derived from the distribution and colour of segmented cell nuclei. However, we recognise that some tissue compartments may not have any nuclei, leading to missing feature values. To prepare our data for machine processing, missing values are imputed with the mean value from the rest of the datasets. Some slide-level information, such as the total area of the biopsy, is appended to the feature vector of each individual tissue. In total, this resulted in 98 unique handcrafted features for the PAS-stained slides and 40 features for the SR slides.

In addition to handcrafted features, we also experimented with features from several established deep neural networks. Deep features are obtained from the same patch (with surroundings blurred) as the handcrafted features at 0.44mpp. The crops were not resized before we fed them into neural networks as we want objects of the same physical size to elicit the same filter responses. Fully-connected layers of neural networks were replaced by adaptive average pooling, resulting in a single 1D feature vector for each tissue. Each feature is normalised to unit variance with zero mean over our datasets to speed up convergence during training.

6.2.5 Predicting Slide Level Labels

Several different physiologically relevant measurements are available for our datasets.

A subset of PAS-stained slides has been assessed by an experienced pathologist (blinded to the donor characteristics and outcome) to determine the extent of histological changes. Slides are graded according to the Remuzzi criteria [4] based on the severity of Tubule Atrophy (Remuzzi TA), arterial and arteriolar narrowing (Remuzzi A), glomerular global sclerosis (Remuzzi G), and interstitial fibrosis (Remuzzi IF). In addition, as part of the assessment routine, biopsies are also graded for Acute Tubular Injury (ATI). The distribution of the assessed grades can be found in Section 3.3.1.

For labels with insufficient cases for training or testing, we regrouped the most severe cases until there were enough donors for cross-validation to reduce the class imbalance. As a result, Remuzzi G, TA, and IF become a binary classification task, whereas labels for ATI are regrouped to either two or three (0-2) grades instead of four grades from the original assessment.

Apart from eGFR, the QUOD dataset also contains binary labels regarding whether the recipient has suffered from DGF. DGF may have origins in a variety of diagnoses. While ATI is one of the known leading culprits, there are also other causes such as T cell-mediated rejection, antibody-mediated rejection, and acute calcineurin toxicity [47, 119]. While it may not be possible to detect some recipient-related factors (such as rejection) or surgical causes (such as anastomosis) from histology, subtle or localised acute lesions may be within small regions of some biopsies. A method based on multi-instance learning may have a chance of detecting these localised changes missed by human inspectors or not meeting histological thresholds of established grading criteria.

Combinations of handcrafted histological features and deep features are used as input for our multi-instance soft attention model (Figure 6.1). The length of training is determined using the validation AUC at 40 epochs after the metric becomes stagnant. A weighted sampling approach was used to increase the frequency of sampling the rarer classes.

Our attention model's hyperparameters are tuned based on the following trivial task to reduce bias: Slides are labelled by whether they contain enough glomeruli for assessment. Slides with < 7, $7 - 9$, and ≥ 10 unique glomeruli are separated into three classes. Hyperparameters are searched using HyperOpt [126] and remain fixed for other tasks in order to be able to compare different featuresets fairly. This task is trivial because glomeruli can easily be distinguished from tubules just by several handcrafted features (e.g. area). Setting hyperparameters this way helps to avoid bias when we compare different models' performance on the main tasks. Examples of hyperparameters explored include the choice of gated/un-gated attention network, size of network layers, and regularisation weight.

Because the QUOD slides were obtained from real transplant settings, there were several challenges to developing an algorithmic workflow based on these slides. Firstly, these slides consist mostly of kidneys with little chronic damage - the dataset is biased due to pre-transplant donor screening, and the distribution of biopsies with pathological changes is highly imbalanced. Secondly, the majority of biopsies are inadequately small in that they do not meet the Banff criteria [30]. For predicting assessment grades given by pathologists, we only included slides with enough tissues for each grade. As for the prediction of DGF, we found that it was necessary to include slides of all sizes in order to achieve at least one donor per label per cross-validation. Cross-validation splits are subjected to the constraint where all slides from each donor remain in the same training/validation/test set.

6.3 Results and Discussion

Experimental results are presented in four sections. Section 6.3.1 briefly describes the segmentation results on the QUOD dataset. In Section 6.3.2, features are extracted based on segmented tissue instances and are used to predict slide labels. We also compare classification performance between different featuresets across several tasks. Section 6.3.3 shows a pilot visualisation scheme of the proposed workflow. Finally, Section 6.3.4 presents results showing that segmentation uncertainty can be used to improve prediction performance.

6.3.1 Segmentation Results

As per earlier experiments, we found that UNets that were trained on the same magnification tend to produce false-positive segments in the same areas of the test data regardless of how the dataset splits, weight parameters, input tile size, and regularisation are initialised. These false positives were likely caused by tissues with morphology not present in the training data. Combining multiple magnifications according to Equation 6.3 has helped to remove most of these false positives. Figure 6.2 shows the segmentation examples from individual models and the combined soft prediction. It can be seen that most artefacts from individual UNets (b-e) are no

longer visible once combined (f). Although it may be sufficient to use fewer UNets in the ensemble to reduce the amount of computational work, for this experiment, we are interested in the rest of the workflow where segmentation is not a limiting factor.

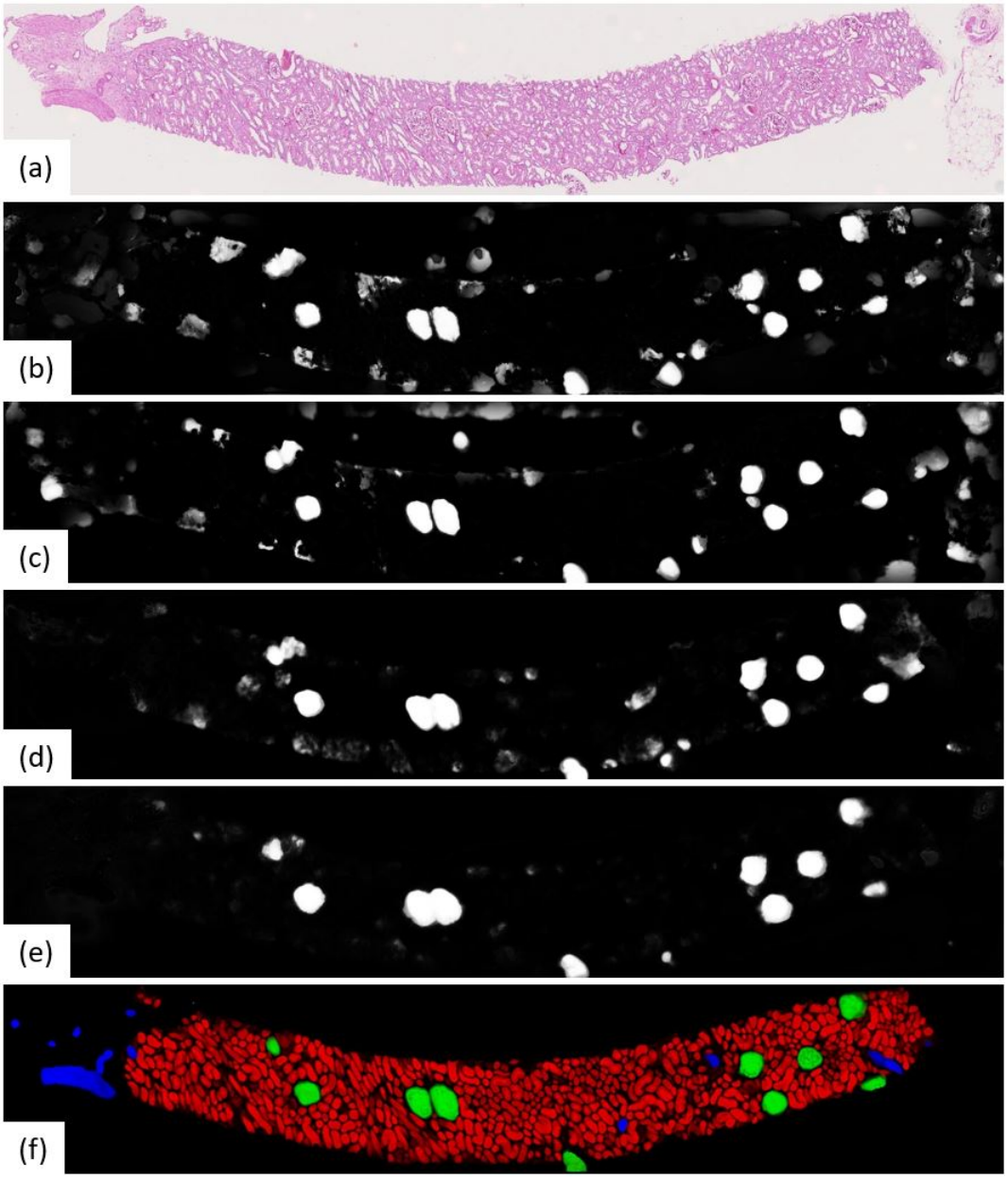


Figure 6.2: Output from the UNet ensemble. (a): Original PAS-stained slide; (b)-(e): Softmax predictions of glomeruli from single UNets at different magnifications segmenting specific tissue classes. (b) 0.44mpp, all tissue classes; (c) 0.44mpp, tubules + glomeruli; (d) 0.88mpp, glomeruli + vessels; (e) 1.76mpp, glomeruli+vessels; (f) Ensemble-combined predictions showing tubules, glomeruli, and vessels in red, green, and blue respectively.

Ideally, we would want the soft values from the UNet ensemble to scale linearly

with the probabilities for correct class prediction. However, we find this is not the case for data-limited tissue classes such as glomeruli.

Figure 6.3 shows segmentation results on the QUOD-PAS slides. The plots show the soft values of the combined segmentation scale with A, the multiplier of σ in Equation 6.3. For each value of A, we calculate a histogram binning all pixels predicted a certain value $\tilde{p}$ by the UNet ensemble. The predicted probabilities $\tilde{p}$ is plotted against the actual probabilities in (b) and (d) for different values of A. Then, we compute the L2 difference between the array $\tilde{p}$ against the actual probabilities p as shown in (a) and (c).

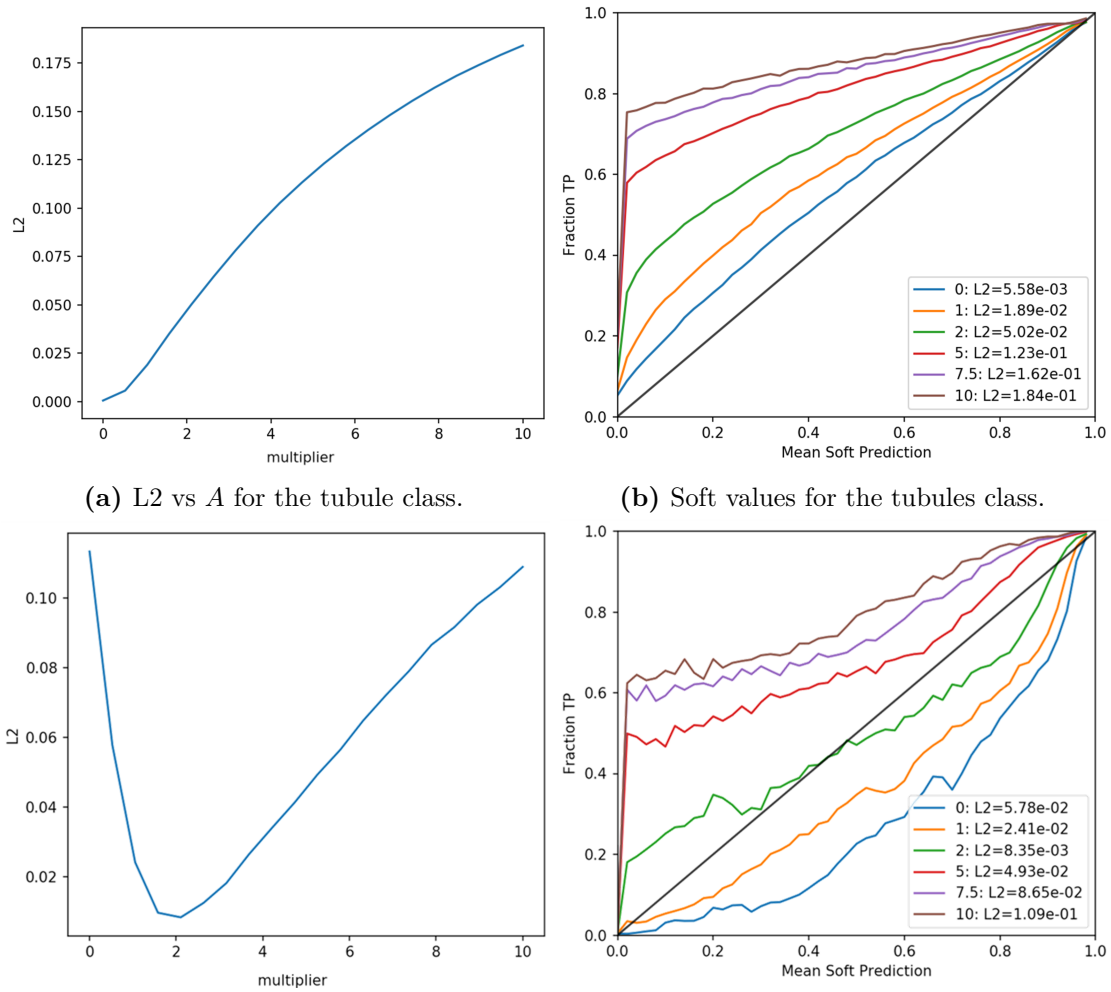
It can be seen that the optimal values are $A = 0$ for tubules (Figures 6.3a-b) and $A = 2$ for glomeruli (Figure 6.3c-d). These results suggest that while outputs from the Bayesian network ensemble scale linearly with class probabilities when class labels are abundant, this linear relationship breaks down when uncertainties are data-limited, and some empirical corrections might be needed.

Note that Equation 6.3 serves to remove areas that are overconfident but would not add to under-confident areas. Thus, the final number of tissues detected is likely to be underestimated. This bias is introduced to offset the asymmetrical consequences of false-positive detection compared to a false-negative: while a false-negative may simply result in fewer tissues being processed, a false-positive detection would lead to misleading information being introduced into the workflow. The former case is far easier to deal with as we can simply flag a slide as "Needs Review" if we detect too few tissues.

6.3.2 Performance of Different Featuresets

In this section, the soft attention models are implemented directly without accounting for the segmentation quality (Figure 6.1) of the tissues.

We compared predictive performance using a variety of featuresets as the input to the soft attention models. In order to avoid the need to set up an arbitrary threshold, we reported results in the form of ROC and PR curves (Tables 6.3, 6.2,



(a) L2 vs A for the tubule class.

(b) Soft values for the tubules class.

(c) L2 vs A for the glomerulus class.

(d) Soft values for the glomerulus class.

Figure 6.3: Class Probabilities vs Ensemble Averaged Segmentation Predictions. (a) and (c) show the L2 distance between values predicted by the UNet ensemble and the actual probabilities a pixel belongs to the tubule/glomerulus class at different values of A. (b) and (d) show how the calibrated soft values compare with the actual probabilities a pixel belongs to a class.

6.1). Curves are weighted by the number of tissues/patches in each slide[1] to reduce the noisy impact from biopsies that do not meet the Banff criteria.

To conclude the optimal methodology, we have calculated not just AUCs but also their variability. Reported ROC-AUC values of the soft attention model in all tables in this paper averaged across 5 fold cross-validation test set (3:1:1 training/validation/testing) and 5 different seeded weight initialisation (total 25

[1]Computed using "roc_curve" and "precision_recall_curve" in Scikit-learn Python package where the "sample_weight" array is the number of tissue compartments in each slide.

models). Multi-class models are macro-class-averaged. (In macro-class-averaging, an ROC curve is plotted in one-vs-other form for each class and we take the mean AUC of all classes. In contrast, in micro-class averaging, each validation example contributes equally to the AUC. We find that in our case micro-averaging often generates overly-optimistic results.) We reported the unbiased standard error of the mean AUC to ensure that comparisons are meaningful and to account for data and model noise. Standard errors are calculated as if the different prediction tasks are independent. In addition, because the training was performed on mixed datasets, neural networks may have learned to look for "shortcuts" (e.g. classification based on staining protocol rather than pathological changes) instead of performing the main task. Thus, all AUC values are evaluated based on only the QUOD slides.

Deep features are extracted from a number of neural networks, including several networks pre-trained with ImageNet (ResNet50 [81], VGG16 [95], InceptionV3 [189]), two networks trained using cropped image patches (ResNet50 and a Variational AutoEncoder [90]), and ScatterNet [190]. Details of these features are summarised in Table 6.1. Features extracted from the segmented tissues are given the prefix "Tissue". We have also extracted features using fixed-sized rectangular tiles - these are given the prefix "Tiles" in the table. While handcrafted features can be calculated for individual tissues, there is no intuitive way to do so for rectangular tiles as they may contain a varying number of tissues. A breakdown of the predictive performance of some of these featuresets for different tasks is shown in Table 6.2.

To compare the general utility of the different methodologies, we averaged the AUCs from different tasks. ROC-AUCs are averaged with inverse variance weighting, so tasks with higher prediction consistencies are weighted more. Precision-Recall (PR) curve AUCs are also given in the results (Table 6.1) as a complementary metric to ROC-AUCs. PRs could be insightful for tasks with highly imbalanced labels. However, variances of PR-AUCs tend to be small for the more challenging tasks, so we only reported the arithmetic mean rather than the inverse variance weighted mean in the table.

The mean ROC-AUC values in Table 6.1 demonstrate the progress towards improving performance by using features extracted from tissue compartments versus the simplistic model that uses only features from rectangular tiles. When only ResNet50 features are used at 0.44mpp, features extracted from tissues are significantly better than features from rectangular tiles ($AUC = 0.545 \pm 0.012$ vs 0.507 ± 0.014 in rows 11 and 9), though the performance of both cases is very poor. Using tissue compartments enables the extraction of handcrafted features and this has led to a significant improvement in AUC (row 1, 0.598 ± 0.011). It would be difficult to extract these handcrafted features from rectangular tiles as some features depend on the compartment boundaries. In contrast, a common trick to extract more features from rectangular tiles is to run the feature extractor at multiple scales over a larger receptive field. We found that concatenating ResNet50 features at 1.76mpp from concentric tiles led to an improvement (row 7, 0.532 ± 0.012) over a single scale but is still inferior to row 1. These results show that our proposed approach is superior in most cases. If we look at the results at a more granular level, we will find that there are some tasks where tile features may have the potential to outperform tissue features. From the prediction task breakdown for these two featuresets (corresponding columns in Table 6.2), we see that tiles perform better for Remuzzi G (0.556 ± 0.049 vs 0.631 ± 0.074), DGF/PAS (0.504 ± 0.025 vs 0.525 ± 0.018), and DGF/SR (0.649 ± 0.031 vs 0.656 ± 0.032). However, the differences for these individual tasks have not yet reached statistical significance, which could be down to noise in the data.

A second observation to note is that performance is better when deep and handcrafted features are combined compared to using only deep features or only handcrafted features ($AUC = 0.598 \pm 0.011$, 0.545 ± 0.012, and 0.542 ± 0.011 in respective order as seen from rows 1, 11, and 12 in Table 6.1). While the overall AUC values for "Tissue ResNet50 Only" and "Tissue HC" are not significantly different, from the task breakdown in Table 6.2 we can see that AUCs are more variable for the predictions based on handcrafted features. This could be because handcrafted features are specialised in specific tasks. For example, in our dataset,

most slides that have been graded for Remuzzi A had only the minimum number of one artery, many of which are partially truncated. So it may be hard for our attention model to learn to predict Remuzzi A grades based on only deep features. Handcrafted features can help to supply complementary information based on domain knowledge. These results suggest that both types of features may have their respective advantages. Thus, implementing a hybrid approach in the workflow may be optimal for general tasks.

Furthermore, we have also attempted to compare our soft attention model with other multi-instance methods such as CLAM [181] and MIL with max pooling [183, 186], as shown in rows 14 and 15 in Table 6.1. While soft attention has consistently outperformed MIL, a comparison is more challenging for CLAM due to the extra hyperparameters involved. An extensive search over several hyperparameters using HyperOpt's Bayesian optimisation algorithm [126] with the Ray Tune platform [127] has shown that the extra clustering step in CLAM has not benefited our prediction tasks.

In additional to neural networks pre-trained with ImageNet, we have modified a ResNet50 architecture to predict ATI scores (0-3) based on localised tissue patches (Image patches with ATI distribution are shown in Figure 3.14). This modified ResNet was trained and validated on 731 images of proximal tubules with a 4:1 split. The purpose of this model is to test whether convolution filters would better capture histopathological changes if it has prior exposure to such images. The results ($AUC = 0.598 \pm 0.011, 0.598 \pm 0.009$ from rows 1 and 2 in Table 6.1) show no significant difference in performance regarding how the networks were trained, suggesting the convolution filters learned from ImageNet may already be adequate for capturing diversity in renal histology.

6.3.3 Attention Visualisation

Multi-instance learning on whole slide images is not commonly trained end-to-end due to their large size. Features are usually saved onto the disk before being processed by the MIL model. Gradients from neural network feature extractors

Table 6.1: Overview of the featuresets. Most featuresets are made from concatenating deep neural network (DNN) and handcrafted (HC) features. We have also tested adding donor and recipient metadata (Section A.2) as feature vectors, as shown in row 13 (Meta). Mean AUC values over multiple tasks are also shown. Detail breakdown of AUC values of some of these featuresets is shown in Table 6.2.

#	Featureset Name	Mean ROC AUC	Mean PR AUC	Features (PAS)			Features (SR)			Description
				DNN	HC	Meta	DNN	HC	Meta	
1	Tissue ResNet50	0.598 ±0.011	0.279 ±0.009	1024	98	0	1024	40	0	Handcrafted features combined with DNN features from ResNet50, pretrained with images from ImageNet
2	Tissue ResNet50 (ATI)	0.598 ±0.009	0.283 ±0.008	1024	98	0	1024	40	0	Handcrafted features combined with DNN features from ResNet50, trained to classify acute tubular injuries at the tissue level.
3	Tissue VGG16	0.596 ±0.011	0.280 ±0.009	1000	98	0	512	40	0	DNN features from VGG16, pretrained with images from ImageNet.
4	Tissue VAE	0.559 ±0.009	0.252 ±0.007	200	0	0	200	0	0	DNN features from a simple Variational AutoEncoder trained on tissue patches.
5	Tissue InceptionV3	0.553 ±0.011	0.269 ±0.009	768	98	0	768	40	0	DNN features from InceptionV3, pretrained with images from ImageNet.
6	Tissue ScatterNet	0.551 ±0.012	0.269 ±0.008	1029	0	0	1029	0	0	Features from ScatterNet (2nd order).
7	Tiles (2 Levels) ResNet50	0.532 ±0.012	0.258 ±0.007	2048	0	0	2048	0	0	Features from ResNet50 (ImageNet pretrained) based on 256x256 tiles @0.44mpp and @1.76mpp. Features from concentric tiles are concatenated together.
8	Tiles (2 Levels) ScatterNet	0.514 ±0.01	0.275 ±0.009	1029	0	0	1029	0	0	Features from ScatterNet (2nd order) based on 256x256 tiles @0.44mpp and @1.76mpp. Features from concentric tiles are concatenated together.
9	Tiles (1 Level) ResNet50	0.507 ±0.014	0.252 ±0.005	1024	0	0	1024	0	0	Features from ResNet50 based on 256x256 tiles @0.44mpp.
10	Tiles (2 Levels) VAE	0.460 ±0.012	0.226 ±0.007	200	0	0	200	0	0	Features from the same VAE as (4) based on 256x256 tiles @0.44mpp.
11	Tissue Resnet50 Only	0.545 ±0.012	0.259 ±0.011	1024	0	0	1024	0	0	DNN features from ResNet50, pretrained with images from ImageNet.
12	Tissue HC	0.542 ±0.011	0.259 ±0.007	0	98	0	0	40	0	Handcrafted features only.
13	Tissue Resnet50 Metadata	0.607 ±0.01	0.280 ±0.008	1024	98	299	1024	40	270	ResNet50 pretrained with ImageNet; Feature vectors are appended with donor metadata such as age and cause of death. Categorical metadata are cast into one-hot format.
14	Tissue Resnet50 CLAM	0.597 ±0.011	0.259 ±0.011	1024	98	0	1024	40	0	Same featureset as #1 but uses CLAM model.
15	Tissue Resnet50 MIL	0.553 ±0.008	0.259 ±0.007	1024	98	0	1024	40	0	Same featureset as #1 but uses MIL model with max pooling instead of soft attention.

Table 6.2: Overview of AUC values of predictions based on different featuresets (columns) and prediction tasks (rows). The second last row shows the inverse-variance weighted mean of the various tasks above. Columns are arranged in descending order of the mean ROC-AUC. It can clearly be seen that features extracted from tissues (column prefix "Tissue") perform better than features extracted from fixed-sized rectangular tiles (column prefix "Tiles"). Note also that combining deep features with handcrafted features (Tissue ResNet50, $AUC = 0.598$) resulted in better performance than using only either deep features (Tissue ResNet50 Only, $AUC = 0.545$) or handcrafted features (Tissue HC, $AUC = 0.542$).

Label/Stain	n (Donor / Slides) (QUOD Dataset)	Tissue ResNet50	Tissue ResNet50 (ATI)	Tissue VGG16	Tissue Inception V3	Tissue ScatterNet	Tissue ResNet50 Only	Tissue HC	Tiles (2 Levels) ResNet50	Tiles (2 Levels) ScatterNet	Tiles (1 Level) ResNet50
ATI/PAS	170/316 (135/145)	0.673 ±0.021	0.651 ±0.013	0.645 ±0.02	0.530 ±0.017	0.593 ±0.017	0.535 ±0.026	0.525 ±0.019	0.482 ±0.035	0.532 ±0.015	0.424 ±0.039
DGF/PAS	283/321	0.504 ±0.025	0.512 ±0.023	0.534 ±0.024	0.527 ±0.031	0.512 ±0.025	0.521 ±0.027	0.463 ±0.035	0.525 ±0.018	0.496 ±0.021	0.502 ±0.028
DGF/SR	143/143	0.649 ±0.031	0.634 ±0.021	0.588 ±0.029	0.655 ±0.044	0.532 ±0.028	0.607 ±0.03	0.579 ±0.039	0.656 ±0.032	0.527 ±0.031	0.674 ±0.037
Remuzzi A/ PAS	89/95	0.590 ±0.028	0.581 ±0.024	0.549 ±0.026	0.593 ±0.033	0.532 ±0.05	0.539 ±0.024	0.594 ±0.027	0.466 ±0.03	0.497 ±0.042	0.461 ±0.029
Remuzzi G/ PAS	137/163 (133/143)	0.556 ±0.049	0.566 ±0.047	0.568 ±0.034	0.588 ±0.039	0.555 ±0.096	0.575 ±0.05	0.580 ±0.038	0.631 ±0.074	0.628 ±0.089	0.604 ±0.073
Remuzzi IF/ PAS	170/311 (135/145)	0.427 ±0.029	0.526 ±0.036	0.509 ±0.075	0.531 ±0.026	0.404 ±0.058	0.430 ±0.075	0.521 ±0.035	0.469 ±0.06	0.451 ±0.036	0.466 ±0.058
Remuzzi TA/ PAS	135/145	0.571 ±0.047	0.588 ±0.034	0.638 ±0.053	0.479 ±0.064	0.538 ±0.075	0.528 ±0.056	0.569 ±0.05	0.556 ±0.044	0.489 ±0.043	0.482 ±0.045
Mean ROC-AUC (weighted)		**0.598 ±0.011**	**0.598 ±0.009**	**0.596 ±0.009**	**0.553 ±0.009**	**0.551 ±0.012**	**0.545 ±0.012**	**0.542 ±0.011**	**0.532 ±0.012**	**0.514 ±0.01**	**0.507 ±0.014**
Mean PR-AUC		**0.279 ±0.009**	**0.283 ±0.008**	**0.28 ±0.009**	**0.269 ±0.009**	**0.269 ±0.008**	**0.259 ±0.011**	**0.258 ±0.007**	**0.258 ±0.007**	**0.275 ±0.009**	**0.252 ±0.005**

could take up an enormous amount of space, so they are discarded in real-life implementations during the feature extraction process. As a result, the most straightforward parameters that could be directly visualised are the attention values attributed to the different instances. However, the visualisation of attention parameters could often be ambiguous. In the case where rectangular tiles are chosen, the resolution of the attention map is limited by the size of the tile. Localisation of the diagnostic will be poor if the tile size is too large. If the tile size is too small, there would be a limited receptive field, and we may risk truncating meaningful tissue structures; a hybrid approach that uses tiles at multiple scales may be complicated to visualise as interpolations may be required to fuse them together.

On the other hand, soft attention mechanism based on individual tissues allows for improved diagnostic interpretability strictly confined to the anatomical boundaries of these tissues. Figure 6.4 shows how our visualisation scheme (left)

compares to the standard approach (right), which uses deep features from rectangular tiles. The models producing these exemplary overlay images were both trained to predict Remuzzi G grade. Both overlays show that the networks have learned to attend to the Glomeruli. In the standard approach, we see that the tiles around most glomeruli are highlighted in red, but in many cases, it may not be clear to an inexperienced observer which tissues within the tiles triggered the prediction. Conversely, our approach clearly pinpointed the tissues of interest as the glomeruli are mostly highlighted in red, whereas irrelevant tissues, including those adjacent to the glomeruli, are shaded in blue.

Saliency maps within individual instances can be produced if greater clarity is desired at the cost of processing time. In implementations where only deep features are used, it may be possible to produce saliency maps directly by chaining up feature extractors with the soft attention model while keeping track of gradients in only a small number of instances. However, if handcrafted features are included, direct localisation to the image space would not be possible using the original network. In these cases, heatmaps can be produced with an ad-hoc network trained using only the relevant tissues (as given by the soft attention model) to predict slide-level labels.

We demonstrate the overlay of saliency map based on a soft attention model trained on ResNet50+handcrafted features to predict binary ATI grades as this is a case with high AUC (corresponding model in Figure 6.7). Using the attention values and slide labels, we trained a ResNet18 model to predict ATI slide labels using cropped individual tissue compartments as inputs. The crops used here are the same as those discussed in Section 6.2.3, but have been resized to 128x128 for input into the ResNet18 with adaptive average pooling. The network is trained using $Loss = \sum_{k \in B} a_k(a_k - \tilde{a}_k)^2/|B|$, where a_k is the attention value (Equation 6.2) of an instance k which is used as the label here; $\tilde{a}_k$ is the output of the ResNet18; batch size $|B|$ is set to 24. Since $a_k \in [0, 1]$, the first term in the loss means that instances with higher attention scores would contribute more to the overall loss. Figure 6.5 shows the result of this attempt - a saliency map within each tissue is produced using an occlusion-based approach [191] using the Python package Captum [192],

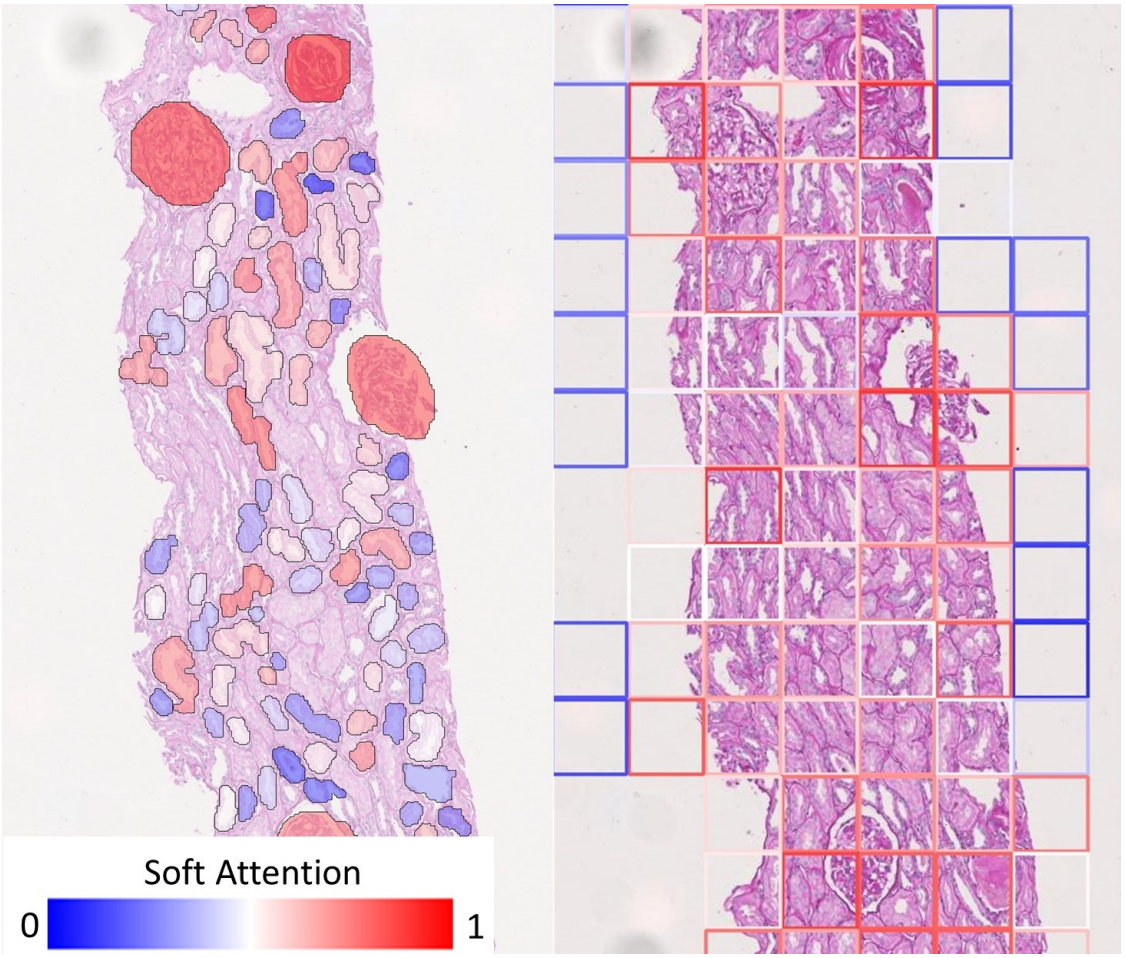


Figure 6.4: Comparison of visualisation of attention map. (Left) Our visualisation scheme highlights individual tissue instances relevant to the prediction. (Right) If rectangular tiles are used, diagnostic could be ambiguous as the tile boundaries do not generally convey any diagnostic meaning. This makes it hard to pinpoint the relevant tissue if they are much larger or smaller than the tile. While it may be possible to produce a smoother heatmap by using overlapping tiles, this will merely be a visual gimmick - the resolution of the attention map cannot be improved because using a large sliding window to compute attention over different parts of the slide is equivalent to a low pass filter. We won't be able to localise small relevant objects this way.

whereas the outline of the tissues is colour-coded by the instance-level attention. We can see that the attention model has learned to focus on the proximal tubules in the cortical regions of the biopsy (outlined in red). Within these tissues, there is some evidence that the saliency maps highlight areas close to the boundaries where the epithelial cells are located. Apart from the occlusion-based approach, we also attempted to use Integrated Gradients [193] and Noise Tunnel [194], but visualisation was found to be less intuitive as the saliency maps generated were

too rough for the magnification we worked on.

6.3.4 Confidence-Weighted Predictions

In the previous sections, we presented results whereby tissue features were combined using soft attention models without accounting for the quality of the segmentation. We argue that the soft attention vectors learned by the models may not always be meaningful. This will happen if the instances do not resemble those in the training data, in which case the presence of the instance would only contribute to noise. We propose a scheme to consider the quality of the instances in the form of Equations 6.2 and 6.4. Under the proposed scheme, $\mathbf{g}$ is added as a feature to the original featureset and each instance k is weighted by g_k during both training and testing time.

Figures 6.6, 6.7, and 6.8 show ROC/PR curves on three different tasks, all of which were weighted by segmentation quality in the models. The solid plot lines show the median values, and the shaded regions show the range of five bootstraps. Figure 6.6 is a plot of the "Trivial" task where models are trained using only ResNet50 features to classify whether a slide contains an adequate number of glomeruli. This task differs from counting glomeruli as there are duplicated tissue sections in 268 out of 612 slides (some slides contain multiple cut-throughs of the same biopsy). A naive model that only counts glomeruli would overestimate the number. We tuned our model based on this task as all slides are labelled so that the dataset-induced noise would be minimal.

Figure 6.7 shows ROC/PR curves for ATI grades where each solid line (Grade 0-2) represents the performance in terms of one-vs-other classification with models trained using ResNet50 combined with handcrafted features. If the grading is instead re-formulated as a binary task (labelled "Binary" in the plot), grouping all > 0 grades together and re-training the models, we get a mean $AUC = 0.874 \pm 0.016$. The optimal operating threshold can be found using the curve's intercept with the maximum iso-accuracy line. At this threshold, we report a performance of $tpr = 0.985$ and $fpr = 0.495$.

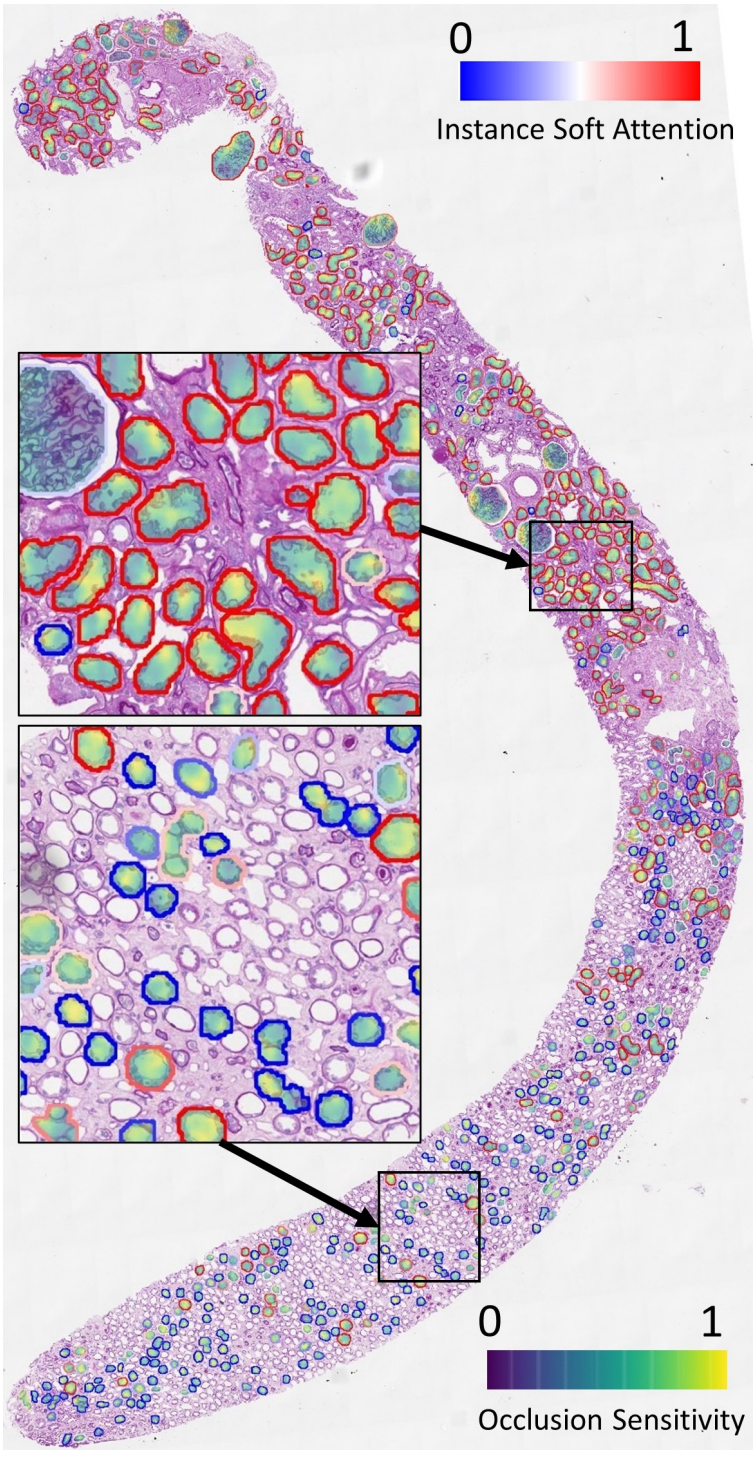
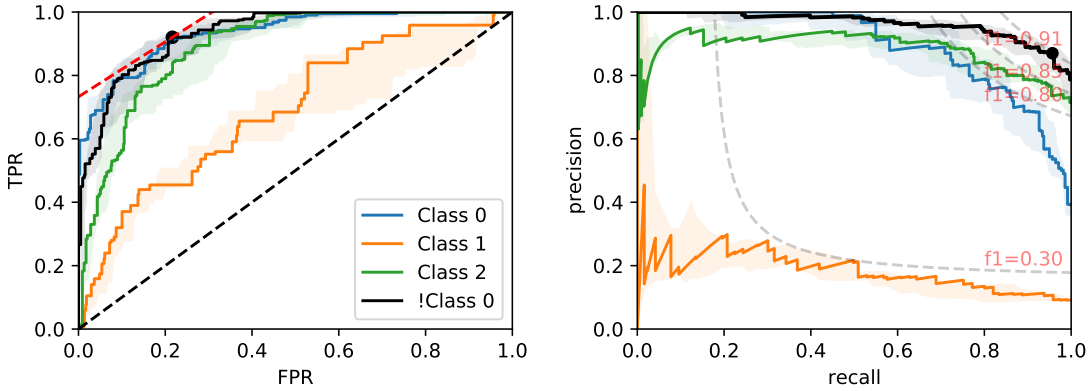


Figure 6.5: Visualisation of instance-level attention combined with saliency maps from individual tissues. Tissue boundaries are coloured by the learned instance-level soft attention value, indicating its relevance to the slide-level prediction. In this example, the slide-level label is ATI. The image clearly shows that our network can attend to proximal tubules (mostly outlined in red). The saliency maps are produced from an occlusion-based approach using an additional network trained on individual tissues to predict the slide label.

As for DGF, we find that performances are generally higher in SR slides. Figure 6.8 shows the performance of the models trained on ResNet50 and handcrafted features. Based on the optimal threshold from the ROC curve, we get $tpr = 0.645$ and $fpr = 0.277$.

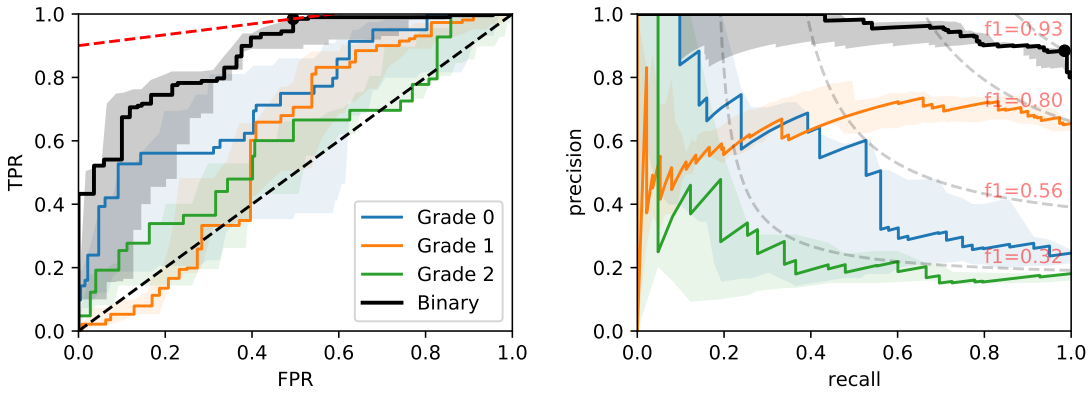


(a) *!Class 0*: Median ROC-AUC=0.934 **(b)** *!Class 0*: Median PR-AUC=0.963

Figure 6.6: ROC/PR plots from models weighted by segmentation quality predicting whether a PAS-stained slide has an adequate number of unique glomeruli for assessment (3 classes). *Class 0*: < 7; *Class 1*: $7 - 9$; *Class 2*: ≥ 10; *!Class 0*: ≥ 7 glomeruli. The curves with the median AUC from 5 bootstraps are shown as solid lines. There are duplicated tissue sections in approximately 268 slides, so models are likely to overestimate the number of unique glomeruli. The marker on the plots shows the optimal operating threshold for the respective class. ROC/PR curves for *!Class 0* represent the performance when *Class 1-2* are grouped as a single class.

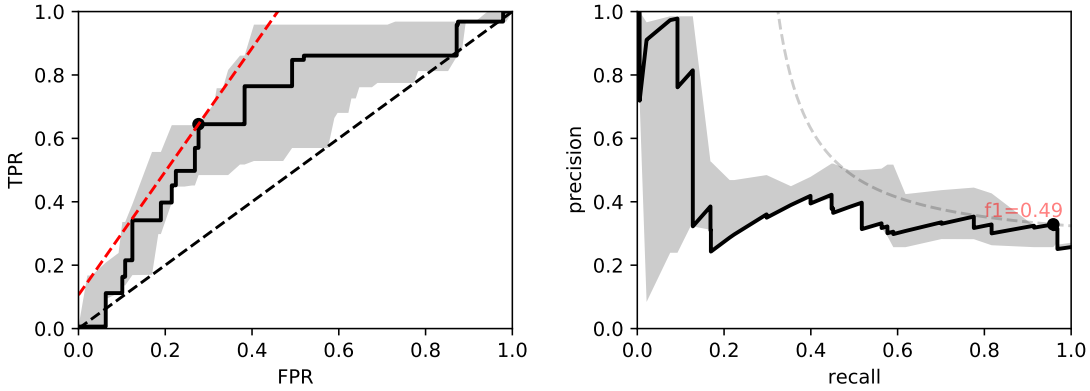
Table 6.3 shows the mean AUCs for different tasks with the proposed addition where instances are weighted by segmentation quality. Entries in this table correspond to those in Table 6.2. Again, comparing individual predictive tasks is not always possible due to the limited size of the labelled data, but an improvement can be seen when we average over different tasks. For example, when we use ResNet50 and handcrafted features (column "Tissue ResNet50"), implementation of g has led to an increase in ROC-AUC and PR-AUC from $(0.598 \pm 0.011, 0.279 \pm 0.011)$ to $(0.618 \pm 0.010, 0.313 \pm 0.014)$.

While weighting tissues by segmentation quality has provided a performance boost for most prediction tasks, the boost is largest for the prediction of Remuzzi IF (average ROC-AUC improvement: 0.101 ± 0.023). One possible explanation is that interstitial fibrosis is a histological change not so well captured by our



(a) *Binary*: Median ROC-AUC=0.873 **(b)** *Binary*: Median PR-AUC=0.963

Figure 6.7: ROC/PR plots from models weighted by segmentation quality predicting ATI grades based on PAS slides. Models are trained using ResNet50 and handcrafted features. Curves with the median AUC from five bootstraps are shown as solid lines. (*Grade 0-2*): Models trained to predict ATI grades 0, 1, and > 2; *Binary:* ROC/PR curves from models specifically trained to predict only two grades: $(0, > 0)$.



(a) ROC: Median AUC=0.676 **(b)** PR: Median AUC=0.406

Figure 6.8: ROC/PR plots from models weighted by segmentation quality predicting the presence of DGF after transplant based on SR slides. 37 out of 143 cases exhibited DGF. Models are trained using ResNet50 and handcrafted features. Curves with the median AUC from 5 bootstraps are shown as solid lines.

featuresets as the changes reside between, rather than within, the segmented tissues. Fibrosis also correlates strongly with tubular atrophy, which is characterised by changes in the basement membrane near tissue boundaries. Neither of these will be visible if objects are under-segmented. In our case, most instances with poor segmentation quality are also the ones that are under-segmented due to how the segmentation results were combined (Equation 6.3). Consequently, these tissues also become less relevant to the prediction task.

There may be a second reason why the proposed weighting improves predictions. As biopsies tend to contain many tissues, it is often not possible to delineate every tissue within a slide. The annotator may have the tendency to choose regions where tissues are legible and create training sets based on these examples. Even though the delineation task was not performed by an expert, the choice of tissues on its own may still contain information indicative of what constitutes as "good quality" or what counts as "relevant".

6.4 Limitations

Some limitations need to be addressed in our experiments. Firstly, it is not clear whether the advantage of combining handcrafted features with deep features will still hold for larger datasets. The choice of the optimal model may change as the size of our dataset grows.

Secondly, there are also some limitations to our visualisation scheme. When there are multiple relevant instances in a slide, soft attention would only focus on a small number of instances in most cases. Many relevant instances are predicted low values. Therefore, it remains challenging to quantify the number of relevant tissue compartments in a slide. Furthermore, if there are several correlated histological changes or commonly co-occurring diseases, soft attention may learn to highlight instances associated with either of these diseases. This would result in correct slide-level prediction but a wrong focus of attention. In our case, TA is highly correlated with IF and it is possible to predict IF without segmenting any interstitial tissues. Isolating highly correlated pathological changes may be possible through

Table 6.3: Segmentation Quality-Weighted Performance. Primary performance is in ROC-AUC. Mean PR-AUC over different prediction tasks is also given in the bottom row. Comparison between the weighted (bottom half of the table) and unweighted (top half of the table) show that segmentation quality weighting instances significantly boosts performance.

Label/Stain	Tissue Resnet50	Tissue Resnet50 (ATI)	Tissue VGG16	Tissue InceptionV3	Tissue Resnet50 Only	Tissue HC
	Unweighted					
ATI/PAS	0.673 ±0.021	0.651 ±0.013	0.645 ±0.02	0.530 ±0.017	0.535 ±0.026	0.525 ±0.019
DGF/PAS	0.504 ±0.025	0.512 ±0.023	0.534 ±0.024	0.527 ±0.031	0.521 ±0.027	0.463 ±0.035
DGF/SR	0.649 ±0.031	0.634 ±0.021	0.588 ±0.029	0.655 ±0.044	0.607 ±0.03	0.579 ±0.039
Remuzzi A/ PAS	0.59 ±0.028	0.581 ±0.024	0.549 ±0.026	0.593 ±0.033	0.539 ±0.024	0.594 ±0.027
Remuzzi G/ PAS	0.556 ±0.049	0.566 ±0.047	0.568 ±0.034	0.588 ±0.039	0.575 ±0.05	0.580 ±0.038
Remuzzi IF/ PAS	0.427 ±0.029	0.526 ±0.036	0.509 ±0.075	0.531 ±0.026	0.43 ±0.075	0.521 ±0.035
Remuzzi TA/ PAS	0.571 ±0.047	0.588 ±0.034	0.638 ±0.053	0.479 ±0.064	0.528 ±0.056	0.569 ±0.05
Mean *ROC-AUC*	**0.598 ±0.011**	**0.598 ±0.009**	**0.596 ±0.009**	**0.553 ±0.009**	**0.545 ±0.012**	**0.542 ±0.011**
Mean PR-AUC	**0.279 ±0.011**	**0.283 ±0.009**	**0.28 ±0.011**	**0.269 ±0.011**	**0.259 ±0.017**	**0.258 ±0.014**
	Weighted					
ATI/PAS	0.652 ±0.021	0.644 ±0.015	0.644 ±0.013	0.567 ±0.012	0.654 ±0.012	0.581 ±0.019
DGF/PAS	0.591 ±0.022	0.543 ±0.022	0.507 ±0.024	0.539 ±0.022	0.569 ±0.022	0.489 ±0.027
DGF/SR	0.642 ±0.036	0.635 ±0.037	0.653 ±0.028	0.638 ±0.045	0.613 ±0.027	0.562 ±0.033
Remuzzi A/ PAS	0.589 ±0.028	0.601 ±0.027	0.578 ±0.028	0.575 ±0.026	0.589 ±0.028	0.625 ±0.024
Remuzzi G/ PAS	0.569 ±0.033	0.560 ±0.048	0.551 ±0.035	0.526 ±0.031	0.562 ±0.038	0.506 ±0.038
Remuzzi IF/ PAS	0.648 ±0.025	0.600 ±0.027	0.613 ±0.021	0.553 ±0.029	0.589 ±0.021	0.548 ±0.035
Remuzzi TA/ PAS	0.585 ±0.073	0.596 ±0.045	0.597 ±0.055	0.608 ±0.072	0.562 ±0.056	0.538 ±0.06
Mean *ROC-AUC*	**0.618 ±0.01**	**0.609 ±0.009**	**0.610 ±0.009**	**0.562 ±0.009**	**0.617 ±0.008**	**0.563 ±0.011**
Mean PR-AUC	**0.313 ±0.014**	**0.291 ±0.013**	**0.290 ±0.012**	**0.271 ±0.015**	**0.292 ±0.012**	**0.261 ±0.014**

arithmetic operations with soft attention maps, but this may require a sufficiently large training dataset.

Thirdly, we realise that the ROC-AUC derived from these experiment appears to be rather low for clinical or research purposes. Obviously, the performance of predictions depends on the quantity and quality of the available datasets and the exact prediction task. This work will also benefit if it is developed on a dataset with more disease cases available. For example, there are currently only several dozen of sclerosed glomeruli from all datasets combined. Hence the segmentation performance may not generalise well to diseased instances. We expect ROC-AUC values to improve if we have access to more and larger slides, possibly with labels stratified by inter-observer variation.

In some cases, we recognise that performance of multi-instance learning may still not match models trained by fully-supervised approaches. If full supervision is needed, our proposed platform can be used as a guide for a human-in-a-loop labelling scheme to bootstrap a project. For example, we can choose to label specifically only tissue compartments with high attention values from slides that have been given a wrong prediction. The soft attention model can then be repeatedly re-trained with labelled instances excluded from the slide in each iteration. This may help to reduce the time needed for pathologists to go through the entire dataset when we want to add a new diagnosis.

6.5 Conclusion

We propose a scheme to improve the performance of multi-instance prediction tasks based on soft attention models by incorporating weak labels at the local level. This approach could make projects that are currently bottle-necked by expert annotations more scalable. In our case, the weak labels are the tissues' outlines delineated into coarse classes. These delineated tissue instances provided several advantages over featuresets extracted from rectangular tiles.

Firstly, having features extracted from tissues allow us to design handcrafted features inspired by domain knowledge. We show that the generalised performance

improves significantly when handcrafted features are combined with deep features compared to models trained using only deep or handcrafted features.

Secondly, features from rectangular tiles do not generally conform to the functional boundaries of tissue compartments. Using features based on tissues in soft attention models allows an intuitive visualisation scheme pinpointing relevant tissues for transparent diagnostics.

Thirdly, we argue that the attention values predicted could only be meaningful if the images resemble the training data. Instances significantly different from the training set could contribute to noise at the bag-level prediction, so we propose incorporating a tissue quality metric derived from an ensemble of BNNs to reduce their impact.

7

Detection of Acute Kidney Injury and Delayed Graft Function from Pre-Implantation Biopsies

Contents

7.1 Introduction 178
7.2 Method 183
7.2.1 Dataset 183
7.2.2 Histological Features and Models 186
7.3 Results and Discussion 187
7.3.1 Classification Performance 188
7.3.2 Association between AKI, ATI, and DGF 190
7.3.3 Association between Algorithmic Histology Score and DGF 193
7.3.4 Utility Value of Algorithmic Histological Scores 199
7.3.5 Visualisation of Relevant Tissue Compartments 200
7.3.6 Limitation of ATI Grades 204
7.4 Limitations and Further Work 207
7.5 Conclusion 208

This chapter discusses the application of the quantitative histology workflow described in Chapter 6 for detecting AKI in pre-implantation biopsies. Several versions of the histology-based models trained on different labels are compared - 1) models trained to mimic pathologists' assessment of acute tubular injury (ATI); 2) models trained purely based on donors' AKI clinical state. The models' predictions

are benchmarked using two cohorts newly introduced in this chapter.

After that, algorithmic scores are correlated against Delayed Graft Function (DGF), a complication that may lead to more extended hospitalisation and potential impact on long term post-transplant function. We propose an algorithmic ensemble score obtained by summing numeric scores from these different models. We find the score associated with DGF (RR=5.94, 95% CI: 1.52-23.3) on the test slides. While, within these test slides, the original AKI and ATI scores are not predictive of DGF.

As part of a feasibility study, we inspected the differences in attention maps between models trained on AKI and ATI and highlighted key discrepancies between the different models.

This chapter contains five main sections. Section 7.1 introduces AKI and its clinical significance with a focus on organ donation and transplantation. Section 7.2.1 describes the datasets used in this experiment. Section 7.2.2 describes the implementation details of the histology-based algorithmic models. Next, Section 7.3 details the results – we demonstrate the performance of the models on two test sets and their ability to generalise to histology (Section 7.3.1). In Section 7.3.2, we show both AKI (RR=1.95, 95% CI 1.3-2.93) and ATI (RR=2.09, 95% CI 1.01-4.32) are associated with an increased risk of DGF when all slides are included. However, when individual cohorts are considered, AKI appears to be a slightly more robust predictor of DGF. Section 7.3.3 proposes an algorithmic ensemble score and found it significantly associated with DGF in the test cohort cases (n=94, p=0.00638, RR=5.94; 95% CI 1.52-23.3), while neither AKIN nor ATI is associated with DGF in these cases. In Section 7.3.4, we find the proposed algorithmic histology score only correlates moderately with the donor variables leading to AKI diagnosis, indicating histological information has the potential to add value to the existing diagnosis of AKI. Section 7.3.5 describes the visualisation of the attention map between models trained to reproduce ATI grades and AKIN scores. Section 7.3.6 discusses the different constraints of the pathologist-assessed ATI scale that limits its predictability of DGF. Section 7.4 discusses the limitations of this study and how to extend this work. Section 7.5 summarises the chapter.

7.1 Introduction

Acute Kidney Injury (AKI) can have donor or recipient origins in transplant settings. Pre-implantation biopsies may reveal AKI in deceased donors, whereas biopsies obtained at early post-transplant stages may reveal injuries accumulated during donor management, transplant operation, and reperfusion following anastomosis. Pre-transplant AKI events have been associated with post-transplantation complications such as Delayed Graft Function (DGF) [15–17]. DGF events have been independently linked to higher rates of short- and long-term graft loss [33]. Post-transplant biopsies with DGF were also found to contain a high proportion of cases with acute tubular injury (ATI) - histological changes relating to AKI [39, 47]. Donors with severe AKI stage 3, as defined by Acute Kidney Injury Network (AKIN) criteria, have also been shown to increase morbidity in recipients [7]. The potential of recovery from AKI may depend on the exact cause, its severity, and donor-related characteristics such as age and comorbidities [36]. Early stratification of AKI severity and causes may improve the prediction of the post-transplant function of donor kidneys and identify those grafts at risk of developing DGF or even primary non-function (PNF).

Currently, there are multiple definitions of AKI [34–36], all of which are clinical states calculated from the serum creatinine and urine output of adult donors. However, multiple causes could lead to a diagnosis of AKI. For example, creatinine and urine levels can be influenced by drugs or changes in fluid intake. AKI can also have pre and post-renal causes in addition to vascular, glomeruli, and tubular diseases [24, 54]. The recording of clinical parameters may also occasionally include errors in entries or missing values that can affect the analysis. All the above reasons may have limited the utility of donor AKI as a predictor of post-transplant complications.

AKI often comes with visible histological changes known as Acute Tubular Injury (ATI) - formerly also known as Acute Tubular Necrosis (ATN) or Acute Renal Failure (ARF). However, the diagnosis of these histological changes has not been standardised; hence inter-observer variability has been reported to be large [39–41] and poses a key challenge to histological assessment. For these reasons,

the assessment of acute lesions is not implemented by many transplant centres as it is possibly regarded as not being worthwhile for pathologists' time. Thus, how well the current histological assessment criteria reflect acute physiological changes remains an open question.

A quantitative platform for detecting pre-transplant AKI from biopsies can benefit nephropathology in several ways. Firstly, algorithms have the potential to remove subjectivity and provide a more reproducible assessment. Secondly, an algorithmic histology score can be used as an additional input to predicting post-transplant complication risks. It may be able to detect focal/segmental/sub-clinical histological changes and summarise findings as a continuous score. Since pathologists routinely score time zero biopsies in many transplant centres, the extra assessment of acute injuries can come without additional cost. Thirdly, an interpretable output for detecting AKI may improve our understanding of the underlying physiological mechanisms through visualisation and stratification of donors with and without notable histological changes.

Finally, in the long term, such a tool can help us with fundamental research in the understanding of kidney diseases and the development of treatments. It may be possible to use these artificial intelligence tools to help us develop better guidance for organ preservation, procurement, and histological assessment. For example, studying AKI may help us understand the progression of chronic kidney diseases (CKD), as repeated episodes of AKI are the precursor of CKD [49, 51–54]. This rare opportunity to study biopsies in transplant settings may also lead to an impact in the context beyond transplantation by enhancing our knowledge of the progression of AKI-related physiological changes. This is important as the global burden of AKI-related deaths far exceeds breast cancer, heart failure or diabetes [49].

A schematic of the association between AKI, ATI, and DGF under transplant settings is shown in Figure 7.1. In the schematic, associations based on our initial cohort (QUOD-PAS) are framed in red. We also show several larger studies associating the three variables and long term transplant outcomes. Some statistics based on our datasets are discussed later in Section 7.3.2. Analysis based on this

initial cohort suggests donor AKI is a significant risk factor for DGF, whereas ATI only becomes significant when different cohorts are combined. Association between AKI and ATI means histological changes can also be seen in most cases with AKI (as defined by AKIN) scores greater than 0.

The limitations of using either ATI or AKIN scores to predict DGF risk are summarised in Table 7.1. In particular, we hypothesise seven different factors that limit the usability of ATI grades given by pathologists to predict DGF risks:

1. **Scores given by pathologists are discrete** and may not reflect continual visual changes. Some histological changes might be visible but do not meet the thresholds of the grading criteria. Visual grading scales might have also been defined at levels that do not scale linearly with clinical changes.

2. **Features not already described may exist** in addition to the current visual changes for assessing ATI - loss of brush borders in proximal tubules, drop-out of the epithelial cell cytoplasm, and cell necrosis.

3. **Lack of coherent guidelines on how to represent localised histological changes** (focal/segmental) as a single score. Visual changes along different sections of the nephron may also have different significance. It is hard to tell apart distal from proximal tubules when they are damaged. It is also hard to tell apart different sub-sections of proximal tubules under routine stains.

4. **Subjectivity and lack of reproducibility** from human pathologists.

5. **Biopsy area** may be too small for accurate assessment.

6. **Non-biopsy AKI causes**. Changes might have occurred in other regions of the kidney other than the cortex.

7. **Recipients/surgical risk factors**. Recipients are age-matched to donors. This means donor kidneys with AKI are more likely to be transplanted to recipients who may develop post-transplant AKI. Surgical complications may also lead to DGF.

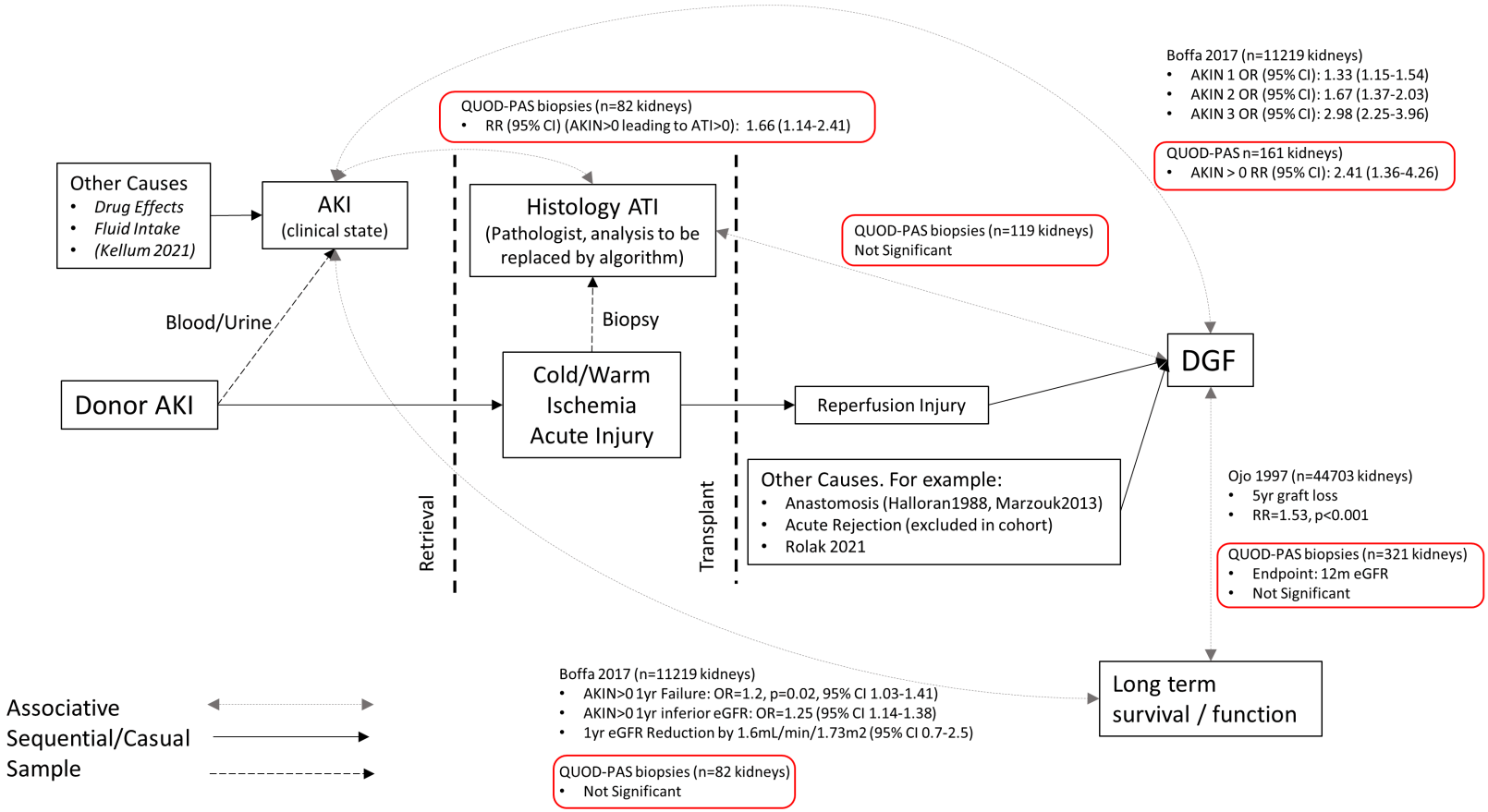


Figure 7.1: Association between AKI, ATI, DGF, and long term graft function in existing studies and the main dataset (QUOD-PAS) used in this thesis.

8. **Recipients/surgical risk factors**. Transplant centre factors have been shown to contribute to an additional 42% DGF risk [44] after accounting for donor factors. Transplant centre factors include organ procurement or surgical variables (e.g. anastomosis). Injuries added to kidneys at the procurement stage may be visible in biopsies, but surgical factors would not be detectable. It is not clear what the proportional contributions are between procurement and surgery. Recipient-related DGF risk factors may be small, as evident from the low incidence of DGF in living donors [195] and studies of deceased-donor pairs [44]. However, there is the possibility that recipient-related factors may be underestimated as donors are often age-matched to recipients.

An algorithmic-based histological analysis tool may help to assess and reduce the impact of contributions 1-4. The importance of contribution 5 may be studied by excluding the dataset according to biopsy size. Contributions 6-7 are fundamental limitations in histopathology and are also limited by how cohorts were selected. Donor AKI status can be used to reduce contribution 6. A robust histological analysis tool can be used to estimate the upper-bound proportion of post-transplant DGF causes.

Table 7.1: Limitations of AKI diagnosis using clinical variables vs histology

	Advantages	Disadvantages
Donor AKIN Score	• Accounts for non-biopsied and pre/post renal cause such as cardiac failure.	• Non-specific of the exact causes of high creatinine / low urine output.
ATI histology score based on pathologists	• Improved ability to determine physiological cause of AKI. • Ability to account for patient's clinical background.	• High inter-observer variability and subjectivity. • Difficulty summarising focal/segmental acute lesions into an ordinal grade. • Non-biopsy causes not available. • Sample volume of biopsy could be small.
Histology score based on algorithms	• Improved ability to determine physiological cause of AKI. • Outputs a continuous score, meaning that histological changes below clinical thresholds can also be captured. • Objective and quantitative.	• Non-biopsy causes not available. • Model performance issues when trained under weakly-supervised settings with limited data. • Sample volume of biopsy could be small.

7.2 Method

In Chapter 6, we describe the use of a soft attention multi-instance learning model to predict slide-level labels. The proposed implementation uses a combination of handcrafted and deep features extracted from various tissue functional compartments from renal biopsies and is weighted by the segmentation quality of the instances. Not only does the workflow output a slide-level prediction, but it is also capable of displaying an attention map overlay on individual functional compartments, highlighting instances it considers important for contributing to the slide's prediction. In this chapter, we demonstrate the use of this platform to detect cases of AKI based on histology. We train different sets of models using the QUOD-PAS slides to reproduce the ATI labels (given by the pathologist) and AKIN labels. The performance of the models is validated on two datasets substantially different from the QUOD-PAS slides. We first describe the datasets used in this experiment in Section 7.2.1. After that, we describe the implementation details in Section 7.2.2. For statistics tests, the Mann-Whitney-Wilcoxon (MWW) test is used to compare numeric variables in pairs of categorical labels, and the Kruskal-Wallis (KW) H test is used to compare numeric labels in more than two sets of categorical labels. Categorical-categorical associations of tables larger than 2x2 are evaluated with Cramer's V test. As a criterion for chi-square statistics, if more than 20% of cells in a table with ordinal scores are less than 5, higher grades are regrouped into the next lower grade. Fisher's Exact Test is used for 2x2 tables, including tables that have been regrouped as a result of not meeting the criteria for chi-square statistics. As this thesis is centred around the established theory that higher ATI/AKI grades lead to an increased risk of post-transplant complications, we report one-sided p-values.

7.2.1 Dataset

Three sets of slides were analysed in this study. All slides consist of pre-implantation biopsies preserved as paraffin sections. The first two sets of slides (QUOD-PAS, and QUOD-HE AKI) were from the QUOD biobank. The first set of QUOD slides (373 slides) was PAS-stained and assessed by a pathologist who provided ATI scores.

The second set of slides (39 slides) was also from QUOD but was specifically selected for AKI characteristics for molecular studies [122] and was stained in H&E – these had not been assessed for ATI. The third set of slides contains 151 PAS-stained slides that were obtained from the University of Amsterdam (AMC) [1]. A different pathologist assessed the AMC-PAS slides according to the same qualitative criteria defined in Chapter 3, Section 3.3.1. The AMC-PAS slides consist of 91 living donors and 60 cadaver donors. Sets two and three were new to this project. A large proportion of slides in all three sets contain information on whether the donor has AKI (as defined using the AKIN [35] criteria) and DGF (present/absent). DGF is defined as the need for at least one session of dialysis the first-week post-transplant with hyperacute rejection, vascular and urinary tract complications, and hyperkalemia excluded [42]. The distributions of ATI, AKI, and DGF status of the three sets of slides are shown in Table 7.2.

Table 7.2: Distribution of ATI, AKIN, and DGF class

# of Slides	ATI Scores (0-3)	AKIN Scores (0-3)	DGF (Absent/Present)
QUOD-PAS	47, 77, 19, 2	125, 19, 3, 39	244, 77
QUOD-HE AKI	-	19, 12, 5, 3	30, 9
AMC-PAS	13, 61, 58, 19	143, 5, 1, 2 (52,5,1,2)**	117, 23 (34, 21)**
NMP-FakePAS	4, 117, 23, 3	-	-

** Living Donors Excluded

A Venn diagram showing the number of slides with different labels simultaneously available is shown in Figure 7.2. Note that the set of slides where all three labels are available is not very large (n=126 cases from all three cohorts). Hence for statistical analysis comparing how well ATI and AKIN predict DGF (Section 7.3.2), we use the largest available set for each analysis (ATI $\cap$ DGF; AKIN $\cap$ DGF). For associative analysis comparing our algorithm against ATI and AKIN in predicting DGF, we use the set ((ATI $\cup$ AKIN) $\cap$ DGF) so that p-value comparisons are slightly fairer when put side-by-side against the original ATI/AKIN labels.

[1] All experiments were done in accordance with the Declaration of Helsinki and were approved by the local ethics and privacy committees at Amsterdam UMC location AMC (19.260).

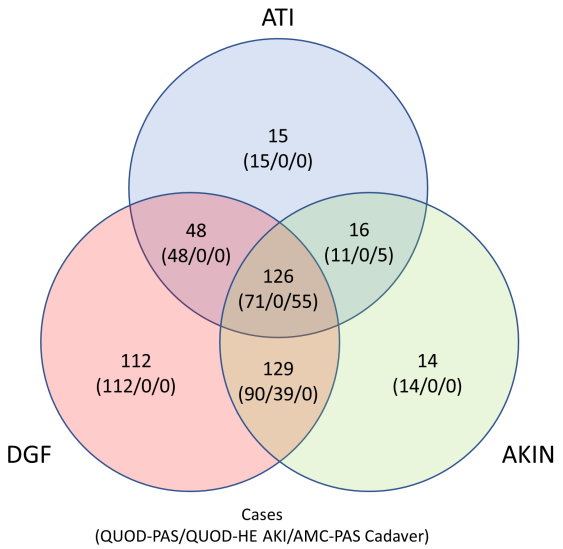


Figure 7.2: Venn diagram showing the number of cases where labels are available. The number of cases in each cohort is shown in brackets.

Note that a small number of QUOD slides contain kidneys from the same donor. Although cases from QUOD were selected based on recipients with similar recipient eGFR12 outcomes, for the current analysis, it would still be beneficial to include as many recipients as possible to average-out post-transplant causes of DGF. In the results section, the performance of the multi-instance learning models is tested against the QUOD-HE AKI and AMC-PAS slides. We shall refer to the combination of these two sets of slides as 'test slides'. However, some interpretations are performed on the QUOD-PAS slides as the test slides only contain five AKIN grade 3 cases altogether, and the AMC-PAS slides do not have donor serum creatinine/urine data available. Characteristics of the datasets are as follow:

QUOD-PAS. These are pre-implantation biopsies from 348 donors and 373 recipients. More details of this dataset are in Chapter 3, Section 3.1.1. ATI grades are available in 145 cases (135 donors), AKIN grades are available in 186 cases (149 donors). There are 82 cases (74 donors) where both scores are available. DGF labels are available for this in 321 cases (283 donors).

QUOD-HE AKI. These are pre-implantation biopsies from 39 donors and 39 recipients. These are new slides obtained close to the end of this project. Hence statistics of this cohort have not been previously described in Chapter 3. While these

slides are also from the QUOD biobank, they were selected to have a quarter of cases with and without AKI (as defined by KDIGO [36]) and with good/bad outcomes (as defined by 1-year eGFR cutoff at 45 $mL/min/1.73m^2$). The cohort was selected for molecular studies [122]. Slides were stained using H&E and scanned using Aperio ScanScope scanner at 0.253mpp - a different scanner from the QUOD-PAS slides. To process these slides with our workflow, they were algorithmically converted to PAS for processing using a CycleGAN (Chapter 3, Section 3.5.1). These slides contain a median number of 5 glomeruli (standard deviation 4.88).

AMC-PAS. These are pre-implantation biopsies from 151 donors and 151 recipients recruited by the University of Amsterdam. Slides were stained in PAS and scanned using a Philips scanner at 0.25mpp. ATI was assessed by a different specialist renal pathologist from the QUOD-PAS slides under the same qualitative criteria. The data contain 91 living donors and 60 cadaver donors. These slides are much larger than the QUOD slides, with a median number of 19 glomeruli (standard deviation 9.72).

NMP-FakePAS. Details of this dataset are described in Chapter 3, Section 3.1.2. These slides were not used for analysis and were used to develop the feature extraction workflow 4. They were also used to train one of our four sets of models (ATI2).

7.2.2 Histological Features and Models

The overall workflow and histological features used in this experiment are the same as proposed in Chapter 6. In this workflow, tissue compartments – tubules, glomeruli, vessels, and nuclei are segmented from the slide. A combination of ResNet50 features and handcrafted features are extracted from each tissue compartment. These features, along with the segmentation quality of different instances (as derived by Equation 6.2 in Chapter 6) are used to train a multi-instance soft attention model.

We wish to see what proportion of cases with AKI can be detected from histology and whether using ATI assessment results from pathologists may help guide AKI classification. To this end, we trained four different sets of models. Each set consisted of binary classifiers to predict whether a slide has a score ATI > 0, ATI > 1, AKIN > 0, and AKIN > 1. We refer to these four sets of models as ATI1,

ATI2, AKIN1, and AKIN2. The models of ATI1, AKIN1, and AKIN2 were trained using 5-fold cross-validation at a ratio of 3:1:1 training/validation/testing split (subjected to constraints in Donor ID) on the QUOD-PAS slides. For the training of ATI2, we find there is an insufficient number of labels from the QUOD-PAS slides. Hence the models for ATI2 are trained using a mixture of QUOD-PAS and NMP-FakePAS slides with 5-fold cross-validation. Predictions of the QUOD-PAS slides are reported based on the test split. Predictions of the other two cohorts are reported based on the mean numeric values from the 5-fold cross-validation models. In addition to the four sets of models, we propose an algorithmic ensemble score by summing up their numeric scores - we call this 'AlgoSum'. When evaluated on the test data, this score is effectively the sum of twenty models (4 sets of models and 5-fold cross-validation). We ensembled these models because AKIN and ATI scores are both known risk factors for DGF, and models trained on different grade boundaries may learn to attend to different features. As models trained on ATI and AKIN each have their limitations, the purpose here is to see whether using an ensemble can lead to a more robust prediction of DGF risks.

7.3 Results and Discussion

In Section 7.3.1 we validate the classification performance of the 4 models on the QUOD-PAS slides and two independent datasets. After that, we describe the association between the different donor scores, whether given (Section 7.3.2) or predicted by an algorithm from histology (Section 7.3.3), against DGF. Section 7.3.4 provides further evidence for the value of histological score as an addition to donor clinical variables for predicting DGF. The proposed workflow also allows the display of tissue compartments the algorithms regard as important for the slides' predictions. As part of a feasibility study, we displayed some examples of attention maps in Section 7.3.5. These attention maps may help pathologists check the validity of predictions. Section 7.3.6 discusses the different constraints of the pathologist-assessed ATI scale that limits its predictability of DGF.

7.3.1 Classification Performance

The purpose of the model evaluation was to show that the models can capture meaningful visual features from histology. Models with decent classification performance may help answer our core clinical question: To what extent can histology, when free from the constraints of human assessment, reflect pre-transplant clinical state and predict post-transplant complications?

Classification performance evaluated on the QUOD-PAS slides was based on the test split from 5-fold cross-validation, where a single model classified each slide. Predictions on the QUOD-HE AKI and AMC-PAS slides were obtained after averaging the numerical probabilities across the 5 cross-validation models. A summary of the performance is shown in Table 7.3. Performance is presented as Receiver Operating Characteristic (ROC) and Precision-Recall (PR) Area-under-the-curve (AUC). To reduce the proportion of recipient and surgical-caused DGF cases and to create a more balanced test set, living donors are excluded from this whole chapter except for AUC benchmarks of the ATI1 and ATI2 models in Table 7.3.

Table 7.3: ROC-AUC and PR-AUC on the three different cohorts. *ATI1* classifies ATI grade 0 vs > 0; *ATI2* classifies ATI grade ≤ 1 vs > 1; *AKIN1* classifies AKIN grade 0 vs > 0; *AKIN2* classifies AKIN grade ≤ 1 vs > 1 from histology.

Task \ Dataset	AUC	QUOD-PAS	Test Data 1 QUOD-HE-AKI	Test Data 2 AMC-PAS	Test Data 1+2
ATI1	ROC	0.874	N/A	0.721	0.721
	PR	0.956		0.967	0.967
ATI2*	ROC	0.521		0.631	0.631
	PR	0.133		0.653	0.653
AKIN1	ROC	0.776	0.839	0.663**	0.713
	PR	0.703	0.861	0.180**	0.345
AKIN2	ROC	0.754	0.804	0.652**	0.744
	PR	0.484	0.736	0.120**	0.417
DGF1	ROC	0.555	0.601	0.638**	0.626
	PR	0.264	0.275	0.442**	0.408

*NMP-FakePAS slides added for training to provide training examples.
** Living Donors Excluded

For the classification of ATI1, performances are (ROC-AUC=0.874, PR-AUC=0.956) on the QUOD-PAS slides and (ROC-AUC=0.721, PR-AUC=0.967) on the AMC-

PAS test set. A drop in ROC-AUC could be caused by 4 main factors: a) Different lab protocols; b) Slides were scored by different pathologists; c) Models overfitted the training data. d) Variations in class distribution between the training and test set. ROC-AUCs alone may lead to misleading conclusions in imbalanced datasets [196]. However, as it can be seen, PR-AUC remains high on the test set, suggesting (d) might be a major cause, and the model remains reliable on the test set.

For the classification of ATI2, performances are (ROC-AUC=0.512, PR-AUC=0.133) on the QUOD-PAS slides and (ROC-AUC=0.631, PR-AUC=0.653) on the AMC-PAS test set. NMP slides that are added during training are excluded from evaluation. Mixing datasets with different label distributions during training may inevitably lead to models learning meaningless dataset-specific shortcuts and impact performance – a practice that could be avoided if larger datasets were available. Despite this, performance on the test set shows that the model maintains some predictability.

For the classification of AKIN1 and AKIN2, performances are (ROC-AUC=0.713, PR-AUC=0.345) and (ROC-AUC=0.744, PR-AUC=0.417) respectively on the combined test data. Two possible causes attributed to the drop in performance: 1) some bias in the training data might be present; 2) only five AKIN grade 3 cases are present in the combined test cohort, so class balance might have influenced AUCs. AUCs are higher in the QUOD-HE AKI test slides than in the training data, even though these slides were originally stained with H&E and algorithmically converted to resemble PAS. The colours of these slides may not functionally resemble true PAS stain. The contrast of nuclei and basement membrane can be poor in some tissue patches on close inspection. Decent AUCs on these fake PAS slides suggest that morphological changes, rather than stain colour changes, are more important for detecting acute injury features.

We plot algorithmic predictions of AKIN grades in Figure 7.3 as predicted by the (a) AKIN1 and (b) AKIN2 models on all slides combined (QUOD-PAS, QUOD-HE AKI, and AMC-PAS cadaver). Each slide that pathologists have assessed has been colour-coded by ATI score. Identical boxplots on just the test slides (QUOD-HE

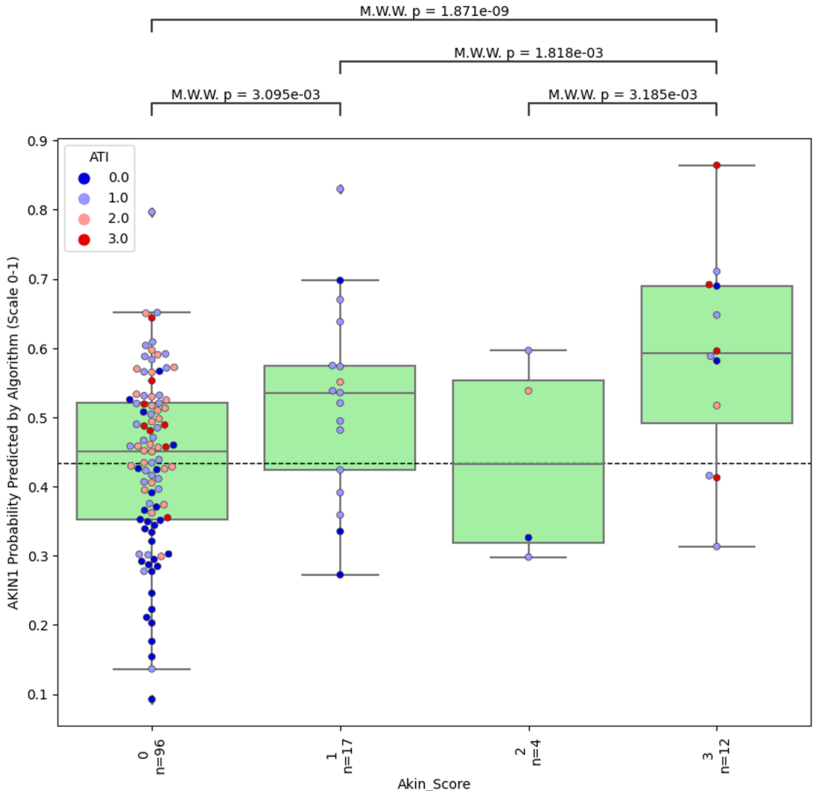
AKI, AMC-PAS cadaver) are shown in Figure 7.4, but these have fewer AKIN grade 3 cases, and a larger proportion of slides have not been assessed for ATI. Slides here are coloured by whether the transplant resulted in DGF in the recipient.

The plots' y-values show the numeric output from the binary classifier models and can be interpreted as algorithmic histology scores. Higher values mean a slide is more likely to be classed as AKIN > 0 and AKIN > 1 by the two respective models. Unlike the models trained on pathologist ATI grades, these models are trained directly from clinical states, so they may have the potential to learn a more linear association between histology and physiology. Significant pair-wise p-values (Mann-Whitney-Wilcoxon test) between different AKIN clinical scores are displayed above the plots. In both models, slides with AKIN grade 1 are given significantly higher algorithmic scores (Figure 7.4: $p = 8.22 \cdot 10^{-3}$ and (b) $p = 1.96 \cdot 10^{-3}$) compared to donors with AKIN grade 0. This suggests that our algorithm can detect subtle histological changes caused by donor AKI. Algorithmic scores for AKIN grade 3 are also significantly higher than grade 0 (Figure 7.4: (a) $p = 5.02 \cdot 10^{-3}$ and (b) $p = 1.08 \cdot 10^{-2}$) and visually higher than grade 1, suggesting our models can detect severe AKI from histology in the majority of cases.

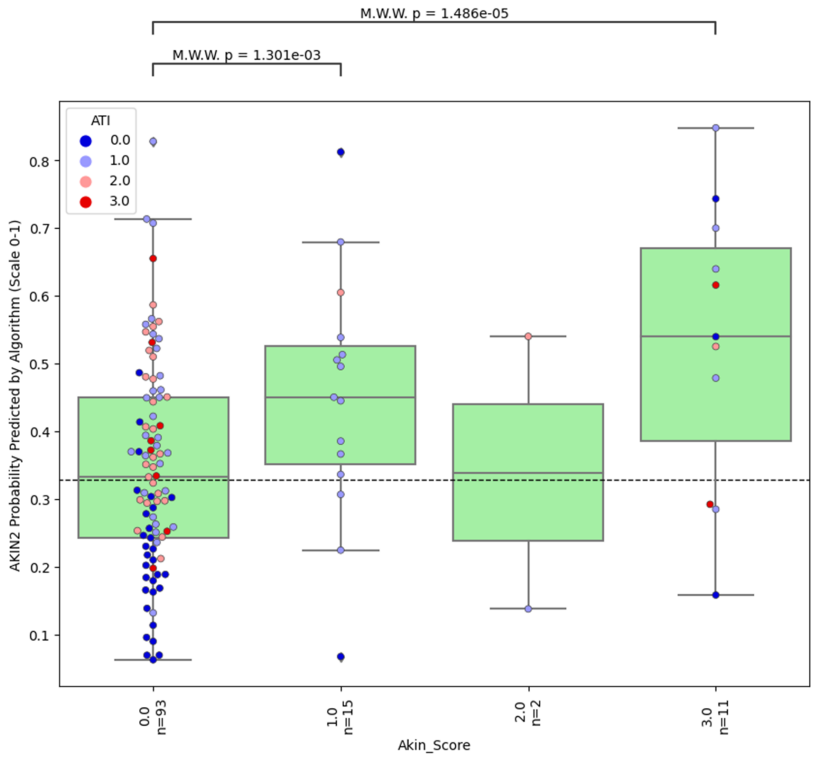
In Figure 7.3a (a), only slides assessed by the pathologist are plotted (p-values above the plot include all slides with AKIN information available - whether assessed or otherwise). The colour of the markers indicates ATI grades given by the pathologist. We find that slides with higher ATI grades are likely to be given a higher score by the models. When only test slides are considered, the mean difference in AKIN1 algorithmic score between ATI grade 0 and >0 is 0.0947, MWW $p = 0.00390$. Association between ATI grades and algorithmic AKIN1 score remains strong when the slides are conditioned to only those with AKIN score 0, indicating the models are capable of generalising to histology and learning continual visual changes.

7.3.2 Association between AKI, ATI, and DGF

Here we show some dataset statistics between the original labels - ATI, AKI, and DGF. As detailed in Table 7.4 (a)-(b), we break down transplant cases according

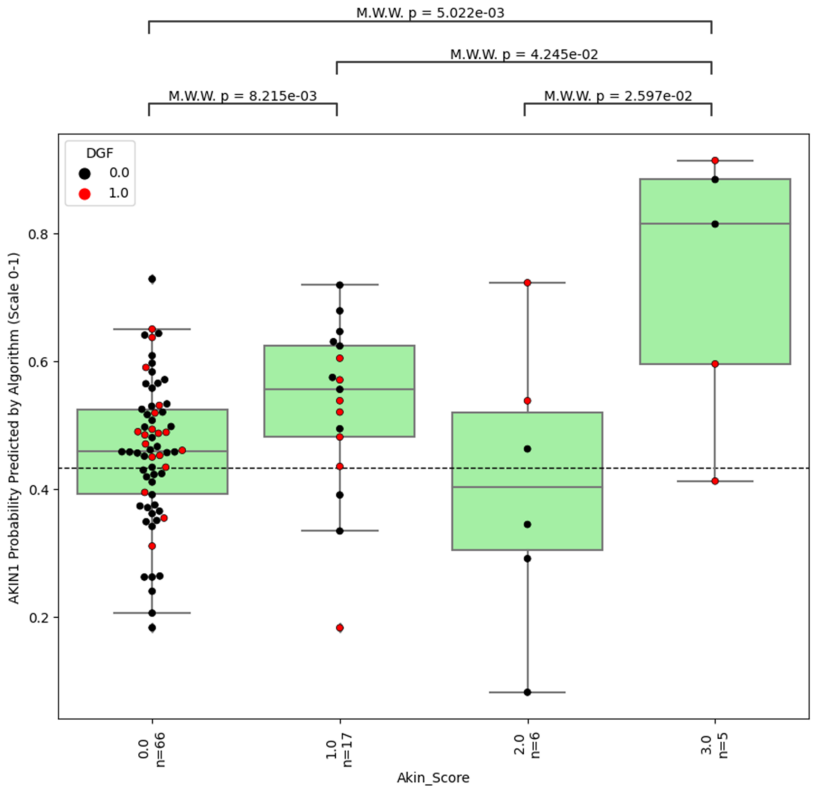


(a) AKIN1

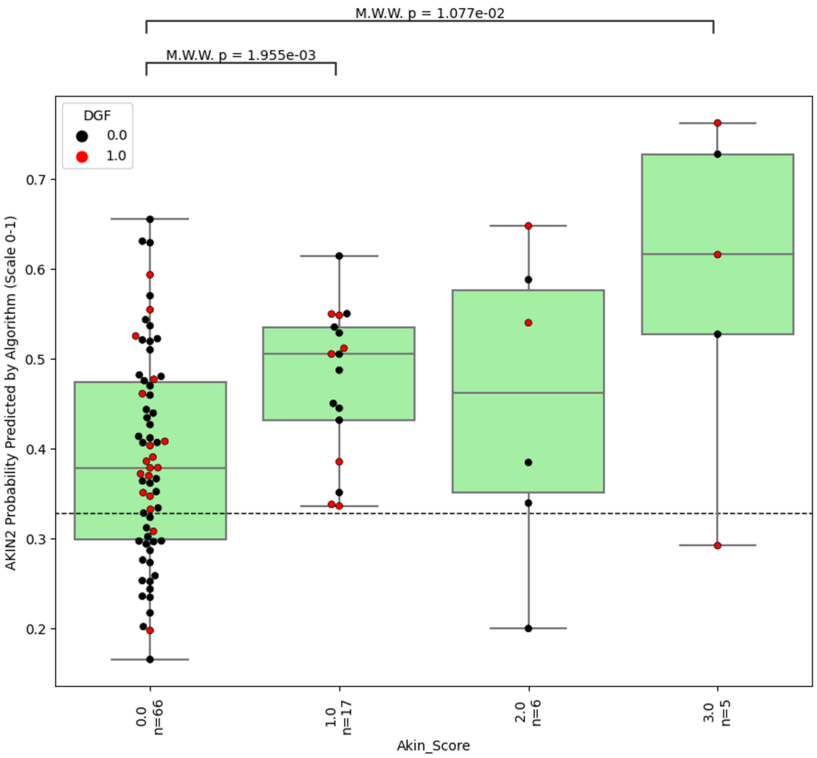


(b) AKIN2

Figure 7.3: Algorithmic numeric score trained on AKIN labels (a) AKIN1 on all slides; (b) AKIN2 on all slides.



(a) AKIN1



(b) AKIN2

Figure 7.4: Algorithmic numeric score trained on AKIN labels (a) AKIN1 on test slides; (b) AKIN2 on test slides.

to different donor AKIN/ATI scores and whether the transplant resulted in DGF (absent/present). Based on all cohorts, we find that both AKIN (Cramer's $V = 0.267$, $p = 4.14 \cdot 10^{-4}$) and ATI ($V = 0.266$, $p = 6.45 \cdot 10^{-3}$) scores are associated with DGF. AKIN grade >0 is associated with DGF risk (RR=1.95, 95% CI 1.3-2.93), and ATI grade >0 is associated with DGF risk (RR=2.09, 95% CI 1.01-4.32). However, the association between ATI is weaker than AKIN. When we break down the analysis to individual cohorts, we find AKIN ($V = 0.371$, p=$1.23 \cdot 10^{-}4$) significant on the QUOD-PAS slides but not ATI (p=0.343). Neither AKIN nor ATI is associated with DGF in the combined test cohort. A comparison between ATI grades and AKIN scores is shown in Table 7.4 (c). We find a significant association between ATI and AKIN scores (Cramer's $V = 0.220$ $p = 8.4 \cdot 10^{-3}$) in all cohorts combined.

7.3.3 Association between Algorithmic Histology Score and DGF

7.3.3.1 Ordinal Scores vs DGF

To compare the association between algorithmic scores against pathologist ATI, clinical AKIN scores, and DGF, we discretise the algorithmic numeric scores into binary grades. Table 7.5 (a)-(f) shows the odds and p-values (Fisher's Exact Test) of the binarised scores against DGF in the different combinations of cohorts. Entries with significant association (p<0.05) are in bold. For a fairer analysis comparing algorithmic scores with the original AKIN/ATI labels, slides with missing AKIN scores have been excluded in these tables. We shall focus the discussion on the test cohorts only (Table 7.5 (e)) in this subsection.

In Table 7.5 (e), we compare the algorithmic scores (rows AKIN1 to AlgoSum), pathologist score (rows ATI>0 to ATI>2), and donor clinical scores (rows AKIN>0 to AKIN>2) binarised at different thresholds against the columns (ATI>0, AKIN>0, DGF). Thresholds of algorithmic scores are defined to maximise the lower bound 95% confidence interval (CI) of relative risk (RR). Algorithmic predictions from AKIN1 (p=0.0473) and AKIN2 ($p = 4.89 \cdot 10^{-3}$) are significantly associated with the donors' AKIN scores. However, neither of these is predictive of DGF. Similarly,

(a) AKIN vs DGF
All Slides (QUOD-PAS/QUOD-HE AKI/AMC-PAS Cadaver)

Slides	AKIN Scores			
	0	1	2	3
DGF 0	138 (90/32/16)	26 (16/2/8)	6 (2/0/4)	19 (17/0/2)
DGF 1	34 (16/15/3)	9 (2/3/4)	2 (0/1/1)	21 (18/2/1)

(b) ATI vs DGF
All Slides (QUOD-PAS/AMC-PAS Cadaver)

Slides	ATI Scores			
	0	1	2	3
DGF 0	38 (33/5)	60 (47/13)	23 (10/13)	4 (1/3)
DGF 1	7 (7/0)	21 (14/7)	14 (6/8)	7 (1/6)

(c) AKIN vs ATI
All Slides (QUOD-PAS/AMC-PAS Cadaver)

Slides	AKIN Scores			
	0	1	2	3
ATI Score 0	34 (29/5)	3 (3/0)	1 (1/0)	4 (4/0)
ATI Score 1	34 (17/17)	14 (9/5)	2 (2/0)	7 (7/0)
ATI Score 2	28 (6/22)	1 (1/0)	1 (0/1)	1 (1/0)
ATI Score 3	8 (0/8)	0 (0/0)	0 (0/0)	4 (2/2)

Table 7.4: Association between different labels in our datasets. Numbers in the brackets in the above tables represent the number of slides in the different datasets as indicated by the respective table headings. Cramer's V test is used as these are categorical tables larger than 2x2. When all slides are considered, the statistics are: (a) AKIN vs DGF ($V = 0.267$, $p = 4.14 \cdot 10^{-4}$); (b) ATI vs DGF ($V = 0.266$, $p = 6.45 \cdot 10^{-3}$); (c) AKIN vs ATI ($V = 0.220$ $p = 8.4 \cdot 10^{-3}$).

neither the original AKIN clinical scores nor ATI pathologist scores are predictive of DGF. We find predictions made by the larger ensemble of models to be the only significant predictor of DGF ($r = 5.94, p = 6.38 \cdot 10^{-3}$).

7.3.3.2 Numeric Algorithmic Scores vs DGF

Table 7.6 (b) shows correlation statistics, as calculated using the Mann-Witney-Wilcoxon Test, between the numeric algorithmic scores and DGF. The ensemble score 'AlgoSum' here appears strongly significant with DGF ($p = 4.35 \cdot 10^{-3}$), while the statistical significance of the constituent components is weaker.

Table 7.5: Association between ordinal scores vs labels. The columns represent the manual and clinical labels. Rows *AKIN1-AlgoSum* are algorithmic scores; *ATI>0-ATI>2* are binarised manual grades; *AKIN>0-AKIN>2* are binarised donor clinical grades. Numbers (z, p) in the table are the odds ratio and p-values obtained from Fisher's exact test as we are comparing pairs of binary labels.

(a) *QUOD-PAS (n=373 recipients)*

Fisher's z,p	ATI>0	AKIN>0	DGF
AKIN1	2.59, 0.0719	**3.48, 0.00854**	**3.84, 0.0144**
AKIN2	**3.29, 0.0334**	**2.98, 0.0175**	2.68, 0.059
ATI1	2.31, 0.103	2.8, 0.0188	1.4, 0.36
ATI2	2.31, 0.103	**2.54, 0.0188**	2.53, 0.0703
AlgoSum	**3.1, 0.0351**	**3.03, 0.00619**	**3.32, 0.0351**
ATI>0	-	**3.47, 0.00953**	1.71, 0.192
ATI>1	-	1.18, 0.534	2.42, 0.0896
ATI>2	-	inf, 0.131	3.33, 0.417
AKIN>0	**3.47, 0.00953**	-	**3.21, 0.00238**
AKIN>1	2.33, 0.117	-	**5.58, 4e-05**
AKIN>2	2.36, 0.142	-	**6.35, 1.37e-05**

(b) *Test Slides 1: (QUOD-HE AKI, n=39)*

Fisher's z,p	ATI>0	AKIN>0	DGF
AKIN1	N/A	2.18, 0.13	2.23, 0.237
AKIN2	N/A	2.22, 0.136	4.08, 0.0886
ATI1	N/A	1.93, 0.201	3.62, 0.0917
ATI2	N/A	0.611, 0.275	1.8, 0.34
AlgoSum	N/A	2, 0.168	3.62, 0.0917
ATI>0	-	N/A	N/A
ATI>1	-	N/A	N/A
ATI>2	-	N/A	N/A
AKIN>0	N/A	-	2.29, 0.252
AKIN>1	-	-	1.14, 0.607
AKIN>2	-	-	1.75, 0.556

(c) *Test Slides 2: AMC-PAS (Cadaver, n=60)*

Fisher's z,p	ATI>0	AKIN>0	DGF
AKIN1	1.47, 0.301	3.6, 0.0955	1.98, 0.182
AKIN2	1.3, 0.35	3.6, 0.0955	2.92, 0.0901
ATI1	1.16, 0.425	3, 0.145	1.93, 0.164
ATI2	1.13, 0.445	3, 0.173	1.72, 0.213
AlgoSum		3.62, 0.1	3.56, 0.0835
ATI>0	-	inf, 0.476	inf, 0.08
ATI>1	-	0.44, 0.246	2.25, 0.127
ATI>2	-	1.83, 0.4	4.13, 0.0625
AKIN>0	inf, 0.476	-	**6.4, 0.0284**
AKIN>1	inf, 0.767	-	inf, 0.0507
AKIN>2	inf, 0.839	-	inf, 0.141

(d) *QUOD slides (a)+(b)*

Fisher's z,p	ATI>0	AKIN>0	DGF
AKIN1	2.59, 0.0719	2.89, 0.00482	**2.93, 0.0128**
AKIN2	**3.29, 0.0334**	**2.46, 0.0208**	2.62, 0.0591
ATI1	2.31, 0.103	**2.08, 0.0203**	1.55, 0.216
ATI2	2.31, 0.103	1.4, 0.188	2.15, 0.0638
AlgoSum	**3.1, 0.0351**	**2.47, 0.0046**	**3.95, 0.00466**
ATI>0	-	**3.47, 0.00953**	1.71, 0.192
ATI>1	-	1.18, 0.534	2.42, 0.0896
ATI>2	-	inf, 0.131	3.33, 0.417
AKIN>0	**3.47, 0.00953**	-	**2.96, 0.00143**
AKIN>1	2.33, 0.117	-	**4.16, 0.000141**
AKIN>2	2.36, 0.142	-	**5.23, 2.59e-05**

(e) *Test slides only (b)+(c)*

Fisher's z,p	ATI>0	AKIN>0	DGF
AKIN1	1.47, 0.301	**2.97, 0.0473**	2.46, 0.0656
AKIN2	1.3, 0.35	**3.39, 0.00489**	3.09, 0.0562
ATI1	1.16, 0.425	**2.5, 0.0465**	3.78, 0.0536
ATI2	1.13, 0.445	1.07, 0.581	1.94, 0.1
AlgoSum		1.54, 0.217	**5.94, 0.00638**
ATI>0	-	inf, 0.476	inf, 0.08
ATI>1	-	0.44, 0.246	2.25, 0.127
ATI>2	-	1.83, 0.4	4.13, 0.0625
AKIN>0	inf, 0.476	-	2, 0.108
AKIN>1	inf, 0.767	-	1.93, 0.243
AKIN>2	inf, 0.839	-	3.44, 0.184

(f) *All Slides (a)+(b)+(c)*

Fisher's z,p	ATI>0	AKIN>0	DGF
AKIN1	**3.69, 0.00643**	**3.48, 0.000835**	**3.01, 0.00286**
AKIN2	**5.91, 0.00793**	**2.83, 0.00723**	**3.1, 0.00459**
ATI1	**2.51, 0.0446**	**2.28, 0.00493**	**2.54, 0.0389**
ATI2	**2.42, 0.00933**	1.07, 0.502	**2.25, 0.0171**
AlgoSum	**4.4, 0.00395**	**2.01, 0.0148**	**4.35, 0.000345**
ATI>0	-	1.82, 0.127	**2.62, 0.0204**
ATI>1	-	**0.427, 0.0463**	**2.72, 0.00489**
ATI>2	-	1.41, 0.405	**5.04, 0.0124**
AKIN>0	1.82, 0.127	-	**2.55, 0.00131**
AKIN>1	1.31, 0.423	-	**3.51, 0.000203**
AKIN>2	1.3, 0.459	-	**4.18, 7.44e-05**

Algorithmic Histology Scores | Pathologist Score | Clinical Score

(a) QUOD Slides

MWW	ATI>0	AKIN>0	DGF
AKIN1	$\mathbf{2.78 \cdot 10^{-3}}$	$\mathbf{3.53 \cdot 10^{-5}}$	$\mathbf{3.09 \cdot 10^{-3}}$
AKIN2	$\mathbf{1.13 \cdot 10^{-3}}$	$\mathbf{2.63 \cdot 10^{-5}}$	$\mathbf{1.49 \cdot 10^{-2}}$
ATI1	$\mathbf{2.40 \cdot 10^{-3}}$	$\mathbf{9.55 \cdot 10^{-3}}$	$2.53 \cdot 10^{-1}$
ATI2	$\mathbf{5.22 \cdot 10^{-3}}$	$2.22 \cdot 10^{-1}$	$7.99 \cdot 10^{-2}$
AlgoSum	$\mathbf{8.89 \cdot 10^{-4}}$	$\mathbf{3.45 \cdot 10^{-4}}$	$\mathbf{1.07 \cdot 10^{-2}}$

(b) Test Slides

MWW	ATI>0	AKIN>0	DGF
AKIN1	$\mathbf{3.90 \cdot 10^{-3}}$	$\mathbf{7.96 \cdot 10^{-3}}$	$1.07 \cdot 10^{-1}$
AKIN2	$\mathbf{7.54 \cdot 10^{-3}}$	$\mathbf{3.89 \cdot 10^{-4}}$	$1.27 \cdot 10^{-1}$
ATI1	$\mathbf{2.27 \cdot 10^{-2}}$	$1.61 \cdot 10^{-1}$	$7.73 \cdot 10^{-2}$
ATI2	$1.60 \cdot 10^{-1}$	$\mathbf{1.68 \cdot 10^{-4}}$	$\mathbf{2.89 \cdot 10^{-2}}$
AlgoSum	$\mathbf{9.76 \cdot 10^{-3}}$	$3.80 \cdot 10^{-1}$	$\mathbf{4.35 \cdot 10^{-3}}$

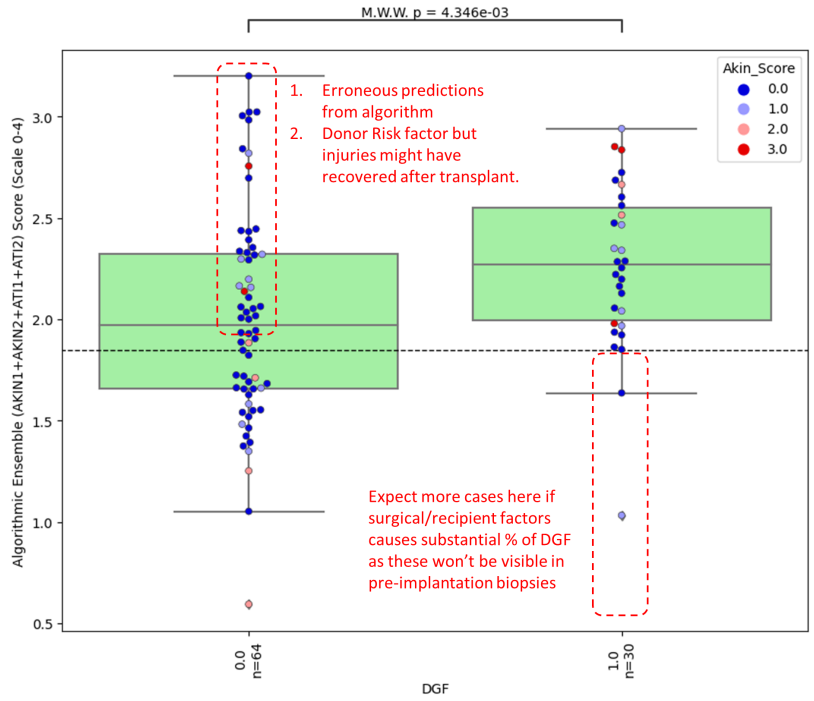
(c) All Slides

MWW	ATI>0	AKIN>0	DGF
AKIN1	$\mathbf{5.93 \cdot 10^{-6}}$	$\mathbf{5.85 \cdot 10^{-5}}$	$\mathbf{1.40 \cdot 10^{-3}}$
AKIN2	$\mathbf{5.76 \cdot 10^{-6}}$	$\mathbf{2.82 \cdot 10^{-6}}$	$\mathbf{1.61 \cdot 10^{-2}}$
ATI1	$\mathbf{3.31 \cdot 10^{-5}}$	$\mathbf{6.09 \cdot 10^{-3}}$	$1.523 \cdot 10^{-1}$
ATI2	$\mathbf{1.56 \cdot 10^{-7}}$	$\mathbf{4.89 \cdot 10^{-2}}$	$\mathbf{1.31 \cdot 10^{-3}}$
AlgoSum	$\mathbf{1.56 \cdot 10^{-7}}$	$\mathbf{5.14 \cdot 10^{-3}}$	$\mathbf{4.46 \cdot 10^{-4}}$

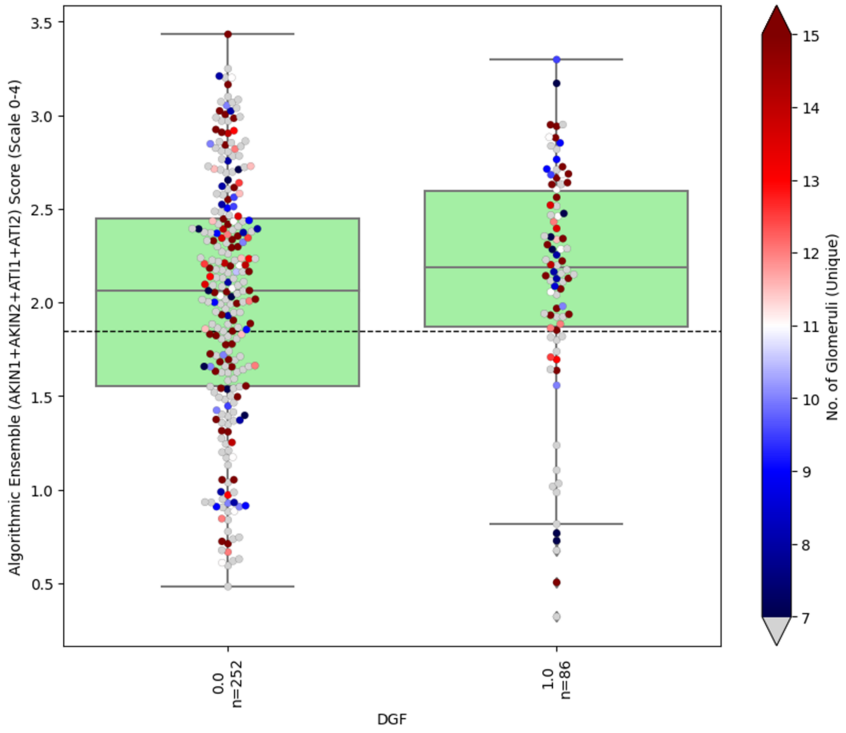
Table 7.6: Association between histological scores and ATI/AKIN/DGF. Here we compare numeric algorithmic scores and binarised manual/clinical scores hence the MWW test is used. Values in the tables show p-values. The proposed algorithmic score is significantly associated with DGF ($p = 4.35 \cdot 10^{-3}$) in the test slides (b).

7.3.3.3 Surgical/Recipient Related DGF

The ensemble score is plotted against DGF in Figure 7.5. The boxplot on the test slides (Figure 7.5a) shows that the algorithmic score has a negative predictive value (NPV) of 0.929 at the chosen threshold. DGF incidences caused by surgical or recipient factors would not be detectable in pre-implantation biopsies and will likely appear as false negatives. Hence the absence of false negatives suggests that most DGF incidence may be donor-related. In contrast, the false-positive slides may be either caused by i) erroneous predictions from the algorithm or ii) donor risk factors may be present, but the injuries might have recovered after transplantation.



(a) Test Slides



(b) All Slides

Figure 7.5: Algorithmic ensemble score vs DGF (a) Test slides, colour-coded by AKIN grades; (b) All slides, colour-coded by the number of glomeruli.

7.3.3.4 Biopsy Size

We are interested in investigating how excluding smaller biopsies affect predictions of DGF. Figure 7.5b shows the box plot of the ensemble score on all slides with DGF status available. Note that the QUOD slides are included in the swarm plot - these are, on average, much smaller than the AMC-PAS slides. Slides are colour-coded by the number of glomeruli. A slightly larger proportion of small slides are in the false-negative category. We get NPV=0.851 if all slides are included and NPV=0.896 if slides with <6 glomeruli are excluded. This shows that the effect of biopsy size is moderate.

Table 7.7 shows how the relative risk of DGF calculations changes as slides with fewer than a certain number of glomeruli is excluded. We can see that RR increases initially up to 6 glomeruli. However, as fewer slides remain available for analysis, the confidence interval widens. RR of pathologist ATI grade >0 against DGF is also shown in the table. Overall, we find there is little benefit in the analysis to excluding the smaller biopsies as a small number of proximal tubules is likely to be sufficient for assessing acute changes.

Table 7.7: Relative Risk of DGF excluding slides below a minimum number of glomeruli (All cohorts)

Minimum	AlgoSum		ATI	
Glomeruli	n	RR (95% CI)	n	RR (95% CI)
0	185	4.35 (1.95-9.73)	126	2.52 (1.46-4.36)
1	160	4.21 (1.75-10.1)	126	2.52 (1.46-4.36)
2	150	6.17 (2.00-19.0)	124	2.48 (1.44-4.28)
3	146	6.23 (2.02-19.2)	123	2.34 (1.29-4.24)
4	137	5.9 (1.92-18.2)	120	2.36 (1.3-4.29)
5	131	5.35 (1.74-16.4)	117	2.38 (1.3-4.34)
6	121	6.39 (1.61-25.4)	114	2.4 (1.31-4.4)
7	108	5.44 (1.37-21.5)	106	2.5 (1.35-4.65)
8	104	4.83 (1.22-19.1)	93	2.22 (1.11-4.44)
9	95	4.81 (1.21-19)	88	2 (0.908-4.41)
10	88	4.19 (1.06-16.7)	81	1.53 (0.755-3.11)

7.3.3.5 DGF Relative Risk vs Algorithmic Numeric Score

In Table 7.5 (e), thresholds between algorithmic scores and DGF were determined by maximising the lower bound 95% CI Relative Risk for DGF. At the optimal threshold, we find our algorithmic ensemble score associated with a relative risk of 5.94 (95% CI: 1.52-23.3) of developing DGF on the test data. A plot of how relative risk varies with threshold is shown in Figure 7.6. There is some bias in this reported value as the threshold is determined from the test data. If the threshold is taken from the training data and applied to the test data, we get RR=4.80 (95% CI 1.24-18.7). Overall, these results suggest that learning directly from the AKIN clinical labels and ATI labels may each provide complementary advantages in detecting slides that are likely to be free from DGF.

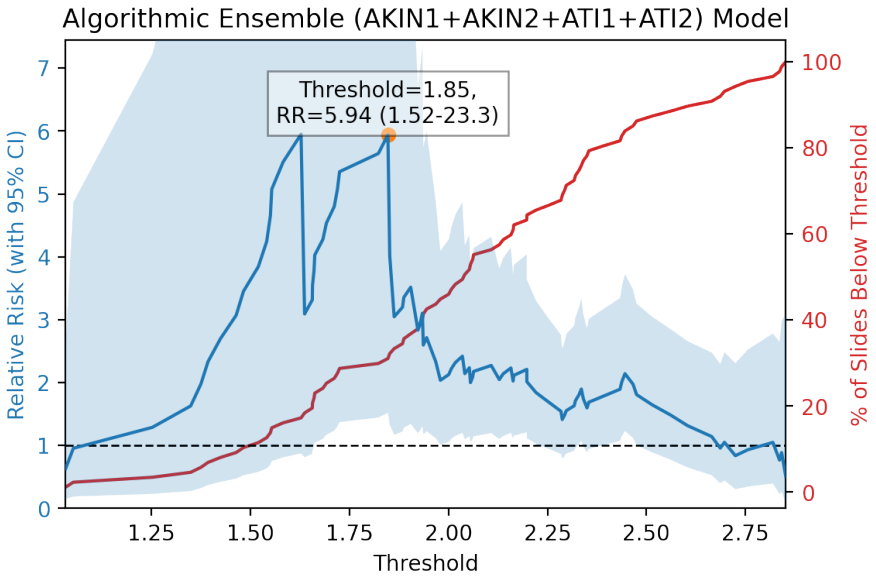


Figure 7.6: DGF RR vs numeric score threshold on the test slides

7.3.4 Utility Value of Algorithmic Histological Scores

Although clinical and histological grades are categorical, scores assigned by algorithms can be continuous and may help pathologists establish more precise definitions of thresholds linking visual changes to risks of post-transplant complications. If biopsies are readily available, as in the case of transplantation, histological

information can be incorporated without extra costs to determine the risk of DGF. While a clinically-scrutinised version of DGF risk score may require a larger independent cohort, we can perform a feasibility check by plotting and confirming the individual risk factors of DGF. Firstly, to create a combined risk score for DGF from different variables, it will be advantageous for these input variables to be not too strongly correlated. Otherwise, we will just be replicating information. Table 7.8 shows the Pearson correlation matrix of the algorithmic numeric scores, the constitutes of AKIN score, and Donor Age. We can see that AlgoSum only weakly correlates with creatinine fold change (r=0.0553) and creatinine increase (r=0.08), suggesting it may be a valuable independent addition for predicting DGF risk.

The thresholds of various risk factors can be determined using plots of the continuous values against relative risk (RR), as shown in Figures 7.7. These plots are based on the QUOD slides. The blue curve shows RR at varying thresholds with the 95% confidence interval shaded in each plot. We define the orange markers where the lower bound 95% CI RR is maximum. Figure 7.7a shows how well the algorithmic ensemble histology score scales with RR, (b-d) constitute AKIN clinical score, and (e) show how RR varies with donor age. From (a), it can be seen that at a threshold of 1.73, RR= 4.35 (95% CI 1.95-9.73), slightly better than the next best variable (c) Creatinine Increase with RR=2.77 (95% CI 1.88-4.08). At the proposed threshold, the algorithmic histology score can predict 39% of cases as low risk for DGF. This may be used complementary to (b) creatinine fold change (RR=2.18; 95% CI 1.39-3.42; 8.1% slides above threshold) and (c) creatinine increase (RR=2.77; 95% 1.88-4.08; 8.2% slides above threshold) which appears to be more optimal for detecting high-risk cases. The evidence above suggests that an algorithmic-based histology score may prove helpful as a potent and independent source of information as a predictor for DGF.

7.3.5 Visualisation of Relevant Tissue Compartments

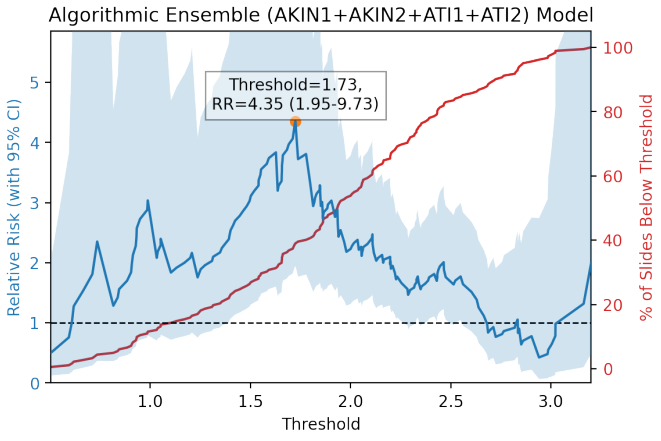
As described in Chapter 6, our workflow allows an intuitive way of highlighting tissue areas it considers important for predictions through a compartment-based attention

Table 7.8: Correlation between different DGF risk factors

Pearson r	Creatinine Increase	Urine Output	Donor Age	AKIN1	AKIN2	ATI1	ATI2	AlgoSum
Creatinine Fold Change	0.726	-0.171	-0.031	0.17	-0.007	0.049	-0.026	0.0553
Creatinine Increase		-0.158	-0.085	0.162	0.0781	0.154	0.0273	0.08
Urine Output			-0.075	0.0107	0.0235	-0.04	-0.01	-0.00352
Donor Age				-0.021	0.0573	0.0269	-0.025	0.0108
AKIN1					0.897	0.429	0.3	0.826
AKIN2						0.42	0.364	0.861
ATI1							0.21	0.698
ATI2								0.639

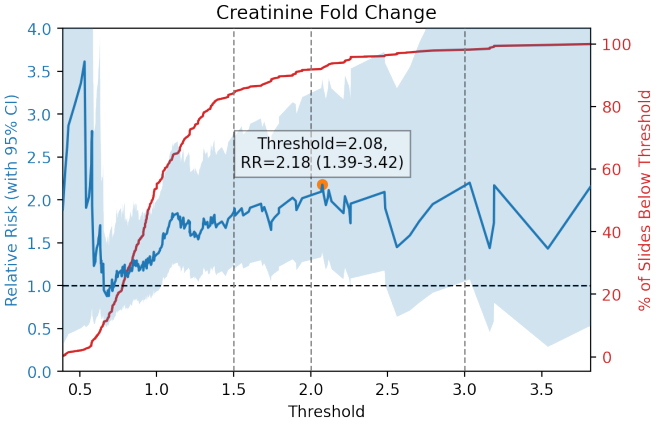
map. The proposed visualisation scheme can help shed light on the differences between what different models look at. In particular, we want to understand better why models trained on pathologists' scores correlate with DGF to a different extent than those trained on clinical states.

Logits of attention values for each tissue compartment are extracted from the models to produce a visualisation. In the test data, logits for each compartment are averaged across five models. Colour scales are normalised with the top and bottom 2% clipped. An example of where the ATI1 and AKIN2 models have learned to look is shown in Figure 7.8. Tissues the models regard as important

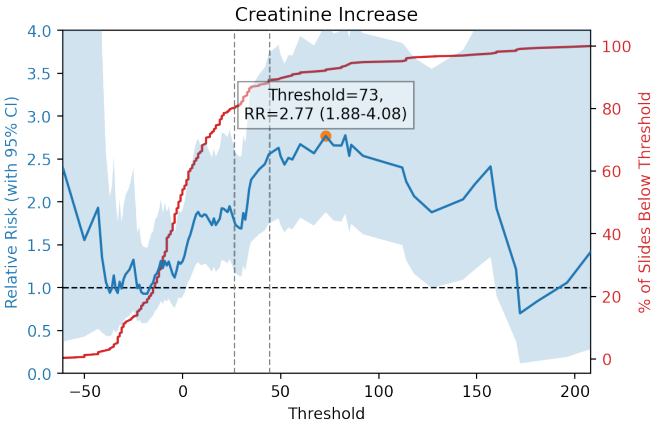


(a) DGF RR vs AlgoSum Threshold

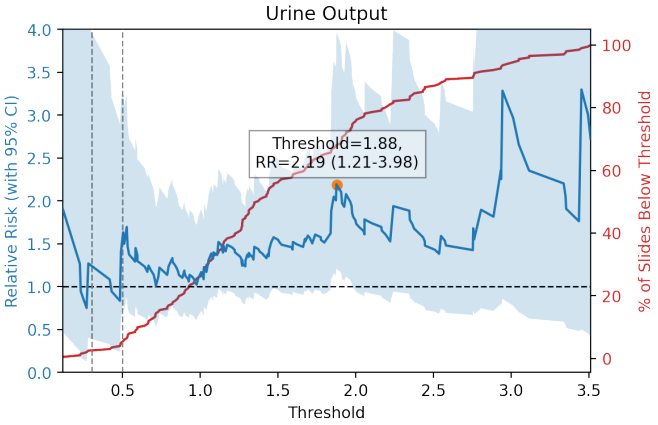
Figure 7.7: DGF RR vs risk factors. Plots are based on the QUOD slides.



(b) DGF RR vs Creatinine Fold Change Threshold

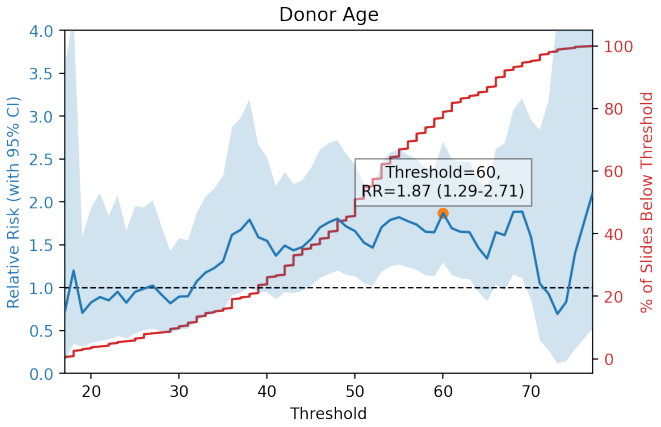


(c) DGF RR vs Creatinine Increase Threshold



(d) DGF RR vs Urine Output Threshold

Figure 7.7: DGF RR vs risk factors. Plots are based on the QUOD slides. (cont.)



(e) DGF RR vs Donor Age Threshold

Figure 7.7: DGF RR vs risk factors. Plots are based on the QUOD slides. (cont.)

for prediction are highlighted in red, whereas compartments not utilised by the models are shaded in blue. From the figure, it is apparent that AKIN1 (bottom) has a tendency to attend to the smaller tubules while ATI1 (top) attends to the larger tubules. A caveat of the visualisation is that the attention values assigned to individual compartments are noisy. Since multiple relevant tissue compartments may be present in each slide, models only need to be capable of detecting a small number of relevant instances. We plot histograms of attention values versus tubule compartment area on the AMC-PAS slides. Figure 7.9 (a) shows the AKIN1 models' tendency to attend to smaller tubules; Figure 7.9 (b) shows the ATI1 models' tendency to attend to larger tubules; Figure 7.9 (c) shows the histogram of the attention difference from individual tubules. All plots confirm the observation of the difference between the two models' behaviour.

Multiple possibilities may explain the difference between the two models. One feasible explanation is that ATI1, trained to mimic the pathologist's assessment, may have learned biases from human observers who are more likely to attend to the larger objects in a slide. A second feasible explanation is that acute tubular changes on different subsections (S1-S3) of proximal tubules may indicate varying progressive stages of injuries due to their different susceptibility to ischemic or hypoxic damage. The S3 proximal tubule segments are generally seen as more susceptible to ATI due to the section's higher oxygen demand, but the theory has been controversial

[24]. On PAS slides, no definitive features are known to distinguish the different proximal tubules (S1-S3), but these segments may have a different cross-section area distribution when sliced into thin sections due to their different degree of convolution and the insertion angle of the sampling needle. It is possible that with further work involving immunohistochemistry stain, this proposed workflow may be used to improve our understanding of the physiology underpinning AKI.

7.3.6 Limitation of ATI Grades

Section 7.1 described seven constraints of the pathologist-assessed ATI scale that make it less predictive of DGF. Preliminary evidence in this study is used to discuss the relative importance of these factors. Overall, models trained on pathologists' ATI grades may inherit some of the assessment biases. Our proposed ensemble includes models that also learn from clinical AKIN grades and can partially circumvent some human biases and discontinuity in grading scales.

1. **Visual grading scales may not scale linearly with clinical changes**. We believe the mismatch between visual histological grading scales and clinical changes to be the major limitation of ATI assessment. The first evidence of this can be seen in Figure 7.3. Here, each case is colour-coded according to the ATI grade assigned by pathologists. Cases with AKIN grade 0 contain a mix of ATI grades, with those of higher grades scored more highly by our algorithm (trained on AKIN labels). The second evidence supporting this can be seen in Table 7.5 (f). we find that ATI1 correlate with AKIN labels (p=0.00493) more strongly than ATI vs AKIN (p=0.0463). ATI1 would have learned some biases from pathologists, but its ability to output a numeric score is advantageous.

2. **Possibility of other features not known to pathologists**: From Table 7.5 (f), the correlation between AKIN1 vs ATI is (r=3.69, p=0.00643) and AKIN2 vs ATI is (r=5.91, p=0.00793). The strong association between algorithms trained on clinical labels and ATI grades suggests that pathologists

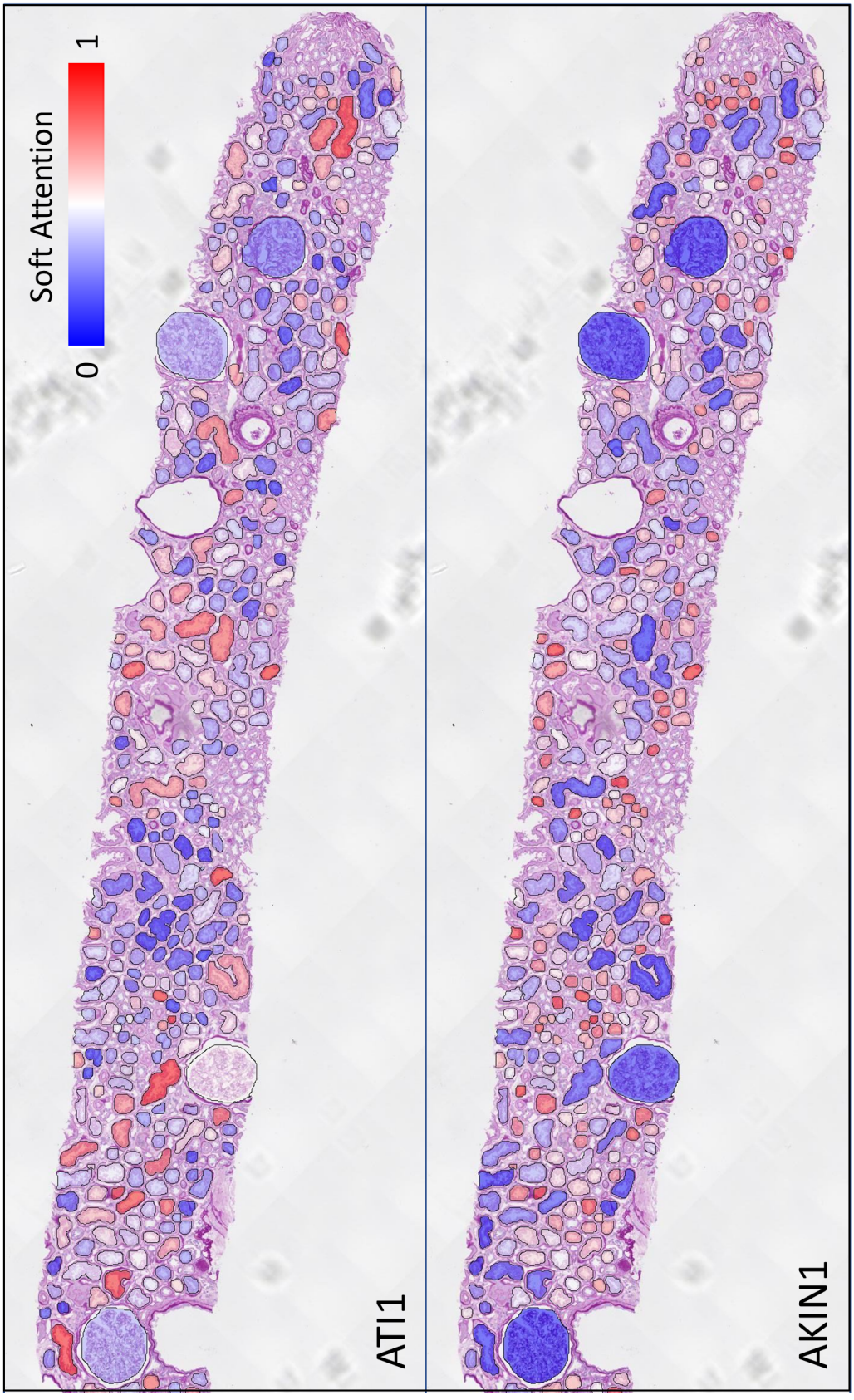


Figure 7.8: Visualisation of tissue compartment attention map (Top) ATI1; (Bottom) AKIN1

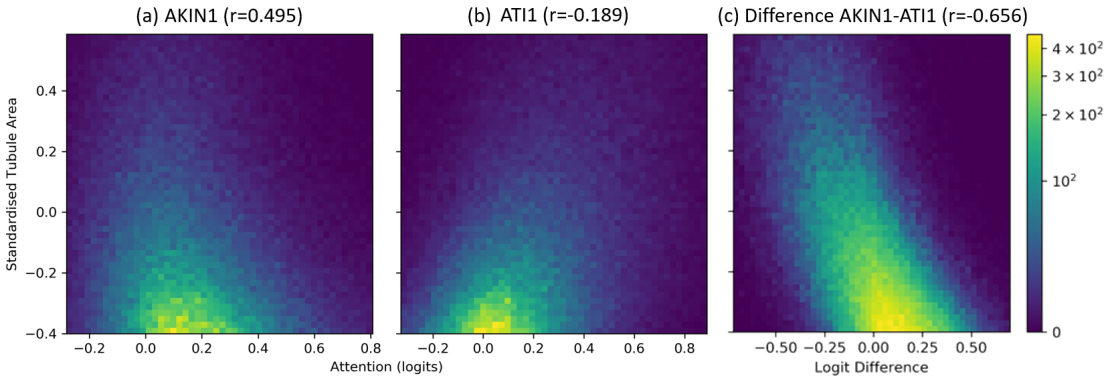


Figure 7.9: Histogram of attention values vs tissue area (a) The model AKIN1 appears to attend to smaller tubules; (b) The model ATI1 appears to attend to larger tubules; (c) Histogram of the difference in model attention values vs compartment area.

mostly look at the correct visual features. There is no evidence of any features unknown to pathologists.

3. **Lack of coherent guidelines on representing localised changes as a single score**: Early evidence from Section 7.3.5 suggests that models trained on clinical labels attend to different tissue compartments than models trained to mimic pathologists.

4. **Subjectivity and lack of reproducibility from human pathologists**: Regrettably, we do not have slides scored by multiple pathologists. However, when we trained the ATI1 and ATI2 models and tested them on slides assessed by different pathologists, PR-AUCs were higher on the test sets (Table 7.3). This indirectly suggests that inter-observer variability is not a major issue. This seems to contrast with existing literature [39–41] but may also be explained by both of our pathologists being renal specialists.

5. **Biopsy area may be too small for accurate assessment**: This was discussed in Section 7.3.3. We find only a moderate increase in effect size for predicting DGF when slides with <6 glomeruli are excluded. As proximal tubules are usually abundant in the cortex, there is little benefit in having slides with more than a handful of glomeruli.

6. **Non-biopsied AKI causes**: these cases would contribute to poorer apparent performance in the AKIN1 and AKIN2 models. As discussed in Section 7.3.1 Figure 7.4, most cases with AKIN grade 3 are detectable from biopsies. We estimate non-biopsied AKI causes to contribute to a small fraction of cases.

7. **Recipient/Surgical DGF Risk Factors**: From Figure 7.5a (discussed in Section 7.3.3), we find our algorithm has a Negative Predictive Value of 0.929 - cases predicted a low score rarely suffers from DGF. If surgical/recipient causes are a major source of DGF, they would not be visible in biopsies; hence we would expect more false-negative cases. Hence we expect the majority of DGF cases to be donor-related. This finding is coherent with earlier studies involving deceased-donor pairs [44].

7.4 Limitations and Further Work

The main limitation of this study is the lack of a large independent test cohort. Although our models' performance was validated on independent slides, some interpretation was performed on the original QUOD-PAS slides as these have serum creatinine and urine data available. We also could not fully account for recipient/surgical related DGF causes. These could be more easily identified if the test cohort contains paired recipients from the same donor. For clinical implementation, the development of the workflow may require frozen rather than paraffin-fixed biopsies. There are also gaps in our data. As shown in the Venn diagram in Figure 7.2, only a small number of slides have all DGF, ATI, and AKIN labels available hence some comparisons have been performed across slightly different subsets of data and may not have been fair.

There are also restrictions on the biopsies used for training. Although we showed biopsy size to have only a limited impact on evaluation, the same may not apply to the training stage. Currently, slides with few glomeruli have not been excluded during training as this may otherwise shrink the number of training data. Slides containing very little cortical tissue might have misguided our models during

training. Furthermore, there are other biopsy-visible changes pertaining to AKI (such as glomeruli and vascular diseases). Our models are unlikely to have learned these changes due to the small number of glomeruli and vessels in the QUOD slides.

We find using an ensemble of histology-based algorithms trained on multiple labels (ATI, AKIN) produces a more robust association with DGF. We may try a larger ensemble trained on more labels for future work. For example, additional models can be trained using different AKI definitions (KIDGO, RIFLE), ATI grades given by multiple pathologists, or ATI defined under different semi-quantitative definitions [39]. In fact, we experimented with including an additional model trained directly on DGF labels (we call this DGF1 in Table 7.3) to the existing ensemble 'AlgoSum'. This resulted in stronger significance (Fisher's Exact Test: p=0.00402) and higher lower-bound Relative Risk (RR=3.83 95% CI 1.73-8.51) on the test data. This is part of ongoing research and has not yet been described in the main text. Additional results by combining algorithmic scores and the original AKIN/ATI scores are available in the supplementary material Section A.4, but this will need some interpretation.

There is also room to explore different ways of summing the scores of the ensemble models instead of using just a linear sum. For example, models can be weighted according to their performance on the training data. Other future work includes more robust validation of tissue attention visualisation involving more pathologists or alternative tissue staining assays. If a larger cohort is available, we can also test the construction of a DGF risk score by combining both histological and clinical information.

7.5 Conclusion

Acute injury in donor kidneys may increase DGF risks and morbidity in transplant recipients. Renal biopsy can distinguish the cause of AKI more specifically and assist recipient management. However, semi-quantitative histology scores appear to have a limited association with both donor and recipient clinical states. A quantitative workflow may eliminate several constraints of histological assessment

and has the potential of being more predictive of DGF risks. It can also help us answer questions regarding the limits of histology.

An ensemble of deep neural network models was developed to detect histological changes pertaining to AKI. This ensemble consisted of models trained to reproduce either ATI grades given by pathologists or donor AKIN clinical scores. The former models trained on ATI grades have the possibility of inheriting the subjectivity of humans, while the later models trained on AKIN may be more capable of discovering features independent of current histological assessment guidelines but could be misguided by AKI cases with non-biopsied causes.

We found the ensemble score predictive of DGF in the test cohort (n=94, RR=5.94, 95% CI: 1.52-23.3; Fisher p=0.00638). Neither AKIN nor ATI scores are predictive of DGF in these cases due to the small sample size. We attribute the main success of the algorithm to its ability to identify low-risk kidneys (NPV=0.929). Furthermore, the numeric scores from the algorithm only weakly correlate with serum creatinine and urine output - components of AKIN scores, indicating histology can add value to donor clinical variables. Finally, the visualisation of tissue compartment attention maps produced by different models may also help to shed light on the differences between where humans and machines look. This may help to refine our understanding of histological changes and limitations of current practice for scoring acute lesions.

8 Conclusion

Contents

8.1 Summary of Thesis 210
8.2 Further Work 211

8.1 Summary of Thesis

In this thesis, we have developed the methodology and workflow for the automated extraction of visual features from various tissue compartments. We examined the use of both handcrafted histological measurements and deep features and proposed a way to improve these measurements on biopsies which are small in area. Finally, we show that the algorithm can be used to detect acute injuries from pre-implantation biopsies and has the potential to be combined with other clinical variables to predict the onset of post-transplant complications.

Chapter 3 describes the primary datasets used throughout the thesis, an annotation platform to assess individual tissue compartments, and some pre-processing work.

Chapter 4 details the main quantitative analysis workflow that serves as the backbone of the upcoming research chapters. The workflow is built on convolutional

neural networks (CNN) and automatically segments different tissue compartments in renal biopsies: tubules, glomeruli, vessels, and nuclei. We design handcrafted features and qualitatively show that some of these features can capture pathological changes related to chronic and acute injuries. We also explore some of the immediate extensions of this workflow: (i) Detection of anomalous tissue compartments; (ii) Computational quantification of Sirius Red slides; (iii) Correlation of visual features with tissue proteomic signatures.

Chapter 5 describes an approach to overcome a challenge in processing small biopsies with artefacts. We discuss Bayesian Neural Networks (BNNs) to quantify uncertainties in image segmentation maps and propose a way to improve histological estimates using confidence measures derived from these BNNs.

Chapter 6 describes tissue compartment-based visual features to improve classification performance and clinical interpretability in a multi-instance learning setting. In addition to learned attention, tissue compartments are weighted by segmentation quality to reduce the impact of instances that bear little resemblance to the training data.

Chapter 7 combines different techniques developed in this project to detect acute kidney injury (AKI) - a known cause of Delayed Graft Function (DGF) - in pre-implantation biopsies. We propose an algorithmic ensemble score consisting of models trained on pathologists' assessments and donors' clinical states. The algorithm is tested on two cohorts newly introduced towards the end of the project and is found to be significantly associated with DGF. The results show that our workflow has the potential to be combined with clinical variables for a more accurate prediction of DGF risk and better post-transplant recipient management.

8.2 Further Work

The work developed so far can obtain some histological features related to chronic and acute injuries in pre-implantation biopsies. Many of these morphometrics measurements are hard to assess by eye. This work can help advance the translational aspect of renal histopathology if it is deployed to other projects using the QUOD

biobank data. In Chapter 7, we find this work particularly useful for detecting complications caused by acute injuries. Further validation and methodological development will be needed to use this work in clinical settings. We believe significant advances will come from the lateral exploration involving additions to the analysis established in this thesis.

Linking Histology Features to Other Assays/Biomarkers The QUOD biobank has expanded significantly since the start of this DPhil project. As of January 2022, 31,807 biobank items had been issued to 71 research project applications. Approximately half of these projects involve the kidney. Research efforts based on molecular signatures and other assays will become increasingly more assessable. Our workflow may be able to discover histological parameters associated with these assays.

We can train weakly supervised models using these assays as labels to guide the discovery of histological features in place of, for example, the AKIN and ATI grades in Chapter 7. Models that manage to converge and learn from the assay may indicate a strong linkage between histology and the assay. Discoveries of association between histology and other biomarkers may help justify or enhance our understanding of the pathways of injury and repair in the kidney transplant process. Since our workflow has shown some promise in detecting acute injuries, it may be deployed in large scale studies of organ preservation to reduce incidences of ischemic injuries.

To detect chronic lesions, we may train models to detect fibrotic tissue from PAS slides based on pseudo predictions generated from adjacent sections stained in Sirius Red (see Chapter 4, Section 4.4.2). This will allow us to bypass limitations such as the availability of pathologists and the rounding errors caused by semi-quantitative assessment.

Exploring Alternative Workflow. In the main flow of this thesis, we focused on a single workflow based on segmentation and feature extraction from tissue compartments. Alternatives or additions to this workflow have also been explored. Most importantly, the workflow would benefit from a working classifier separating

tissues of the cortex from the medulla. This had been initially implemented in Chapter 4 Section 4.4.1 (Figure 4.14), where we showed that healthy and diseased tubules form distinct clusters in a TSNE visualisation once tissues from the medulla were excluded. However, while this classifier only worked on the NMP dataset, it has not yet been able to generalise to the QUOD-PAS slides as a pathologist had labelled only 104 tiles. Since the predictions made by the multi-instance soft attention model in Chapter 6 are noisy, a cortex classifier can pre-filter relevant instances and improve slide-level predictions.

Another direction is to develop a workflow optimised at combining clinical data with histology. While we had attempted to include donor clinical variables in our multi-instance learning model in Chapter 6 Section 6.3.2, fully connected neural networks are typically not a favourable approach for handling donor clinical variables. Since categorical variables are involved, decision tree-based methods may be more appropriate. Individual histological features and clinical variables can be used as decision nodes to predict kidney function. For the prediction of DGF, we have shown that the optimal decision thresholds for histology differ from clinical variables (serum creatinine and urine output) as the former has a high negative predictive value (Chapter 7, Section 7.3.3.3). Hence it may be possible to produce better predictions using both histology and numeric donor variables using decision trees with boosting.

One more challenge that requires further work revolves around how to derive histological measurements from tissue compartments at the edge of the biopsy. For small needle biopsies, a significant proportion of the larger compartments such as arteries and glomeruli have been partially cut off. Experienced pathologists can often assess these compartments, but these can often be troublesome for basic feature extraction algorithms. A possible workaround is to explore the use of impainting techniques to “complete” these compartments.

Appendices

A
Appendices

Contents

A.1 Impact of Covid19 215
A.2 Donor and Recipient Metadata 216
A.3 Handcrafted Features 218
A.4 DGF Risk Predicted by Different Algorithmic Scores 222
A.5 TCGA Slides 224

A.1 Impact of Covid19

Before the pandemic, there were only 144 slides available in the QUOD-PAS dataset. However, due to these biopsies' small size, only 42 could be fully assessed. Since most slides received the same score, it was challenging to produce publishable results. Additional slides were initially due to be stained in April 2020. However, the closure of labs meant that I could only receive these slides in late August. Thus, from March to August, I worked on the quantification of uncertainty, which resulted in the publication of a MICCAI paper (Chapter 5). As I have not been able to access the batch scanner in the Biomedical Engineering department, the new QUOD slides (268 PAS + 175 SR) and other slides received after the pandemic had to be digitised using a single-slide scanner. This took an extra three weeks of working

time compared to what it would have taken on the batch scanner. It took additional time for the pathologist to annotate/assess these slides. Remuzzi scores and ATI grades of these QUOD-PAS slides were finally received on 1 Feb 2021.

The pandemic also meant workload changes for pathologists. I had not been able to receive any advice from the pathologist for an extended period between Feb – Nov 2021. This has led to some loose ends in the project, as pathologists had only been able to review a small portion of tissue compartments with chronic lesions. Given the situation, I decided to focus on the assessment of ATI/AKI instead. Thankfully, pathologist Dr Jesper Kers had agreed to collaborate with this project in early 2021. However, due to the administrative paperwork involved, file exchange only started in Nov 2021. This meant some key results in Chapter 7 were obtained last minute in April 2022.

Other tolls caused by the pandemic include: i) not being able to cross-pollinate ideas between colleagues in my research group on a casual basis; ii) being stuck overseas for an extra month while visiting family due to flight bans after the unexpected emergence of the covid alpha variant. During which I could not establish a smooth connection with the computing cluster back in Oxford.

A.2 Donor and Recipient Metadata

A number of metadata are available in the QUOD dataset. Based on metadata alone, the strongest predictors for the onset of DGF are donor and recipient age. These two metadata are correlated as kidneys from older donors are generally matched to older recipients due to ethical reasons. Predicting the presence of DGF based on these variables alone would give us ROC-AUC of 0.584 and 0.529, respectively. Other metadata used in the main paper are summarised in Table A.1. Note that some metadata are categorical. To utilise these in neural networks, we converted them into one-hot representations. Metadata is concatenated to each tissue's feature vector, resulting in 299 extra features. Missing values and normalisation are processed in the same way as other features.

Table A.1: Description of metadata available for transplantation.

#	Description
1	Calculated Reaction Frequency at Transplant
2	Dialysis Status at Transplant
3	Donor (History of) Hypertension
4	Donor (History of) Hypotension
5	Donor Age
6	Donor Blood Group
7	Donor Blood Rhesus
8	Donor Body Mass Index
9	Donor Cause of Death
10	Donor Cytomegalovirus Test Results
11	Donor Diabetes
12	Donor Ethnicity
13	Donor Family History of Cardiovascular Disease
14	Donor Family History of Diabetes
15	Donor Gender
16	Donor Gender
17	Donor Height
18	Donor Hepatitis C Virus Results
19	Donor History of Cardiovascular Disease
20	Donor History of Liver Disease
21	Donor Homozygous/Heterozygous at A Locus
22	Donor Homozygous/Heterozygous at B Locus
23	Donor Homozygous/Heterozygous at DR Locus
24	Donor Kidney Estimated Glomerulus Filtration Rate (eGFR)
25	Donor Number of Occasions with Hypotension
26	Donor Number of Occasions with Hypertension
27	Donor Type (DBD/DCD)
28	Donor Weight
29	HLA Mismatch Groups
30	Kidney Cold Ischemic Time
31	Machine Reperfusion (None/Normothermic/Hypothermic)
32	Matchability
33	Match Grade
34	Perfusate Used
35	Perfusion Quality
36	Points Score Based on Current Matchability Points Band
37	Primary Renal Disease (Categorical)
38	Recipient Age
39	Recipient Body Mass Index
40	Recipient Cytomegalovirus Test Results
41	Recipient Ethnicity

Continued on next page

Table A.1 – continued from previous page

#	Description
42	Recipient Gender
43	Recipient Height
44	Recipient Height
45	Recipient Hepatitis C Virus Results
46	Recipient Homozygous/Heterozygous at A Locus
47	Recipient Homozygous/Heterozygous at B Locus
48	Recipient Homozygous/Heterozygous at DR Locus
49	Recipient Waiting Time
50	Status of Dialysis Prior to Transplant
51	Time between admission and aorta flushing
52	Time between admission and circulatory arrest
53	Time between admission and in-situ cold perfusion
54	Time between admission and ventilation ceased
55	Time between admission and support withdrawal
56	Time between aortic perfusion and circulatory arrest
57	Time between aortic perfusion and systolic BP below 50mmhg
58	Time between circulatory arrest and in-situ cold perfusion
59	Time between circulatory arrest to retrieval (DCD)
60	Time between 2nd brain stem death to retrieval (DBD)
61	Total Warm Ischaemic Time
62	Whether Recipient was Highly Sensitised

A.3 Handcrafted Features

A list of handcrafted features is shown in Table A.2.

Table A.2: List of handcrafted features extracted from tissues.

#	Name : Description	n
1	area: *Area of the segmented tissue*	1
2	n_glom: *# of Glomeruli*	1
3	n_tub: *# of Tubules*	1
4	n_ves: *# of Vessels*	1
5	slide_area: *Area of slide in mmˆ2*	1
6	biopsy_area: *Total area of biopsy tissues in mmˆ2*	1

Continued on next page

Table A.2 – continued from previous page

#	Name : Description	n
7	max_dist: *Maximum value of distance transform of the tissue — (max(D))*	1
8	nuclei_density: *Nuclei density; # of nuclei / area of tissue*	1
9	nuclei_moments_centre_max_dist: *Maximum distance of nuclei measured from centre of tissue*	1
10	nuclei_moments_centre_mean_dist: *Mean distance of nuclei measured from centre of tissue*	1
11	nuclei_moments_centre_min_dist: *Minimum distance of nuclei measured from centre of tissue*	1
12	nuclei_moments_centre_norm_max_dist: *Maximum distance of nuclei measured from centre of tissue, normalised by max(D) for each tissue*	1
13	nuclei_moments_centre_norm_mean_dist: *Mean distance of nuclei measured from centre of tissue, normalised by max(D) for each tissue*	1
14	nuclei_moments_centre_norm_min_dist: *Minimum distance of nuclei measured from centre of tissue, normalised by max(D) for each tissue*	1
15	nuclei_moments_kurtosis: *Kurtosis of nuclei distribution from edge*	1
16	nuclei_moments_max_dist: *Maximum distance of nuclei measured from edge of tissue*	1
17	nuclei_moments_mean_dist: *Mean distance of nuclei measured from edge of tissue*	1
18	nuclei_moments_min_dist: *Minimum distance of nuclei measured from edge of tissue*	1
19	nuclei_moments_norm_max_dist: *Maximum distance of nuclei measured from edge of tissue, normalised by max(D) for each tissue*	1
20	nuclei_moments_norm_mean_dist: *Mean distance of nuclei measured from edge of tissue, normalised by max(D) for each tissue*	1
21	nuclei_moments_norm_min_dist: *Minimum distance of nuclei measured from edge of tissue, normalised by max(D) for each tissue*	1
22	nuclei_moments_norm_variance: *Variance of nuclei distance from edge, normalised by max(D) for each tissue*	1
23	nuclei_moments_skewness: *Skewness of nuclei distance from edge*	1

Continued on next page

Table A.2 – continued from previous page

#	Name : Description	n
24	nuclei_moments_variance: *Variance of nuclei distance from edge*	1
25	nuclei_nnuclei: *Number of nuclei per tissue*	1
26	nuclei_nuclei_area_050percentile: *Area of nuclei*	10
27	nuclei_nuclei_col_b_050percentile: *Nuclei colour pixel values, blue channel*	10
28	nuclei_nuclei_col_g_050percentile: *Nuclei colour pixel values, green channel*	10
29	nuclei_nuclei_col_r_050percentile: *Nuclei colour pixel values, red channel*	10
30	shape_Ixx: *sum(M_x * M_x)*	1
31	shape_Ixx_norm: *sum(M_x * M_x) / count(M)*	1
32	shape_Iyy: *sum(M_y * M_y)*	1
33	shape_Iyy_norm: *sum(M_y * M_y) / count(M)*	1
34	shape_Izz: *Moment of inertia of tissue — sum(M_x * M_x + M_y * M_y)*	1
35	shape_Izz_norm: *Moment of inertia of tissue, normalised by max(D). Larger value = more elongated — sum(M_x * M_x + M_y * M_y) / count(M)*	1
36	shape_aspect: *Minor / Major Axis ratio*	1
37	shape_ax1: *Major axis of tissue*	1
38	shape_ax1_norm: *Major axis of tissue, normalised by max(D)*	1
39	shape_ax2: *Minor axis of tissue*	1
40	shape_ax2_norm: *Minor axis of tissue, normalised by max(D)*	1
41	shape_convex: *Ratio of the tubule's mask over the convex hull of the mask*	1
42	shape_moment_mask: *np.sum(r1 * dist * mask) / np.sum(mask); r1 is radial distance from Centre of Mass of M*	1
43	shape_moment_mask_dist: *np.sum(r2 * dist * mask) / np.sum(mask); r2 is radial distance from D*M Centre of Mass of M*	1

Continued on next page

Table A.2 – continued from previous page

#	Name : Description	n
44	shape_moment_mask_dist_norm: *np.sum(r2 * dist * mask) / np.sum(mask) / max(D); r2 is radial distance from D*M Centre of Mass of M*	1
45	tissue_bbar: *Blue-channel mean value in tissue's cytoplasm*	1
46	tissue_bstd (b^{σ}): *Blue-channel std value in tissue's cytoplasm*	1
47	tissue_gbar: *Green-channel mean value in tissue's cytoplasm*	1
48	tissue_gstd: *Green-channel std value in tissue's cytoplasm*	1
49	tissue_rbar: *Red-channel mean value in tissue's cytoplasm*	1
50	tissue_rstd: *Red-channel std value in tissue's cytoplasm*	1
51	tissue_moment_mean_dist (y^{γ}): *Mean distance of cytoplasm pixel values, as measured from edge of tissue — (mean((255 - img_1d[:, 0]) * distance_1d))*	1
52	tissue_moment_norm_mean_dist: *Mean distance of cytoplasm pixel values, as measured from edge of tissue, normalised by size of tissue — (mean((255 - img_1d[:, 0]) * distance_1d)/ max(D))*	1
53	tissue_moment_kurtosis: *4th order statistics of cytoplasm pixel values, as measured from edge of tissue*	1
54	tissue_moment_norm_variance: *Spatially-weighted (D / max(D)) variance of cytoplasm pixel values, as measured from edge of tissue, normalised by size of tissue*	1
55	tissue_moment_skewness: *Spatially-weighted (D) skewness of cytoplasm pixel values, as measured from edge of tissue, normalised by size of tissue*	1
56	tissue_moment_variance (y^{σ^2}): *Spatially-weighted (D) variance of cytoplasm pixel values, as measured from edge of tissue, normalised by size of tissue*	1
57	tissue_moments_centre_mean_dist ($(y^{-\gamma})$): *Mean distance of cytoplasm pixel values, as measured from centre of tissue — (255 - img_1d[:, 0]) * (max(D) - D)*	1
58	tissue_moments_centre_norm_mean_dist: *Mean distance of cytoplasm pixel values, as measured from centre of tissue, normalised — (255 - img_1d[:, 0]) * (max(D) - D) / (max(D)*	1
59	glom_bm_capsule_area: *Area of urinary space in glomerulus*	1

Continued on next page

Table A.2 – continued from previous page

#	Name : Description	n
60	glom_bm_capsule_area_ratio: *(Area of Urinary Space) / (Area of Glomeruli)*	1
61	ves_lumen_area: *Lumen area in vessels*	1
62	ves_lumen_ratio: *Ratio of lumen to total area in vessels*	1
	Total number of features	98

A.4 DGF Risk Predicted by Different Algorithmic Scores

Table A.3 shows DGF risk as predicted by different algorithmic scores based on the models described in Chapter 7. (a-c) shows the original AKIN and ATI labels' AUC and RR predicting DGF where the ordinal grades are treated as if they are numeric. (d-j) shows how different combinations of algorithmic scores predict DGF. (k-n) shows AUCs and RR after combining algorithmic scores with AKIN and ATI through direct summation. Higher AUC can be seen in (n), which uses all algorithmic models plus ATI and AKIN.

Table A.3: RR of DGF based on different scores

	Score	All Slides				Test Slides			
		ROC-AUC	PR-AUC	RR	MWW p	ROC-AUC	PR-AUC	RR	MWW p
a	ATI	0.685	0.511	2 (1.17-3.41)	0.000334	0.66	0.602	1.67 (0.799-3.48)	0.0182
b	AKIN	0.626	0.483	2.51 (1.69-3.72)	0.000119	0.579	0.442	1.98 (0.905-4.32)	0.0634
c	ATI+AKIN	0.733	0.577	3.48 (2.26-5.37)	1.31E-05	0.713	0.673	2.25 (1.24-4.08)	0.00261
d	ATI1	0.56	0.279	2.29 (1.11-4.73)	0.0747	0.592	0.356	3.78 (0.984-14.5)	0.0773
e	ATI2	0.624	0.383	2.09 (1.32-3.33)	0.00158	0.622	0.457	1.94 (1.11-3.42)	0.0286
f	AKIN1	0.652	0.382	2.83 (1.63-4.92)	0.000235	0.58	0.372	2.46 (1.04-5.81)	0.107
g	AKIN2	0.623	0.305	3.54 (1.76-7.12)	0.00366	0.573	0.388	3.09 (1.03-9.26)	0.127
h	DGF1	0.618	0.329	2.23 (1.35-3.68)	0.00558	0.64	0.39	2.71 (1.23-6.01)	0.0147
i	AlgoSum (d+e+f+g)	0.663	0.334	4.35 (1.95-9.73)	0.000446	0.669	0.403	5.94 (1.52-23.3)	0.00435
j	AlgoSum5 (d+e+f+g+h)	0.678	0.355	3.94 (1.75-8.84)	0.00037	0.713	0.483	3.83 (1.73-8.51)	0.000461
k	AlgoSum+AKIN	0.685	0.41	2.86 (1.82-4.49)	7.55E-05	0.65	0.487	2.45 (1.43-4.19)	0.00985
l	AlgoSum5+AKIN	0.694	0.4	4.67 (1.75-12.4)	0.000122	0.668	0.498	2.14 (1.23-3.73)	0.00456
m	AlgoSum+AKIN+ATI	0.74	0.575	3 (1.74-5.17)	0.000158	0.711	0.642	2.46 (1.37-4.41)	0.00456
n	AlgoSum5+AKIN+ATI	0.752	0.574	3.11 (1.72-5.62)	0.000221	0.73	0.661	2.66 (1.45-4.87)	0.00231

A.5 TCGA Slides

Table A.4: Eleven slides were chosen from the TCGA [80] repository to serve as training examples in our segmentation workflow. These slides are publicly available at https://portal.gdc.cancer.gov/ and can be searched by entering the prefix of the following filenames.

Case Filename
1. TCGA-CJ-4891-01Z-00-DX1.5E9F7E8F-110A-4A9F-BCD0-9CFBA14E1419
2. TCGA-CJ-4923-01Z-00-DX1.ADD3D7FE-46D1-49AF-B6D4-1FD07B761EF1
3. TCGA-CJ-6027-01Z-00-DX1.ADA07C08-C97A-4F41-9467-8F4F49F96EDA
4. TCGA-G6-A8L8-01Z-00-DX1.32E022BA-9959-4E2F-9D9D-AEEB5B7E7E9F
5. TCGA-MM-A564-01Z-00-DX1.F8F976F5-500F-4B3F-BF0E-EC3D126AF7D3
6. TCGA-MM-A564-01Z-00-DX2.C54B9A26-7221-4B29-8331-6D8F8F52B058
7. TCGA-BP-4994-01Z-00-DX1.87afea97-277f-4eac-a82c-64570d27cea7
8. TCGA-CJ-4643-01Z-00-DX1.50F38125-1825-4B66-A171-4C92279E306D
9. TCGA-CJ-4870-01Z-00-DX1.38991588-D25A-48F6-80B5-51ED34AA8D35
10. TCGA-CJ-4874-01Z-00-DX1.614D8D7D-E9FA-464D-A90F-C5C789664287
11. TCGA-CJ-4882-01Z-00-DX1.C39D94E4-192C-4B0F-93E6-6A18AA9F11A8

References

[1] C Rudge et al. “International practices of organ donation”. In: *British journal of anaesthesia* 108.suppl_1 (2012), pp. i48–i55.

[2] NHS Blood et al. *Annual Report on Kidney Transplantation (2017/18)*. Sept. 2018. URL: https://nhsbtdbe.blob.core.windows.net/umbraco-assets-corp/12256/nhsbt-kidney-transplantation-annual-report-2017-2018.pdf.

[3] H Liapis et al. “Banff histopathological consensus criteria for preimplantation kidney biopsies”. In: *American Journal of Transplantation* 17.1 (2017), pp. 140–150.

[4] Giuseppe Remuzzi et al. “Long-term outcome of renal transplantation from older donors”. In: *New England Journal of Medicine* 354.4 (2006), pp. 343–352. URL: https://doi.org/10.1056/nejmoa052891.

[5] Stanley J Robboy et al. “Pathologist workforce in the United States: I. Development of a predictive model to examine factors influencing supply”. In: *Archives of Pathology and Laboratory Medicine* 137.12 (2013), pp. 1723–1732. URL: https://doi.org/10.5858/arpa.2013-0200-OA.

[6] CJ Wang et al. “The donor kidney biopsy and its implications in predicting graft outcomes: a systematic review”. In: *American Journal of Transplantation* 15.7 (2015), pp. 1903–1914.

[7] Catherine Boffa et al. “Transplantation of kidneys from donors with acute kidney injury: friend or foe?” In: *American Journal of Transplantation* 17.2 (2017), pp. 411–419.

[8] Ka Ho Tam et al. “Improving Pathological Distribution Measurements with Bayesian Uncertainty”. In: *Uncertainty for Safe Utilization of Machine Learning in Medical Imaging, and Graphs in Biomedical Image Analysis*. Springer, 2020, pp. 61–70. URL: http://doi.org/10.1007/978-3-030-60365-6_7.

[9] KH Tam et al. “Predicting Clinical Endpoints and Visual Changes with Quality-Weighted Tissue-based Renal Histological Features”. In: (2022).

[10] KH Tam et al. “Deep-Learning Based Prediction of Clinical Endpoints in Renal Pre-Implantation Biopsies using Slide-Level Labels”. In: ESOT Congress 2021, Aug. 31, 2021. URL: https://esotcongress2021-esot.web.indrina.com/dailyprogramme/programelement/60d0515d9929d2001032573d.

[11] Hanna T Sjoberg et al. “Tumour irradiation combined with vascular-targeted photodynamic therapy enhances antitumour effects in pre-clinical prostate cancer”. In: *British journal of cancer* 125.4 (2021), pp. 534–546. URL: https://doi.org/10.1038/s41416-021-01450-6.

[12] H. McGivern et al. “Research Biopsies Kidneys For Transplant Audit To Compare Clinical Safety With Tissue Quality”. In: ESOT Congress, 2021. URL: https://www.postersessiononline.eu/173580348_eu/congresos/ESOT2021/aula/-POS_105_ESOT2021.pdf.

[13] Hongqing Han et al. “Self-Supervised Voxel-Level Representation Rediscovers Subcellular Structures in Volume Electron Microscopy”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 1874–1883.

[14] OpenStax. *The Kidneys and Osmoregulatory Organs*. https://bio.libretexts.org/Bookshelves/Introductory_and_General_Biology/General_Biology_1e_(OpenStax)/7%3A_Animal_Structure_and_Function/41%3A_Osmotic_Regulation_and_Excretion/41.2%3A_The_Kidneys_and_Osmoregulatory_Organs. [Online; last accessed 9 Apr 2022]. 2022.

[15] Andrew Siedlecki et al. “Delayed graft function in the kidney transplant”. In: *American journal of transplantation* 11.11 (2011), pp. 2279–2296.

[16] Reka Korbély et al. “Molecular biomarker candidates of acute kidney injury in zero-hour renal transplant needle biopsies”. In: *Transplant International* 24.2 (2011), pp. 143–149.

[17] Caroline Liu et al. “Association of deceased donor acute kidney injury with recipient graft survival”. In: *JAMA network open* 3.1 (2020), e1918634–e1918634.

[18] Winfred W Williams et al. “Clinical role of the renal transplant biopsy”. In: *Nature Reviews Nephrology* 8.2 (2012), pp. 110–121.

[19] Maarten Naesens. “Zero-time renal transplant biopsies: a comprehensive review”. In: *Transplantation* 100.7 (2016), pp. 1425–1439.

[20] Katrien De Vusser et al. “The predictive value of kidney allograft baseline biopsies for long-term graft survival”. In: *Journal of the American Society of Nephrology* 24.11 (2013), pp. 1913–1923.

[21] Mark Haas et al. “Arteriosclerosis in kidneys from healthy live donors: comparison of wedge and needle core perioperative biopsies”. In: *Archives of pathology & laboratory medicine* 132.1 (2008), pp. 37–42.

[22] Zsuzsanna Bago-Horvath et al. “The cutting (w) edge—comparative evaluation of renal baseline biopsies obtained by two different methods”. In: *Nephrology Dialysis Transplantation* 27.8 (2012), pp. 3241–3248.

[23] Abdulla K Salahudeen et al. “Involvement of the mitochondrial pathway in cold storage and rewarming-associated apoptosis of human renal proximal tubular cells”. In: *American journal of transplantation* 3.3 (2003), pp. 273–280.

[24] David P Basile et al. “Pathophysiology of acute kidney injury”. In: *Comprehensive Physiology* 2.2 (2012), p. 1303.

[25] RB Munivenkatappa et al. “The Maryland aggregate pathology index: a deceased donor kidney biopsy scoring system for predicting graft failure”. In: *American Journal of Transplantation* 8.11 (2008), pp. 2316–2324.

[26] D Anglicheau et al. “A simple clinico-histopathological composite scoring system is highly predictive of graft outcomes in marginal donors”. In: *American Journal of Transplantation* 8.11 (2008), pp. 2325–2334.

[27] M Antonieta Azancot et al. “The reproducibility and predictive value on outcome of renal biopsies from expanded criteria donors”. In: *Kidney international* 85.5 (2014), pp. 1161–1168.

[28] MGJ Snoeijs et al. “Histological assessment of preimplantation biopsies may improve selection of kidneys from old donors after cardiac death”. In: *American Journal of Transplantation* 8.9 (2008), pp. 1844–1851.

[29] José António Lopes et al. “Evaluation of pre-implantation kidney biopsies: comparison of Banff criteria to a morphometric approach”. In: *Kidney international* 67.4 (2005), pp. 1595–1600.

[30] Lorraine C Racusen et al. “The Banff 97 working classification of renal allograft pathology”. In: *Kidney international* 55.2 (1999), pp. 713–723. URL: https://doi.org/10.1046/j.1523-1755.1999.00299.x.

[31] Reza Abdi et al. “BASELINE GLOMERULAR SIZE AS A PREDICTOR OF FUNCTION IN HUMAN RENAL TRANSPLANTATION”. In: *Transplantation* 66.3 (1998), pp. 329–333.

[32] Kevin Chen et al. “Deceased-Donor Kidney Biopsy Scoring Systems for Predicting Future Graft Function: A Comparative Study”. In: *Transplantation Proceedings*. Vol. 53. 3. Elsevier. 2021, pp. 906–912.

[33] Akinlolu O Ojo et al. “Delayed Graft Function: Risk Factors and Implications for Renal Allograft Survival1”. In: *Transplantation* 63.7 (1997), pp. 968–974.

[34] Rinaldo Bellomo et al. “Acute renal failure–definition, outcome measures, animal models, fluid therapy and information technology needs: the Second International Consensus Conference of the Acute Dialysis Quality Initiative (ADQI) Group”. In: *Critical care* 8.4 (2004), pp. 1–9.

[35] Ravindra L Mehta et al. “Acute Kidney Injury Network: report of an initiative to improve outcomes in acute kidney injury”. In: *Critical care* 11.2 (2007), pp. 1–8.

[36] John A Kellum et al. “Kidney disease: improving global outcomes (KDIGO) acute kidney injury work group. KDIGO clinical practice guideline for acute kidney injury”. In: *Kidney international supplements* 2.1 (2012), pp. 1–138.

[37] K Solez et al. “The morphology of" acute tubular necrosis" in man: analysis of 57 renal biopsies and a comparison with the glycerol model.” In: *Medicine* 58.5 (1979), pp. 362–376.

[38] Kim Solez et al. “Morphology of ischemic acute renal failure, normal function, and cyclosporine toxicity in cyclosporine-treated renal allograft recipients”. In: *Kidney international* 43.5 (1993), pp. 1058–1067.

[39] Tobias T Pieters et al. “Histological characteristics of Acute Tubular Injury during Delayed Graft Function predict renal function after renal transplantation”. In: *Physiological reports* 7.5 (2019), e14000. URL: https://dx.doi.org/10.14814%5C%2Fphy2.14000.

[40] Neil J Glassford et al. “Interobserver agreement for post mortem renal histopathology and diagnosis of acute tubular necrosis in critically ill patients”. In: *Critical Care and Resuscitation* 19.4 (2017), pp. 337–343.

[41] Yumeng Wen et al. “A systematic review of clinical characteristics and histologic descriptions of acute tubular injury”. In: *Kidney international reports* 5.11 (2020), pp. 1993–2001. URL: https://doi.org/10.1016/j.ekir.2020.08.026.

[42] Bernd Schröppel et al. “Delayed kidney graft function: from mechanism to translation”. In: *Kidney international* 86.2 (2014), pp. 251–258.

[43] Christophe Legendre et al. “Factors influencing long-term outcome after kidney transplantation”. In: *Transplant international* 27.1 (2014), pp. 19–27.

[44] Daniel W Louvar et al. ““Nature versus nurture” study of deceased-donor pairs in kidney transplantation”. In: *Journal of the American Society of Nephrology* 20.6 (2009), pp. 1351–1358.

[45] Karim Marzouk et al. “The impact of vascular anastomosis time on early kidney transplant outcomes”. In: *Transplantation research* 2.1 (2013), pp. 1–5.

[46] P Halloran et al. “Factors influencing early renal function in cadaver kidney transplants. A case-control study.” In: *Transplantation* 45.1 (1988), pp. 122–127.

[47] Stacey Rolak et al. “Outcomes of Delayed Graft Function in Kidney Transplant Recipients Stratified by Histologic Biopsy Findings”. In: *Transplantation Proceedings*. Vol. 53. Elsevier. 2021, pp. 1462–1469.

[48] Steen Olsen et al. “Primary acute renal failure (" acute tubular necrosis") in the transplanted kidney: morphology and pathogenesis.” In: *Medicine* 68.3 (1989), pp. 173–187.

[49] Andrew JP Lewington et al. “Raising awareness of acute kidney injury: a global perspective of a silent killer”. In: *Kidney international* 84.3 (2013), pp. 457–467. URL: https://doi.org/10.1038/ki.2013.153.

[50] Jeffrey R Schelling. “Tubular atrophy in the pathogenesis of chronic kidney disease progression”. In: *Pediatric nephrology* 31.5 (2016), pp. 693–706.

[51] Manjeri A Venkatachalam et al. “Acute kidney injury: a springboard for progression in chronic kidney disease”. In: *American Journal of Physiology-Renal Physiology* 298.5 (2010), F1078–F1094.

[52] Asif A Sharfuddin et al. “Pathophysiology of ischemic acute kidney injury”. In: *Nature Reviews Nephrology* 7.4 (2011), pp. 189–200.

[53] Prasad Devarajan. “Update on mechanisms of ischemic acute kidney injury”. In: *Journal of the American Society of Nephrology* 17.6 (2006), pp. 1503–1520. URL: https://doi.org/10.1681/ASN.2006010017.

[54] John A Kellum et al. “Acute kidney injury”. In: *Nature Reviews Disease Primers* 7.1 (2021), pp. 1–17.

[55] Lakhmir S Chawla et al. “The severity of acute kidney injury predicts progression to chronic kidney disease”. In: *Kidney international* 79.12 (2011), pp. 1361–1369.

[56] LK Kayler et al. “Outcomes and utilization of kidneys from deceased donors with acute kidney injury”. In: *American Journal of Transplantation* 9.2 (2009), pp. 367–373.

[57] Isaac E Hall et al. “Associations of deceased donor kidney injury with kidney discard and function after transplantation”. In: *American Journal of Transplantation* 15.6 (2015), pp. 1623–1631.

[58] K Amann et al. “Kidney transplant pathology”. In: *Der Pathologe* 32.2 (2011), pp. 124–134.

[59] Alton B Farris et al. “What is the best way to measure renal fibrosis?: A pathologist’s perspective”. In: *Kidney international supplements* 4.1 (2014), pp. 9–15.

[60] Desley Neil. *Interpreting the Remuzzi score - The Pithia Trial*. Nov. 2016. URL: http://www.pithia.org.uk/wp-content/uploads/2016/11/Interpreting-the-Remuzzi-Score_Desley-Neil.pdf.

[61] Mark A Lusco et al. “AJKD atlas of renal pathology: tubular atrophy”. In: *American Journal of Kidney Diseases* 67.6 (2016), e33–e34. URL: https://doi.org/10.1053/j.ajkd.2016.04.007.

[62] Agnes B Fogo et al. “AJKD Atlas of Renal Pathology: Kidney Transplant Interstitial Fibrosis/Tubular Atrophy”. In: *American Journal of Kidney Diseases* 69.5 (2017), e23–e24. URL: https://doi.org/10.1053/j.ajkd.2017.02.002.

[63] Agnes B Fogo et al. “AJKD atlas of renal pathology: toxic acute tubular injury”. In: *American Journal of Kidney Diseases* 67.6 (2016), e31–e32. URL: https://doi.org/10.1053/j.ajkd.2016.04.003.

[64] Sept. 2019. URL: https://quod.org.uk/.

[65] URL: https://www.odt.nhs.uk/odt-structures-and-standards/key-strategies/archived-strategies/taking-organ-transplantation-to-2020/.

[66] Jürgen Schmidhuber. “Deep learning in neural networks: An overview”. In: *Neural networks* 61 (2015), pp. 85–117. URL: https://doi.org/10.1016/j.neunet.2014.09.003.

[67] Christopher M Bishop et al. *Pattern recognition and machine learning*. Vol. 4. 4. Springer, 2006.

[68] Antonio Criminisi et al. *Decision forests for computer vision and medical image analysis*. Springer Science & Business Media, 2013.

[69] Navneet Dalal et al. “Histograms of oriented gradients for human detection”. In: *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR’05)*. Vol. 1. Ieee. 2005, pp. 886–893.

[70] David G Lowe. “Object recognition from local scale-invariant features”. In: *Proceedings of the seventh IEEE international conference on computer vision*. Vol. 2. Ieee. 1999, pp. 1150–1157.

[71] Dong-Chen He et al. “Texture unit, texture spectrum, and texture analysis”. In: *IEEE transactions on Geoscience and Remote Sensing* 28.4 (1990), pp. 509–512.

[72] Serge Beucher. “Use of watersheds in contour detection”. In: *Proceedings of the International Workshop on Image Processing*. CCETT. 1979.

[73] Xiaofeng Ren et al. “Learning a classification model for segmentation”. In: *Computer Vision, IEEE International Conference on*. Vol. 2. IEEE Computer Society. 2003, pp. 10–10.

[74] Kurt Hornik et al. “Multilayer feedforward networks are universal approximators”. In: *Neural networks* 2.5 (1989), pp. 359–366.

[75] Jia Deng et al. "Imagenet: A large-scale hierarchical image database". In: *2009 IEEE conference on computer vision and pattern recognition*. Ieee. 2009, pp. 248–255.

[76] David E Rumelhart et al. "Learning representations by back-propagating errors". In: *nature* 323.6088 (1986), pp. 533–536. URL: https://doi.org/10.1038/323533a0.

[77] David H Hubel et al. "Shape and arrangement of columns in cat's striate cortex". In: *The Journal of physiology* 165.3 (1963), p. 559.

[78] Kunihiko Fukushima et al. "Neocognitron: A self-organizing neural network model for a mechanism of visual pattern recognition". In: *Competition and cooperation in neural nets*. Springer, 1982, pp. 267–285.

[79] Yann LeCun et al. "Gradient-based learning applied to document recognition". In: *Proceedings of the IEEE* 86.11 (1998), pp. 2278–2324.

[80] *The Cancer Genome Atlas Program*. URL: https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga.

[81] Kaiming He et al. "Deep residual learning for image recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.

[82] Ross Girshick. "Fast r-cnn". In: *Proceedings of the IEEE international conference on computer vision*. 2015, pp. 1440–1448.

[83] Kaiming He et al. "Mask r-cnn". In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 2961–2969. URL: https://doi.org/10.48550/arXiv.1703.06870.

[84] Shaoqing Ren et al. "Faster r-cnn: Towards real-time object detection with region proposal networks". In: *Advances in neural information processing systems* 28 (2015).

[85] Joseph Redmon et al. "YOLOv3: An Incremental Improvement". In: *CoRR* abs/1804.02767 (2018). arXiv: 1804.02767. URL: http://arxiv.org/abs/1804.02767.

[86] Jonathan Long et al. "Fully convolutional networks for semantic segmentation". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 3431–3440.

[87] Olaf Ronneberger et al. "U-net: Convolutional networks for biomedical image segmentation". In: *International Conference on Medical image computing and computer-assisted intervention*. Springer. 2015, pp. 234–241. URL: https://doi.org/10.48550/arXiv.1505.04597.

[88] Jun-Yan Zhu et al. "Toward multimodal image-to-image translation". In: *Advances in neural information processing systems*. 2017, pp. 465–476.

[89] Phillip Isola et al. "Image-to-image translation with conditional adversarial networks". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 1125–1134.

[90] Diederik P. Kingma et al. "Auto-Encoding Variational Bayes". In: *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*. Ed. by Yoshua Bengio et al. 2014. URL: http://arxiv.org/abs/1312.6114.

[91] Irina Higgins et al. "beta-vae: Learning basic visual concepts with a constrained variational framework". In: (2016).

[92] Alex Krizhevsky et al. "Imagenet classification with deep convolutional neural networks". In: *Advances in neural information processing systems* 25 (2012).

[93] Dan Ciresan et al. "Deep neural networks segment neuronal membranes in electron microscopy images". In: *Advances in neural information processing systems* 25 (2012).

[94] Dan C Cireşan et al. "Mitosis detection in breast cancer histology images with deep neural networks". In: *International conference on medical image computing and computer-assisted intervention*. Springer. 2013, pp. 411–418.

[95] Karen Simonyan et al. "Very Deep Convolutional Networks for Large-Scale Image Recognition". In: *International Conference on Learning Representations*. 2015. URL: https://doi.org/10.48550/arXiv.1409.1556.

[96] Christian Szegedy et al. "Going deeper with convolutions". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 1–9.

[97] Ian Goodfellow et al. "Generative adversarial nets". In: *Advances in neural information processing systems* 27 (2014).

[98] Jun-Yan Zhu et al. "Unpaired image-to-image translation using cycle-consistent adversarial networks". In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 2223–2232. URL: https://doi.org/10.48550/arXiv.1703.10593.

[99] Andrew Janowczyk et al. "Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases". In: *Journal of pathology informatics* 7 (2016).

[100] Pooya Mobadersany et al. "Predicting cancer outcomes from histology and genomics using convolutional networks". In: *Proceedings of the National Academy of Sciences* 115.13 (2018), E2970–E2979.

[101] Manan Shah et al. "Deep learning assessment of tumor proliferation in breast cancer histological images". In: *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE. 2017, pp. 600–603.

[102] Korsuk Sirinukunwattana et al. "Novel digital signatures of tissue phenotypes for predicting distant metastasis in colorectal cancer". In: *Scientific reports* 8.1 (2018), pp. 1–13.

[103] Yan Xu et al. "Deep learning of feature representation with multiple instance learning for medical image analysis". In: *2014 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE. 2014, pp. 1626–1630.

[104] Angel Cruz-Roa et al. "Automatic detection of invasive ductal carcinoma in whole slide images with convolutional neural networks". In: *Medical Imaging 2014: Digital Pathology*. Vol. 9041. SPIE. 2014, p. 904103.

[105] Geert Litjens et al. "A survey on deep learning in medical image analysis". In: *Medical image analysis* 42 (2017), pp. 60–88.

[106] Michael Gadermayr et al. "CNN cascades for segmenting sparse objects in gigapixel whole slide images". In: *Computerized Medical Imaging and Graphics* 71 (2019), pp. 40–48.

[107] Jaime Gallego et al. "Glomerulus classification and detection based on convolutional neural networks". In: *Journal of Imaging* 4.1 (2018), p. 20. URL: https://doi.org/10.3390/jimaging4010020.

[108] Olivier Simon et al. "Multi-radial LBP features as a tool for rapid glomerular detection and assessment in whole slide histopathology images". In: *Scientific reports* 8.1 (2018), pp. 1–11. URL: https://doi.org/10.1038/s41598-018-20453-7.

[109] Jon N Marsh et al. "Deep learning global glomerulosclerosis in transplant kidney frozen sections". In: *IEEE transactions on medical imaging* 37.12 (2018), pp. 2718–2728. URL: https://doi.org/10.1109/tmi.2018.2851150.

[110] David Ledbetter et al. "Prediction of kidney function from biopsy images using convolutional neural networks". In: *arXiv preprint arXiv:1702.01816* (2017). URL: https://doi.org/10.48550/arXiv.1702.01816.

[111] Yoshimasa Kawazoe et al. "Faster R-CNN-based glomerular detection in multistained human whole slide images". In: *Journal of Imaging* 4.7 (2018), p. 91. URL: https://doi.org/10.3390/jimaging4070091.

[112] Michael Gadermayr et al. "Generative adversarial networks for facilitating stain-independent supervised and unsupervised segmentation: a study on kidney histology". In: *IEEE transactions on medical imaging* 38.10 (2019), pp. 2293–2302.

[113] M Tarek Shaban et al. "Staingan: Stain style transfer for digital histological images". In: *2019 Ieee 16th international symposium on biomedical imaging (Isbi 2019)*. IEEE. 2019, pp. 953–956.

[114] Michael Gadermayr et al. "Which way round? a study on the performance of stain-translation for segmenting arbitrarily dyed histological images". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2018, pp. 165–173.

[115] Michael Gadermayr et al. "Unsupervisedly training GANs for segmenting digital pathology with automatically generated annotations". In: *arXiv preprint arXiv:1805.10059* (2018).

[116] Meyke Hermsen et al. "Deep learning–based histopathologic assessment of kidney tissue". In: *Journal of the American Society of Nephrology* 30.10 (2019), pp. 1968–1979. URL: https://doi.org/10.1681/asn.2019020144.

[117] Nassim Bouteldja et al. "Deep learning–based segmentation and quantification in experimental kidney histopathology". In: *Journal of the American Society of Nephrology* 32.1 (2021), pp. 52–68. URL: https://doi.org/10.1681/ASN.2020050597.

[118] Brandon Ginley et al. "Neural network segmentation of interstitial fibrosis, tubular atrophy, and glomerulosclerosis in renal biopsies". In: *arXiv preprint arXiv:2002.12868* (2020). URL: https://doi.org/10.48550/arXiv.2002.12868.

[119] Jesper Kers et al. "Prediction models for delayed graft function: External validation on the Dutch prospective renal transplantation registry". In: *Nephrology Dialysis Transplantation* 33.7 (2018), pp. 1259–1268.

[120] Annemarie Weißenbacher. "Normothermic Kidney Preservation". PhD thesis. University of Oxford, 2018.

[121] Annemarie Weissenbacher et al. "Twenty-four–hour normothermic perfusion of discarded human kidneys with urine recirculation". In: *American Journal of Transplantation* 19.1 (2019), pp. 178–192. URL: https://doi.org/10.1111/ajt.14932.

[122] F Neri et al. "Kidneys from AKI Donors with Poor Long-Term Outcomes Present an Impaired Protein Expression for Cellular Stress and Repair". In: *AMERICAN JOURNAL OF TRANSPLANTATION*. Vol. 20. WILEY 111 RIVER ST, HOBOKEN 07030-5774, NJ USA. 2020, pp. 348–348.

[123] LK Kayler et al. "Influence of CIT-induced DGF on kidney transplant outcomes". In: *American Journal of Transplantation* 11.12 (2011), pp. 2657–2664.

[124] Ozan Sener et al. "Active learning for convolutional neural networks: A core-set approach". In: *arXiv preprint arXiv:1708.00489* (2017). URL: https://doi.org/10.48550/arXiv.1708.00489.

[125] Zhou Wang et al. "Image quality assessment: from error visibility to structural similarity". In: *IEEE transactions on image processing* 13.4 (2004), pp. 600–612.

[126] James Bergstra et al. "Hyperopt: A python library for optimizing the hyperparameters of machine learning algorithms". In: *Proceedings of the 12th Python in science conference*. Vol. 13. Citeseer. 2013, p. 20.

[127] Richard Liaw et al. "Tune: A Research Platform for Distributed Model Selection and Training". In: *arXiv preprint arXiv:1807.05118* (2018). URL: https://doi.org/10.48550/arXiv.1807.05118.

[128] Peter Bankhead et al. "QuPath: Open source software for digital pathology image analysis". In: *bioRxiv* (2017). eprint: https://www.biorxiv.org/content/early/2017/03/06/099796.full.pdf. URL: https://www.biorxiv.org/content/early/2017/03/06/099796.

[129] Hsin-Ying Lee et al. "Diverse image-to-image translation via disentangled representations". In: *Proceedings of the European conference on computer vision (ECCV)*. 2018, pp. 35–51.

[130] Mehul P Sampat et al. "Complex wavelet structural similarity: A new image similarity index". In: *IEEE transactions on image processing* 18.11 (2009), pp. 2385–2401.

[131] Diederik P. Kingma et al. "Adam: A Method for Stochastic Optimization". In: *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*. Ed. by Yoshua Bengio et al. 2015. URL: http://arxiv.org/abs/1412.6980.

[132] Yaroslav Ganin et al. "Domain-adversarial training of neural networks". In: *The journal of machine learning research* 17.1 (2016), pp. 2096–2030.

[133] Takeru Miyato et al. "Virtual adversarial training: a regularization method for supervised and semi-supervised learning". In: *IEEE transactions on pattern analysis and machine intelligence* 41.8 (2018), pp. 1979–1993.

[134] Konstantinos Kamnitsas et al. "Unsupervised domain adaptation in brain lesion segmentation with adversarial networks". In: *International conference on information processing in medical imaging*. Springer. 2017, pp. 597–609.

[135] Markus A Degel et al. "Domain and geometry agnostic CNNs for left atrium segmentation in 3D ultrasound". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2018, pp. 630–637.

[136] Maxime W Lafarge et al. "Domain-adversarial neural networks to address the appearance variability of histopathology images". In: *Deep learning in medical image analysis and multimodal learning for clinical decision support*. Springer, 2017, pp. 83–91.

[137] Leslie N Smith. "Cyclical learning rates for training neural networks". In: *2017 IEEE winter conference on applications of computer vision (WACV)*. IEEE. 2017, pp. 464–472.

[138] Tsung-Yi Lin et al. "Focal loss for dense object detection". In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 2980–2988.

[139] Yuri Boykov et al. "Graph cuts and efficient ND image segmentation". In: *International journal of computer vision* 70.2 (2006), pp. 109–131. URL: https://doi.org/10.1007/s11263-006-7934-5.

[140] Neeraj Kumar et al. "A multi-organ nucleus segmentation challenge". In: *IEEE transactions on medical imaging* 39.5 (2019), pp. 1380–1391.

[141] Quirino Lai et al. "Donor Quality Scoring Systems and Early Renal Function Measurements in Kidney Transplantation". In: *Understanding the Complexities of Kidney Transplantation*. Ed. by Jorge Ortiz et al. Rijeka: IntechOpen, 2011. Chap. 9. URL: https://doi.org/10.5772/17845.

[142] Bernhard Schölkopf et al. "Nonlinear component analysis as a kernel eigenvalue problem". In: *Neural computation* 10.5 (1998), pp. 1299–1319.

[143] Laurens Van der Maaten et al. "Visualizing data using t-SNE." In: *Journal of machine learning research* 9.11 (2008).

[144] Abubakar Abid et al. "Exploring patterns enriched in a dataset with contrastive principal component analysis". In: *Nature communications* 9.1 (2018), pp. 1–7.

[145] Alton B Farris et al. "Morphometric and visual evaluation of fibrosis in renal biopsies". In: *Journal of the American Society of Nephrology* 22.1 (2011), pp. 176–186.

[146] Montserrat M Diaz Encarnacion et al. "Correlation of quantitative digital image analysis with the glomerular filtration rate in chronic allograft nephropathy". In: *American Journal of Transplantation* 4.2 (2004), pp. 248–256.

[147] L Cx U Junqueira et al. "Picrosirius staining plus polarization microscopy, a specific method for collagen detection in tissue sections". In: *The Histochemical journal* 11.4 (1979), pp. 447–455.

[148] Ståle Sund et al. “Computerized image analysis vs semiquantitative scoring in evaluation of kidney allograft fibrosis and prognosis”. In: *Nephrology Dialysis Transplantation* 19.11 (2004), pp. 2838–2845.

[149] Paul C Grimm et al. “Computerized image analysis of Sirius Red–stained renal allograft biopsies as a surrogate marker to predict long-term allograft function”. In: *Journal of the American Society of Nephrology* 14.6 (2003), pp. 1662–1668.

[150] Jonathan M Street et al. “Automated quantification of renal fibrosis with Sirius Red and polarization contrast microscopy”. In: *Physiological reports* 2.7 (2014), e12088.

[151] Lars Pape et al. “Computer-assisted quantification of fibrosis in chronic allograft nephropaty by picosirius red-staining: a new tool for predicting long-term graft function1”. In: *Transplantation* 76.6 (2003), pp. 955–958.

[152] Ludovic C Gillet et al. “Targeted data extraction of the MS/MS spectra generated by data-independent acquisition: a new concept for consistent and accurate proteome analysis”. In: *Molecular & Cellular Proteomics* 11.6 (2012). URL: https://doi.org/10.1074/mcp.O111.016717.

[153] Yarin Gal et al. “Dropout as a bayesian approximation: Representing model uncertainty in deep learning”. In: *international conference on machine learning*. PMLR. 2016, pp. 1050–1059.

[154] Charles Blundell et al. “Weight uncertainty in neural network”. In: *International conference on machine learning*. PMLR. 2015, pp. 1613–1622.

[155] Kumar Shridhar et al. “Uncertainty estimations by softplus normalization in bayesian convolutional neural networks with variational inference”. In: *arXiv preprint arXiv:1806.05978* (2018).

[156] Simon Kohl et al. “A probabilistic u-net for segmentation of ambiguous images”. In: *Advances in Neural Information Processing Systems*. 2018, pp. 6965–6975.

[157] Corina Gurau et al. “Dropout distillation for efficiently estimating model confidence”. In: *arXiv preprint arXiv:1809.10562* (2018).

[158] Yongchan Kwon et al. “Uncertainty quantification using Bayesian neural networks in classification: Application to biomedical image segmentation”. In: *Computational Statistics & Data Analysis* 142 (2020), p. 106816.

[159] Alex Kendall et al. “Bayesian segnet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding”. In: *arXiv preprint arXiv:1511.02680* (2015). URL: https://doi.org/10.48550/arXiv.1506.02142.

[160] Jishnu Mukhoti et al. “Evaluating bayesian deep learning methods for semantic segmentation”. In: *arXiv preprint arXiv:1811.12709* (2018).

[161] José Ignacio Orlando et al. “U2-net: A bayesian u-net model with epistemic uncertainty feedback for photoreceptor layer segmentation in pathological oct scans”. In: *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*. IEEE. 2019, pp. 1441–1445.

[162] Terrance DeVries et al. “Leveraging uncertainty estimates for predicting segmentation quality”. In: *arXiv preprint arXiv:1807.00502* (2018).

[163] Matthew Ng et al. *Estimating Uncertainty in Neural Networks for Segmentation Quality Control*. Tech. rep. Technical Report. URL: https://www. doc. ic. ac. uk/˜ bglocker/public . . ., 2018.

[164] Dimity Miller et al. "Dropout sampling for robust object detection in open-set conditions". In: *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE. 2018, pp. 1–7.

[165] Dimity Miller et al. "Evaluating merging strategies for sampling-based uncertainty techniques in object detection". In: *2019 International Conference on Robotics and Automation (ICRA)*. IEEE. 2019, pp. 2348–2354.

[166] D. Morrison. "Uncertainty-aware Instance Segmentation using Dropout Sampling". In: 2019.

[167] Iftikhar Ahmad. "Biopsy of the transplanted kidney". In: *Seminars in interventional radiology*. Vol. 21. 04. Copyright© 2004 by Thieme Medical Publishers, Inc., 333 Seventh Avenue, New. 2004, pp. 275–281.

[168] Durk P Kingma et al. "Variational dropout and the local reparameterization trick". In: *Advances in neural information processing systems*. 2015, pp. 2575–2583.

[169] Kumar Shridhar et al. "A comprehensive guide to bayesian convolutional neural network with variational inference". In: *arXiv preprint arXiv:1901.02731* (2019).

[170] Nitish Srivastava et al. "Dropout: a simple way to prevent neural networks from overfitting". In: *The journal of machine learning research* 15.1 (2014), pp. 1929–1958.

[171] Balaji Lakshminarayanan et al. "Simple and scalable predictive uncertainty estimation using deep ensembles". In: *Advances in neural information processing systems*. 2017, pp. 6402–6413.

[172] Alex Kendall et al. "What uncertainties do we need in bayesian deep learning for computer vision?" In: *Advances in neural information processing systems*. 2017, pp. 5574–5584. URL: https://doi.org/10.48550/arXiv.1505.05424.

[173] Thomas A Nketia et al. "Towards quantifying the impact of cell boundary estimation on morphometric analysis for phenotypic screening". In: *2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI)*. IEEE. 2015, pp. 781–784.

[174] Anders Boesen Lindbo Larsen et al. "Autoencoding beyond pixels using a learned similarity metric". In: *International conference on machine learning*. 2016, pp. 1558–1566.

[175] Andrew L Maas et al. "Rectifier nonlinearities improve neural network acoustic models". In: *Proc. icml*. Vol. 30 (1). 2013, p. 3.

[176] Andrea Chatrian et al. "Towards the Identification of Histology Based Subtypes in Prostate Cancer". In: *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*. IEEE. 2019, pp. 948–952.

[177] Aaditya Ramdas et al. "On wasserstein two-sample testing and related families of nonparametric tests". In: *Entropy* 19.2 (2017), p. 47.

[178] Zhengzi Yi et al. "Deep learning identified pathological abnormalities predictive of graft loss in kidney transplant biopsies". In: *Kidney International* 101.2 (2022), pp. 288–298.

[179] Richard C Davis et al. "Deep Learning Segmentation of Glomeruli on Kidney Donor Frozen Sections". In: *medRxiv* (2021). URL: https://doi.org/10.1101/2021.09.16.21263707.

[180] Jesper Kers et al. "Deep learning-based classification of kidney transplant pathology: a retrospective, multicentre, proof-of-concept study". In: *The Lancet Digital Health* 4.1 (2022), e18–e26. URL: https://doi.org/10.1016/S2589-7500(21)00211-9.

[181] Ming Y Lu et al. "Data-efficient and weakly supervised computational pathology on whole-slide images". In: *Nature Biomedical Engineering* 5.6 (2021), pp. 555–570. URL: https://doi.org/10.1038/s41551-020-00682-w.

[182] Osamu Iizuka et al. "Deep learning models for histopathological classification of gastric and colonic epithelial tumours". In: *Scientific Reports* 10.1 (2020), pp. 1–11. URL: https://doi.org/10.1038/s41598-020-58467-9.

[183] Gabriele Campanella et al. "Clinical-grade computational pathology using weakly supervised deep learning on whole slide images". In: *Nature medicine* 25.8 (2019), pp. 1301–1309. URL: https://doi.org/10.1038/s41591-019-0508-1.

[184] B. Li et al. "Dual-stream Multiple Instance Learning Network for Whole Slide Image Classification with Self-supervised Contrastive Learning". In: *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Los Alamitos, CA, USA: IEEE Computer Society, June 2021, pp. 14313–14323. URL: https://doi.ieeecomputersociety.org/10.1109/CVPR46437.2021.01409.

[185] Sri G Yarlagadda et al. "Marked variation in the definition and diagnosis of delayed graft function: a systematic review". In: *Nephrology Dialysis Transplantation* 23.9 (2008), pp. 2995–3003.

[186] Oded Maron et al. "A framework for multiple-instance learning". In: *Advances in neural information processing systems* (1998), pp. 570–576.

[187] Maximilian Ilse et al. "Attention-based deep multiple instance learning". In: *International conference on machine learning*. PMLR. 2018, pp. 2127–2136. URL: https://doi.org/10.48550/arXiv.1802.04712.

[188] Kim Solez et al. "International standardization of criteria for the histologic diagnosis of renal allograft rejection: the Banff working classification of kidney transplant pathology". In: *Kidney international* 44.2 (1993), pp. 411–422. URL: https://doi.org/10.1038/ki.1993.259.

[189] Christian Szegedy et al. "Rethinking the inception architecture for computer vision". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 2818–2826.

[190] Joan Bruna et al. "Invariant scattering convolution networks". In: *IEEE transactions on pattern analysis and machine intelligence* 35.8 (2013), pp. 1872–1886.

[191] Matthew D Zeiler et al. "Visualizing and understanding convolutional networks". In: *European conference on computer vision*. Springer. 2014, pp. 818–833. URL: https://doi.org/10.48550/arXiv.1311.2901.

[192] Narine Kokhlikyan et al. *Captum: A unified and generic model interpretability library for PyTorch*. 2020. arXiv: 2009.07896 [cs.LG]. URL: https://doi.org/10.48550/arXiv.2009.07896.

[193] Mukund Sundararajan et al. "Axiomatic attribution for deep networks". In: *International Conference on Machine Learning*. PMLR. 2017, pp. 3319–3328. URL: https://doi.org/10.48550/arXiv.1807.05118.

[194] Julius Adebayo et al. "Sanity checks for saliency maps". In: *Advances in neural information processing systems* 31 (2018).

[195] Shruti N Tapiawala et al. "Delayed graft function and the risk for death with a functioning graft". In: *Journal of the American Society of Nephrology* 21.1 (2010), pp. 153–161.

[196] Takaya Saito et al. "The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets". In: *PloS one* 10.3 (2015), e0118432. URL: https://doi.org/10.1371/journal.pone.0118432.