IMPERIAL

MEng Individual Project

Imperial College London

Department of Computing

KidneyGrader: An Interpretable Deep Learning Pipeline for Automated Kidney Transplant Grading

Supervisor:
Professor Bernhard Kainz

Author:
Abrar Rashid

Second Marker:
Professor Ben Glocker

Clinical Supervisor:
Dr. Candice Roufosse

June 13, 2025

Abstract

A major cause of end-stage renal disease worldwide is kidney transplant failure. Despite this, transplant biopsy grading is still performed manually, which is inevitably subject to inter-rater variability and pathologist workload. We propose KidneyGrader, the first fully automated deep learning pipeline for fine-grained Banff tubulitis scoring - the principal marker of acute rejection. Our system comprises (i) a fully interpretable modular pipeline of deep learning models that performs structure segmentation of tubules, detection of inflammatory cells and subsequent grading based on tubular inflammation, and (ii) a end-to-end, tranformer-based TransMIL regressor model enhanced by segmentation-guided feature extraction. Our end-to-end model surpasses the state-of-the-art for coarse binary tubulitis classification (AUC 0.95 vs 0.83) on held-out data. It achieves a substantial reliability κ_w with expert labels, over twice the gold standard (0.75 vs <0.3), and a Pearson correlation of $r = 0.81$ with expert labels, after training on just 75 slides. Our modular interpretable pipeline offers clinically auditable structure-level quantification, mirroring the gold standard for tubulitis scoring and facilitating transparent AI deployment in clinical pathology. By reducing transplant grading variability and pathologist workload through automation, KidneyGrader marks a step towards closing the care gap for kidney transplant patients.

Acknowledgements

I would like to express my deep gratitude to my supervisor, Professor Bernhard Kainz, whose support, valuable insights, and encouragement to go above and beyond have not only shaped the development of this work, but also left a lasting impact on my approach to future endeavours. I want to also thank Professor Ben Glocker, for his thoughtful advice and insights early on during the interim review for the project in its early stages. I am also very grateful to both of them for their courses in Deep Learning and Machine Learning for Imaging, which have equipped me with the strong foundations to take on this work.

My sincere thanks goes to Dr Candice Roufosse from the Department of Immunology and Inflammation, who has very generously contributed her time and expertise to help shape the clinical relevance of this work. Her encouragement, advice and the time she dedicated to our discussions have added significant value and meaning to this endeavour.

I am also greatly thankful to Vishal Jain, whose unwavering support and presence since the very beginning has been invaluable to this work. His contributions to the patch extraction and detection inference scripts have also been a great help to the project. I would also like to thank Sarah Cechnicka, who supported me during the early stages of the project and shared with me valuable resources. I wish both Vishal and Sarah the very best for the remainder of their PhDs and I look forward to their future contributions.

I want to give a special thank you to my family, who have been with me every step of the way. Words cannot express how grateful I am for their love, sacrifices and support, through both good and difficult times. Their support and belief in me have been the foundation of everything I have achieved, and I would not be here without them.

Above all, Alhamdulillah (praise be to God).

Contents

Chapter 1

Introduction

1.1 Motivation

With the world's ageing population and the increasing rates of conditions such as high blood pressure and diabetes, kidney diseases such as Chronic Kidney Disease have been on the rise [1] and affect more than 10% of the world's population [2] as of 2022. For many patients at the later stages of kidney disease, kidney transplantation is the recommended option as it has a significantly higher survival rate and quality of life [3] than dialysis. These transplants require biopsies to determine the health of the graft and identify potential symptoms of rejection, through grading the amount of damage to structures such as scarring or inflammation by a renal pathologist [4]. However, this manual procedure poses two problems. Firstly, scoring a graft is labour-intensive, and requires a considerable amount of consultant time which could be better used elsewhere such as contact hours with patients. Secondly, the manual nature of the procedure means that it is difficult to reproduce as it is prone to inter-observer variability. There is an increasing need for a scalable, standardised solution in the healthcare industry to overcome these limitations in kidney transplant pathology.

The evolution of deep learning in the recent years offers a promising approach to these challenges. It seeks to leverage the capabilities of the recent shift to digital methods in pathology, and automate these routine and time-consuming tasks such as those concerning segmentation of images, classification of diseases and detecting regions of interest. Deep learning has had a lot of success in the recent years, having been deployed to clinical settings in fields like radiology, opthamology and dermatology [5]. It has also shown promising work in kidney pathology [6], which leads to this project's motivation of exploring the application of deep learning techniques for kidney transplant pathology to automate the analysis and grading process for kidney transplants.

This study aims to develop a novel interpretable kidney transplant biopsy grading pipeline to assess the degree of tubulitis, the primary indicator of acute allograft rejection [7], from a biopsy slide. We hypothesise that developing an interpretable system of deep-learning models to quantify the presence of inflammatory cells in tubules will be sufficient to predict a tubulitis score to pathologist-level reliability and will have considerable agreement with expert labels. To test this hypothesis, our approach will be in a modular fashion: we will first train a segmentation model to identify structure classes, followed by an instance labelling algorithm to extract tubules. We then use a detection model to identify inflammatory cell presence, and finally use an algorithm to count the presence of inflammatory cells within the tubule masks and derive a rule-based score based on the Banff classification.

We also hypothesise that, due to the recent advancements in deep learning with the attention mechanism and transformers [8], similar prediction performance and a degree of explainability comparable to the interpretability of the modular approach can be achieved with a single end-to-end model. The attention mechanism may well be analogous to the gold-standard approach, identifying key focal areas in the biopsy and prioritising them to make the final decision on the tubulitis score. With sufficient training, a model can potentially learn to detect domain-specific features and perhaps even capture new abstract patterns that suggest novel indicators for kidney allograft rejection. We will therefore also train two different attention-based models to benchmark their performance and explainability against the modular pipeline.

1.2 Novel Contributions

We present KidneyGrader - the first fully automated deep learning pipeline for fine-grained tubulitis grading of kidney transplant biopsies. As part of this work, we have developed two novel approaches for tubulitis scoring:

1. A **modular, fully interpretable multi-stage pipeline** based on inflammatory tubule quantification, providing pathologists with full visual and statistical verifiability of the decisions made, for clinical alignment with the gold-standard Banff grading rubrics.

2. A set of **end-to-end attention-based** regressor and classifier models that directly predict tubulitis score from whole slide images, achieving state-of-the-art (SOTA) performance and significantly exceeding inter-rater reliability of the gold standard, and already showing promising, though currently limited, explainability.

Our key technical contributions include:

- **State-of-the-Art Tubulitis Prediction Performance:** Our novel weakly supervised transformer-based (TransMIL) regressor model, trained on a small dataset of 75 slides, exceeded the existing best model [9] performance on coarse binary classification of tubulitis with an AUC of 0.95 (vs 0.83 in SOTA), attained a clinically significant accuracy of 83.3% of predictions within 1 grade of the expert label, and a high Pearson correlation of 0.81 with expert labels (section 4.3).

- **Substantial Reliability with E2E model:** The TransMIL-based end-to-end model achieves an inter-rater reliability κ_w of 0.75, more than double the gold standard (expert labelling) value of <0.3 (section 4.3). This value of 0.75 falls within the 'substantial' threshold [10], showing great potential for clinical integration (section 4.3).

- **Expert-Level Agreement with Interpretable Pipeline:** Our modular interpretable pipeline with the best setup has attained an inter-rater reliability κ_w of 0.29, at the upper end of the gold-standard's value of <0.3. It also achieves an impressive accuracy of 82.1% of predictions within 1 grade of the expert label (section 4.3).

- **Improved Segmentation Model:** For the modular pipeline, we introduced attention gates and an EfficientNet encoder into the U-Net segmentation model, improving performance for almost all classes, but most notably, for the vein/indeterminate vessel class, with a jump from 0.0352 to 0.432 for IoU (section 4.1.1). More importantly, we achieved an order-of-magnitude speedup in runtime due to the lightweight encoder, allowing for potential clinical application.

- **Introduction of New Data-Driven Metrics for Tubulitis Quantification:** During experimentation, we discovered greater correlations between the top-percentile of inflamed tubules and the expert labels compared to the max-count method from the gold-standard, giving up to a $\sim$10% increase in the Pearson r value (section 4.1.4).

- **Improved Instance Labelling in Histopathology on Unlabelled Data:** Previous studies have made use of connected component labelling to identify structures like glomeruli from semantic masks, but suffered from poor segmentations of touching structures like tubules (section 2.5.2). We designed a new two-pass, patch-level, watershed-based approach which addressed this problem (section 3.3).

- **Pathologist Collaboration:** Throughout the development of KidneyGrader, we have been in close collaboration with a leading consultant transplant pathologist at the Department of Immunology and Inflammation at Imperial College London to align our methods with clinical workflows.

Chapter 2

Background

2.1 Kidney Allograft Pathology

2.1.1 Digital Pathology

Digital pathology refers to set of practices surrounding the digitisation of the pathology imaging pipeline, comprising image acquisition, storage, processing and visualisation [11]. The primary procedure conducted is Whole Slide Imaging (WSI), which is the practice of using ultra high resolution scanners to capture a digital image of a glass slide, which can then be stored and annotated. This digital form of annotation opens the door for automation of the pathology workflow, and is fertile ground for application of advancements in artificial intelligence for image analysis, such as deep learning.

The introduction of WSI has provided many improvements over traditional microscopy. It eliminates the burden of physically storing slides, as well as the risk of deterioration of stains and breakage of slides [12]. It also provides the convenience of remote access over a virtual storage mechanism such as a cloud platform. Under the right conditions, it can greatly improve diagnostic efficiency and allow greater extents of observation of the slides for diagnosis, and the acquisition and reconstruction of slides can be further expedited by 80% and 50x respectively with specialised GPU-accelerated algorithms [13]. However, some disadvantages of WSI include a barrier to adoption due to the high costs of the infrastructure such as digital scanners and the high capacity local or cloud storage costs. Additionally, time and capital need to be invested in educating, convincing and training professionals to transition [14]. However, with the recent evolutions in automated analysis of WSIs, the emergence of classification and segmentation models as well as the rise of more affordable digital pathology scanners, adoption is gradually increasing in the space.

2.1.2 Banff Grading

Patients with end-stage renal disease (ESRD) suffer from permanent kidney failure, and a kidney transplant, or allograft, is often the preferred course of action. Graft function, however, needs to be closely monitored to identify symptoms suggesting acute rejection.

To achieve the above, a grading system known as the Banff Classification was developed in 1991, and is the gold standard for pathologists and clinicians [4]. It is comprised of calculating a series of lesion scores corresponding to specific types of damage to structures, such as tubulitis, glomerulitis and interstitial inflammation. The Banff Classification has many advantages, providing a standardised framework for pathologists to report the graft biopsies, enabling consistent patient care and research. Future improvements for this system aim to address drawbacks such as the complexity and reproducibility of the procedure through new diagnostic algorithms and web-based resources. Additionally, advancements in deep learning for digital pathology provide exciting grounds for identification and quantification of these structures to calculate scores in order to automate the grading process.

2.1.3 Inter-rater Reliability

Manual expert labelling of the individual scores in the Banff classification poses a fundamental limitation common in the medical field: low inter-rater reliability, which is the primary motivation for our development of automated Banff scoring. Inter-rater reliability is the level of agreement between different observers when evaluating the same data. It is measured in terms of Cohen's Kappa, which is defined as

$$\kappa = \frac{p_o - p_e}{1 - p_e}$$

where p_o is the *observed agreement* among raters, and p_e is the *expected agreement* by chance.

Quadratic Weighted Kappa

More specifically, in the context of the Banff classification, due to the ordinal nature of metrics (ordered scores such t0, t1, etc), *quadratic weighted kappa* is used [15]. This score is used to more severely penalise greater differences between scores (e.g., severely misclassifying a t0 as a t3 should be penalised much more than misclassifying the t0 as t1), allowing for a more nuanced representation of the inter-rater reliability.

$$\kappa_w = 1 - \frac{\sum_{i=1}^{N} \sum_{j=1}^{N} w_{ij} O_{ij}}{\sum_{i=1}^{N} \sum_{j=1}^{N} w_{ij} E_{ij}}$$

with the quadratic weight:

$$w_{ij} = \frac{(i-j)^2}{(N-1)^2}$$

where

- N is the number of ordinal categories
- i, j are indices for category levels
- w_{ij} is the weight assigned to disagreement between category i and j; more disagreement means higher weight
- O_{ij} is the matrix of observed agreement, which is proportion of times Rater A gave score i and Rater B gave score j
- E_{ij} is the matrix of expected agreement matrix, meaning the expected frequency (by chance) that Rater A gives score i and Rater B gives score j.

2.1.4 Quantifying Renal Allograft Rejection: Tubulitis Scoring

We aimed to identify the most clinically significant metric to determine renal allograft failure. We studied a subset of the lesion scores from the Banff Classification [4], and have chosen the tubulitis score metric for our pipeline, finding that it is the main lesion for diagnosing acute rejection [7]. Crucially, the tubulitis score is also a prime candidate for standardisation, as its inter-observer reliability is very low, reported to have a quadratic-weighted κ_w value of 0.1 - 0.3 in a study on Banff scoring inter-observer reliability [16], as show in figure 2.1. The specification for tubulitis scoring is as follows:

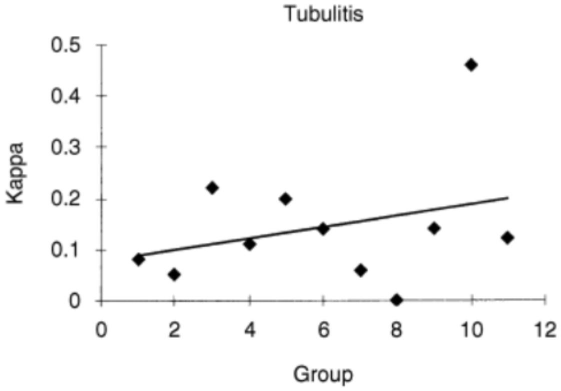


Figure 2.1: Kappa values for tubulitis across different groups by Furness *et al.* [16]

Description characterised by mononuclear cell presence in the basolateral (around the circumference or edge) parts of (non-atrophic/moderately atrophic) tubules, within the membrane. The score considers the count of mononuclear cells per 10 tubular epithelial cells (10 is the average number of epithelial cells per tubular cross-section). (Shown below in figure 2.2)

Scoring

- **t0**: no mononuclear cells in tubules or single focus of tubulitis only
- **t1**: at least 2 areas with 1 to 4 mononuclear cells/tubular cross section (or 10 tubular cells) in the most affected tubule of the focus
- **t2**: at least 2 areas with tubulitis, and one or more areas with 5 to 10 mononuclear cells/tubular cross section (or 10 tubular cells) in the most affected tubule of the focus
- **t3**: at least 2 areas with tubulitis, and at least one of those foci with more than 10 mononuclear cells.

For our study, we will consider the *tubular cross-sections* method rather than *per 10 tubular cells* for simplicity, and have therefore chosen the approach of segmenting out tubule instances in the first part of the multi-stage pipeline.

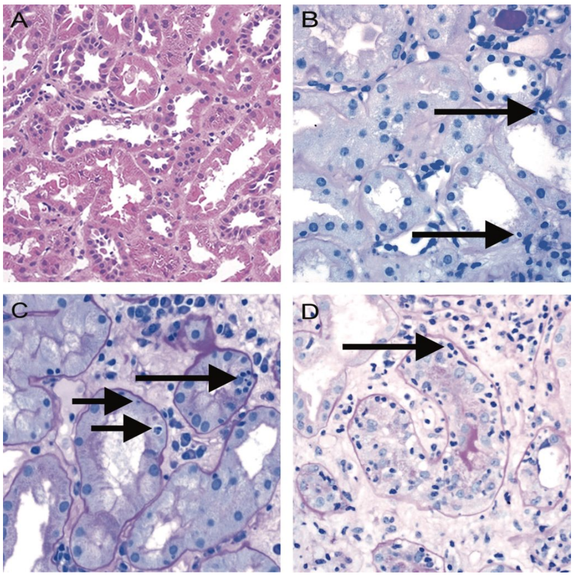


Figure 2.2: Degrees of Tubulitis. A is an example of lesion score t0. B shows t2, and arrows show mononuclear cells. C is a case of t2, with long arrows showing epithelial cells and short arrows indicate tubule with tubulitis. D is a cross-section of a tubule at t4, with long arrow pointing to one of the >10 mononuclear cells in the cross-section. All images are taken from samples stained with Periodic acid–Schiff (PAS) and at 400x original magnification. [17]

2.2 Foundations of Deep Learning in Medical Imaging

Deep learning is the subset of machine learning that is characterised by multi-layered artificial neural networks used to learn and predict features at varying scales, automating feature extraction. This has many use cases in medical image analysis as biological structures are highly complex, and can be adapted for different types of task. For example, Computer Aided Diagnosis (CAD) utilises image classification for identifying lesions and malignancies, predicting binary risk patterns of diseases like cancer, and has reportedly achieved better performance than conventional techniques by medical professionals [18]. Another key application is segmentation, for uses such as analysing organs, tumours, but also histopathology, which is the study of tissue and cells at the microscopic level. It is therefore becoming an increasingly mainstream option for digital pathology.

2.2.1 Convolutional Neural Networks

A fundamental component of the deep learning architecture is feature extraction, which is the process of retrieving hierarchical information about patterns and structures in data [19]. There have been notable attempts in the past to accomplish this, such as Fully Connected Networks (FCNs), which consists of each input in a layer being connected to each node in the next layer. However, this method falls subject to the curse of dimensionality, which outlines that as the dimensions of the feature space increases, data becomes sparser and tends to be closer to the boundary of the data space. This results in exponentially more data points being needed to learn meaningful patterns. In the case of FCNs, this leads to an exponentially increasing number of parameters, which is highly computationally expensive, and the potential of overfitting and loss of generalisation.

Convolutional neural networks (CNNs) were a key breakthrough in computer vision to address this problem. They consist of convolutional layers, which detect local patterns in small receptive fields, and perform special operations on previous layers to extract hierarchically higher level features. Convolutional layers consist of filters, or kernels, that move across the input image and perform a dot product between the filter and the pixels of the image in the receptive field, as follows:

$$y(k,i,j) = \sum_{c=1}^{C} \sum_{m=1}^{M} \sum_{n=1}^{N} x(c, i+m, j+n) \cdot w_k(c,m,n) + b_k \tag{2.1}$$

where:

- $y(k,i,j)$ is the output pixel in the resulting feature map
- (m,n) is the current position in the kernel
- $x(c, i+m, j+n)$ is the input value at colour channel c and position (i+m, j+n)
- $w_k(c,m,n)$ is the weight of the k^{th} filter for channel c at position (m,n)
- b_k is the bias term.

The above equation encapsulates some important advantages of CNNs. The small $M \times N$ receptive field of the kernel focuses on local regions in the image, accounting for spatial locality, which is the relationship between nearby pixels. This is complemented by pooling layers, which help reduce the dimensionality whilst retaining important information and extracting higher level features. In addition, the fact that the same kernel is used across the whole image is an example of parameter sharing, which results in much greater computational efficiency due to the small set of trainable parameters. Thus, CNNs have since become a widely adopted method for computer vision tasks.

In the medical imaging domain, a special variant of the CNN called the U-Net [20] was introduced in 2015, for the task of semantic segmentation. The U-Net presented a novel architecture (see figure 2.3), consisting of a contracting path (encoder) to capture context and a corresponding symmetric expanding path (decoder) to enable precise localisation. The contracting path consists of repeated convolutions and max pooling to downsample the image and obtain increasingly high level features. The expanding path then upsamples the feature map, and at each step concatenates the corresponding feature map from the contracting path; this is called a skip connection, and allows retention of spatial information lost during downsampling. This architecture allowed the U-Net to outperform existing state-of-the-art methods, such as CNNs, in segmentation tasks for various competitions such as the ISBI cell tracking challenge, as mentioned in [20]. It has presented vast benefits for biomedical imaging, due to its use of techniques such as data augmentation, which resulted in superior performance even with a limited amount of labelled training data - crucial for medical applications where it is notoriously difficult to obtain annotated data.

Since then, many variants of U-Net have been introduced, such as Attention U-Net, which consists of attention gates to retain important features and suppress background noise [21], UNet++, which contains nested skip pathways (multiple levels of skip connections) to better obtain multi-scale features and improve generalisation [22], and 3D U-Net for 3D volumetric data such as CT scans [23]. In the context of renal allograft WSI segmentation, work from previous students has made use of the nnU-Net architecture, which is a self-adapting segmentation pipeline that adjusts to a dataset's properties such as image size and label distribution, in order to standardise pre-processing.

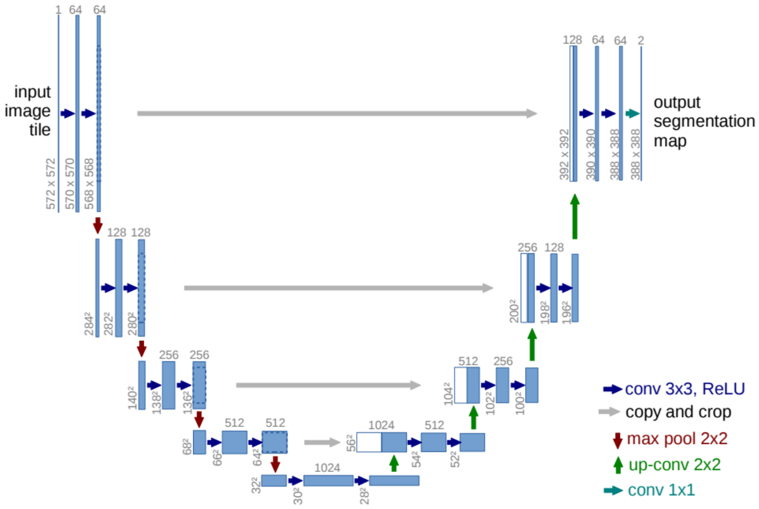


Figure 2.3: Original U-Net architecture [20]

2.2.2 Attention Mechanisms

Despite the widespread success of CNNs and their numerous applications, they suffer from a fundamental limitation. Convolution operations have inherent locality due to the small context window, which fails to capture long range dependencies, thus limiting performance. An equivalent problem was also suffered by the existing state-of-the-art for natural language processing (NLP) tasks at the time, namely recurrent neural networks (RNNs), which are sequential predictor networks based on feedback loops that take in sequential inputs as well as the previous output. The relationship between distant temporal inputs gets weakened due to the vanishing gradient problem during back propagation. This led to the introduction of attention mechanisms (figure 2.4) to dynamically focus on desired parts of the input and model global dependences. The function is as follows:

$$Attention(Q,K,V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{2.2}$$

where Q, K and V are linear projections of input data such that:

- Q represents the queries, or focus points, for attention
- K denotes the keys that the queries are compared against
- V contains the information obtained based on the attention scores from queries and keys.

2.2.3 Transformers

The attention mechanism primary focused on modelling parts of input sequences for generating output sequences. However, it did not address dependencies within a single sequence. This led to the birth of the transformer [8], which introduced self-attention mechanisms to compute relationships between any two parts of the input sequence. The transformer architecture also introduced multi-head attention, which solved a fundamental problem of single-head attention, which was that it could only focus on one type of relationship at a time. Multi-head attention allows for modelling of multiple complex relationships at once, which would be especially beneficial for applications like whole slide imaging, due to the complex patterns and signals found in tissue.

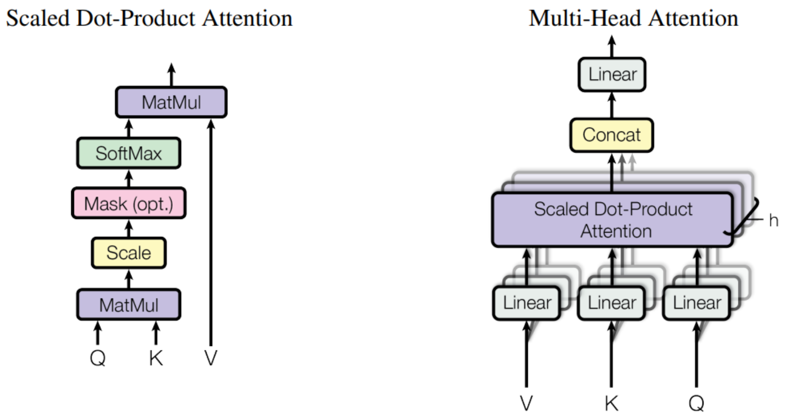


Figure 2.4: Attention mechanism [8]

The transformer architecture (figure 2.5) comprises of an encoder to process the input sequence and map it to continuous representations, while the decoder generates output symbols sequentially, using both the input encodings and the previously generated symbols as input. It utilises feed forward networks consisting of fully connected layers, applied separately to each token after the attention mechanism to introduce non-linearity, as well as positional encoding to retain information about token ordering. In addition, skip connections are in place to address the vanishing gradient problem and allow faster training convergence. This led to state-of-the-art performance and significantly less training time compared to CNNs and RNNs.

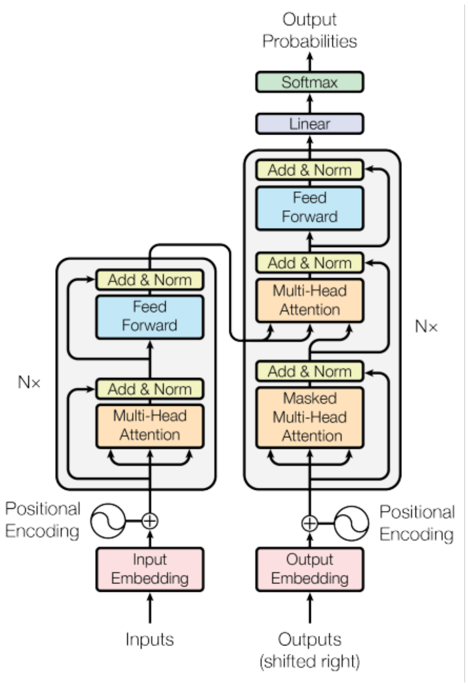


Figure 2.5: Transformer architecture [8]

Transformers have been revolutionary for NLP tasks, but have required adaptation for computer vision tasks. The Vision Transformer (ViT) [24] introduces support for image classification, by splitting images into patches to behave as tokens. The paper mentions that, when pre-trained on large amounts of training data, the ViT exceeds the performance of state-of-the-art CNNs while having better computational efficiency during training.

However, in the context of medical imaging, transformers like the ViT are less applicable due to the lack of abundant medical training data, but also the fact that transformers have limited localisation due to the lack of low-level details captured. This gave rise to architectures such as the Swin transformer [25], a pure transformer that creates hierarchical feature maps to capture features at multiple scales and uses a shifted windowing scheme of computing self-attention for computational efficiency, which led to it surpassing the state-of-the-art performance on the ADE20K and COCO (Common Objects in Context) datasets. For medical image segmentation, TransUNet [26] was a hybrid architecture developed to take advantage of the offerings of both the U-Net and the transformer. In the encoder, a CNN is used to generate a feature map, which is then used by the transformers to extract global context. It has a similar shape to the U-Net, where the self-attentive features are encoded by the transformer layers are upsampled and combined with CNN features from the skip-connections to allow precise localisation - which is critical for medical image segmentation.

We hypothesise that transformers have promising applications in explainable, automated evaluation of renal allograft quality due to their ability to model long-range dependencies and visualisability of the attention mechanism. Thus, in our study, we will explore the use of a transformer in the context of end-to-end tubulitis prediction.

2.3 Image Segmentation in Digital Pathology

In computer vision, image segmentation refers to the set of tasks of grouping regions in the image with similar properties and associating them to a class label or instance (an object identified in the image). In order to group regions with similar properties, features such as colour and intensity were primarily used in the past through algorithms such as clustering and thresholding. However, the progression of deep learning has enabled us to look into more complex features due to richer and deeper neural networks composed of multiple layers. Convolutional neural networks (CNNs) have enabled us to determine things such as structural features and patterns from images, which is beneficial for medical applications where structures and patterns can be difficult to predict beforehand.

In the field of digital pathology, deep learning has uses such as object detection and counting for mitotic events, segmentation for microscopic structures such as nuclei, and classification for diseases like cancer [27]. In our study, we perform segmentation for WSIs, in order to identify and quantify histological features. This will help derive basic metrics corresponding to the Banff Classification in order to be able to predict graft outcomes.

2.3.1 Semantic and Instance Segmentation

Semantic segmentation groups regions of pixels into classes and does not distinguish between instances of the same class. It consists of an input image $I \in \mathbb{R}^{H \times W \times C}$, where H and W are the height and width of the image respectively and C is the number of colour channels. The output segmentation mask will then be $M \in \mathbb{R}^{H \times W \times K}$, where K is the number of classes. Each pixel in the mask M is therefore mapped to a class label l_i, where $l_i \in \{l_0, l_1, ..., l_K\}$. Semantic segmentation has many uses, ranging from medical imaging for identifying tumour or cell regions, to autonomous vehicles to identify lanes and pedestrians. Previous work in kidney transplant pathology has applied semantic segmentation to WSIs of allograft biopsies to identify key histological structures [28].

Instance segmentation extends semantic segmentation by providing a distinct identification between instances, or objects of the same class in the segmentation mask. This means that the output is composed of two components, for example M_{cls} and M_{inst}, where $M_{cls} \in \mathbb{R}^{H \times W \times K}$ and $M_{inst} \in \mathbb{R}^{H \times W \times N}$, where N is the number of instances. This allows each pixel to simultaneously be associated with a class and an instance. Applications include those of the domain of semantic segmentation, but having extra requirements such as identifying and quantifying individual biological structures rather than simply segmenting the group of structures. For our study, the ideal approach to identify histopathological structures would be instance segmentation, but, as described in the methods section, our data is limited to semantic labels and we therefore take an alternative approach to achieve pseudo-instance segmentation.

2.3.2 Classical Instance Segmentation Algorithms

CCL Algorithm

In the absence of labels for segmentation, classical methods in computer vision such as connected component labelling (CCL) allow for instance segmentation, based on graph-based approaches. CCL works by identifying disjoint regions of foreground pixels with special connectivity rules, and assigning instance labels to those 'connected' sets of pixels. It is highly efficient due to its linear complexity and is therefore commonly used in instance labelling for disjoint structures in WSIs, such as the work by Hermsen *et al.* detailed in section 2.5.2.

Watershed Algorithm

A fundamental limitation of CCL is the inability to split touching instances. The watershed algorithm seeks to solve this by identifying foreground and background areas through traditional filtering, such as grayscale and Sobel filtering to find intensity changes to correspond to edges. It then applies markers to the foreground and background regions, and then 'floods' from the markers, assigning pixels to the closest marker that floods to it and assigning boundaries where floods from different markers meet. Moreover, enhancements can be applied prior to flooding, such as erosion to clean up noise, dilation to merge fragmented areas and distance transform to enhance foreground seed finding. We will explore the use of watershed for our tubule instance labelling approach in our pipeline, and compare it with CCL.

2.3.3 Modern Instance Segmentation Architectures

Instanseg

The tubulitis score, and many other scores in the Banff classification, measure levels of inflammation in histological structures of the kidney. In order to quantify this level of inflammation, we require deep learning models that are specialised to detect instances of cells in high-density regions. This is a complex problem, due to the morphological variability of these inflammatory cells, frequent overlapping of structures and domain shift due to factors such as stain variation. For our study, we require a method that is accurate and robust enough to be able to generalise to these factors without additional fine-tuning (due to our lack of ground truth instance labels), and simultaneously efficient enough for deployment in a clinical setting. InstanSeg [29] is a novel framework that addresses these through its use of predicting flow fields from each pixel towards the likely centroids that it belongs to, and grouping pixels that belong to the same nucleus into a compact region of feature space. Through leveraging spatial directionality instead of appearance-based features, it allows for greater generalisability to stain and morphology variations. Moreover, Goldsborough *et al* reported in the paper that InstanSeg achieved SOTA performance on six public brightfield and fluorescence datasets.

The limitations of this model, however, are that it was predominantly trained on other tissues such as skin, and on non-PAS stained slides. However, it shows promising signs of generalisability to our domain, having recently won second place in the MONKEY [30] challenge for its detection performance of inflammatory cells in kidney transplant biopsies. Therefore, we proceed to explore how InstanSeg generalises to our in-house PAS-stained dataset to detect inflammatory cells in our multi-stage pipeline, in order to test the hypothesis mentioned in the introduction.

2.4 Learning Paradigms for Medical Image Analysis

2.4.1 Supervised and Weakly Supervised Methods

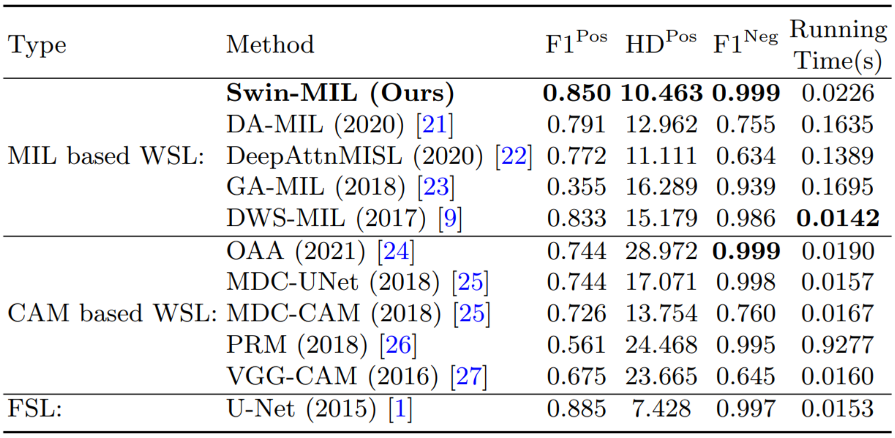
Supervised learning requires training with annotated, labelled datasets, in order to be able to make predictions on future data. However, in medical applications such as digital pathology, it is often difficult to get access to annotated and labelled datasets. This is due to various reasons, such as the time taken by an expert to manually annotate large images, and data protection concerns relating to patients' medical data. For this reason, these applications require models that are less supervised, with minimal requirement of annotated data.

Unsupervised learning involves leaving the model to train on completely unlabelled data, in order to discover patterns and structures in the data on its own. However, this comes with many concerns and challenges, such as the ambiguity caused by the absence of a ground truth, which makes evaluation subjective and application-dependent. For example, in a medical application, a model might produce a mask that erroneously segments non-diagnostic artifacts, which could cause clinically irrelevant results. Despite this, recent attempts such as the zero-shot unsupervised models CutLer [31] and ZUTIS [32] are impressive, but are still far from achieving comparable performance to supervised methods. However, it is worth noting that these recent methods have been trained on predominantly common object datasets such as COCO, so this poses another concern for their performance on WSI data. Therefore, while unsupervised techniques hold promise, further advancements are needed to address their limitations and ensure clinically accurate and meaningful results.

Weakly and semi-supervised methods propose a promising middle ground between fully unsupervised and supervised methods. Semi-supervised learning combines a small amount of labelled data and a larger set of unlabelled data to train the model, while weakly supervised learning characterises the use of incomplete or coarse labels during training. In the domain of digital pathology, a popular weakly supervised technique is multiple instance learning (MIL), and a recent transformer-based MIL method shows promising results for histopathology image segmentation [33].

Multiple Instance Learning

MIL collects training data into bags, with each bag containing multiple instances. Labels are only given at the bag level, rather than per instance. In the case of WSIs, the WSIs are the bags and the individual patches or pixels are the instances; the model can then learn to associate the presence of the bag-level label with the patterns of the individual instances. The paper above presents the novel use of a transformer to address the problem of instance independence, where typical MIL treats instances independently in a bag, which restricts the ability to capture relationships between instances. The integration of the Swin transformer resolved this by introducing self-attention into the architecture, to give high attention weights to similar features and low attention weights to dissimilar ones. This resulted in state-of-the-art performance comparable to fully supervised methods like the U-Net (2.6).



Type	Method	$F1^{Pos}$	HD^{Pos}	$F1^{Neg}$	Running Time(s)
MIL based WSL:	**Swin-MIL (Ours)**	**0.850**	**10.463**	**0.999**	0.0226
	DA-MIL (2020) [21]	0.791	12.962	0.755	0.1635
	DeepAttnMISL (2020) [22]	0.772	11.111	0.634	0.1389
	GA-MIL (2018) [23]	0.355	16.289	0.939	0.1695
	DWS-MIL (2017) [9]	0.833	15.179	0.986	**0.0142**
CAM based WSL:	OAA (2021) [24]	0.744	28.972	**0.999**	0.0190
	MDC-UNet (2018) [25]	0.744	17.071	0.998	0.0157
	MDC-CAM (2018) [25]	0.726	13.754	0.760	0.0167
	PRM (2018) [26]	0.561	24.468	0.995	0.9277
	VGG-CAM (2016) [27]	0.675	23.665	0.645	0.0160
FSL:	U-Net (2015) [1]	0.885	7.428	0.997	0.0153

Figure 2.6: Swin Transformer-based MIL vs other architectures including U-Net. WSL = weakly supervised learning, FSL = fully supervised learning. [33]

2.4.2 CLAM - A Weakly Supervised and Data-Efficient Attention-Based DL Framework for WSIs

Manual labelling of gigapixel WSIs is a primary limitation in the field of deep learning for pathology. Pixel-level annotations are time-consuming to produce and are thus scarce, so fully supervised machine learning pipelines may not have enough data to achieve human-level performance. Standard MIL methods attempt to solve this by using techniques like max-pooling to find the most informative patch in the slide and train the model, which can often lead to poor data efficiency and high label noise. Clustering-constrained Attention Multiple Instance Learning (CLAM) [34] is a Resnet-50 based framework designed to reduce dependence on pixel-level annotations and simultaneously make efficient use of data, by using slide-level labels with attention-based pooling.

The attention mechanism allows the model to learn to weight the diagnostic importance of different patches, and instance-level clustering of the high and low-attention patches can be used to behave as pseudo-labels to replace manual annotations. The attention mechanisms allow for greater explainability of these models, allowing the generation of attention heatmaps to visualise diagnostically relevant signals. CLAM has achieved high performance even with limited training data, such as its SOTA performance of >0.9 AUC on Renal Cell Carcinoma detection with 25% of the training data used by the other methods. It has also shown notable adaptability to domain shift and image modalities such as resection and biopsy samples. We therefore incorporate a CLAM-based approach as the baseline in our end-to-end model approach for automated kidney biopsy tubulitis scoring.

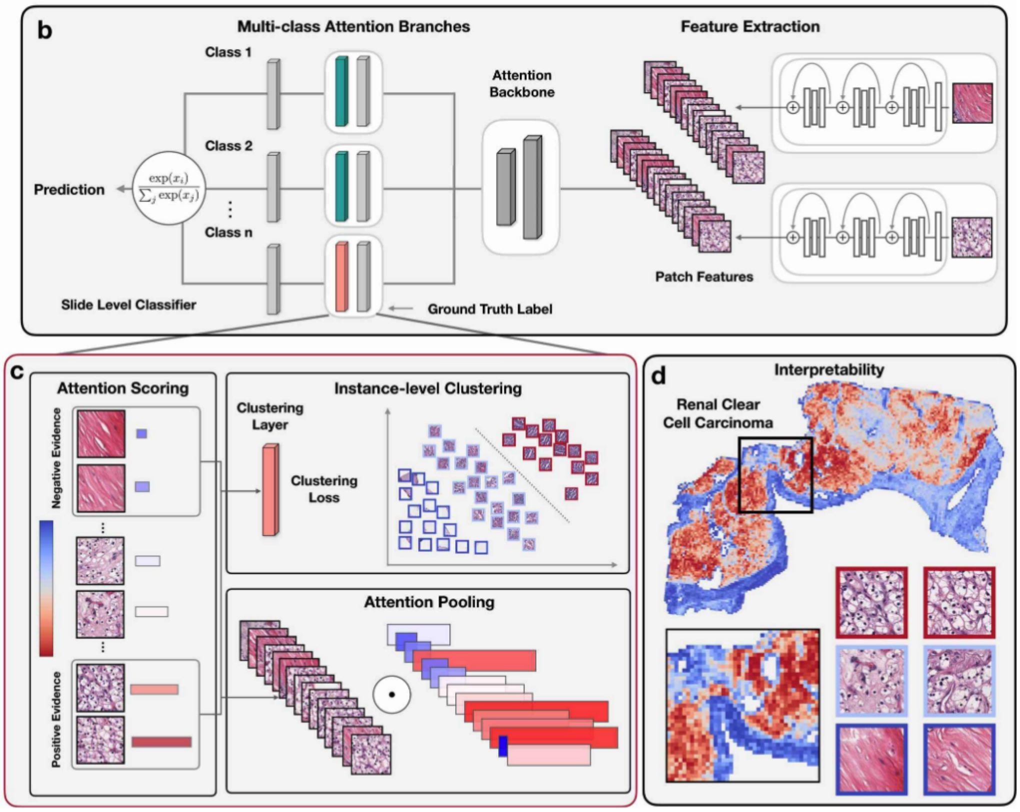


Figure 2.7: CLAM Architecture Overview [34]

2.4.3 TransMIL: Correlated MIL for WSIs with Transformers

Another key limitation in MIL methods is the reliance on the independent and identically distributed assumption for instances, failing to acknowledge any spatial correlations between them. TransMIL is another ResNet50-based method that proposes a correlated MIL framework, based on self-attention and a special Token Pyramid Transformer to model these correlations. In contrast to CLAM's attention based pooling, which does not take into account spatial relationships, TransMIL uses a Pyramid Position Encoding Generator (PPEG) (rather than a sinusoidal position encoder), to model spatial relationships at multiple levels. To investigate the importance of these spatial relationships and context-dependence of patches that characterise a WSI with tubulitis, we leverage TransMIL in our experimentation for the end-to-end model approach.

2.4.4 Foundation Models

Data scarcity is a limitation for many applications of deep learning, particularly in our domain of research, as mentioned previously. Transfer learning is an approach that has recently gained a lot of traction, due to its diverse range of applications and adaptations. It is defined as the reuse of an existing model trained on a specific task as the starting point for the new model, in order to transfer the existing 'knowledge' gained (hence the name). This is beneficial for data-scarce applications as the existing models are generally trained on large datasets to provide generalised outputs, and can then be fine-tuned with domain-specific training data to specialise the model.

Foundation models are a recent emergence based on transfer learning, and are general purpose models trained on massive (often multi-modal) datasets that can then be fine-tuned and adapted for specific applications. UNI is a general-purpose foundation model

[35] used for computational pathology, that has been pre-trained on over 100 million tissue patches from over 100,000 H&E slides across 20 different tissue types. It is therefore a strong candidate to apply to our dataset, due to the relatively small domain shift required to generalise to our PAS-stained data. For this study, we will take advantage of UNI to enhance feature extraction and facilitate better training of our end-to-end models, as part of our efforts to validate the research hypothesis.

2.5 Related work

2.5.1 Deep Learning for Coarse-Label Classification of Transplant Rejection

A recent study published in 2022 by Kers *et al.* [36] was one of the first to explore the application of deep learning to automate classification in kidney allograft histopathology. It aimed to classify kidney allograft biopsies into three categories: *normal*, *rejection* and *other diseases*, through using a two-step CNN approach, with the first CNN classifying the biopsy as *normal* or *diseased*, and the second then classifying the *diseased* biopsies as either *rejection* or *other diseases*. They trained the models on 5844 WSIs stained with PAS, H&E and Jones Silver, from 1948 patients from three different centres, and performed three-fold cross validation, with 101 of the patients' slides held out for testing. The pipeline achieved impressive performance, with an AUROC (measure of a model's ability to distinguish between classes) score of 0.87 for normal and diseased biopsy classification (where 1.0 is a perfect classifier score), and 0.75 for rejection and other disease classification. This study demonstrated great promise in the clinical application of deep learning models in kidney transplant pathology.

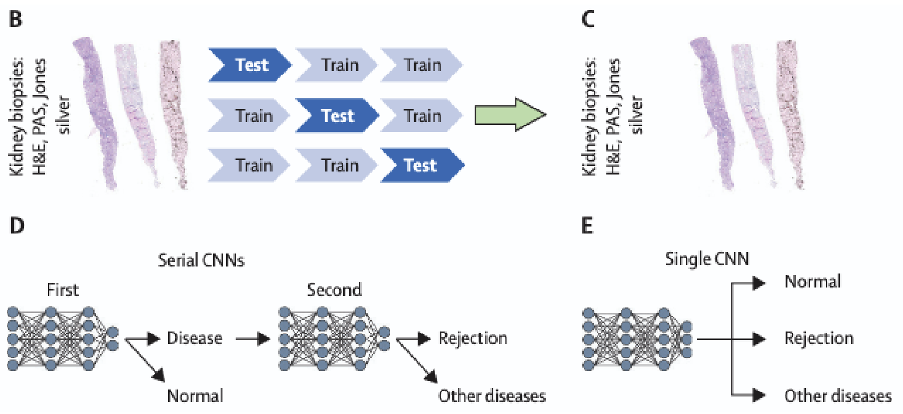


Figure 2.8: Two-step CNN workflow by Kers *et al.* [36].

However, while achieving great performance, the classifier is limited to coarse labels, which were originally assigned based on finer Banff lesion labels that characterised antibody-mediated rejection (ABMR), T-cell mediated rejection (TCMR), and mixed rejection. This ultimately leads to lack of explainability and limited clinical applicability due to granular Banff scoring being the clinically-accepted standard for rejection diagnosis. A model that is instead based on individual Banff lesion classification, would correlate better with the rubrics used by pathologists to calculate individual Banff scores and determine rejection. This underpins one of the key motivations for our study, which is to develop a novel pipeline to predict a granular Banff lesion score, namely tubulitis, due to its clinical significance in acute rejection.

2.5.2 CNNs for Correlation with Banff Lesion Scores

Hermsen *et al.* [37] aimed to take on this challenge of building a prediction system for individual Banff lesion scores for renal biopsies. They took the approach of computing quantitative tissue metrics, namely healthy and sclerotic glomeruli, tubular atrophy, interstitial fibrosis and inflammation within atrophic and non atrophic tubuli. This was done by implementing a structure segmentation CNN to semantically segment the structure classes, followed by a lymphocyte detection CNN to quantify inflammation, and trained on 125 WSI pairs of PAS and CD3-stained slides. The pipeline achieved notable performance, with many metrics correlating strongly with mean scores from five transplant pathologists, the highest being for total inflammation (r=0.84).

While this study shows commendable progress in automating quantification of features for Banff scoring, it has fallen short of predicting scores, which is the fundamental step towards clinical integration. This is understandable due to their pipeline, as only the glomeruli were successfully segmented at the instance level, but the other fundamental structures such as tubuli failed to adequately segment into instances. Their correlation plots confirm this, as the plot with the weakest correlation is the one for highest cell count per tubule (figure 2.9). In an ideal pipeline with perfect tubular instance segmentation, the highest inflammatory cell count per tubule should give perfect correlation and identical T-score results to the gold standard tubulitis scoring method, assuming the pathologist has made no mistakes either. We aim to address this challenge in our pipeline through a (to our knowledge) novel instance labelling approach on histological structures.

The approach by Hermsen *et al.* shows promise for clinical relevance of deep learning for Banff classification, and we aim to advance the field with our multi-stage pipeline by taking their study a step further, addressing the challenge of turning these structure level statistics into an explainable Banff score.

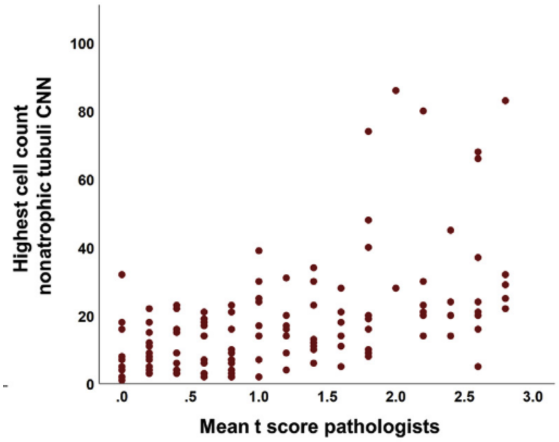


Figure 2.9: Correlation between expert T-scores and tubule inflammation quantification by Hermsen *et al.* [37].

Chapter 3

KidneyGrader: Methods and Experimentation

3.1 Overview

We develop KidneyGrader - an interpretable pipeline for automated analysis and prediction of renal transplant biopsy grading based on the Banff Tubulitis score (T-score). We aim to address a critical challenge in the assessment of renal biopsies among pathologists, namely, reproducibility, or inter-rater reliability. While there has been previous work to automate biopsy analysis with deep learning (please see section 2.5), the work has fallen short of deriving predictions for individual Banff classification scores, the fundamental metrics used in the gold standard approach. We aim to make a contribution in the field by addressing this challenge and, as mentioned in the hypothesis in 1.1, aim to make this pipeline explainable to facilitate potential for clinical integration. To test the hypotheses mentioned in the introduction (1.1), we consider two novel explainable approaches to derive a T-score:

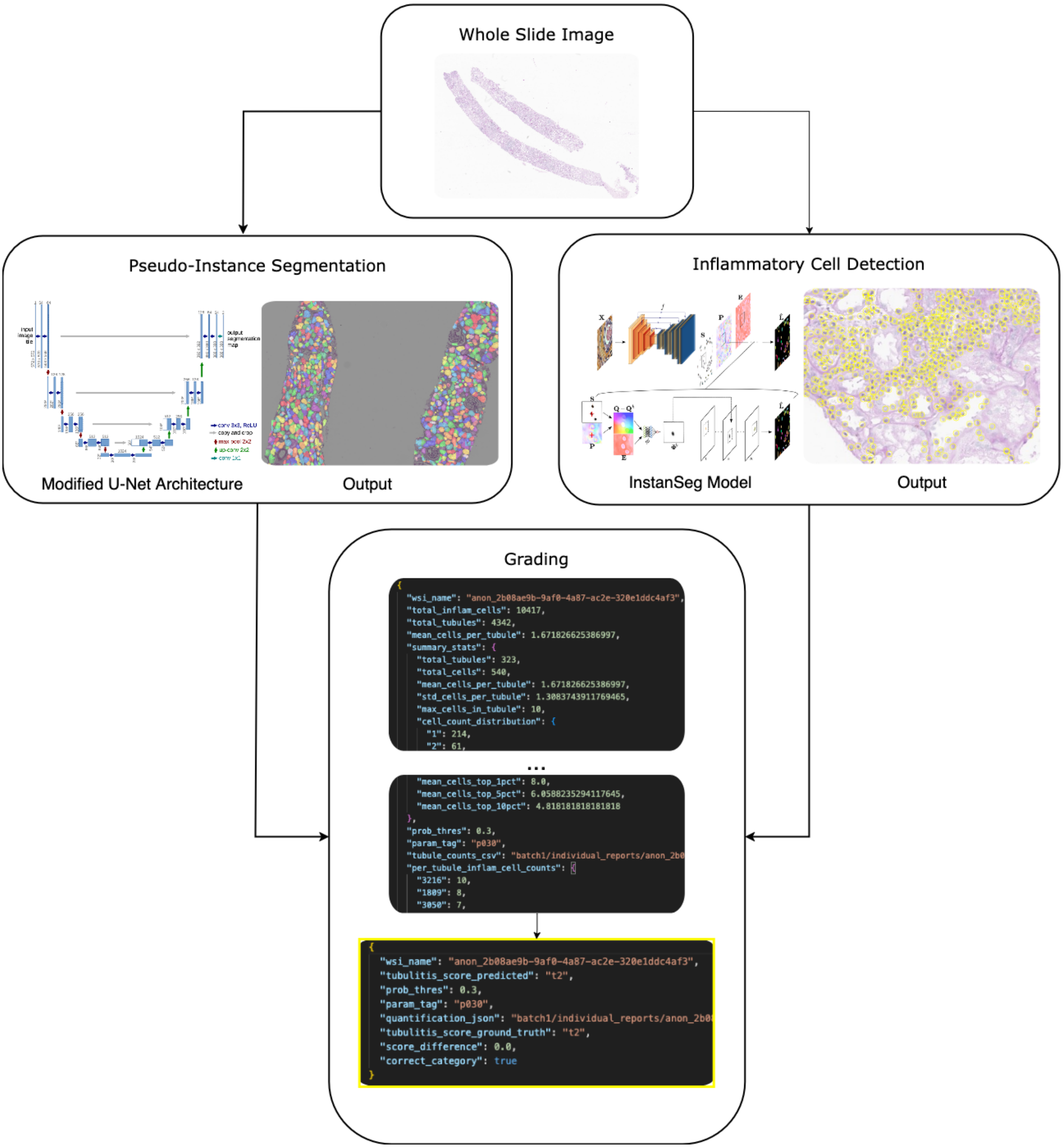
1. A modular, multi-stage pipeline that quantifies the presence of inflammatory cells in tubules in the WSI. This consists of three stages: (1) pseudo-instance segmentation of tubules from the WSI, (2) detection of inflammatory cells, (3) counting of inflammatory cells inside tubules and subsequent calculation of T-score. These stages can be executed individually or all at once, in order for pathologists to be able to interpret and visualise results between stages.

2. An end-to-end attention-based model that directly predicts the T-score from a WSI. Its explainability comes from the ability to visualise the attention mechanism's heatmaps, highlighting which regions in the slide had the greatest contributions to the final prediction. We compare two different models for this approach, aiming to assess how different attention mechanisms capture histological patterns and how they contribute to the overall prediction.

This chapter will detail the architectures, experimentation, engineering challenges and implementation strategies taken in building KidneyGrader.

3.2 Modular Approach Pipeline

A WSI is passed into the pipeline, and is segmented into tubule (and optionally, glomerulus, artery and indeterminate vessel) instances in the pseudo-instance segmentation stage. This consists of creating semantic masks for the class regions via the segmentation model and subsequently splitting these masks into instances via an instance labelling procedure detailed in the section on Stage 1 (see coloured instances in the figure). Stage 2 is independent of stage 1, and takes in the original WSI to perform detection of inflammatory cells (circled in blue in the figure) using a SOTA nuclei instance segmentation model, which is further described in the section on Stage 2. Stage 3 receives the outputs from stage 1 and 2, namely the tubule instance masks and inflammatory cell nuclei coordinates, and quantifies the extent of inflammation in each instance, and calculates a T-score based on the Banff specification [17].


Whole Slide Image
Pseudo-Instance Segmentation
Modified U-Net Architecture
Output
Inflammatory Cell Detection
InstanSeg Model
Output
Grading
...

```
{
  "wsi_name": "anon_2b08ae9b-9af0-4a87-ac2e-320e1ddc4af3",
  "total_inflam_cells": 10417,
  "total_tubules": 4342,
  "mean_cells_per_tubule": 1.671826625386997,
  "summary_stats": {
    "total_tubules": 323,
    "total_cells": 540,
    "mean_cells_per_tubule": 1.671826625386997,
    "std_cells_per_tubule": 1.3083743911769465,
    "max_cells_in_tubule": 10,
    "cell_count_distribution": {
      "1": 214,
      "2": 61,
```

```
    "mean_cells_top_1pct": 8.0,
    "mean_cells_top_5pct": 6.0588235294117645,
    "mean_cells_top_10pct": 4.818181818181818
  },
  "prob_thres": 0.3,
  "param_tag": "p030",
  "tubule_counts_csv": "batch1/individual_reports/anon_2b0
  "per_tubule_inflam_cell_counts": {
    "3216": 10,
    "1809": 8,
    "3050": 7,
```

```
{
  "wsi_name": "anon_2b08ae9b-9af0-4a87-ac2e-320e1ddc4af3",
  "tubulitis_score_predicted": "t2",
  "prob_thres": 0.3,
  "param_tag": "p030",
  "quantification_json": "batch1/individual_reports/anon_2b0
  "tubulitis_score_ground_truth": "t2",
  "score_difference": 0.0,
  "correct_category": true
}
```

Figure 3.1: The modular pipeline for T-score prediction.

3.3 Datasets and Preprocessing

3.3.1 Data Sources and Annotations

The training and testing of the segmentation model in stage 1 has made extensive use of ethically sourced, pseudonymised patient data, consisting of 444 scanned renal transplant biopsy slides from 501 patients [38], held at the Imperial College Tissue Bank. The slides were stained with PAS and H&E and scanned at a magnification of 40x, and were subsequently manually annotated by an expert cellular pathologist from the Department of Immunology and Inflammation at Imperial College London, via the QuPath software [39].

To test the overall performance of the modular pipeline and provide training and testing data for the end-to-end models, we also use a held-out curated dataest of n = 93 renal transplant biopsy WSIs, each labelled with 7 corresponding Banff lesion scores, including the T-score (t0-t3). These were also scanned at 40x magnification and with the PAS stain, with anonymisation.

3.3.2 Preprocessing

The gigapixel scale of WSIs make direct analysis of these images in stages 1 and 2 computationally impractical. It is therefore crucial to preprocess the data by splitting the image into manageable segments called 'patches', and filtering for relevant tissue regions for more efficient and accurate training and inference.

The `patch_extractor.py` script takes a WSI as an input using OpenSlide and extracts patches based on desired parameters such as the percentage of tissue content to consider it a valid patch, the patch size, and the amount of overlap between patches. The enhanced script was developed by a collaborating PhD candidate [40], as an extension of the extraction code we originally developed for this pipeline. This new script enables the configuration of parameters such as contiguous or random patch extraction, region bounding and hierarchical downsampling. The core functions of the patch extractor are as follows:

1. Patch size and overlap - the patch size determines the granularity of detail, and the complexity of features learnt by the model. From investigation, a patch size of 512 works well to encapsulate structures during segmentation such as tubules and glomeruli at a convenient size, allowing fast model convergence. The overlap parameter was set at 25%, allowing for shared context between patches to help reduce the chance of the model missing features on boundaries, as well as being a form of data augmentation.

2. Tissue detection - tissue regions are extracted by converting patches from RGB to HSV (Hue, Saturation and Value) and filtering the mask based on saturation and value (brightness) thresholds. A tissue percentage for the patch is then calculated based on the mask, and the patch is subsequently included or excluded from the patches.

3. Patch extraction modes - patches can be randomly sampled from the tissue regions or contiguously taken to ensure coverage. Due to the requirement of stitching back the patches in the segmentation stage in order to perform instance labelling on the objects, we need the patch structure to be kept intact and thus use the contiguous extraction mode for the purpose of this project.

3.3.3 Engineering Decisions and Alternatives

A key decision to make when designing the patch extractor was approach to take for tissue filtering. One such approach was grey-level thresholding, which was computationally the fastest, but it failed to account for factors such as stain intensity variability and lighting. In addition, stains like PAS may be incompatible with grey-level thresholding due to their high contrast colourings, which may cause lightly stained tissue to be mistaken for background. HSV was instead chosen, as it provided a balance between speed and accuracy. The hue channel enabled capturing of colour information, and saturation channels enabled the distinguishing of faintly and highly stained regions, making it more robust to variations in stain environments and background noise in slides.

3.3.4 Challenges

Despite the difficulty of obtaining and annotating WSIs, the efforts of our pathologists at Imperial College London have brought forth an abundance of annotations for training the segmentation model, resulting in impressive performance for identifying common structures like tubuli. However, the scarcity of structures like arteries and indeterminate vessels in biopsy WSIs have made it difficult to achieve notable performance for those structures. We may therefore need even more data or domain shift techniques on external sources of data to address the data imbalance and achieve sufficient performance to enable calculation of other important Banff scores such as Intimal Arteritis (Banff Lesion Score v [4]), which is a key indicator of acute rejection.

The end to end dataset with the 93 WSIs with Banff score labels, though sufficient to test the modular pipeline, is vastly low for training the end-to-end prediction models. This led to the cross validation folds having considerable inter-fold variation in performance and the use of heavy data augmentation and regularisation techniques to prevent overfitting.

3.3.5 Data Augmentation Strategies

In order to enhance generalisability of the segmentation and end-to-end models to factors such as variability of kidney tissue regions, and reduce overfitting to the training set, various augmentations were applied:

- Spatial augmentations - random 90° rotations, horizontal and vertical flips, and affine and elastic transformations, in order to improve model robustness to variation in tissue and structure orientation across samples.

- Colour augmentations specific to PAS and H&E stained tissue - allowing for simulation of stain variability between slides in histology.

- Artifact simulation - introducing Gaussian noise and blur to simulate random artifacts present in WSI scans, in order to make it more robust to degraded images and able to distinguish meaningful features.

- Contrast Limited Adaptive Histogram Equalization (CLAHE) [41] - used to increase the local contrast of images, in order to help highlight the more subtle tissue structures and boundaries.

- Normalisation - scales pixel values to a range to help the model have more stable gradient updates and have faster convergence, as well as improving generalisation.

3.4 Stage 1: Structure Segmentation

3.4.1 Problem Formulation

The foundational stage of the pipeline is to accurately segment histopathological structures from the WSI, in order to facilitate later quantification of metrics to arrive at Banff lesion scores. This includes structures such as tubuli, glomeruli, arteries and indeterminate vessels (incuding veins, peritubular capillaries and lymphatics)[28]. The subsequent identification of individual instances of these structures is required to derive structure-level metrics relating to inflammation and other properties. This stage must therefore generate the following:

1. A semantic mask $M_s \in \{0, \ldots, 4\}^{H \times W}$, which assigns each pixel to a label: 0 = background, 1 = tubuli, 2 = indeterminate vessel, 3 = artery and 4 = glomeruli.
2. An instance mask M_i which assigns pixels an additional ID, thus mapping all individual structures like tubules to unique integer labels.

3.4.2 Engineering Considerations

The limitations of our training data meant that we did not have instance labels to train a dedicated instance segmentation model for this dataset. Hence, we have chosen the approach of semantic segmentation followed by, to our knowledge, a novel instance labelling strategy to perform pseudo-instance segmentation.

In addition, the accuracy and precision requirements of such a segmentation model mean significant computational power will be required for a user to run inference. A prior project [28] had successfully developed a U-Net architecture that achieved strong performance for the segmentation of key histopathological structures from renal biopsy WSIs. However, in order to meet the strict latency and reproducibility requirements of a pipeline made for deployment, we explored alternative architectures to try to reduce inference time while simultaneously improving performance, in order to be able to facilitate more accurate downstream calculation of Banff lesion scores.

3.4.3 Experimentation During Architecture Selection

To achieve our goal of reducing inference time while improving semantic segmentation accuracy, we explored improvements on top of the previous best U-Net architecture for this dataset. As a result, we developed our own U-Net architecture that introduced architectural innovations such as an EfficientNet encoder and attention-modulated skip connections (further details in 3.4.4).

We conducted a comparative experiment between our new model and the previous best-performing U-Net model. The experimental setup involved running evaluation using a suite of semantic segmentation performance metrics, including distance metrics across all the structure classes. The full results are included in section 4.1.1. We found that our model reduced inference time by an order of magnitude and had greater segmentation performance for almost all structure classes. We thus proceed with this architecture for stage 1 of the pipeline.

3.4.4 Final Architecture

Our model architecture is shown in table 3.1. It is made of a U-Net backbone, and consists of the following:

Layer Group	**Layer Type**	**Shape/Details**	**Params**
Class Weights	Parameter	[5]	5
Backbone Conv Stem	Conv2d	[32, 3, 3, 3]	864
Backbone BN1	BatchNorm2d	[32]	32
Backbone Blocks (Grouped)	Conv2d	[96, 16, 1, 1]	1536
	BatchNorm2d	[96]	96
	SqueezeExcite	[8, 32, 1, 1]	256
Bottleneck Conv	ResidualConv	[512, 320, 3, 3]	1,474,560
Upsamplers	ConvTranspose2d	[512, 320, 2, 2]	655,360
	ConvTranspose2d	[320, 112, 2, 2]	143,360
Attention Blocks	Attention	[160, 320, 1, 1]	51,200
Decoders	ResidualDoubleConv	[320, 432, 3, 3]	1,244,160
Final Decoder	ResidualDoubleConv	[64, 24, 3, 3]	13,824
Total Trainable Parameters	-	-	11,376,375

Table 3.1: Summary of the segmentation model architecture, with key layer types and parameter counts.

- An EfficientNet-B0 feature extractor as the encoder, pretrained on ImageNet-1K and provides a five-level pyramid with channel depths [24, 40, 80, 112, 1280]. EfficentNet [42] provides SOTA accuracy and efficiency on datasets like ImageNet [43], with the B0 variant having better accuracy than ResNet50 with 1/5 the parameters.

- A bottleneck layer consisting of a residual double-conv block, which compresses features to 512 channels and triggers a dropout of p = 0.2 for regularisation.

- A decoder path with standard transpose-up convolutions to restore spatial resolution, as well as skip connections between encoder outputs and corresponding decoder stages for direct gradient flow.

- Attention gates [44], which modulate the skip connections by suppressing background noise. These are made of attention blocks, made of an attention mechanism that processes both the gating signal from the decoder and the skip connection from the encoder, and subsequently calculates an attention map for modulation.

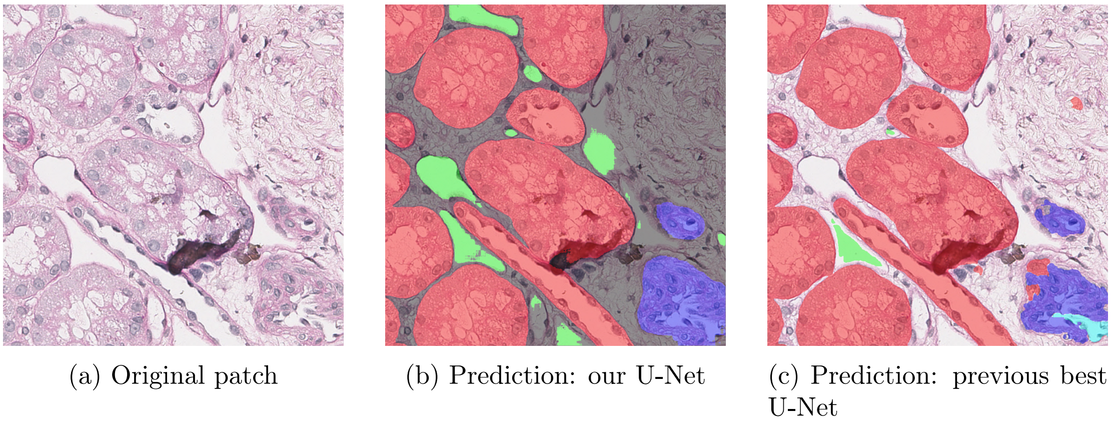
- In order to prevent vanishing gradients, the gradient flow is maintained by residual double-conv blocks.

- Deep supervision layers in the intermediate decoder stages implemented as 1x1 convolutions, generating auxiliary outputs which are then upsampled to the final resolution to participate in the loss. This promotes the network to learn meaningful features and class-specific representations in the earlier layers, improving overall gradient flow and segmentation performance.

- Dynamic class weights, consisting of a learnable parameter vector $w \in \mathbb{R}^5_{>0}$, which re-weights logits before soft-max. This gives the network the capacity to address data imbalance during traning instead of relying on a static heuristic.

3.4.5 Training Strategy

The model was trained on the dataset mentioned in 3.3.1, which was preprocessed with the methods in 3.3.2. The setup in table 3.2 shows the training parameters. We initially set up the model to train for 200 epochs, but found that the model converged at around



(a) Original patch

(b) Prediction: our U-Net

(c) Prediction: previous best U-Net

Figure 3.2: Comparison of original patch, segmentation mask overlay from our U-Net, and segmentation mask overlay from previous best U-Net. Red represents tubule tissue, dark blue/purple represents glomeruli, green represents vein/indeterminate vessel and cyan represents artery tissue.

60 epochs, and triggered early stopping. After comparison of results with the existing best U-Net model for this dataset, we found that our model had outperformed it (further details in 4.1.1).

3.4.6 Instance Labelling Strategies

We hypothesise that an intelligent implementation of a classical instance labelling algorithm applied on the semantic masks can provide sufficient instance segmentation quality for performing downstream quantification and grading. The latter part of stage 1 is therefore to take the semantic segmentation masks and extract individual instances, outputting a 16-bit TIFF file for each class, with each tubule having a unique label $L_i > 0$ and background $= 0$. To design our instance labelling strategy, we had to take special care to maintain accuracy of the instances while being computationally efficient to support gigapixel-size inputs. For example, during our experimentation, splitting the segmentation mask into patches like that of the preprocessing stage and running classical instance segmentation algorithms, namely connected component labelling and marker-controlled watershed on the patches led to reduced accuracy. This was likely due to the large proportion of structure instances spreading across patch borders, causing a single structure to be identified as multiple instances, losing spatial continuity between patches. We therefore devised three different approaches to compare:

1. Connected component labelling (CCL) on the slide level: a classical approach that relies purely on topological connectivity, and does not suffer from the over-segmentation issue at the patch level.

2. Watershed + Distance Transform on the slide level: a morphology-based method that uses shape information to split overlapping objects, and has the same advantage as above.

3. Patch-based Two-Pass Labelling (proposed) on the patch level: a novel algorithm to overcome the patch-level limitations of the classical methods, by performing instance labelling on overlapping patches with marker-controlled watershed, and then globally stitching back the patches and merging adjacent/touching tubules.

Various factors motivated this comparison, such as the variability of structures such as the shapes, sizes, overlapping and degradation of tubules. CCL, while fast and accurate

Parameter	**Value**
Batch size	32
Number of epochs	200
Initial learning rate	0.0003
Learning rate scheduler	ReduceLROnPlateau
Weight decay	1×10^{-4}
Optimizer	AdamW
Scheduler factor	0.5
Scheduler patience	5 epochs
Early stopping patience	10 epochs
Validation split	20%
Validation frequency	Every epoch
Patience for early stopping	10
Gradient accumulation	1 step
Number of workers	4
Device	GPU (CUDA 1, 2)
Loss function	Combined Loss (Dice + CE)
Alpha (Dice loss weight)	0.5
Beta (CE loss weight)	0.5
Gamma (Focal loss gamma)	2.0

Table 3.2: Training Hyperparameters for Segmentation Model

for non-touching instances, tend to merge touching instances and dramatically reduce in accuracy. Watershed, on the other hand, deals with this but introduces a new problem of over-segmenting and fragmenting a single object into multiple. The memory requirements of CCL linearly increase with image size, while pure watershed requires multiple full-resolution arrays in memory (both the distance map and label image) at once, which is very computationally expensive for WSIs.

The proposed patch-level two-pass approach seeks to address these challenges. The algorithm works by first performing local watershed-based instance labelling on 512*512 patches, and suppressing noisy peaks with a h-maxima transform and distance filtering. Crucially, it uses 50% overlapping patches to track continuity of structures across borders, which the Union-Find algorithm uses to construct an equivalence class of instances, and merge connected instances on the global WSI scale - thus mitigating the over-segmentation issue stated above. The second pass then reprocesses the patches and writes globally unique instance IDs to a memory-mapped output file. This approach therefore offers the accuracy of watershed, while maintaining the scalability of patch-level methods. We compare the performance of the different approaches in 4.1.1.

3.4.7 Evaluation Metrics

To rigorously benchmark performance of our semantic segmentation model against the existing best for the dataset, we employ a selection of pixel-based and boundary-based metrics, in order to quantify both its general and histopathology-specific performance, namely coherence between tissue structures. These include:

- Intersection over Union (IoU), measuring overlap between the predicted and expert annotation masks, based on the equation IoU $= \frac{|\hat{Y} \cap Y|}{|\hat{Y} \cup Y|}$.
- Precision, recall and F1 score, which measure pure pixel-level agreements per-class.

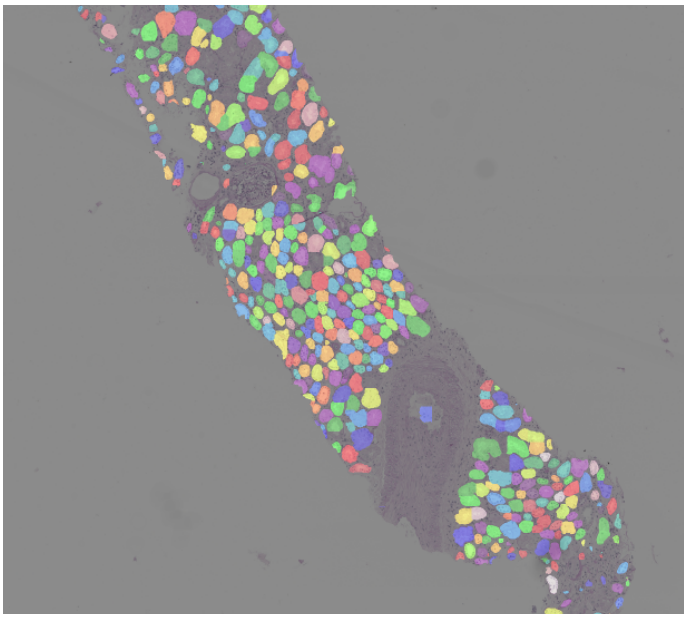


Figure 3.3: The instance labelling stage output with the two-pass method, visualised by the pipeline. The coloured shapes show instances of tubules identified on a section of a biopsy.

- Overall accuracy and Mean IoU: global pixel-level metrics that are more appropriate for segmentation tasks with data imbalance.
- Boundary IoU: a class-level metric assessing the ratio of correctly predicted boundary pixels to total number of boundary pixels.
- Hausdorff Distance (HD): a boundary metric that measures structural alignment, by finding the maximum difference between predicted and expert annotation mask contours, or outlines. $\text{HD}(A, B) = \max\left\{\sup_{a \in A} \inf_{b \in B} \|a - b\|, \sup_{b \in B} \inf_{a \in A} \|b - a\|\right\}$, where A and B represent sets of boundary points. The HD is rigorous but sensitive to outliers as it considers the worst-case segmentation error.
- Average Symmetric Surface Distance (ASSD): a more stable boundary deviation computation than HD. Finds average distance between boundary points in both directions. $\text{ASSD}(A, B) = \frac{1}{|A|+|B|}\left(\sum_{a \in A} \min_{b \in B} \|a - b\| + \sum_{b \in B} \min_{a \in A} \|b - a\|\right)$.

3.5 Stage 2: Inflammatory Cell Detection

3.5.1 Problem Formulation

Inflammation of histopathological structures is a key indicator of renal allograft rejection. We aim to be able to examine the extent of inflammation of tubuli in this project, as described in 2.1.4. This stage therefore requires a model to detect the counts of inflammatory cells across WSIs, in order to be able to later derive structure-level inflammation metrics in the quantification and grading stage. More specifically, the outputs of this stage should consist of a set of points $\mathcal{P} = \{(x_i, y_i, p_i)\}_{j=1}^{N}$, denoting the centroid positions of the detected inflammatory cells and the corresponding confidence levels of a true positive detection.

3.5.2 Baseline Detector: InstanSeg

As mentioned in 2.3.3, we employ InstanSeg as the baseline detector, using a frozen pre-supplied weights from the creator's MONKEY challenge submission. While we do not have ground truth labels of inflammatory cells for our dataset, we anticipate that transfer learning will be sufficient for our problem, and test this in (4.1.2). We leverage the bounding box-based inference script developed by a collaborating PhD candidate [40], and use their patch extractor to get bounding boxes that are regions of interest. We then perform inference with the following parameters:

Parameter	Value
Destination pixel size	0.5µm
Tile size	1024x1024
Seed threshold	0.1

Table 3.3: Parameters for InstanSeg-based inference. Each 1024x1024 tile is rescaled to 0.5µm per pixel to produce the nuclei instance labels. The seed threshold is used to calculate a seed entropy score $\mathcal{H} = -\frac{1}{n_c}\sum_{i=1}^{n_c} p_i \log p_i$, where p_i is the class probability of pixel i being a nucleus seed.

The labels that the model produces undergo extraction of centroids and a 128x128 RGB patch and binary mask channel are produced for each centroid. An ensemble of three EfficientNet-B0 classifiers are then used to get mean probabilities of a cell being inflammatory, as well as more granular probabilities of being a monocyte or leukocyte. Since the Banff tubulitis scores are agnostic of inflammatory cell type, we only require the combined probability of the cell being inflammatory. This classification is also supplemented with test-time augmentation consisting of vertical flips and 90° rotations to improve prediction robustness. The final confidence score based on the logits $\ell^{(i)}$ is $p = \text{softmax}\left(\frac{1}{3}\sum_{i=1}^{3} \ell^{(i)}\right)_{\text{inflam}}$.

3.5.3 Probability Thresholding

The inflammatory cell candidates produced by the detection stage each have a confidence probability p_i. In order to accept a candidate as an inflammatory cell, we performed filtration based on a minimum confidence threshold, or probability threshold. We hypothesised that this threshold was a critical parameter and would have direct impact on the quantification results, and that increasing the threshold would weaken correlation with final T-score (see 3.4 for example). To assess this impact, we conducted experimentation applying different probability thresholds ranging from $p = 0$ to $p = 0.5$. At probability thresholds above 0.3, the detections become too few, as we show in 4.1.2. We subsequently evaluate overall pipeline performance based on these thresholds, as shown in 4.1.3.

3.6 Stage 3: Grading

The third stage of the pipeline takes in the following inputs:

1. Tubule instance masks from stage 1, where each instance has a unique label $L_i > 0$, and background $= 0$.

2. Inflammatory cell detections from stage 2, with the millimetre-based coordinates of the detected inflammatory cells. These are converted to pixel space before counting cells in tubules.

3. A confidence probability threshold for cell detections.

It then generates per-tubule statistics such as cell counts and mean cell count per tubule, and finally outputs a grading report containing the predicted tubulitis grade.

3.6.1 Cell-to-tubule Assignment

Once we obtain the integer pixel coordinates of the inflammatory cell nuclei, we assign them to tubules by indexing the tubule instance mask at each location. The procedure is shown below.

```
Algorithm 1: Count cells per tubule
  Input: Cell coordinates $\mathcal{C}$, tubule instance mask $M$
  Output: Table of per-tubule cell counts and centroids
  if $\mathcal{C}$ is empty then
      return Empty table with columns [tubule_id, x, y, cell_count]
  Convert $\mathcal{C}$ to integer coordinates $(y_i, x_i)$;
  Filter coordinates to ensure $0 \leq y_i < \text{height}(M)$ and $0 \leq x_i < \text{width}(M)$;
  Retrieve tubule labels $t_i \leftarrow M[y_i, x_i]$;
  Filter out cells where $t_i \leq 0$;
  Count occurrences of each tubule ID $\rightarrow$ map $t_i \mapsto$ cell_count;
  Initialize empty result list $\mathcal{R}$;
  foreach region in regionprops($M$) do
      $t \leftarrow$ region label;
      if $t$ not in count map then
          continue
      $(y, x) \leftarrow$ centroid of region;
      Add row $\{t, x, y, \text{count}[t]\}$ to $\mathcal{R}$
  return table $\mathcal{R}$ sorted in descending order by cell_count
```

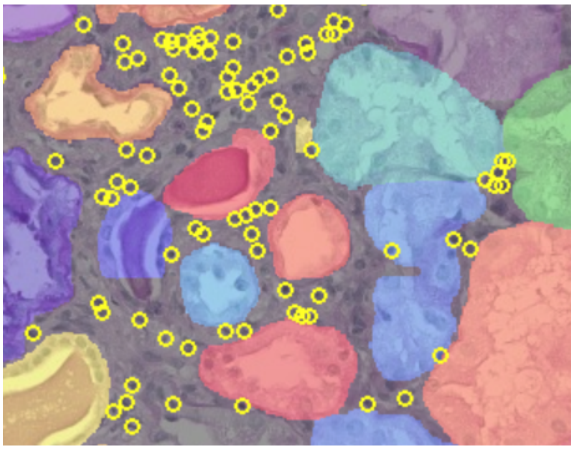


Figure 3.4: Detection of inflammatory cells (yellow circles) with tubules (multi-coloured shapes) during cell-to-tubule assignment, on a cell detection probability threshold of 0.3, visualised. Note how not all cells are being detected due to this confidence thresholding.

3.6.2 Tubulitis Scoring

The final component of the pipeline is the scoring, or grading, of the WSI based on the Banff 2019 Tubulitis score criteria [17], the defining marker for acute renal transplant rejection. The logic for this part is surprisingly simple, shown by the algorithm 2, which follows from 2.1.4.

Algorithm 2: Determine Tubulitis Score

```
Input: Table of per-tubule cell counts $\mathcal{T}$
Output: Tubulitis grade t0, t1, t2, or t3
Filter $\mathcal{T}$ to rows where cell_count $> 0 \rightarrow \mathcal{I}$;
if $\mathcal{I}$ is empty or $|\mathcal{I}| < 2$ then
    return t0
$M \leftarrow \max(\mathcal{I}[\texttt{cell\_count}])$;
if $M > 10$ then
    return t3
else
    if $M \geq 5$ then
        return t2
    else
        return t1
```

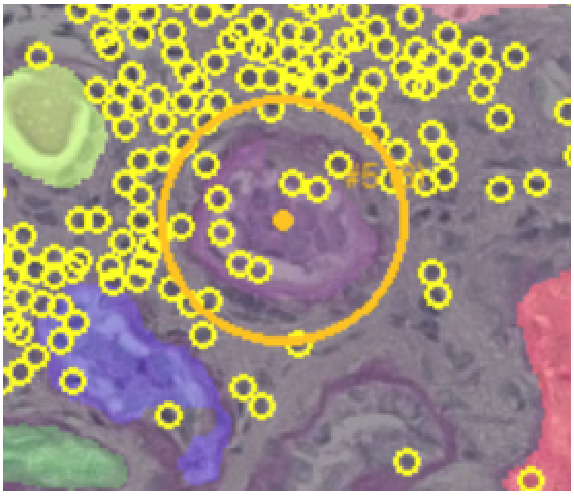


Figure 3.5: The 5th most highly inflamed tubule identified and visualised by the pipeline. Since only the highest inflamed tubule is taken into account for tubulitis scoring, this feature is only for interpretability purposes for the pathologist, and calculation of further statistics, such as top 5% of inflamed tubules as we discuss in the evaluation.

3.7 Alternative Approach: An End-to-End Model

Due to the recent success of multi-instance transformers in digital pathology, with methods such as CLAM [34] and TransMIL [45], we wanted to test whether global context of the biopsy WSI can capture tubulitis severity without explicit segmentation or quantification. This, albeit less granular than the former approach, can retain explainability due to the visualisability of the attention mechanism. It offers speed advantages, greater scaling potential to other Banff lesion scores, and maintains reproducibility for pathologists. This single stage, end-to-end model takes the WSI as the input and directly outputs a predicted T-score, improving on top of the work mentioned in 2.5.1.

We carried out extensive experimentation with this novel approach, through exploring different dimensions of this problem space: (1) problem type - trying both a regression model and classification model approach, (2) architectural choices between attention-based (CLAM) vs transformer-based (TransMIL) MIL frameworks, (3) held-out vs cross-validation settings, and (4) the effect of optimisation/generalisation strategies. We systematically evaluate these methods, comparing these strategies on the baseline CLAM model, and finally comparing against TransMIL, and results are discussed in 4.2.

3.7.1 Dataset Considerations and Preprocessing

For this experiment, we utilised the dataset mentioned in 3.8. We carried out multiple strategies to capture the global morphological patterns across the entire biopsy images in this direct prediction approach. Given the very limited size of the dataset, we employed various augmentation and regularisation strategies to mitigate overfitting and improve generalisability which we outline in the following sections.

Patch Extraction Strategy

Similar to the segmentation model from the modular pipeline, we carried out contiguous patch extraction with 25% overlap to maximise coverage of regions and reduce the risk of missing rare regions with important features. A 512x512 patch size was used, and approximately 1000-8000 patches per WSI were yielded. Patch-level augmentations were applied to maximise data utilisation, including noise injection ($\sigma = 0.2$) and feature dropout (p=0.1).

Foundation Model Feature Extraction

We utilise the UNI model (described in 2.4.4) for feature extraction, to represent each 512x512 patch as a 1024-dimensional embedding. Despite UNI being trained on H&E stains, it may well be highly advantageous to our model, helping identify semantically meaningful and domain-specific features such as histopathological structures and spatial relationships in pathology images. These features are stored in HDF5 format for each WSI.

3.7.2 Architecture

CLAM Framework

Given that our dataset consists of labels at a high level of abstraction i.e. slide-level lesion scores, we required a weakly supervised paradigm for this problem. However, a fundamental challenge of learning from weakly-supervised labels is the fact that few instances in the bag, or in our case, patches in the WSI, are informative, due to the fact that inflammatory tubules may be present in only a few focal regions. We chose Clustering-constrained Attention Multiple Instance Learning (CLAM) as our baseline architecture for this experiment as it addresses this challenge, through using attention mechanisms to identify diagnostically relevant regions. The architecture is composed of:

- Patch-level feature extraction via UNI embeddings, processed through patch-level MLPs.

- Gated attention mechanisms with top k attention for the instance-level predictions, to increase generalisation and computational efficiency. We outline our experimentation with the hyperparameter k in 3.7.4.

- Bag level prediction, through the aggregation of features to the T-scores by the final MLP layers.

3.7.3 Model Formulations: Regression vs Classification

We developed and evaluated models for two different problem formulations, each with their own advantages:

- Regression: a continuous score from 0.0-3.0 using MSE loss. The continuity preserves the natural ordering and the granularity enables intermediate predictions to reflect pathologist uncertainty.

- Ordinal classification: a 4-class classification problem (t0, t1, t2, t3) that preserves order to represent inherent ordering of Banff scores while having useful class probabilities. The loss functions are ordinal and contain cumulative logit monitoring to respect the score ordering.

3.7.4 Experimental Procedure

Held-Out vs Cross-Validation Setup

To investigate the tradeoffs between maximising dataset usage and maximising performance on unseen data, we systematically compared a held-out and a cross-validation test set setup under the same conditions (training setup outlined below, and k=32 for top-k attention). The held-out setup involved holding out 18 slides ($\sim$ 20%) as a completely unseen test set, and performing 5-fold stratified cross-validation internally for training and validation on the remaining 75 slides, with a 60/15 train/val slides per fold. This was to make evaluation completely unbiased and objectively evaluate the model's overall performance without any data leakage during training. On the other hand, the cross-validation (non-held-out) setup involved doing 5-fold stratified cross-validation on all 93 slides, each fold offering a 65/14/14 slide (75%/15%/15%) train/val/test split. The goal of the cross-validation setup was to maximise data usage due to the small size of the dataset. Importantly, each fold had no exposure to the test set for that fold, but eventually, all slides were used as test data across the different folds. In both approaches, we stratified the datasets based on the T-scores to ensure a balanced split of scores in each fold and had the same random seeds for the model for reproducibility.

Training Configuration

We employed aggressive regularisation techniques to increase model generalisability, as shown below.

Table 3.4: Regularisation and optimisation parameters during CLAM training

Parameter	Value
Dropout	50%
L2 weight decay	0.01
Label smoothing	10%
Gradient Clipping	1.0 norm
Epochs	300
Cosine-annealing learning rate	0.00005
Early Stopping	Patience: 30

Loss functions

Different loss functions were used for the instance-level and bag level. The bag-level learning objective simply uses Mean Squared Error (MSE), to find the difference between the prediction and expert labelled score. At the instance level, we convert the regression problem into a classification subproblem, creating binary pseudo-labels based on bag labels. We therefore employed SVM-style hinge loss for the instance-level, as it encourages discriminative separation between relevant and irrelevant features, and works well with weakly supervised settings.

Top-k Attention

We also experimented with the value of k in the top-k attention sampling mechanism, which takes the top k patches for attention when performing inference at the instance-level (while at the bag-level, all patches are taken into account), in order to investigate its effects on model generalisation and computational efficiency. This included using:

- Fixed k values of 32 and 64, which we chose based on research such as the original paper [46].
- Adaptive k values, proportional to the number of patches in the WSI. More specifically, we use $k = 0.01 \cdot |P|$, and $0.02 \cdot |P|$, clamped to $16 \leq k \leq 64$, and $|P|$ ranges from 1000-8000 patches per WSI.

3.7.5 Comparative Architecture: TransMIL

We compared the CLAM-based approach with TransMIL, a transformer-based method that captures spatial relationships between instances, in order to investigate whether accounting for spatial relationships and patterns of tubule-infiltrating lymphocytes is diagnostically meaningful and improves grading performance. TransMIL is composed of a multi-layer transformer architecture with bi-directional attention blocks, and, similar to CLAM, uses instance-level pseudo-labelling based on bag-level supervision.

Training

Based on our learnings from the CLAM-based experimentation, we used an identical dataset split to the held-out test setting used for CLAM, with 18 held-out test slides and a stratified 5-fold cross-validation on the remaining 75 slides, and used the same UNI feature extractor to extract features.

Novel Semantic Guidance Approach

Due to the sole dependence of the tubulitis score on the level of tubular inflammation, we decided to take a novel approach of first segmenting the input WSI to get the tubule semantic mask (with stage 1 of the modular pipeline) and filtering based on tubule-rich areas. We have set the filtration to require at least 30% of the patch to be tubule tissue to be considered for feature extraction. This enables the model to learn more diagnostically relevant features and reduces the risk of attending to irrelevant tissue.

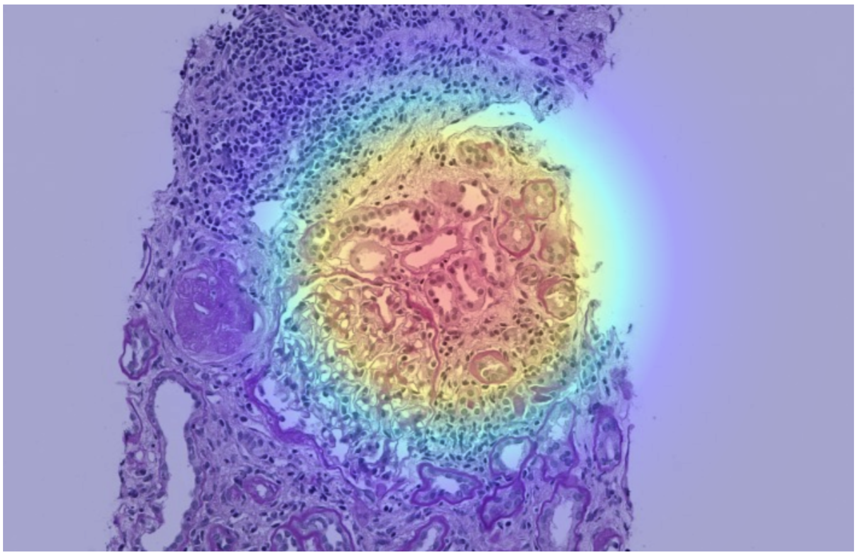


Figure 3.6: An attention heatmap zoomed in, produced by the TransMIL-based model, showing it attends to areas where there are tubules with high inflammation.

3.8 Evaluation Metrics

To assess the grading performance of the modular pipeline and end-to-end models, we compare their performance on the held-out curated dataset mentioned in 3.3.1 and compute a selection of agreement and error metrics, including our own custom metric that is clinically relevant to determine performance against the gold standard.

- **Within-1-grade accuracy** - a custom metric, which may be more meaningful than exact accuracy in clinical practice as it accommodates for the fact that another expert rater may have given our data slightly different labels. This metric is defined to be the proportion of predictions that fall within 1 tubulitis grade of the expert label.
- **Exact Accuracy** (for regression, the predictions are rounded to nearest integer to calculate accuracy) - to measure strict agreement between predicted and labelled scores.
- **Mean Absolute Error** - finding the average absolute difference between our predictions and the expert labels
- **Correlation coefficients (Pearson, Spearman)** - to assess linear and non-linear agreement monotonicity with the expert labels.
- **Quadratic weighted kappa** κ_m, as mentioned in section 2.1.3, to compare reliability against the gold standard's κ_m for tubulitis scoring.

3.9 Software Engineering and Technical Specification

3.9.1 Software

Modular Codebase and Configurations

The codebase is implemented entirely in Python, and has been designed with the goal of supporting clinical deployment and research experimentation. We have thus put focus on the codebase's modularity, efficiency, reproducibility and extensibility through separate modules for different stages, configuration files for parameter tuning, efficiency optimisations, and an enhanced visual interface for ease of use and explainability.

Optimisation

We have implemented optimisations such as:

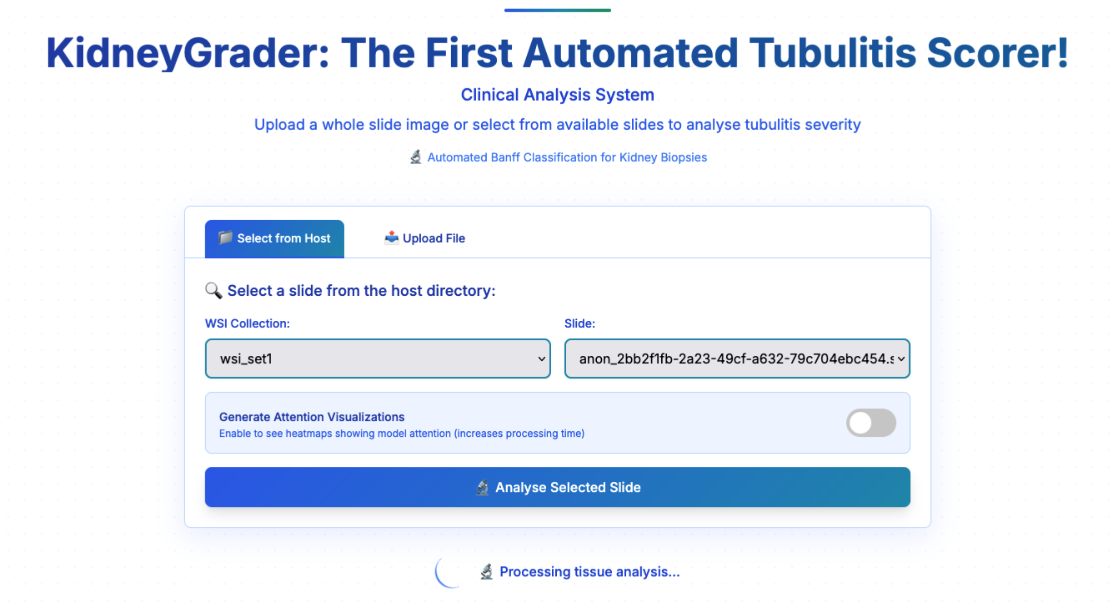
- Caching - results from individual stages do not need to be recomputed if the parameters remain unchanged.
- Storage efficiency - instance masks have been compressed from 20GB per class to under 100MB without loss of accuracy, through using Zstandard [47] compression.
- Memory mapping of masks - large GB-scale output masks are stored on a disk-backed memmory map to prevent running out of RAM.
- Tiling - faster access and visualisation and lower memory burden than full slide image processing.
- GPU resource management - periodically freeing unused GPU memory to avoid memory leaks.
- Vectorised numpy logic - array-based numpy operations applied where possible, to avoid time-consuming Python `for` loops.

3.9.2 Hardware

The segmentation model of our modular pipeline and the end-to-end models have all been trained on a NVIDIA RTX A6000 48GB GPU, with a total training duration of approximately 10 hours. Inference has been run on both NVIDIA A5000 24GB GPUs and NVIDIA A6000 GPUs.

3.9.3 Web App

We deployed KidneyGrader as a user-friendly web application (shown in 3.7) for experimental use by pathologists. A demonstration with our collaborating pathologist is scheduled, in which we aim to gather feedback on its clinical usability and integration potential. The tool currently uses the TransMIL model, but can allow for easy substitution as needed. We may also consider adding KidneyGrader to KidneyCaliper [48] as an experimental feature.



KidneyGrader: The First Automated Tubulitis Scorer!

Clinical Analysis System

Upload a whole slide image or select from available slides to analyse tubulitis severity

Automated Banff Classification for Kidney Biopsies

Select a slide from the host directory:

WSI Collection: wsi_set1

Slide: anon_2bb2f1fb-2a23-49cf-a632-79c704ebc454.s

Generate Attention Visualizations

Enable to see heatmaps showing model attention (increases processing time)

Processing tissue analysis...

Figure 3.7: KidneyGrader Interface

Chapter 4

Evaluation

To test our hypotheses mentioned in the introduction (1.1), we carry out in-depth experimental evaluation of both the modular pipeline approach and the end-to-end model approach. We compare their performance based on the gold standard expert labels, based on the metrics detailed in 3.8.

For the modular pipeline, we benchmark the segmentation performance of our new attention-gated U-Net against the existing U-Net in terms of the metrics in 3.4.7, to reach a decision of which model would be the best for clinical application. We then compare qualitatively, the different instance labelling algorithms we tried, through visual inspection of the instance segmentation quality.

We carry out similar visual inspection of the detection model at various probability thresholds to understand its performance and contribution to downstream grading. We then compare the performance of the overall modular pipeline based on different probability thresholds and propose new tubulitis quantification statistics for automated Banff scoring, and subsequently assess its correlation with the expert labelled T-scores. We test the performance on the 93 PAS-stained slide dataset mentioned in 3.3.1, which is completely held-out from this pipeline.

For the end-to-end models, we begin by determining whether a classification or a regression model would yield better performance and clinical relevance. We compare a CLAM-based regressor and a CLAM-based classifier model on a small 18-slide held-out test set from the 93 slide dataset. We proceed with the CLAM-based regressor as the baseline and then investigate the effect of the top-k attention sampling optimisation on performance of attention-based models. We then apply our learnings from the prior experimentation to a TransMIL-based regressor model, and benchmark performance against the best CLAM-based regressor.

Finally, we critically compare the modular and end-to-end approach, evaluating trade-offs in performance, interpretability and clinical applicability to conclude which approach is the most promising for real-world clinical adoption.

Note on ground truth: For evaluation, we define terms such as accuracy, MAE and kappa to be based on agreement with expert renal pathologist annotations, as manual annotations are (currently) the gold standard in the field. We therefore use the terms ground truth and expert labels interchangeably.

4.1 Modular Pipeline Performance

4.1.1 Structure Segmentation

Semantic Segmentation Performance: Prev. U-Net vs Attention-gated U-Net

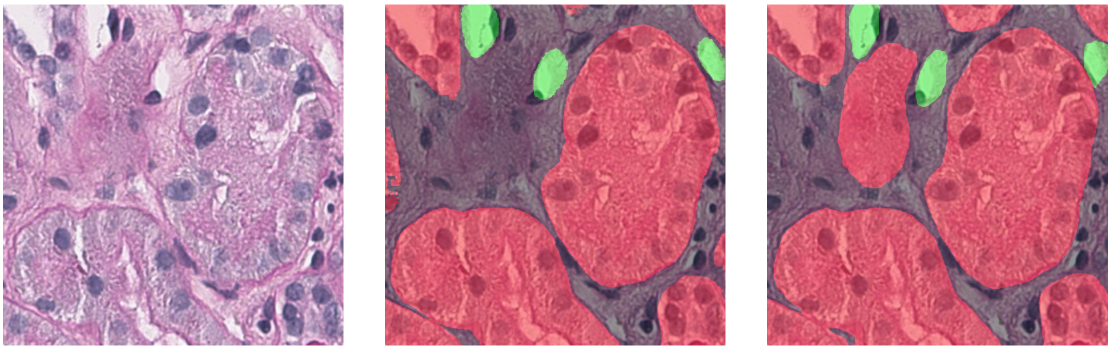
To assess whether our attention-gated semantic segmentation model gave performance enhancements over the existing best model for this dataset, we benchmarked performance using a held-out test set of annotated renal biopsy slides split into 512*512 patches and stored as a h5 file. We demonstrate the results in table 4.1.

Class	**Metric**	**Previous U-Net (Baseline)**	**Our Model**
Background	IoU	0.8091	0.8304
	Precision	0.9049	0.9008
	Recall	0.8843	0.9139
	F1 Score	0.8945	0.9073
	Boundary IoU	0.2829	0.3371
	Hausdorff Distance	175.55	133.45
	ASSD	14.48	11.21
Tubuli	IoU	0.8751	0.8852
	Precision	0.9242	0.9524
	Recall	0.9428	0.9262
	F1 Score	0.9334	0.9391
	Boundary IoU	0.3353	0.3706
	Hausdorff Distance	150.82	120.93
	ASSD	15.89	12.47
Glomeruli	IoU	0.6952	0.7639
	Precision	0.7068	0.7917
	Recall	0.9769	0.9561
	F1 Score	0.8202	0.8661
	Boundary IoU	0.0342	0.0697
	Hausdorff Distance (HD) (pixels)	628.57	280.03
	ASSD (pixels)	193.44	74.20
Overall Accuracy		0.8967	0.9101
Mean IoU		0.5146	0.5913

Table 4.1: Previous Best U-Net Model [28] vs. Our U-Net Segmentation Performance

Key Observations and Discussion Our attention-gated model has outperformed the baseline across nearly all classes and metrics, demonstrating an overall improvement of semantic segmentation performance for this dataset. In particular, the glomeruli segmentation has had the most pronounced improvement, with an increase of IoU from 0.6952 to 0.7639, a drop of HD by over 55% ($\sim$ 628 to $\sim$ 280 pixels) and a similar reduction in ASSD, indicating significantly better structural correspondence. The indeterminate vessels class (shown in A.1) also benefit from a significant improvement. The tubule and background classes had much more marginal performance enhancements, with an increase in IoU, precision and F1 scores by 2-3 % at most, but noticable improvements in HD and ASSD (20 - 40% reduction). Interestingly, the recall has reduced by 2 - 3 % for some classes including tubuli. This may be due to the new model segmenting more conservatively due to attention-based skip connections that suppress noisy evaluation, making the model more selective, thus reducing false positives but increasing false negatives.

Conclusion and Clinical Relevance Despite the modest improvements for our desired class (i.e. tubules), the effects on the overall pipeline performance will likely compound due to the dependence of the instance labelling algorithm on the semantic segmentation quality, and thus may have a considerable effect on the overall tubulitis scoring accuracy. Importantly, the inference time with this model is also an order-of-magnitude faster than the previous ($<$5 minutes vs $\sim$ 60 minutes), due to the highly lightweight EfficientNet encoder. This improvement of both accuracy and computational efficiency is essential for our goal of clinical suitability, and we therefore proceed with our attention-gated U-Net as the default segmentation model for the pipeline.



(a) Original patch

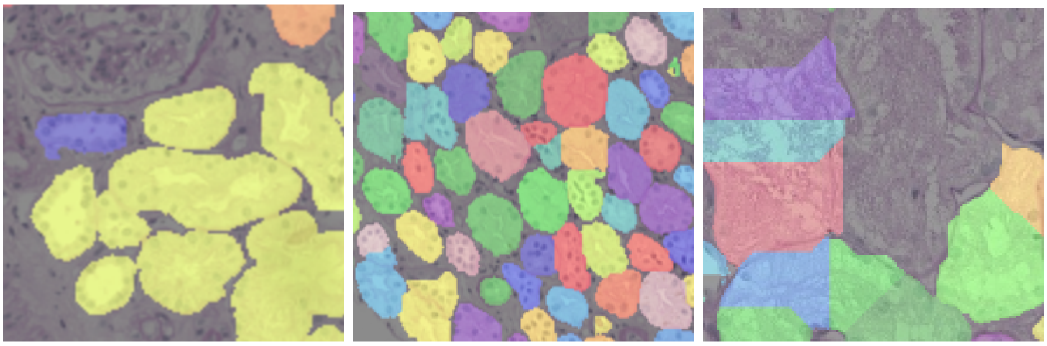
(b) Prediction overlay (Our U-Net)

(c) Expert annotation overlay

Figure 4.1: Comparison of patch, prediction from our U-Net, and expert annotation overlay. Red represents tubule tissue, and green represents vein/indeterminate vessel tissue.

Instance Labelling: Classical vs Patch-Based Algorithms

The evaluation method for this experiment was limited to qualitative analysis, due to the absence of ground truth instance labels for our dataset. We compare performance through visual inspection of the tubule instances generated by the three methods.



(a) CCL Method

(b) Watershed (if memory permits) or 2-pass method

(c) Watershed or 2-pass, with larger heterogeneous tubules

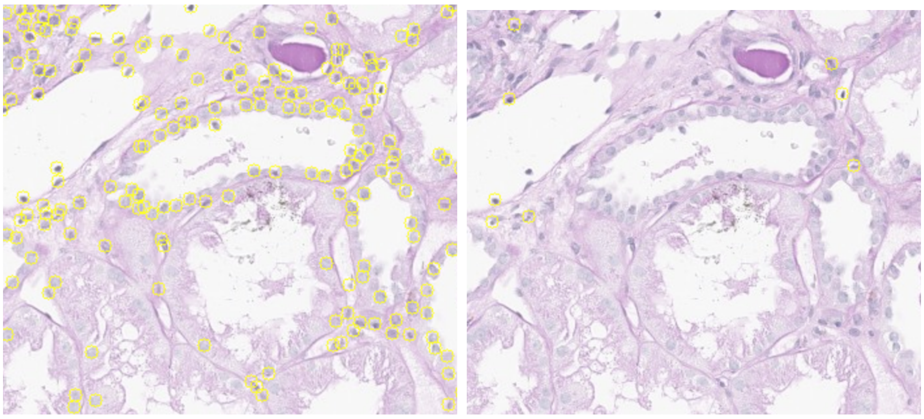
Figure 4.2: Comparison of instance segmentation methods

Despite CCL being the fastest and most memory-efficient to run, it merged many instances together, which was expected, due to the dense presence and frequent overlapping of tubuli across the WSI. Watershed achieved better segmentation despite often splitting single instances, as the merging of overlapping structures via CCL was more frequent and prevalent than the occurrence of single structures being split into multiple via watershed.

One observation was that a footprint radius that was too low would produce a barcode-like instance mask for structures, while an overly high footprint radius under-segmented the structures similar to CCL. This required some tuning to about 20 pixels to eventually achieve a balance between under and over-segmentation, but this was still sensitive to large, heterogenous tubules (see 4.2). Also, watershed failed to execute on some WSIs due to the sheer memory requirements and the variability in WSI size. The patch-level two-pass approach solved this memory issue, while providing similar segmentation quality to the watershed algorithm - processing slides several gigapixels in size with under 2GB peak memory usage. Therefore, we considered the patch-based algorithm to be the best approach to proceed with for downstream quantification and grading.

4.1.2 Inflammatory Cell Detection

Due to the lack of expert annotations of cells on our dataset, we will instead qualitatively evaluate the InstanSeg cell detection model at the different minimum confidence probability thresholds. We compare two thresholds $p = 0.0$ and $p = 0.3$ on the same section of a WSI.



(a) $p = 0.0$

(b) $p = 0.3$

Figure 4.3: Comparison of detections at minimum confidence thresholds $p = 0.0$ and $p = 0.3$

Despite the domain shift, it appears that the transfer learning from the MONKEY data has provided adequate generalisation on our data, but with low confidence probabilities. The visualisations support this, as the number of detections greatly decrease with increased minimum confidence threshold p. This should most likely have an effect on the grading performance of the pipeline, which we will test in the next section. Interestingly, at $p = 0$, there are still cells that are undetected, and simultaneously areas that are falsely detected as cells, such as at boundaries between structures. As we increase p, these false positives decrease but we also lose true positives. This trade-off is particularly concerning for WSIs with weaker cases of tubulitis (i.e. T < 2) as there is a greater chance that the loss of true positives will push down the score. We therefore evaluate the effects of this downstream on the grading to try to find a threshold with an acceptable trade-off.

4.1.3 Grading

The T-score prediction performance of the modular pipeline on the dataset in 3.3.1, based on the confidence thresholds $p = 0.0$, $p = 0.3$, $p = 0.5$ and $p = 0.7$, are presented in table 4.2.

Table 4.2: Modular pipeline performance at different confidence probability (p) thresholds. The best values in each row are **bolded**.

Metric	$p =$ **0.0**	$p =$ **0.3**	$p =$ **0.5**	$p =$ **0.7**
Exact Accuracy	29.55%	**30.23%**	28.24%	28.57%
Within-1-Grade Accuracy	55.68%	74.42%	80.00%	**82.14%**
Mean Absolute Error (MAE)	1.3523	0.9884	0.9529	**0.9167**
Pearson r (p-val)	0.0841 (0.4362)	**0.3006 (<0.005)**	0.2984 (<0.01)	0.3048 (0.0048)
Spearman ρ (p-val)	0.0900 (0.4044)	0.2948 (<0.01)	0.3034 (<0.01)	**0.3175** (0.0033)
Quadratic Weighted Kappa κ_w	0.0176	0.2384	0.2584	**0.2920**

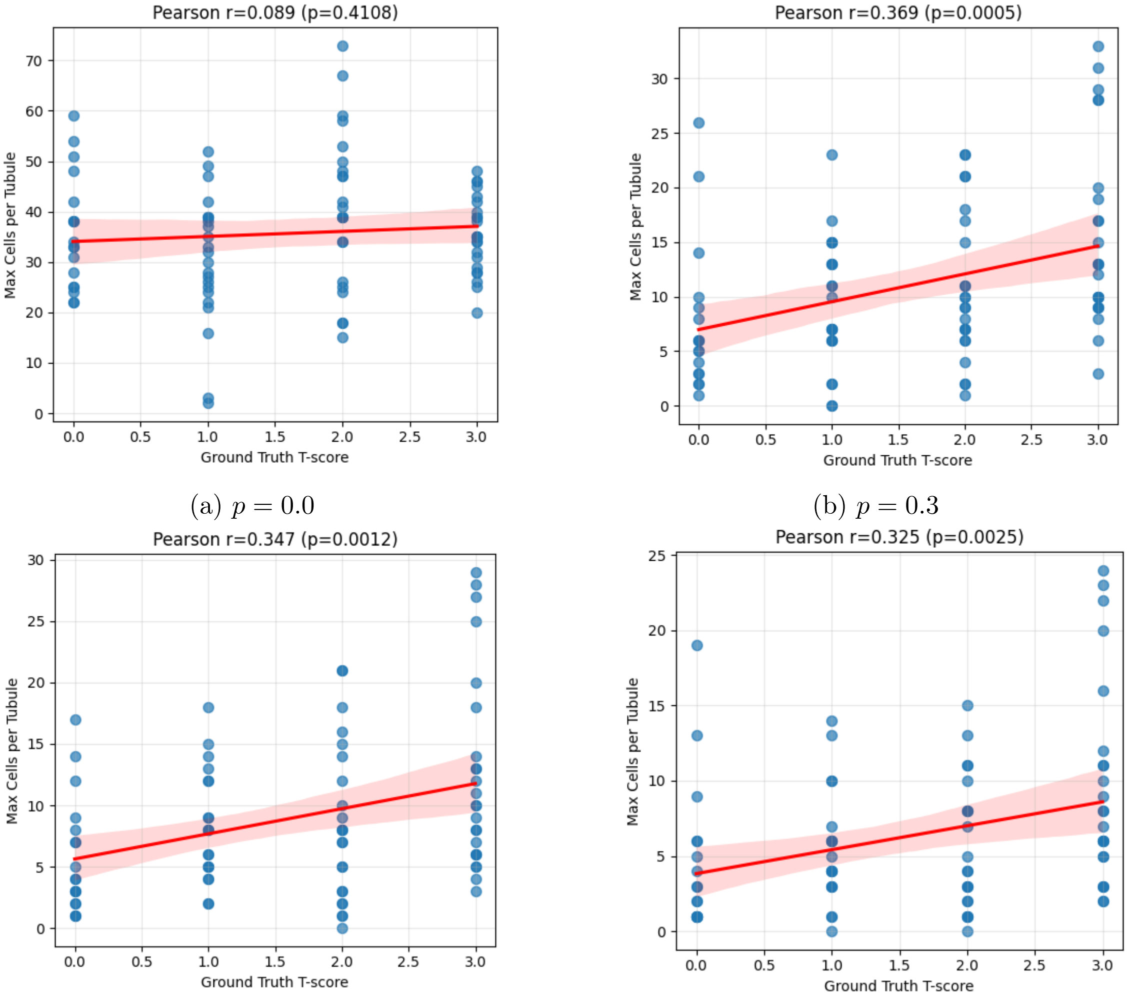
Key findings The results are revealing of the limitations of the detection model on our dataset and the use of classical instance labelling techniques for segmenting tubuli, with moderate within-1-grade accuracy but only weak positive correlation between the predictions and expert labels. Contrary to our expectations, there has been a general trend of improved performance as the minimum confidence threshold increases. While the exact accuracy has remained fairly stagnant, the within-1-grade accuracy has had a large improvement from $p = 0$ to $p = 0.3$, while a more marginal increase from 0.5 to 0.7. The MAE, Pearson and Spearman correlations, and quadratic weighted kappa behaved similarly, with a large improvement from 0 to 0.3, and plateauing from 0.5 to 0.7.

Discussion These results disprove our hypothesis in 3.5.3 that increasing the confidence threshold will reduce performance. It suggests that the reduction in false positive detections has been more significant than the loss of true detections. This may be due to the fact that, since the highest inflamed cells are the sole contributors to the resulting prediction, the confidence probability of the cells detected for them are already high and do not get removed by these low thresholds. This also suggests that the fundamental accuracy bottleneck is false positive detections, and that a single tubule with enough false positive detections (e.g. due to colour variations or dark boundaries) may be enough to change the final score.

Conclusion and Clinical Relevance The modular pipeline results have achieved similar levels of inter-reliability to the gold standard, with κ_w for $p = 0.7$ being 0.29, at the upper end of the cited gold standard range of <0.3, and within-1-grade accuracy of 82.1% for $p = 0.7$. It is, however, important to note that, from the perspective of interpretability and clinical relevance, an increasingly conservative detection model should be approached with increasing caution. As the visualisation in 4.3 shows, there is a concerning reduction in the number of inflammatory cells and this limits the amount of visual evidence provided to a clinician for verifying the model's decision and worsened by the issue of false positives mentioned above. Nevertheless, the pipeline has successfully validated our hypothesis that we can produce a pipeline for automated Banff tubulitis grading that achieves expert-level inter-rater reliability.

4.1.4 Alternative Quantification of Tubulitis: Top Percentile Averaging

As per the gold standard of Banff tubulitis scoring, our modular pipeline finds the tubule with the highest inflammatory cell count to derive the score (figure 4.4). However, based on observation of the detection stage outputs, it suggests we require more robust and stable metrics to account for false and missed detections, and it perhaps may be a good idea to aggregate the inflammatory burden of multiple tubules to arrive at a score. We therefore propose top-percentile average metrics, consisting of the average inflammatory cell count of the top 1%, 5% and 10% of inflamed tubuli respectively. We assess the correlation of



(a) $p = 0.0$

(b) $p = 0.3$

(c) $p = 0.5$

(d) $p = 0.7$

Figure 4.4: Correlation between maximum tubular cell count and expert labels at different p thresholds.

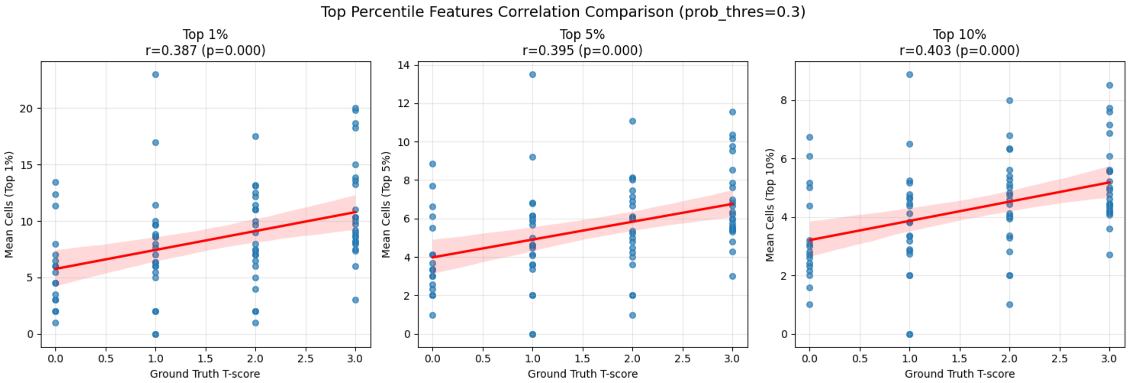


Figure 4.5: Correlations of average inflammatory cell count for top N% of inflamed tubules vs expert labels, for $p = 0.3$.

these metrics with the expert labels on the same dataset, with the probability threshold of 0.3, shown in figure 4.5.

Conclusion We can see that the top percentile metrics indeed had a greater correlation with the expert labels than the volatile the max cells per tubule metric, with the top 10% metric having a 9% higher correlation than that of the max cells for threshold $p = 0.3$. The correlation also improves as a larger percentage of the tubule population is considered (r = 0.403 for top 10%, compared to r = 0.387 for top 1%). This validates our assumption that a more stable quantification technique would be suitable for tubulitis scoring. It also raises an important discussion (see section 5.1), regarding the potential for introducing new data-driven metrics for automated Banff classification.

4.2 End-to-End Model Performance

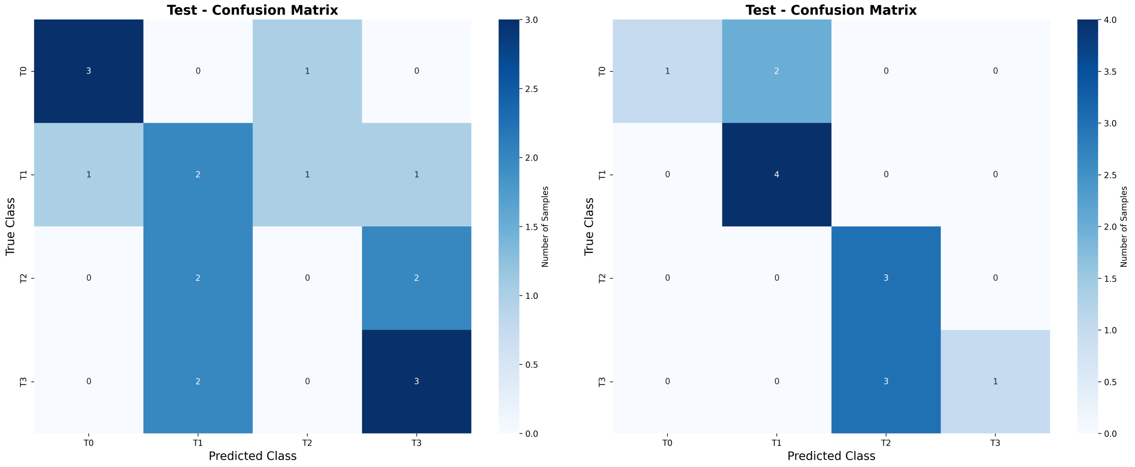
We evaluated the T-score prediction performance of our end-to-end models, comparing it to the expert-labelled T-scores of WSIs, but also assessing the explainability of the end-to-end approach. The visualisations and interpretability analysis can be found in 4.2.6.

4.2.1 Regression vs Ordinal Classification

Table 4.3: Comparison of CLAM-based regression and classification models for Banff tubulitis scoring performance on the held-out test set (18 WSIs). The regressor model's predictions were rounded to calculate exact accuracy, and values for each metric are **bolded**.

Metric	**Regressor**	**Classifier**
Exact Accuracy (%)	38.9	**44.4**
Within-1-Grade Accuracy (±1) (%)	**100.0**	77.8
Mean Absolute Error (1.0 = 1-grade difference)	**0.6567**	0.7778
Quadratic Weighted Kappa κ_w	**0.6621**	0.5308

Discussion Based on this small test set, the models both achieve similar levels of exact accuracy, with the classifier correctly predicting the Banff T-score for 44.4% of cases, compared to the regressor's 38.9%. However, despite classifier's exact accuracy being slightly higher, the regressor's within-1-grade accuracy is considerably higher than the classifier,



(a) CLAM-based Classifier

(b) CLAM-based Regressor

Figure 4.6: Confusion matrices for the CLAM regressor and classifier on the held-out 18-slide test set. The numbers in the squares indicate the number of predictions, with darker squares indicating more predictions. The diagonal from top left to bottom right represents the squares of correct predictions.

giving predictions ± 1 grade of the expert label, 100% of the time. The confusion matrices support this, showing how the error pattern from the classifier is more distributed, while the regressor has concentrated predictions closer to the expert labels and misclassifications that are confined to directly adjacent grades.

Clinical relevance The quadratic weighted kappa is notable for both models, far exceeding the gold standard of <0.3, which shows promise for clinical adoption. The regressor model, however, is likely more clinically suitable due to its higher overall agreement and stability relative to the expert labels. In addition, after discussion with our expert renal pathologist at the Department of Immunology and Inflammation at Imperial College London, they mentioned that a continuous scoring mechanism like regressor's outputs would be, in fact, a *more helpful* metric, as it enhances the granularity of the interpretation. This could potentially aid clinical decision making due to the wider spectrum of expressiveness of tubular inflammation. We therefore proceed with the regressor model for further comparisons in the end-to-end approach.

4.2.2 Held-Out vs Cross-Validation Settings

The performance of the CLAM regressor under the held-out and cross-validation test setups is compared in table 4.4.

Discussion The cross-validation setup performed consistently better than the held-out setup across all error metrics and most of the agreement metrics, with mean MAE, Pearson r value and exact accuracy of 0.51 ± 0.20, 0.85 ± 0.15 and 58.6% respectively, compared to 0.592, 0.753 and 38.9% with the held-out setup. This is expected with cross-validation, due to the possible data leakage and overfitting to the dataset as a whole, meaning that the reported performance may be optimistically biased. However, we can draw some interesting insights from the data, as follows:

- Ensemble performance exceeds individual performance in held-out setup. This is due to the fact that the ensemble cancels out individual errors and smooths noisy predictions through averaging predictions, thus reducing the variance and giving improved overall metrics.

- More stable performance in the held-out setup. The likely reason is that the held-out test set is the same for all models, while the cross-validation test set varies per fold and contains cases of varying difficulty, resulting in the fluctuations of inter-fold performance.

- The held-out setup outperforms cross-validation on the within-1-grade accuracy metric on the ensemble level and on almost all individual folds. This could be for many reasons, such as the held-out test set containing cases that were easier to predict, or more representative of the training data. Another potential reason could be the better stability of the held-out setup (as mentioned above), resulting in less outliers and more conservative predictions.

Conclusion Overall, the results clearly demonstrate the tradeoff between maximising use of data and minimising evaluation bias. However, despite the superior performance by the cross-validation setup, the held-out test setup still has clinically significant performance and is simultaneously a more rigorous test for generalisation. Moreover, with a larger dataset, the performance difference between the two setups would likely reduce and eventually become negligible, as the data efficiency improvements would be less of an advantage to a model that has an abundance of data.

Table 4.4: Comparison of CLAM regressor performance on held-out vs. cross-validation test set settings, under identical conditions. Best values in each column are **bolded**.

Setting	Fold	MAE	Pearson r	Exact Acc. (%)	Within–1 Acc. (%)
Held-Out	0	0.846	0.577	16.7	94.4
	1	0.590	0.700	61.1	88.9
	2	0.701	0.664	33.3	94.4
	3	0.769	0.595	38.9	94.4
	4	**0.549**	**0.762**	**66.7**	88.9
	Ensemble	0.657	0.753	38.9	**100.0**
Cross-Validation	0	0.344	**0.962**	71.4	**100.0**
	1	0.538	0.872	50.0	78.6
	2	0.815	0.599	21.4	71.4
	3	0.535	0.877	57.1	85.7
	4	**0.334**	0.956	**92.9**	**100.0**
	$\mu \pm \sigma$	0.51±0.20	0.85±0.15	58.6±25.2	87.1±10.9

Note. The held-out setup also employs cross-validation on the development set, hence the ensemble. We can report ensemble results for the models on the held-out setting due to the fact that the test set is held-out, and therefore the same, for all models. However, this is not possible for the cross-validation setting due to the test set being different for each model. Thus, we show the mean μ instead.

4.2.3 Effect of Top-k Sampling Strategies

Our CLAM-based models select the top-k highest attention instances, i.e. patches, from each WSI before the forward pass which is at the bag-level. Our investigation of the constant and adaptive k values yielded the following results, having fixed all other hyperparameters and performed five-fold cross-validation (on the cross-validation setup), shown in table 4.5.

Table 4.5: CLAM regressor performance for fixed k configurations on test set across 5 folds, under cross validation setup. The best values in each column are **bolded**.

k	Fold	MAE	Pearson r	Exact Acc. (%)	Within–1 Acc. (%)
32	0	0.344	0.962	71.4	**100.0**
	1	0.538	0.872	50.0	78.6
	2	0.815	0.599	21.4	71.4
	3	0.535	0.877	57.1	85.7
	4	**0.334**	**0.956**	**92.9**	**100.0**
	$\mu \pm \sigma$	0.51±0.18	0.85±0.13	58.6±23.7	87.1±11.4
64	0	**0.327**	**0.963**	**71.4**	**100.0**
	1	0.491	0.893	57.1	92.9
	2	0.809	0.614	21.4	71.4
	3	0.498	0.887	57.1	85.7
	4	0.673	0.747	21.4	78.6
	$\mu \pm \sigma$	0.56±0.18	0.82±0.12	45.7±20.5	85.7±10.1

Observations The mean performance across the different metrics in 4.5 suggest that k=32 performs better, supporting our hypothesis that lower k values will yield better generalisation. However, upon closer analysis, comparing the per-fold metrics shows that k=64 performed better across all metrics in folds 0-3, except the accuracy metrics, where it performed identically. However, k=32's significantly better performance in fold 4 also highlights its greater robustness and more stable performance across all folds. The results demonstrate that a more expressive model with a higher k can perform better in folds with more familiar data but has greater risk of losing performance in more heterogeneous samples - a classic example of bias-variance trade-off. However, given the small size of the dataset, caution must be taken when interpreting from these findings, and we require a larger dataset to draw more definitive conclusions about an optimal choice of k.

Adaptive Top-k Attention Sampling

Table 4.6: Performance of adaptive top-k sampling on a single split. $\bar{k}$ refers to avg number of patches per slide. Best values per column are **bolded**.

Regime	$\bar{k}$	MAE	Pearson r	Exact Acc. (%)	Within–1 Acc. (%)
Adaptive–1 %	39	**0.827**	**0.621**	**21.4**	57.1
Adaptive–2 %	79	0.890	0.552	14.3	**71.4**

Table 4.6 shows the performance of adaptive top-k attention sampling. Setting k to 1% of patch count demonstrates greater performance than 2% across all error and agreement metrics except within-1-grade accuracy. This potentially supports the conclusions made from the fixed-k strategies, which were that large k increases model expressiveness but risks overfitting/reduced robustness, especially on small datasets. Overall, both adaptive strategies showed significantly weaker performance compared to the fixed-k strategies. This could be due to batch-wise variance, meaning that the small slides get too low of a k, which starves the attention module of context and large slides get too large of a k, which could dilute the discriminative signals.

Suggested Improvements

Discriminative features for tubulitis in WSIs will likely be concentrated in regions where tubuli are highly present. Setting k to be a fixed or adaptive value may be too naïve of an approach for our problem, and perhaps the following more intelligent methods might yield better results:

- Semantically guided approach - use the semantic tubule masks from the modular pipeline's semantic segmentation model as a starting point to retrieve discriminative features. This would mean adding the segmentation model as an intermediate step to make the new inputs to the end-to-end module the tubule masks, rather than the raw WSIs.
- Adaptive sampling that selects patches until an attention mass threshold, is met, to ensure adequate discriminative features are taken into account.

4.2.4 Architectural Comparison: CLAM vs TransMIL

We incorporated our learnings from the CLAM-based model into the TransMIL-based model to determine whether it could achieve similar or better performance. We chose the regressor approach due to its greater clinical benefit as our pathologist mentioned (4.2.1), to take advantage of ensembling the outputs from the models in each fold, incorporated a semantically guided approach with tubule masks for better performance, trained with the held-out test set setting to assess generalisation, and avoided using the Top-k patch optimisation to ensure spatial relationships between patches (a key motivation for using TransMIL) were accounted for. Table 4.7 shows the comparison of the TransMIL regressor with the CLAM-based regressor (k=32), also trained on 5-fold cross validation and ensembling, for predictions on the same held-out test set of 18 slides.

Table 4.7: Comparison of T-score predictions of best CLAM-based and TransMIL-based regressor models on held-out tubulitis scoring test set. The best values in each row are **bolded**.

Metric	**CLAM**	**TransMIL**
Mean Absolute Error (MAE)	0.6567	**0.5523**
Pearson Correlation	0.7527	**0.8131**
Pearson *p*-value	0.0003	<**0.0001**
Spearman Correlation	0.7768	**0.8513**
Spearman *p*-value	0.0001	<**0.0001**
Exact Accuracy	0.3889	**0.5556**
Within-1-Grade Accuracy	**1.0000**	0.8333
Prediction Range	[0.63, 2.75]	[0.22, 2.99]
Quadratic Weighted Kappa κ_w	0.6621	**0.7518**

Performance Overview The TransMIL-based model consistently outperformed the CLAM-based model, across nearly all metrics. It achieved lower error from the expert label, with an MSE of 0.5523 compared to 0.6567 for CLAM and attained a higher Pearson correlation with expert labels of 0.81 vs CLAM's 0.75. It is also more expressive, with scores ranging from 0.22 - 2.99 Interestingly, from a clinical perspective, the exact accuracy with TransMIL is notably higher than that of CLAM (55.6% vs 39% accuracy), while the within-1-grade accuracy is slightly lower (83% vs 100% accuracy). Remarkably, TransMIL achieved the best quadratic weighted kappa of the end-to-end models, with an inter-observer reliability of $\kappa_w = 0.7518$.

Discussion A potential explanation for TransMIL's superior performance would be, of course, the improvements stated above in 4.2.4. It may also potentially be due to its preservation of spatial relationships between patches, such as identifying clusters of adjacent inflamed tubules - this spatial grouping is analogous to the first stage of the gold-standard Banff-scoring method, that is, finding focal regions that contain clusters of inflamed tubules. We further see this demonstrated in the attention heatmap visualisations, which often place high attention on tight tubule clusters 4.9 to contribute to its final prediction. Another factor could be TransMIL accounting for the tubule-to-tubule continuity due to its contextualisation of fragmented structures, which the top-k CLAM-based model may have missed.

Conclusion and Clinical Implications Based on our small dataset of 18 WSIs, the results confirm our hypothesis that an end-to-end model can achieve clinically comparable performance to the gold-standard for tubulitis scoring. Both CLAM and TransMIL-based architectures have exceeded our expectations given the limited training dataset size with strong correlations to expert labels and generalisation. Crucially, the potential for explainability of these models due to their attention mechanism holds promise for clinical utility for pathologists, to be able to visualise the decision-making of these otherwise 'black-box' models. We proceed with TransMIL as our flagship end-to-end model to compare with the modular approach and the gold standard.

4.2.5 Class-based Analysis

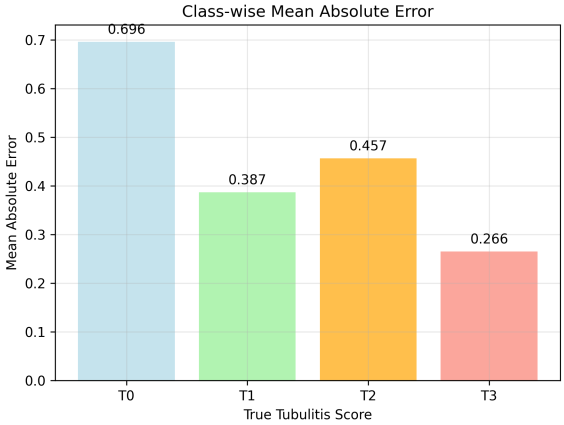


Figure 4.7: Class-specific mean error for the best fold from the TransMIL model.

Despite having a relatively balanced class distribution for training, the end-to-end models tend to achieve lower prediction performance on T0 WSIs, and gradually increase in performance as T-score increases (figure 4.7). This intuitively makes sense, as higher levels of inflammation in slides will most likely be easier for the model to attend to, while a slide with a lack of inflammation may mean the model struggles to identify the discriminatory feature of tubular inflammation and try to capture less relevant properties to correlate with the T-score, such as tubular atrophy.

4.2.6 Explainability and Clinical Applicability

Analysis and Conclusion From the top attended patch visualisations in figures 4.8 - 4.11 and the attention heatmap in 3.7.5, we can see that the model appears to attend to areas with tubules that suffer from infiltration of inflammatory tubules. The top attended patches of WSIs with higher expert-labelled T-scores tend to have a higher concentration of tubules suffering from inflammation. However, upon sharing some visualisations with an expert pathologist, they mentioned that the top attended patches do not directly attend in the way that is analogous to the gold standard, which is finding the tubule with the **highest** number of inflammatory cells. This therefore conveys a limitation of the end-to-end approach, which is that, with the current amount of training on these models, the attention mechanism is not yet able to follow the exact rubric used in Banff classification to calculate tubulitis score, thus limiting the explainability. This therefore only partially confirms our initial hypothesis, as the end-to-end models have achieved notable performance but still lack complete explainability.

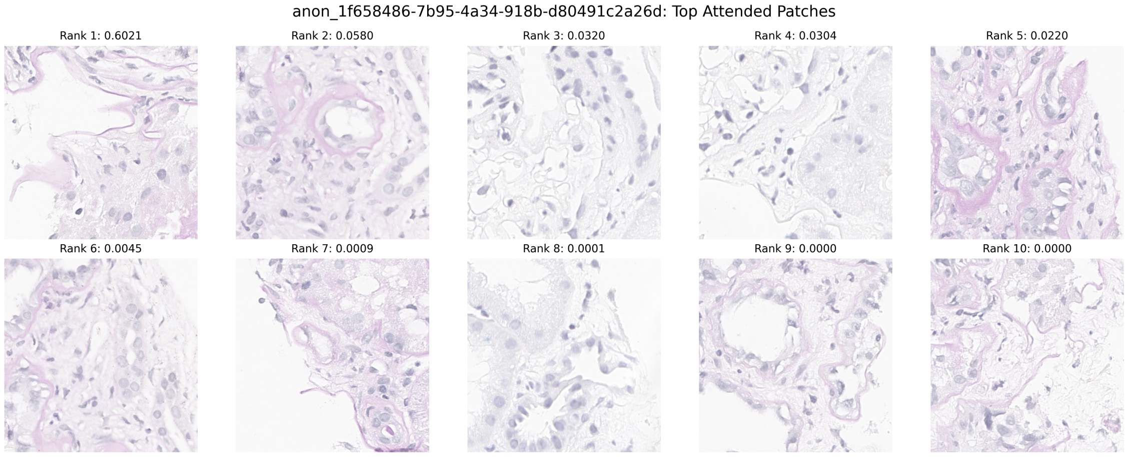


Figure 4.8: Top 10 attended patches for a WSI with expertly labelled T-score of 1 and TransMIL predicted score of 1.330. Here the model mostly attends to non-tubular regions or tubules with less than 5 inflammatory cells, which is characteristic of a slide with a T-score of t1.

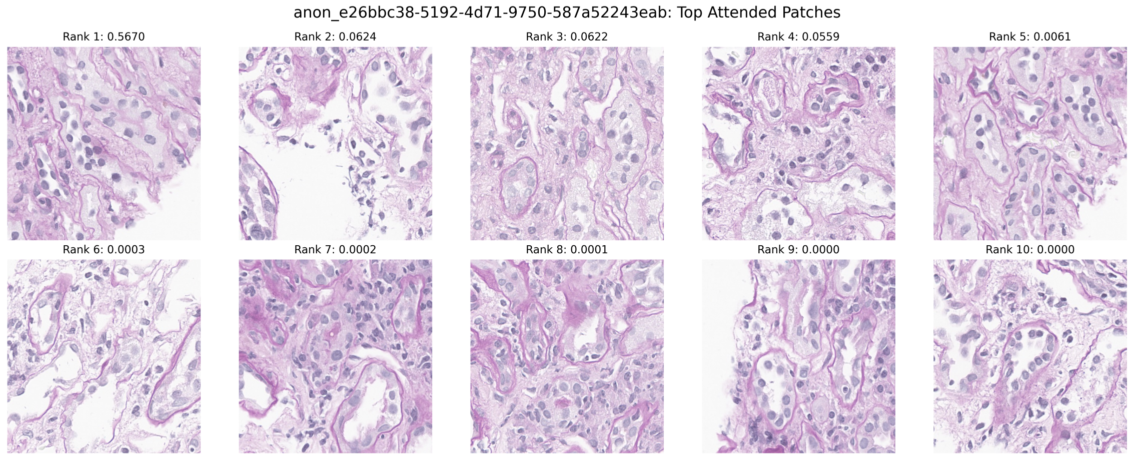


Figure 4.9: Top 10 attended patches for a WSI with expertly labelled T-score of 3 and TransMIL predicted T-score of 2.759. The model seems to have attended to tubules that have high inflammation, with >10 inflammatory cells per tubule, characteristic of a slide with a T-score of t3.

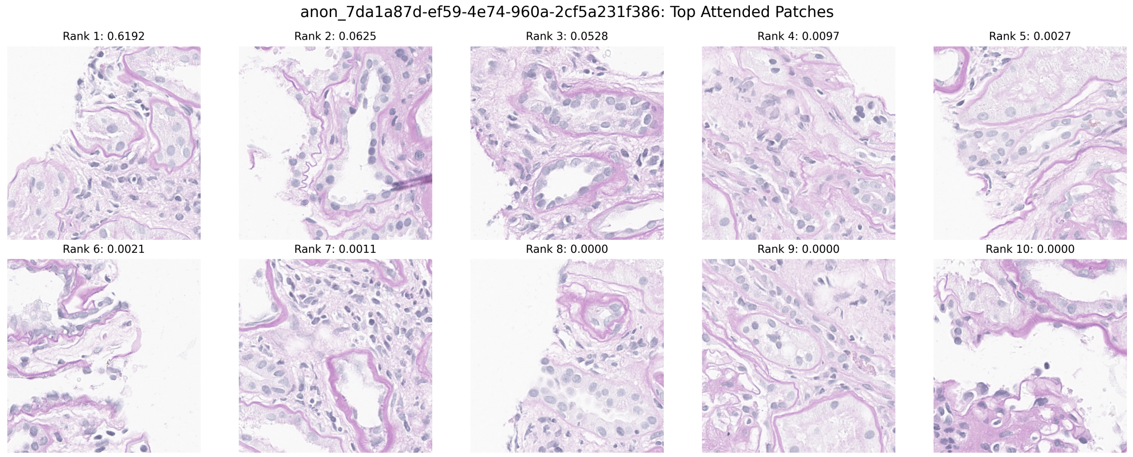


Figure 4.10: Top 10 attended patches for a WSI with expertly labelled T-score of 0 but TransMIL predicted T-score of 1.153. It appears that the model has mistakenly attended to some seemingly inflamed tubules and made a conclusion that it has some level of inflammation.

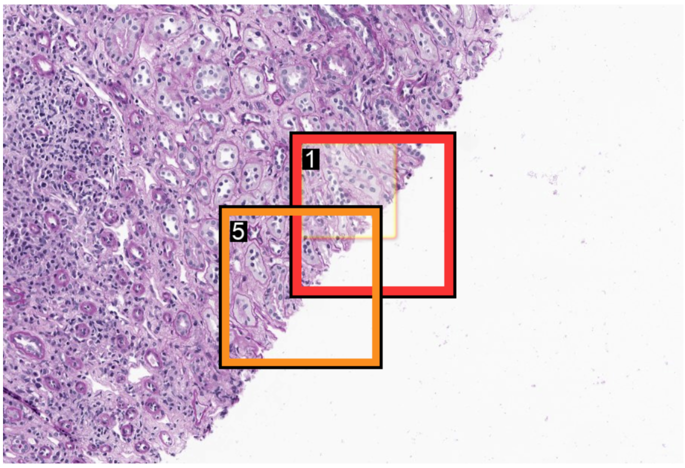


Figure 4.11: Locations of top attended patches on the WSI shown in 4.9. Redder boxes show areas of greater attention.

4.3 Comparison: Modular vs End-to-End

We now compare the results from our best version of the modular pipeline, namely the setup with the confidence threshold $p = 0.7$ with our best end-to-end model, the TransMIL regressor, shown in table 4.8. As mentioned previously, to calculate class-based metrics for the regressor, we round the predictions to the nearest integer.

Table 4.8: Side-by-side comparison of best configurations for each approach. The best results from each row are in **bold**.

Metric	**Modular** **Pipeline** ($p = 0.7$)	**End-to-End** **Model** (TransMIL)	**Gold Standard** (Expert Pathologist)
Exact Accuracy (%)	28.6	**55.6**	–
Within-1-Grade accuracy (%)	82.1	**83.3**	–
MAE	0.917	**0.552**	–
Pearson r	0.305	**0.813**	–
Spearman ρ	0.318	**0.851**	–
Quadratic Weighted Kappa κ_w	0.292	**0.752**	<0.3 [16]
End-to-End Runtime	∼2 hours	**<5 min**	–

Prediction Performance Conclusion The TransMIL end-to-end model delivers remarkable performance, and far exceeds the metrics achieved by the modular pipeline for tubulitis prediction as well as the inter-observer reliability of the gold standard. We discuss these results and their implications in section 5.1.

4.3.1 Comparison with State-of-the-Art

Since a state-of-the-art (SOTA) model does not yet exist for automated *granular* tubulitis scoring (i.e., T0, T1, T2, T3), we compare TransMIL with the existing SOTA for *binary* classification performance, presented at the American Transplant Congress [9]. We therefore turn our granular predictions into coarse binary labels, to contextualise the performance of TransMIL (table 4.9).

Metric	**TransMIL (T≥2)**	**SOTA**
AUC	**0.95**	0.831
Sensitivity	**0.7**	0.51
Specificity	**1.0**	0.843

Table 4.9: Comparison of binary performance for tubulitis scoring with the current SOTA in binary tubulitis classification. The best rows are **bolded**.

Observations TransMIL impressively exceeded the SOTA across all their metrics, despite being trained on a dataset $\sim 1/10^{\text{th}}$ of their training set size. Notably, TransMIL achieves an AUC of 0.95, where AUC is defined to be the area under the curve of true positive rate against false positive rate.

4.3.2 Explainability and Clinical Applicability

The modular pipeline is intrinsically explainable and, even, interpretable, due to its modular computations. The structure segmentations, instance labelling and detection model all provide clear visual insight (as shown in previous sections) and evidence of the decision-making process of the pipeline to the pathologist. The quantification statistics are also interpretable, providing inflammation information at the tubule level, and grading that mirrors the Banff classification rubric.

The TransMIL end-to-end model, though highly performant, does not achieve this level of explainability. The attention heatmaps appear to identify dense areas of highly inflamed tubules, but, from our discussion with an expert pathologist, fall short of showing patches that show the defining feature of tubulitis: specific tubules that have the highest number of inflammatory cells. We propose a novel strategy to achieve complete explainability for the attention-based end-to-end models in the future work (5.3).

Explainability Conclusion The modular pipeline therefore excels in the domain of explainability.

Chapter 5

Discussion and Conclusions

5.1 Discussion

KidneyGrader was developed to address the challenge of predicting tubulitis, the primary indicator of acute kidney allograft rejection, based on the fine-grained Banff specification. From on our rigorous evaluation of two different approaches, the results demonstrate success from both, with complementary strengths and limitations:

Modular Approach: The modular interpretable pipeline achieved an impressive within-1-grade accuracy of 82.1%, and a quadratic weighted kappa κ_m of 0.29, thereby reaching a reliability level comparable to that of the gold standard for tubulitis scoring (<0.3). It also facilitates for the potential of clinical adoption due to its full interpretability, as mentioned in 4.3.2, allowing for validation of outputs each stage by the pathologist.

However, the other metrics such as Pearson correlation (r=0.3 for $0.3 <= p <= 0.7$) demonstrate that it is suffering from considerable loss of performance at some stage in the pipeline, resulting in a poor positive correlation with the expert labels. As mentioned in the evaluation, the key culprits are the instance labelling and inflammatory cell detection quality. Despite efforts to maximise the quality of the instance segmentations with our two-pass watershed-based algorithm, the tubule instance visualisations (see 4.1.1) highlight the inherent limitations of classical instance labelling algorithms when dealing with structures with high morphological variation. This instability is compounded by the sensitivity-specificity trade-off of the InstanSeg detector (see 4.1.2). While the higher confidence thresholds have actually increased performance, the downstream scoring is inherently highly sensitive to the detections due to the rule-based rubric used at the grading stage. We therefore found that more data driven metrics such as the top percentiles of inflamed tubules led to higher correlation with the expert labels than simply using the max-count method. This raises a fundamental question for the field: if the Banff grading process is to be automated, should it still be constrained by the metrics and heuristics originally made for manual assessment?

End-to-End Approach: The end-to-end models, and in particular, the flagship TransMIL-based model showed remarkable performance on the held-out test set. It achieved a reliability κ_w of 0.75, thus in the 'substantial' range [10], and over 2x higher than the gold standard. It also achieved state-of-the-art performance (4.3.1) across all metrics, for coarse binary classification of tubulitis (though tested on different datasets, both were held-out and we use this to simplify comparison and to contextualise performance). It achieves a Pearson r of 0.81 and a within-1-accuracy of 83.3%, making it the best approach for KidneyGrader in terms of raw prediction performance. It also attains end-to-end processing speed of <5 minutes, compared to the ∼2 hours for the modular pipeline, a critical consideration for the possibility of clinical adoption. Despite this promising performance, the TransMIL model has limited explainability, as discussed in 4.2.6.

5.2 Conclusions

This study was guided by two fundamental hypotheses:

1. An interpretable system of deep-learning models to quantify the presence of inflammatory cells in tubules can achieve gold standard levels of reliability and have considerable agreement with expert labels for Banff tubulitis scores.

2. Similar reliability and agreement, and a degree of explainability comparable to the interpretability of the system of models can be achieved with a single attention-based end-to-end model.

We mostly validate both hypotheses, with both of KidneyGrader's approaches demonstrating distinct strengths while simultaneously showing some limitations.

With respect to the first hypothesis, the modular interpretable pipeline successfully achieved expert-level reliability (κ_w=0.29) and high accuracy within 1 grade of the expert label on specific configurations (82.1%), but failed to achieve high correlation (r=0.3) due to the limitations caused by the absence of expert labels in the instance segmentation and detection stage. It did, however, present complete interpretability and allow for clear visual insight of the decision making process, a property that is aligned with the transparency requirements presented by the recent EU AI Act [49].

Regarding the second hypothesis, our flagship end-to-end TransMIL model achieves notable reliability (κ_w=0.75), over twice that of the gold standard (κ_w<0.3). It has exceptionally high agreement with the expert labels in terms of accuracy within 1 grade (83.3%) and correlation (r=0.81), despite being trained on a small dataset of 75 slides. In terms of explainability, the TransMIL model's attention heatmaps may be capturing more complex patterns that lead to its diagnostic accuracy, but, as our collaborating pathologist mentioned, these heatmaps do not yet attend to the defining feature of tubulitis, which is the tubule with the highest inflammatory cell count.

Overall, the modular pipeline has laid the foundation for fully interpretable Banff scoring, and, with sufficient training labels for each stage, holds promise for a high-performing system ready for clinical deployment. The attention-based end-to-end models have excelled in the domain of raw tubulitis prediction performance, and show great potential for clinical adoption if we can achieve complete explainability that demonstrates that the models attend analogously to the Banff rubrics, and we discuss a novel strategy for this in 5.3.

5.3 Future Work

We highlight key opportunities to work towards clinical adoption of KidneyGrader:

Learnt Instance Segmentation: The inherent limitations of the classical instance labelling algorithms necessitate the use of a machine learning model for the instance labelling of the tubule masks. This will inevitably require expert instance labels, but will almost certainly outperform the existing classical method and improve prediction performance.

Inflammatory Cell Detector Fine-tuning: As mentioned in the discussion and evaluation, the considerable domain shift between the MONKEY dataset and our data requires additional fine-tuning to improve sensitivity and specificity and yield greater downstream prediction performance.

Data Expansion: The end-to-end models in KidneyGrader have been trained with a very small dataset of 75 PAS slides from one cohort. To test and enhance their generalisability, it is a good idea to try larger multi-centre, multi-stain cohorts. To address data-sharing barriers between centres, we can incorporate a federated learning setup. In addition, domain-adaptive stain-normalisation can be employed to address the stain variations between centres.

Expansion to Support All T-Cell Mediated Rejection Scores: The modular interpretable pipeline was designed with the potential for expansion in mind, and the lowest hanging fruit for this expansion would be the other T-Cell Mediated Rejection scores, such as total inflammation, interstitial inflammation and intimal arteritis, which are inflammation based scores that can utilise the detection model. A long-term goal would be to eventually support *all* Banff scores, to fully automate the Banff classification process.

Achieving Explainability for E2E Models with a Hybrid Approach: We propose a novel method to achieve explainability for the current end-to-end models by leveraging the modular pipeline (assuming it achieves good performance once the first two improvements in this section are carried out). Initially, the multi-stage pipeline is to be run on each WSI of the end-to-end model's training set, to identify the single tubules that contain the inflammatory cell count, and label it with a binary '1'. This can be used by the end-to-end model as a weak label, and an auxiliary loss can be added to direct the model's attention to the most inflammatory tubule. The resulting heatmaps from this setup should ideally identify the decisive tubule. Moreover, to verify this, a counterfactual test by masking the decisive tubule and noticing the drop in T-score, assuming the second most decisive tubule has a cell count considerably below the first.

Chapter 6

Declarations

6.1 Use of Generative AI

I acknowledge the use of ChatGPT 4o (https://chatgpt.com/), to generate ideas for structuring my contents, getting feedback and improvement points throughout my report and formatting tables and figures. I also acknowledge the use of ChatGPT 4o and Claude 4 Sonnet (https://claude.ai/) to help with code for the segmentation mask visualisation, inflammatory cell detection visualisation, and attention weight visualisation, debugging, graph plotting code, frontend user interface code and some dataset processing utility code for my end-to-end models. I have also added these acknowledgements as comments in the codebase. I confirm that no content generated by AI has been presented as my own work.

6.2 Ethical Considerations

This research project involves the use of patient medical data in the form of kidney allograft biopsy WSIs, and we have taken into account the ethical concerns relating sharing of sensitive medical data beyond medical personnel. The dataset has been provided by the Imperial College Healthcare Tissue Bank (ICHTB), who are supported by the National Institute of Health Research Imperial Biomedical Research Centre, a partnership between Imperial College London and Imperial College Healthcare NHS Trust. The ICHTB has been approved by the National Research Ethics Service to release human material for research (12/WA/0196), and the samples used in this project are based on research application number R18040.

6.3 Sustainability

Conscious efforts have been made to minimise resource consumption during this project (please see 3.9.1). For instance, caching has been implemented in each stage of the modular pipeline, to avoid needing heavy recomputation of results for each stage. Architectural decisions have prioritised balancing performance with efficiency, such as the use of EfficientNet for the segmentation model encoder. For training, early stopping has been implemented in all model training setups to reduce unnecessary epochs.

6.4 Availability of Data and Materials

The project's materials are accessible as follows:

1. **Project source code:**: `https://github.com/abrar-rashid/kidney-grader`.

2. **Training and test data h5 files** for segmentation model (from stage 1 of the modular pipeline): available on the **mnemosyne** machine of the Biomedic group in path `/data/ar2221/KidneyGrader/data`.

3. **93 anonymised WSI dataset** with correspoding **Banff scores csv** for training and testing of end-to-end models and testing of modular pipeline: available on the **mnemosyne** machine of the Biomedic group, in path `/data/ar2221/all_wsis`.

4. **Results**: in `/data/ar2221/KidneyGrader/`, and can be reproduced by following the README.

5. **Model checkpoints** downloaded from huggingface via the download script in the repo. Please note, downloading the UNI foundation model checkpoint for running training or inference of the end-to-end models requires filling in the form in `huggingface.co/MahmoodLab/UNI` to obtain permission for use.

Appendix A

Additional Tables of Results

Class	Metric	Previous U-Net (Baseline)	Our U-Net
Background	IoU	0.8091	0.8304
	Precision	0.9049	0.9008
	Recall	0.8843	0.9139
	F1 Score	0.8945	0.9073
	Boundary IoU	0.2829	0.3371
	Hausdorff Distance	175.55	133.45
	ASSD	14.48	11.21
Tubuli	IoU	0.8751	0.8852
	Precision	0.9242	0.9524
	Recall	0.9428	0.9262
	F1 Score	0.9334	0.9391
	Boundary IoU	0.3353	0.3706
	Hausdorff Distance	150.82	120.93
	ASSD	15.89	12.47
Glomeruli	IoU	0.6952	0.7639
	Precision	0.7068	0.7917
	Recall	0.9769	0.9561
	F1 Score	0.8202	0.8661
	Boundary IoU	0.0342	0.0697
	Hausdorff Distance	628.57	280.03
	ASSD	193.44	74.20
Indeterminate Vessel	IoU	0.0352	0.4317
	Precision	0.1469	0.5107
	Recall	0.0443	0.7364
	F1 Score	0.0681	0.6031
	Boundary IoU	0.0020	0.0236
	Hausdorff Distance	220.07	212.24
	ASSD	35.13	22.15
Artery	IoU	0.1584	0.0456
	Precision	0.5450	0.8791
	Recall	0.1825	0.0458
	F1 Score	0.2734	0.0871
	Boundary IoU	0.0121	0.0000
	Hausdorff Distance	478.40	291.52
	ASSD	46.55	116.47
Overall Accuracy		0.8967	0.9101
Mean IoU		0.5146	0.5913

Table A.1: Detailed Segmentation Metrics: Previous Best U-Net Model vs. Our Model

Bibliography

[1] National Institute of Diabetes and Digestive and Kidney Diseases. Causes of Chronic Kidney Disease;. Accessed: 05/01/2025. Available from: `https://www.niddk.nih.gov/health-information/kidney-disease/chronic-kidney-disease-ckd/causes`.

[2] Kovesdy CP. Epidemiology of chronic kidney disease: an update 2022. Kidney Int Suppl (2011). 2022 Apr;12(1):7-11.

[3] Matas AJ, Gillingham KJ, Humar A, Kandaswamy R, Sutherland DER, Payne WD, et al. 2202 kidney transplant recipients with 10 years of graft function: what happens next? Am J Transplant. 2008 Nov;8(11):2410-9.

[4] Roufosse C, Simmonds N, Clahsen-van Groningen M, Haas M, Henriksen KJ, Horsfield C, et al. A 2018 Reference Guide to the Banff Classification of Renal Allograft Pathology. Transplantation. 2018 Nov;102(11):1795-814.

[5] Yu KH, Beam AL, Kohane IS. Artificial intelligence in healthcare. Nature Biomedical Engineering. 2018 Oct;2(10):719-31.

[6] Ginley B, Lutnick B, Jen KY, Fogo AB, Jain S, Rosenberg A, et al. Computational segmentation and classification of diabetic glomerulosclerosis. J Am Soc Nephrol. 2019 Oct;30(10):1953-67.

[7] Tsamandas AC, Shapiro R, Jordan M, Demetris AJ, Randhawa PS. Significance of tubulitis in chronic allograft nephropathy: a clinicopathologic study. Clin Transplant. 1997 Apr;11(2):139-41.

[8] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al.. Attention Is All You Need; 2023. Available from: `https://arxiv.org/abs/1706.03762`.

[9] Cooper LA, Farris AB, Allegretti L, et al. Multi-Institutional Validation of Automated AI Banff Lesion Score Predictions from Renal Allograft Biopsies. In: American Transplant Congress. American Society of Transplantation and American Society of Transplant Surgeons; 2023. Conference presentation. Available from: `https://atc.digitellinc.com/p/s/multi-institutional-validation-of-automated-ai-banff-lesion-score-predictions-from-ren`

[10] McHugh ML. Interrater reliability: the kappa statistic. Biochem Med (Zagreb). 2012;22(3):276-82.

[11] Kumar N, Gupta R, Gupta S. Whole Slide Imaging (WSI) in Pathology: Current Perspectives and Future Directions. J Digit Imaging. 2020 Aug;33(4):1034-40.

[12] Kumar N, Gupta R, Gupta S. Whole Slide Imaging (WSI) in Pathology: Current Perspectives and Future Directions. J Digit Imaging. 2020 Aug;33(4):1034-40.

[13] Zhao S, Zhou H, Lin SS, Cao R, Yang C. Efficient, gigapixel-scale, aberration-free whole slide scanner using angular ptychographic imaging with closed-form solution. Biomed Opt Express. 2024 Sep;15(10):5739-55.

[14] Jain E, Patel A, Parwani AV, Shafi S, Brar Z, Sharma S, et al. Whole Slide Imaging Technology and Its Applications: Current and Emerging Perspectives. Int J Surg Pathol. 2023 Jul;32(3):433-48.

[15] Marcussen N, Olsen TS, Benediktsson H, Racusen L, Solez K. Reproducibility of the Banff classification of renal allograft pathology. Transplantation. 1995 Nov;60(10):1083-9.

[16] Furness PN, Taub N. International variation in the interpretation of renal transplant biopsies: Report of the CERTPAP Project^1. Kidney International. 2001 Nov;60(5):1998-2012.

[17] Banff Foundation for Allograft Pathology. Banff Classification for Renal Allograft Pathology, 2022;. Accessed: 11/01/2025. Available from: https://banfffoundation.org/central-repository-for-banff-classification-resources-3/.

[18] Chan HP, Samala RK, Hadjiiski LM, Zhou C. Deep Learning in Medical Image Analysis. Adv Exp Med Biol. 2020;1213:3-21.

[19] Kainz B. Lecture 2: CNNs. Deep Learning Course (70010) at Imperial College London. 2025 Jan.

[20] Ronneberger O, Fischer P, Brox T. U-Net: Convolutional Networks for Biomedical Image Segmentation; 2015. Available from: https://arxiv.org/abs/1505.04597.

[21] Oktay O, Schlemper J, Folgoc LL, Lee M, Heinrich M, Misawa K, et al.. Attention U-Net: Learning Where to Look for the Pancreas; 2018. Available from: https://arxiv.org/abs/1804.03999.

[22] Zhou Z, Siddiquee MMR, Tajbakhsh N, Liang J. UNet++: A Nested U-Net Architecture for Medical Image Segmentation; 2018. Available from: https://arxiv.org/abs/1807.10165.

[23] Özgün Çiçek, Abdulkadir A, Lienkamp SS, Brox T, Ronneberger O. 3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation; 2016. Available from: https://arxiv.org/abs/1606.06650.

[24] Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al.. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale; 2021. Available from: https://arxiv.org/abs/2010.11929.

[25] Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, et al.. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows; 2021. Available from: https://arxiv.org/abs/2103.14030.

[26] Chen J, Lu Y, Yu Q, Luo X, Adeli E, Wang Y, et al.. TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation; 2021. Available from: https://arxiv.org/abs/2102.04306.

[27] Janowczyk A, Madabhushi A. Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases. J Pathol Inform. 2016 Jul;7(1):29.

[28] Sarapata G. Deep Learning–Based Segmentation and Quantification in Kidney Transplant Pathology. Masters Thesis at Imperial College London. 2021.

[29] Goldsborough T, Philps B, O'Callaghan A, Inglis F, Leplat L, Filby A, et al.. InstanSeg: an embedding-based instance segmentation algorithm optimized for accurate, efficient and portable cell segmentation; 2024. Available from: https://arxiv.org/abs/2408.15954.

[30] Studer, Linda and van Midden, Dominique and Ayatollahi, Fazael and van der Laak, Jeroen A W M . MONKEY Challenge: Detection of Inflammation in Kidney Biopsies;. Accessed: 08/06/2025. Available from: https://monkey.grand-challenge.org/.

[31] Wang X, Girdhar R, Yu SX, Misra I. Cut and Learn for Unsupervised Object Detection and Instance Segmentation; 2023. Available from: https://arxiv.org/abs/2301.11320.

[32] Shin G, Albanie S, Xie W. Zero-shot Unsupervised Transfer Instance Segmentation; 2023. Available from: https://arxiv.org/abs/2304.14376.

[33] Qian Z, Li K, Lai M, Chang EIC, Wei B, Fan Y, et al.. Transformer based multiple instance learning for weakly supervised histopathology image segmentation; 2022. Available from: https://arxiv.org/abs/2205.08878.

[34] Lu MY, Williamson DFK, Chen TY, Chen RJ, Barbieri M, Mahmood F. Data Efficient and Weakly Supervised Computational Pathology on Whole Slide Images; 2020. Available from: https://arxiv.org/abs/2004.09666.

[35] Chen RJ, Ding T, Lu MY, Williamson DFK, Jaume G, Chen B, et al.. A General-Purpose Self-Supervised Model for Computational Pathology; 2023. Available from: https://arxiv.org/abs/2308.15474.

[36] Kers J, Bülow RD, Klinkhammer BM, Breimer GE, Fontana F, Abiola AA, et al. Deep learning-based classification of kidney transplant pathology: a retrospective, multicentre, proof-of-concept study. The Lancet Digital Health. 2022;4(1):e18-26. Available from: https://www.sciencedirect.com/science/article/pii/S2589750021002119.

[37] Hermsen M, Ciompi F, Adefidipe A, Denic A, Dendooven A, Smith BH, et al. Convolutional Neural Networks for the Evaluation of Chronic and Inflammatory Lesions in Kidney Transplant Biopsies. The American Journal of Pathology. 2022;192(10):1418-32. Available from: https://www.sciencedirect.com/science/article/pii/S0002944022001985.

[38] Fu J. Mapping Needle Core Biopsies to Kidney Zones using Swin Transformer. Masters Thesis at Imperial College London. 2021.

[39] Bankhead P, Loughrey MB, Fernández JA, Dombrowski Y, McArt DG, Dunne PD, et al. QuPath: Open source software for digital pathology image analysis. Scientific Reports. 2017 Dec;7(1):16878. Available from: https://doi.org/10.1038/s41598-017-17204-5.

[40] Jain V. auto_banff_scoring GitHub repository; 2025. Accessed: 2025-05-22. https://github.com/VishalJ99/auto_banff_scoring.

[41] Zuiderveld K. Contrast Limited Adaptive Histogram Equalization. In: Heckbert PS, editor. Graphics Gems IV. San Diego, CA, USA: Academic Press Professional, Inc.; 1994. p. 474-85.

[42] Tan M, Le QV. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks; 2020. Available from: https://arxiv.org/abs/1905.11946.

[43] Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L. ImageNet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition; 2009. p. 248-55.

[44] Oktay O, Schlemper J, Folgoc LL, Lee M, Heinrich M, Misawa K, et al.. Attention U-Net: Learning Where to Look for the Pancreas; 2018. Available from: https://arxiv.org/abs/1804.03999.

[45] Shao Z, Bian H, Chen Y, Wang Y, Zhang J, Ji X, et al.. TransMIL: Transformer based Correlated Multiple Instance Learning for Whole Slide Image Classification; 2021. Available from: https://arxiv.org/abs/2106.00908.

[46] Gupta A, Dar G, Goodman S, Ciprut D, Berant J. Memory-efficient Transformers via Top-k Attention; 2021. Available from: https://arxiv.org/abs/2106.06899.

[47] Collet Y, contributors. Zstandard: Fast real-time compression algorithm; 2016. Commit version as of 2025-06-08. https://github.com/facebook/zstd.

[48] Wang A, Ball J, Lally J, Kroll P, Radziuk S, Sreeram S. KidneyCaliper: Automated Deep-Learning-Based Workflow for Kidney Pathologists. Imperial College London, Department of Computing; 2025. Supervised by Dr Bernhard Kainz. Clients: Dr Candice Roufosse, Callum Arthurs.

[49] Aboy M, Minssen T, Vayena E. Navigating the EU AI Act: implications for regulated digital medical products. npj Digital Medicine. 2024 Sep;7(1):237.