Transformer-based multiclass segmentation pipeline for basic kidney histology

He, J.L.; Valkema, P.A.; Long, J.M.; Li, J.; Florquin, S.; Naesens, M.; ... ; Kers, J.

scientific reports

Check for updates

OPEN

Transformer-based multiclass segmentation pipeline for basic kidney histology

Junling He[1,2], Pieter A. Valkema[1,4], Jingmin Long[3], Jia Li[3], Sandrine Florquin[4], Maarten Naesens[5], Priyanka Koshy[6], Tri Q. Nguyen[7], Soufian Meziyerh[8,9], Aiko P. J. de Vries[8,9], Onno J. De Boer[4], Fons J. Verbeek[3], Zhan Xiong[3] & Jesper Kers[1,4,9,10]✉

Current applications of deep learning in renal pathology focused on anatomical structures with morphology, yet little research has focused on the performance of models, such as versatility, in regions with severe kidney damage. In this study, we explored the difference in modal/domain shift capabilities between CNN-based and Transformer-based models. Firstly, we adopted two splitting strategies—WSI-level and patch-level—to stimulate sampling on multiple modal data distribution (i.e., renal WSIs collected from multi-centers). Then, we trained multiple CNN- and Transformer-based models on each splitting scheme respectively. We compared cross-splitting performance and analyzed the effective factors of results. For further validation, all models were tested on an independent external dataset for sensitivity analysis on the degree of fibrosis and inflammation. In conclusion, at both the patch- and WSI-level, M2F-Swin-B substantially outperformed UNet-ResNet18 with an average Intersection over Union (A-IoU) and per-class IoU. Notably, M2F-Swin-B outperformed UNet-ResNet18 in areas of a higher degree of fibrosis and inflammation and, a higher IoU score of arteries. In this study, we developed a robust multi-class segmentation pipeline for kidney histology. Moreover, we showed that the attention mechanism in Mask2Former enables visibly crisper and more uniform segmentation, particularly when the data is inadequate.

Keywords Deep learning, Kidney histology, Multiclass segmentation, Semantic segmentation, Transformer

Renal pathology is a subspecialty of general pathology focused on characterizing and diagnosing nephrological diseases. The kidney biopsy is the gold standard for diagnosing and staging kidney diseases. Absolute quantification of every lesion in the renal biopsy, independent of disease entity, is the holy grail of quantitative renal pathology. However, traditional methods relied heavily on manual histologic grading, which often suffered from large international variations and poor reproducibility[1]. Full manual quantification and annotation of abnormalities in repetitive structures like tubules is not feasible within the diagnostic time frame. Consequently, renal pathologists are compelled to score (i.e. estimate) tubular lesions using non-granular, semi-quantitative ranges like 25–50%. Nonetheless, there may be significant differences in prognosis between patients with 26% and 50% tubular injury. Additionally, each underlying disease has its lesion-scoring scheme, despite the general acceptance that kidneys respond to injury in stereotypical patterns that can co-occur within a single structure (e.g., crescents appear morphologically identical across different kidney diseases)[2].

With the advent of digital pathology, the rapid development of big data, and the innovation of artificial intelligence (AI) algorithms, the field of nephrology has experienced significant progress and changes[3,4]. AI-assisted renal pathology offers a promising solution for pathologists to address the tedious process of quantifying thousands of individual lesions. Since the basic microanatomy of the kidney does not change during the progression of any disease, a high-performing and robust multiclass segmentation network could be the first step toward building a community effort for quantitative renal pathology.

[1]Department of Pathology, LUMC, Leiden, The Netherlands. [2]Department of Nephrology, Daping Hospital, Army Medical Center, Army Medical University, Chongqing, China. [3]Leiden Institute of Advanced Computer Science, Leiden University, Leiden, The Netherlands. [4]Department of Pathology, Amsterdam UMC, Amsterdam, The Netherlands. [5]Department of Nephrology and Transplantation, KU Leuven, Leuven, Belgium. [6]Department of Microbiology, Immunology and Transplantation, KU Leuven, Leuven, Belgium. [7]Department of Pathology, UMC Utrecht, Utrecht, The Netherlands. [8]Division of Nephrology, Department of Medicine, LUMC, Leiden, The Netherlands. [9]Leiden Transplant Center, LUMC and Leiden University, Leiden, The Netherlands. [10] Department of Pathology, Erasmus Medical Center, Rotterdam, The Netherlands. ✉email: j.kers@lumc.nl

Most recent research about deep learning (DL) applications on renal histopathology can generally be categorized into two main areas: object detection[5–8] and segmentation[9–14]. As DL methods have advanced, architectures such as U-Net[13–15], DeepLab[12,16], and Cascade Mask R-CNN[17] have been effectively employed to address challenges within both semantic and instance segmentation. Convolutional neural networks (CNNs) are the most widely used DL models for segmenting kidney anatomical compartments. Transformers are alternative architectures based on attention mechanisms[18]. Originally designed for natural language processing tasks, Transformers have since been adopted for various computer vision tasks, including segmentation. Recently, Mask2Former has garnered significant attention for its superior performance in modeling contextual information and segmenting small objects[19].

While DL applications for segmenting renal anatomical compartments are growing, current DL tools in renal pathology perform well in normal anatomical structures, they frequently struggle in areas with severe renal damage[12,20]. Additionally, there is still a lack of solid analysis on model generalization for this fundamental task. Recent studies have shown the potential of high-throughput morphological analysis of kidney tissue, termed Next Generation Morphometry, which aligns with high-throughput techniques like next-generation sequencing used in preclinical kidney disease research[21]. As technology advances, developing a generalized segmentation model could pave the way for future research and drive progress in computational nephropathology.

Our main contributions are listed as follows: 1. We developed a broadly applicable, open-source segmentation pipeline that is robust against the heterogeneity of clinical kidney biopsy presentations. This pipeline is intended to serve as a versatile foundation for future research and drive innovation in computational nephropathology. 2. We investigated model generalization and the explainability of segmentation performance in renal histology whole slide images (WSIs). To minimize the confounding impact of architectural differences, we focus our analysis on two representative models: U-Net (a CNN-based framework) and Mask2Former (a Transformer-based framework). 3. We tested the modal/domain shift performance of these segmentation networks from two respects: (a) multicenter data variability, and (b) renal WSIs exhibiting varying degrees of interstitial inflammation and fibrosis. Therefore, we designed patch-level and WSI-level splitting schemes (Figure S1) to simulate various data distributions. In conclusion, our findings demonstrated that Mask2Former, particularly when equipped with a Swin-B Transformer encoder, consistently outperformed the UNet, especially in challenging scenarios involving inflammation and fibrosis, as well as in the segmentation of smaller, anatomical structures like arterioles. Notably, Mask2Former maintained its robust performance even under data-constrained conditions.

The rest of the paper is organized as follows: In the Related Work section, we reviewed the developments of DL models with summary comments. In the Materials and Methods section, we detailed data preparation (sample generation, staining, annotation, and preprocessing), model structures, and training strategies. In the Results section, we demonstrated our experimental findings and analysis. In the Discussion section, we interpreted the results, highlighting their implications and significance. In the Limitation and Future Research section, we pointed out the limitations and suggested potential improvements to our current study. In the Conclusion section, we summarized our contributions to the community and potential research directions.

Related work

Evolution of deep learning in renal pathology

The application of deep learning in renal pathology has evolved through three distinct architectural paradigms, each addressing specific challenges in histopathological image analysis.

CNNs in renal pathology

CNNs have been the cornerstone of many breakthroughs in medical image segmentation and classification. Early works by Marée et al.[5] and Temerinac-Ott et al.[6] demonstrated the feasibility of using traditional image processing and machine learning techniques for glomeruli detection. However, the introduction of CNNs significantly improved performance. For instance, Bukowy et al.[8] employed region-based CNNs to localize glomeruli in trichrome-stained kidney sections, achieving high accuracy. Similarly, Simon et al.[7] utilized multi-radial LBP features combined with CNNs for rapid glomerular detection.

Further advancements were made with the adoption of U-Net architectures, which became particularly popular for segmentation tasks. Gallego et al.[15] proposed a U-Net-based framework to quantify glomerulosclerosis in PAS and H&E stained tissues, while Ginley et al.[11] and Hermsen et al.[12] applied DL for the classification and assessment of diabetic glomerulosclerosis and other kidney pathologies. The MESCnn pipeline adopted a two-stage deep learning approach, integrating Mask R-CNN for precise glomerular instance segmentation and CNN-based architectures (including EfficientNetV2, MobileNetV2, and ResNet50) to perform Oxford Classification tasks (M, E, S, C)[22]. In addition, the FLASH framework employed CNN-based semantic segmentation for pixel-level analysis, followed by advanced morphometric feature extraction to enable interpretable quantitative pathology[21]. These studies underscored the effectiveness of CNNs in handling spatial hierarchies and local features in histopathological images. While CNN-based models are resource-friendly, our study has demonstrated that they are vulnerable to domain shift, especially when data is scarce.

Transformer-based approaches

While CNNs dominated the field initially, transformer-based models have emerged as powerful alternatives, particularly for tasks requiring global context and long-range dependencies. The foundational work of Vaswani et al.[18] introduced the transformer architecture, which has since been adapted for medical imaging. Cheng et al.[19] proposed a masked-attention mask transformer for universal image segmentation, demonstrating its potential in capturing complex patterns in biomedical images. For instance, gradient-weighted class activation mapping (Grad-CAM) has been integrated with transformers to provide interpretable results in brain MRI

tumor classification[23]. In summary, Transformer-based models can find meaningful information by long-range contextual features, yet at the cost of inefficient resource usage.

The hybrid model structures

CNNs and transformers each offer distinct advantages. CNNs excel in capturing local spatial features and are computationally efficient for high-resolution images, making them ideal for tasks like glomeruli detection and segmentation[8,11,15]. In contrast, transformers provide superior performance in modeling global relationships and are particularly effective for tasks requiring contextual understanding, such as multi-class classification and anomaly detection[18]. Therefore, recent trends indicate a growing interest in hybrid architectures that combine the strengths of CNNs and transformers. For example, attention mechanisms derived from transformers have been integrated into CNN frameworks to enhance feature extraction. These innovations suggest a promising direction for future research, aiming to leverage the complementary strengths of both paradigms. Recent innovations, such as DenseNet channel spatial and Semantic Guidance Attention (DCSSGA-UNet)[24] and Multi-Attention Gated Residual U-Net (MAGRes-UNet)[25], have further enhanced CNN architectures by incorporating attention mechanisms to improve segmentation accuracy and robustness. Similarly, hybrid approaches combining YOLOv8 and Vision Transformers (ViT) have been explored for fracture prediction in X-ray images[26], highlighting the versatility of transformers in medical imaging. Moreover, transformer-based models have shown promise in explainability and multi-class classification tasks. The hybrid models can exploit advantages from both CNN and Transformer, but requires delicate structure design and a complicated training strategy.

While existing studies have established the feasibility of DL in renal pathology, three key limitations persist: (1) performance degradation in cases of severe tissue damage; (2) insufficient evaluation of model generalization across diverse clinical settings; (3) lack of standardized benchmarks for comparing architectural approaches. Building upon these methodological foundations, our study specifically addresses the crucial challenge of evaluating model performance across diverse clinical presentations of kidney biopsies.

Materials and methods

Sample selection

A multi-center retrospective training set was created comprising Jones-silver stained kidney biopsy slides from 147 patients from the Departments of Pathology of the Amsterdam UMC (AUMC), UMC Utrecht (UMCU), and Leiden UMC (LUMC). A pathology staff member anonymized all the slides before further analysis. Kidney WSIs from both transplant and native kidney biopsies were digitized with a Philips UltraFast scanner (AUMC, LUMC) or Hamamatsu XR scanner (UMCU) at a resolution of ~ 0.25 µm/pixel. One to five representative regions of interest (ROIs) on a single WSI were chosen for dense manual annotations, and a total of 261 ROIs were acquired for further annotations. ROIs were chosen based on pathologist review to capture representative anatomical and pathological structures in the WSIs.

An independent external test set was created with an anonymized collection of a total of 24 digital kidney biopsy cases from the Department of Pathology of KU Leuven (KUL), scanned with a Philips UltraFast scanner at a resolution of 0.25 µm/pixel. The external test set was categorized based on Banff ti/IFTA scores: ≤ 25% (easy group), 26–50% (medium group), > 50% (hard group), as per standardized criteria[27]. The challenging areas in the renal WSIs are characterized by significant interstitial inflammation, tubular atrophy, fibrosis, and vascular alterations such as endothelialitis or arteriolar hyalinosis, which obscure anatomical boundaries and introduce morphological variability that complicates segmentation tasks.

All experiments were performed in accordance with relevant guidelines and regulations, and informed consent was obtained from all subjects and/or their legal guardian(s). This study was approved by the Research Ethics Committee of different hospitals (KUL approval number: S64006, LUMC approval number: W2020.031, AUMC approval number: 19.260, and UMCU approval number: 19.482).

Instance object annotations

In each ROI, a total of four classes were densely annotated until every pixel was assigned to a specific class: background, glomeruli, tubules, and arteries (including interlobular, intralobular arteries, and arterioles). The interstitium class, which encompasses interstitium / stroma and peritubular capillaries, was defined as the residual pixels after subtraction of the four aforementioned positive classes (background, glomeruli, tubules, and arteries). This approach ensured that all pixels were assigned a class label. To prevent the predicted tubules from merging due to proximity, particularly in relatively normal kidney biopsies, we introduced an auxiliary boundary class that was generated on the fly during training. Using morphological erosion, we eroded each boundary by 4 pixels along its compartment mask.

All cases were annotated with ASAP version 1.9 (https://computationalpathologygroup.github.io/ASAP) and our in-house custom WSI reader Slidescape (https://github.com/amspath/slidescape). The annotations were performed by trained physicians who are experienced with deep learning pipelines, and all annotations were validated by a nephropathologist with experience in computer vision algorithm training procedures. The label assignment for complex or challenging objects was determined through group discussion.

Pre-processing of WSIs for training

A resolution of 0.5 µm/pixel was chosen to extract ROIs from the annotated WSIs. Patches with fixed dimensions of 512 × 512 pixels were extracted from the ROIs using a sliding window stride of 256 pixels. For the edges of the ROIs, zero padding was added to round their sizes up to a multiple of the patch size and stride. If the patch images were smaller than 512 × 512 pixels, zero padding was used to extend them to the fixed dimensions. A total of 7,858 patch images were generated for the training dataset. During training, the padding regions were ignored. All patch images were randomly split into a training set (90%) and an internal validation set (10%). This

split was performed at both the patch level and the WSI level (case level). The patch-level split may introduce information leakage from the training procedure to the validation procedure, potentially leading to more stable training. In contrast, the WSI-level split creates a larger domain shift from training to validation, which might increase regularization and potentially improve out-of-domain generalization. These patch- and WSI-level train-validation split settings were directly compared using the KUL external test set, which was always left out of the training and validation processes.

Semantic segmentation model training

Two segmentation architectures were adopted in the current study. For CNN-based UNet network, we trained and tested the relatively shallow ResNet18 and the deeper ResNet50 encoder. For Transformer-based Mask2Former, we trained and tested encoders using ResNet50 and Swin-B as backbones. The code was forked from the original UNet (http://lmb.informatik.uni-freiburg.de/people/ronneber/u-net) and Mask2Former (https://github.com/facebookresearch/Mask2Former) implementations. We conducted minimal adjustments to the architectures for the current application. All model training schedules were initialized with ImageNet-pretrained encoder weights. We adopted the default UNet and Transformer decoder, default ADAMW optimizer with momentum 0.9, and default "WarmupPolyLR" learning rate scheduler with learning rate power 0.9. The UNet-ResNet models had a batch size of 16 and were trained for 50 epochs until convergence. Mask2Former models were trained with a batch size of 10 for 64 epochs until convergence. The Intersection over Union (IoU, i.e. Jaccard index) and pixel-level accuracy were used as an outcome metric, as the routine evaluation used by Mask2Former. A single NVIDIA RTX 4090 GPU was used for training all DL models.

Post-processing and generation of new instance segmentation predictions

To reconstruct ROI-level results from patch predictions, the predicted patches were generated with an overlap of 0.5 patch stride. Patch predictions were subsequently stitched by a weighted average of overlapping areas, with weights from the 2-dimensional Hann window function. The peaks near the center of the patch were attenuated to 0 at the edges. Pixel-level class predictions were rescaled using the *argmax()* function. Detectron2's built-in visualizer code was used for visualizations of the predictions. In order to reuse the slide-level segmentation masks for further downstream annotation in Slidescape (e.g. glomerulosclerosis lesions), we propose a process to obtain new ground-truth annotations from the predicted masks and save them in XML format. The process consists of the following functions: (1) For predicted masks of each class, OpenCV *findContours()* is used to extract object boundaries; (2) *approxPolyDP()* with epsilon = 1.0 is further utilized to approximate precise contours; (3) Then, inner contours are merged with outer contours to form a single object; (4) Lastly, objects with areas < 300 square pixels are filtered out. The auxiliary boundary class is ignored in the generation process.

Results

Cohorts and semantic segmentation models

The details of the training, internal validation, and external test cohorts are depicted in Fig. 1A. As shown in Fig. 1B, we employed both CNN-based and Transformer-based models to conduct a comparative analysis of segmentation performance, focusing on robustness against nephropathology heterogeneity. UNet-ResNet18, UNet-ResNet50, M2F-ResNet50, and M2F-SwinB were trained to segment kidney anatomical compartments, including background, glomeruli, tubules, arteries (inter-, intralobular arteries and arterioles altogether), and interstitium as a residual class. The baseline characteristics of the cohorts can be found in Table 1.

Mask2former-based networks outperform UNet-based networks on the internal validation dataset and external test dataset

UNet-based architectures have emerged as the de facto standard models in the vast majority of biomedical computer vision segmentation studies. Interestingly, in UNet-based architectures, when training and internal validation data were split at the patch-level (i.e. patches from the same WSI were shared between training and validation data), we observed that a more shallow ResNet18 encoder outperformed a deeper ResNet50 encoder across all anatomical compartments, achieving an average IoU (A-IoU) 0.84 and 0.72, respectively (Fig. 2A). However, when the training and validation data were split at the WSI level, the IoU scores decreased across all anatomical compartments. In this scenario, the UNet-ResNet18 model achieved a slightly higher A-IoU score of 0.48 compared to the UNet-ResNet50 model, which had A-IoU of 0.41, although the difference was marginal. Moreover, the simpler ResNet18 encoder continued to outperform in individual compartments (Fig. 2B).

On the other hand, at the patch level, both Mask2Former models outperformed the UNet models by a margin, which includes Mask2Former models with the ResNet50 (A-IoU of 0.95) and SwinB (A-IoU of 0.94) backbone (Fig. 2A). This trend was consistent across all histological compartments (Fig. 2A). When the data were split at the WSI level, the IoU scores for the UNet models dropped significantly. In contrast, the Mask2Former models maintained high IoU scores across all histological compartments. Notably, the fully Transformer-based M2F-SwinB model outperformed the CNN-Transformer hybrid M2F-ResNet50 model, with A-IoU of 0.84 compared to 0.79, respectively (Fig. 2B). Of particular interest is the higher performance of the M2F-SwinB model on the difficult-to-segment artery compartment, which is characterized by variable size and relatively low pixel-label prevalence compared to other classes. The M2F-SwinB model achieved an IoU of 0.66 in this compartment, compared to 0.53 for the M2F-ResNet50 model (Fig. 2B). Figure 2E, F presents the confidence intervals and standard deviations of IoU values (overall and individual objects) in the internal dataset with patch-level and WSI-level split strategies. We conclude that Transformer-based models achieved better performance on average for both patch-level and WSI-level split schemes.

Furthermore, we observed that the ResNet18 encoder outperformed the ResNet50 encoder across all histological compartments in the external test dataset, which was split at the patch level, achieving IoU scores

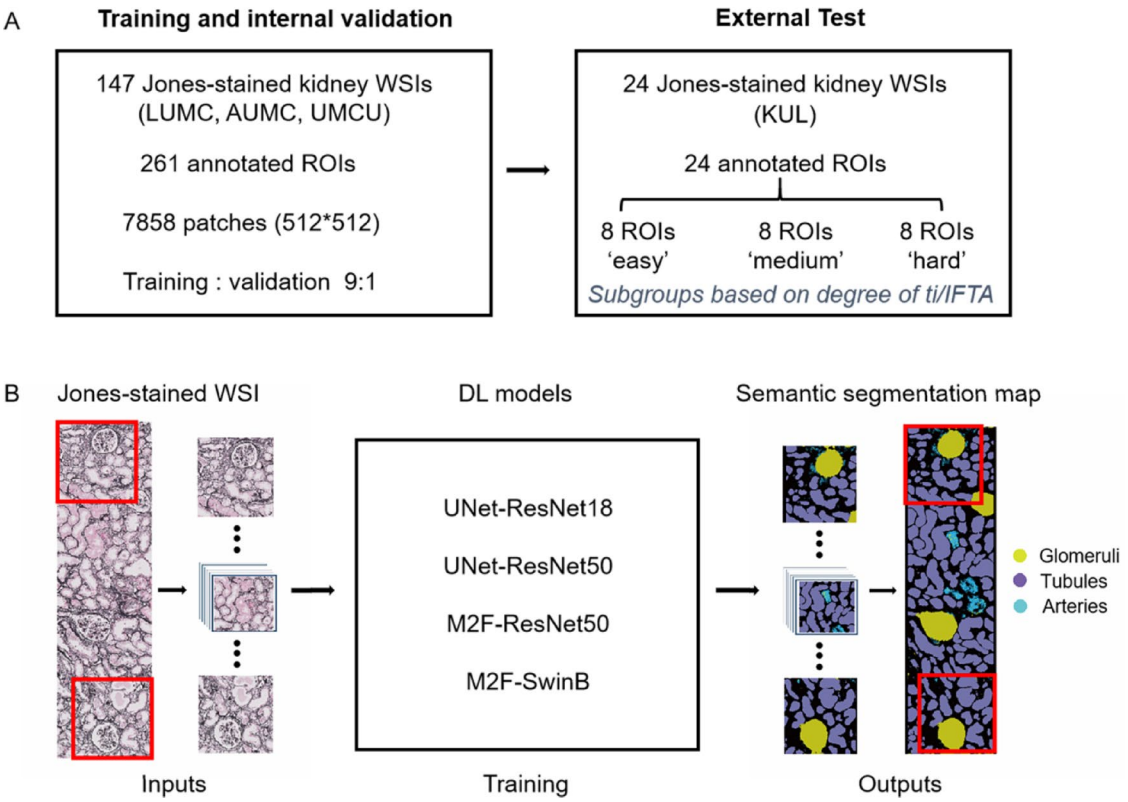


Fig. 1. Study design. (**A**) The details of the training, internal validation and external test datasets used in the current study. (**B**) The workflow of the current study. CNN-based (UNet-ResNet18, UNet-ResNet50) and transformer-based (M2F-ResNet50, M2F-SwinB) architectures were trained to segment the kidney anatomical compartments on Jones-stained renal WSIs. M2F: Mask2Former.

of 0.79 and 0.68, respectively (Fig. 2C). However, With WSI-level splitting of the test data, IoU scores decreased uniformly across all compartments, consistent with the internal validation results (Fig. 2D). The accuracy exhibited a similar trend of IoU scores across both the internal and external datasets (Figure S2). Compared to the WSI-level split, the patch-level split training strategy in both the internal and external datasets exhibited better stability. This suggests that the patch-level split may introduce some information leakage from the training procedure to the validation procedure. Nonetheless, the WSI-level split creates a larger domain shift from training to validation dataset, which might be the main reason for a decline in performance across all compartments.

Mask2Former-based models outperform UNet-based models in challenging areas from inflamed and fibrotic renal biopsies

The external dataset was categorized according to the degree of ti-scores and IFTA-scores. Mask2Former models demonstrated superior performance compared to the UNet models, consistently achieving higher IoU (Fig. 3) and accuracy (Figure S3) scores across all subgroups. Interestingly, in this multiclass segmentation task, glomerular segmentation appeared to be difficult for the UNet models when training and validation datasets were split at the WSI-level. This difficulty was particularly pronounced for the model employing the larger ResNet50 encoder (Fig. 3H). This observation suggests that UNet-based models may be susceptible to a strong inductive bias, potentially due to the between- and within-case pixel-label imbalance for the glomerular class.

The UNet models encountered particular challenges in identifying arteries in biopsies with higher degrees of inflammation and fibrosis. On the external test dataset, the arterial segmentation performance (IoU) of Mask2Former models exhibited a gradual decline under both patch-level (Fig. 3D) and WSI-level (Fig. 3J) splitting strategies as ti-scores and IFTA-scores increased, though this decrease was less pronounced than that observed in UNet models. For arterial segmentation in the hard group, M2F-SwinB (IoU = 0.54) and M2F-ResNet50 ((IoU = 0.50) outperformed UNet-ResNet18 ((IoU = 0.15) and UNet-ResNet50 ((IoU = 0.03) (Fig. 3J). Representative ground truth annotations and model prediction masks for the easy, medium, and hard groups are visualized in Fig. 3M. To better understand the importance of domain shift ability on the performance of semantic segmentation against the challenging areas, we visualized CNN-based and Transformer-based models on the patch-level and WSI-level schemes, as depicted in Figure S4. Note that the black starts represent the challenge areas in the medium and hard group (i.e., the abnormal areas) against the easy group (i.e., normal areas). It is observed that mask2former can transfer learn patterns across different modals. Figure S6 showed the attention heatmaps of glomeruli, tubules, arteries, and interstitium from Mask2Former-Swin.

	Total	AUMC	LUMC	UMCU	KUL	KUL	KUL
Difficulty group	–	–	–	–	Easy	Medium	Hard
Datasets	Train + Val	Train + Val	Train + Val	Train + Val	Test	Test	Test
N (WSIs)	147	49	58	40	8	8	8
Glomeruli, median (IQR)	20 (15–26)	22 (18–32)	19 (13–24)	19 (13–23)	24 (15–30)	27 (22–36)	23 (17–31)
Glomerulosclerosis, median % (IQR)	31(0–29)	11 (3–31)	5 (0–18)	15 (6–32)	10 (4–13)	23 (9–31)	35 (15–41)
TMA, N (%)	30 (20)	11 (22)	16 (28)	3 (8)	0 (0)	0 (0)	0 (0–0)
ATN/ATI, N (%)	83 (56)	28 (57)	40 (69)	15 (38)	2 (25)	4 (50)	4 (50)
g-score, median (IQR)	0 (0–1)	0 (0–2)	0 (0–1)	0 (0–0)	0 (0–0)	0 (0–0)	0 (0–0)
cg-score, median (IQR)	0 (0–1)	1 (0–2)	0 (0–1)	0 (0–0)	0 (0–0)	0 (0–0)	0 (0–1)
mm-score, median (IQR)	0 (0–1)	0 (0–1)	0 (0–0)	0 (0–1)	0 (0–0)	1 (0–3)	0 (0–0)
i-score, median (IQR)	1 (0–2)	0 (0–1)	1 (0–2)	1 (0–1)	0 (0–0)	0 (0–1)	0 (0–2)
t-score, median (IQR)	0 (0–2)	0 (0–1)	1 (0–2)	1 (0–2)	1 (0–1)	1 (0–1)	1 (1–2)
ti-score, median (IQR)	2 (1–3)	2 (1–3)	2 (1–3)	2 (1–3)	1 (0–1)	2 (1–3)	2 (2–2)
IFTA-score, median (IQR)	1 (1–2)	1 (1–3)	1 (0–2)	2 (1–3)	1 (1–1)	2 (2–2)	2 (1–2)
i-IFTA-score, median (IQR)	2 (1–3)	3 (1–3)	2 (0–3)	2 (1–2)	1 (0–1)	3 (2–3)	2 (2–2)
t-IFTA-score, median (IQR)	1 (0–1)	1 (0–2)	1 (0–1)	1 (0–1)	0 (0–0)	2 (1–2)	1 (1–1)
ptc-score, median (IQR)	0 (0–0)	0 (0–0)	0 (0–1)	0 (0–1)	0 (0–0)	0 (0–0)	0 (0–0)
v-score, median (IQR)	0 (0–0)	0 (0–0)	0 (0–1)	0 (0–0)	0 (0–0)	0 (0–0)	0 (0–0)
cv-score, median (IQR)	1 (0–2)	1 (0–2)	1 (0–2)	1 (1–2)	1 (0–1)	2 (1–2)	1 (0–1)
ah-score, median (IQR)	1 (0–2)	1 (0–1)	1 (0–2)	1 (1–2)	1 (0–1)	2 (2–3)	2 (1–2)

Table 1. Baseline Banff lesion scores of the cohorts. WSIs, whole slide images; IQR, interquartile range; TMA, thrombotic microangiopathy; ATN/ATI, acute tubular necrosis/acute tubular injury; g-score, Glomerulitis; cg-score, Transplant glomerulopathy; mm-score, Mesangial proliferation; i-score, Interstitial inflammation; t-score, Tubulitis; IFTA-score, Interstitial fibrosis/tubular atrophy; i-IFTA-score, Interstitial inflammation in in areas of IFTA; t-IFTA-score, Tubulitis in areas of IFTA; ptc-score, Peritubular capillaritis; v-score, Endothelialitis; cv-score, Vascular intima thickening; ah-score, Arterial hyalinosis score.

In Table S1, we included extra metrics and additional benchmark models for a thorough comparison. Under the semantic segmentation task, Transformer-based models achieve better DICE, precision, and recall scores than CNN-based models (including UNet and FLASH) but at the cost of higher FLOP, parameter amounts, and lower inference speed. Likewise, under the instance segmentation for glomeruli, M2F-ResNet50 outperforms MESCnn except for inference speed. Moreover, we conducted an ablation study to show attention mechanism can ensure a more robust modal/domain shift ability than convolution operation, as depicted in Table S2.

Extension of the basic anatomical segmentation model with pathological lesions

Next, we performed an extension experiment to show a proof-of-concept for how researchers can build their own models on top of the basic anatomical segmentation model. Given that the two Mask2Former models exhibited superior performance in segmenting the basic anatomical compartments, we seleScted these models for retraining a multiclass segmentation model that includes tubular atrophy and glomerulosclerosis. Figure 4A, B illustrate the performance of the two Mask2Former models on the 7-class semantic segmentation task. These results indicate a consistent trend of improved performance of the fully Transformer-based M2F-SwinB model (AC-IoU of 0.76) compared to the hybrid M2F-ResNet50 model (AC-IoU of 0.73) in the internal dataset. Specifically, the M2F-SwinB model achieved IoU scores of 0.75 for glomerulosclerosis and 0.55 for tubular atrophy in the internal dataset. In the external dataset, it achieved IoU scores of 0.84 for glomerulosclerosis and 0.38 for tubular atrophy. Examples of regions with tubular atrophy predicted by the M2F-SwinB model are illustrated in Fig. 4C, where there are failure cases with red arrows. The accuracy scores of Mask2Former exhibited a similar trend of IoU scores on segmenting the normal and abnormal kidney anatomical compartments, as shown in Figure S5.

Histological scores that correlate with low segmentation accuracy

To better understand which histological factors negatively impact segmentation performance for the M2F-SwinB model, we performed linear regression analysis on (1—IoU) values calculated for each of the segmented compartments (macro-level, glomeruli, tubules, vessels, interstitium) across the 24 test cases (Table 2). On the macro-level, the presence of acute tubular injury was the only factor significantly correlated with a lower IoU (beta = 0.09, SE = 0.001, $p = 0.002$). For glomerular segmentation, a lower IoU correlated to a lower number of glomeruli in the biopsy (beta = – 0.006, SE = 0.002, $p = 0.01$). For tubular segmentation, a lower IoU correlated to a higher percentage of sclerosed glomeruli (beta = 0.003, SE = 0.001, $p = 0.01$), a higher degree of endothelialitis (beta = 0.21, SE = 0.07, $p = 0.008$) and to a higher degree of vascular intima thickening (beta = 0.04, SE = 0.02, $p = 0.07$). For vascular segmentation, a lower IoU correlated to the presence of acute tubular injury (beta = 0.27, SE = 0.11, $p = 0.02$) and a higher degree of arteriolar hyalinosis (beta = 0.10, SE = 0.05, $p = 0.05$). For interstitium

segmentation, a lower IoU correlated with a lower percentage of sclerosed glomeruli (beta = − 0.003, SE = 0.002, $p = 0.07$), a lower degree of interstitial inflammation (beta = − 0.06, SE = 0.02, $p = 0.03$), a lower degree of tubulitis (beta = − 0.05, SE = 0.03, $p = 0.07$), a lower degree of total inflammation (beta = − 0.06, SE = 0.02, $p = 0.007$), a lower degree of tubulitis in areas of IFTA (beta = − 0.06, SE = 0.03, $p = 0.05$), and a lower degree of arteriolar hyalinosis (beta = − 0.04, SE = 0.02, $p = 0.07$). Notably, the presence of acute tubular injury was the only factor significantly correlated with lower IoU across all four compartments.

Discussion

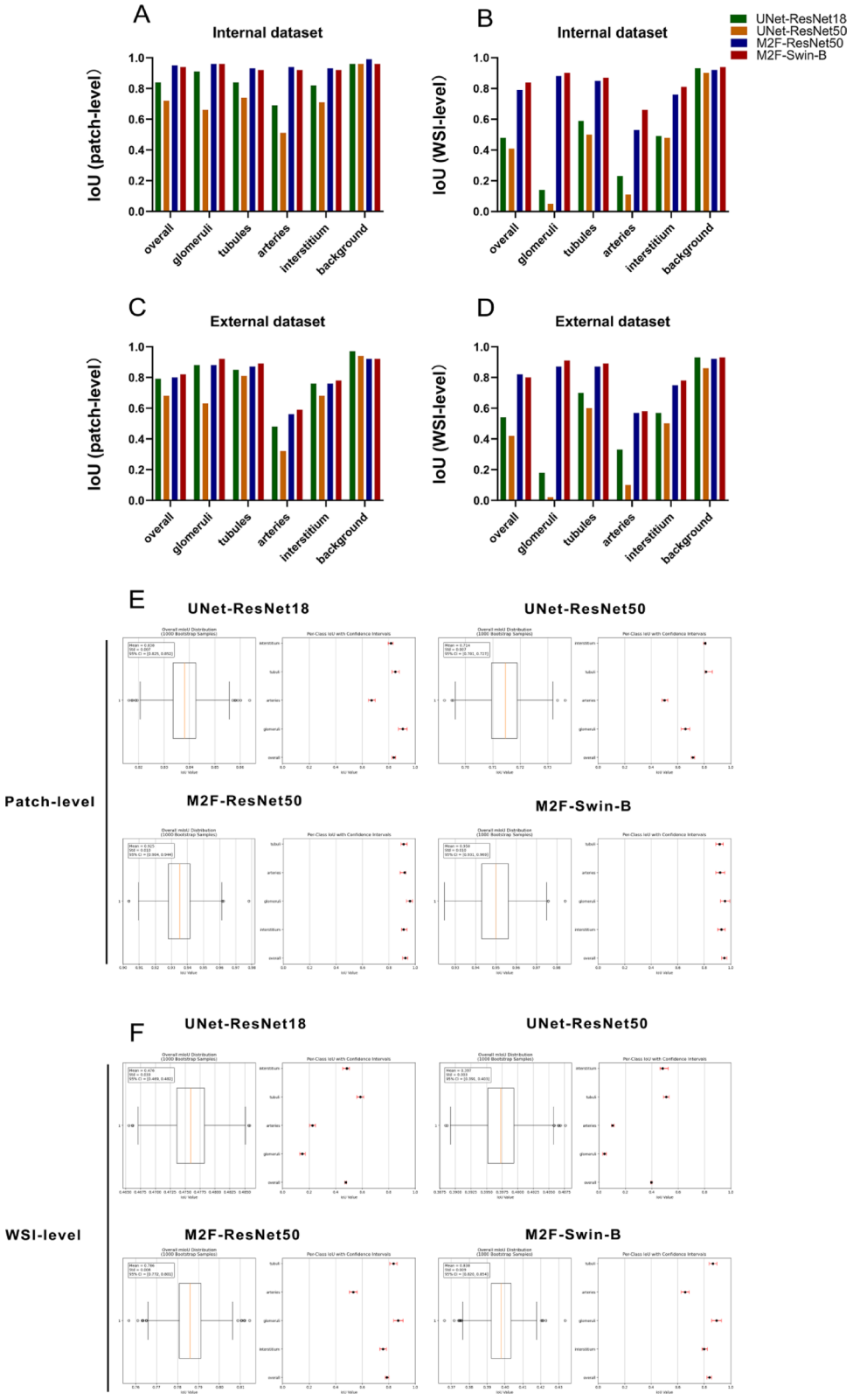
In the current study, we developed a segmentation pipeline for kidney histopathology and compared the modality/domain shift capability between CNN-based and Transformer-based architectures. By employing both WSI-level and patch-level splits across multi-center datasets, we observed that Transformer- based Mask2Former architectures, particularly those equipped with a Swin-B Transformer encoder, consistently outperformed CNN-based UNet models in segmenting challenging pathological features.

Prior to further discussion, it is essential to emphasize that all models were trained and tested on the basis of patches. First, we extracted the patches of interest from all WSIs during our data preparation process. The key difference between the two splitting schemes lies in how to arrange samples (i.e. patches) for training and test phases. As depicted in Figure S1A, the patch-level split scheme uniformly assigns samples across different modals (e.g., various medical centers) into both training and test datasets. That is the reason that we interpret the patch-level split scheme as a case of having adequate data, as a single model can access and learn all potential patterns. However, this scheme can bring false estimation on the versatility of one model as information leakage primarily facilitates pattern recognition across multiple models. In contrast, the WSI-level split scheme ensures that each WSI is exclusively assigned to either the training or validation phase, with all affiliated samples allocated solely to the training or test dataset, as depicted in Figure S1B. Hence, this scheme results in higher variance due to domain or modal shifts, which can be solid evidence for comparing the versatility among different models. That is why we interpret the WSI-level split scheme as a case of having limited data, requiring models to generalize better across domains.

After extensive evaluation using both internal and external datasets, we observed that Transformer-based models significantly outperformed CNN-based counterparts on WSI-level splitting, showing the strong domain transfer ability for large in-group data variations. In contrast, CNN-based models achieved comparable performance on patch-level splitting with lower complexity. This contrastive approach not only illuminates the adaptability of different network architectures but also provides a benchmark for selecting DL models, contingent upon the availability of annotated datasets. In practical applications, data scarcity often poses a significant challenge, particularly within the specialized field of renal pathology. Given these constraints, Transformer-based Mask2Former architecture, with its robust attention mechanism, emerges as a more suitable choice, especially in scenarios where annotated data is limited. In this study, we also found that the Mask2Former models outperformed UNet models in segmenting arteries, which are noted to be small and have low pixel-label prevalence. In brief, Mask2Former leverages global self-attention mechanisms, enabling it to model long-range dependencies and contextual relationships within the data. This capability allows the model can focus on rare, small structures like arteries by integrating information from larger surrounding regions. In contrast, local convolutional operations in U-Net may fail to detect these structures due to their limited receptive field. One limitation of the current study is that our models were trained exclusively on Jones-silver stains. Future work will explore stain normalization and assess the adaptability of our models to other types of stains.

IFTA is a crucial morphologic prognostic indicator of chronic kidney disease progression, regardless of the underlying cause[28–36]. However, manually assessing the extent of atrophic tubules and fibrosis in renal biopsies is impractical in clinical practice. IFTA presents a continuous morphologic spectrum, so grading IFTA is a semi-quantitative, manual process with significant variability among pathologists[1,37–40]. DL models offer a potential solution by providing a quantitative approach to estimate IFTA, thereby reducing inter-observer variability. In this study, a senior nephropathologist reviewed an external dataset comprising 24 WSIs and provided IFTA scores. As anticipated, model performance exhibited a gradual decline with increasing IFTA severity. This trend reflects the challenges that pathologists encounter when assessing IFTA in renal biopsies, where diagnostic consistency often deteriorates as fibrotic changes become more severe. It is important to note that the majority of existing DL models are trained on limited biopsy specimens, lacking the full spectrum of renal pathology. While these models achieve high accuracy in their test cohorts, they often struggle with complex regions such as fibrosis and inflammation. In contrast, the M2F-SwinB model demonstrated high accuracy across different IFTA grades (easy, medium, and hard) in the external dataset, highlighting its robust generalizability. Hence, our model has the potential to be seamlessly integrated into clinical workflows, thereby enabling rapid and quantitative IFTA scoring.

In AI segmentation algorithms, mIoU is a crucial metric that evaluates model performance by averaging the performance across all categories. The contribution of mIoU from each category to the collective outcome is substantial. If the model performs poorly in one or several categories, it can adversely impact the average class mIoU, despite good performance in other categories. In this study, on the macro-level, only acute tubular injury was found to correlate with a lower IoU across all 4 compartments. Acute tubular injury often results in several morphological changes, such as epithelial flattening, loss of brush borders, and the presence of luminal debris. These alterations collectively diminish the visual distinction between tubules and the interstitium. As a result, these morphological changes compromise the ability of both CNNs and Transformers to recognize typical tubular patterns, ultimately leading to lower segmentation performance. Therefore, in practical applications, optimizing the model for the acute tubular injury category could potentially enhance the average class segmentation performance. However, a deeper understanding of the histological parameters that negatively affect segmentation performance might identify case characteristics that can provide an active learning signal



for future iterations of model improvement. These data suggest that in the setting of active learning, including new annotations with high degrees of total inflammation (ti-score) and interstitial fibrosis and tubular atrophy (IFTA-score) could negatively impact the IoU for interstitial segmentation or vice versa.

Furthermore, building on the basic anatomy segmentation model allows us to easily scale the model in complexity by adding classes that represent lesions. In this study, we have developed code that converts the predicted segmentation maps generated from the M2F-Swin-B into XML files, which can be easily edited by the open-source software *Slidescape*. We have built on the concept of the human-AI-loop (HAIL)[41], utilizing the trained segmentation model to generate new annotations, and practically creating instance segmentation

◀**Fig. 2**. The performance of different DL models on the internal validation dataset with patch-level and WSI-level split. (**A–B**) Overall IoU and IoU across all compartments of four models (UNet-ResNet18, UNet-ResNet50, M2F-ResNet50 and M2F-SwinB) in the internal dataset with the patch-level split (**A**) and WSI-level split (**B**). (**C–D**) Overall IoU and IoU across all compartments of four models in the external dataset with the patch-level split (**C**) and WSI-level split (**D**). IoU: Intersection of Union; M2F:Mask2Former. (**E**) The confidence intervals and standard deviations of IoU values (overall and individual objects) in the internal dataset using path-level split strategy. (**F**) The confidence intervals and standard deviations of IoU values (overall and individual objects) in the internal dataset using WSI-level split strategy. IoU: Intersection of Union.

via post-processing of multiclass semantic prediction masks. Such an approach could inspire the community to extend the segmentation models with increased complexity without much effort, similar to the recently open-sourced zero-shot general-purpose Segment Anything Model (SAM) for natural images[42].

Limitations and future work

Several limitations of this study should be noted. First, the exclusive use of Jones-silver stained specimens for model training may limit the generalizability of our findings to other common staining protocols in renal pathology. Second, our framework shows robust anatomical segmentation capabilities. However, it does not yet incorporate disease-specific pathological lesions, which are crucial for comprehensive diagnostic applications.

To address these limitations and advance the field, we propose three key directions for future research. First, curate multi-center, multi-stain datasets representing the full spectrum of renal pathologies to improve model generalizability. Second, augment the framework to segment both anatomical structures and disease-specific pathological lesions (e.g., crescents, hyaline casts). Third, integrate the framework into clinical workflow using HAIL systems that synergize pathologist expertise with automated annotation capabilities, creating a virtuous cycle of model refinement and validation.

Conclusion

In conclusion, we developed an open-source segmentation model for the basic anatomical structures in kidney histology, encompassing both related normal and challenging pathological cases. By releasing the model weights, we provide researchers with a resource to accelerate future investigations in computational nephropathology. Furthermore, we demonstrated that Transformer-based models (particularly Mask2Former with Swin-B backbone) exhibit superior generalizability in data-limited scenarios, while UNet-based models can achieve comparable performance when sufficient training data is available. The developed framework provides a foundation for advancing quantitative renal pathology, offering potential to enhance both research and clinical practice through more precise, standardized, and reproducible tissue analysis. Future extensions of this work could further bridge the gap between computational research and diagnostic pathology by incorporating disease-specific lesion detection and validation in prospective clinical studies.

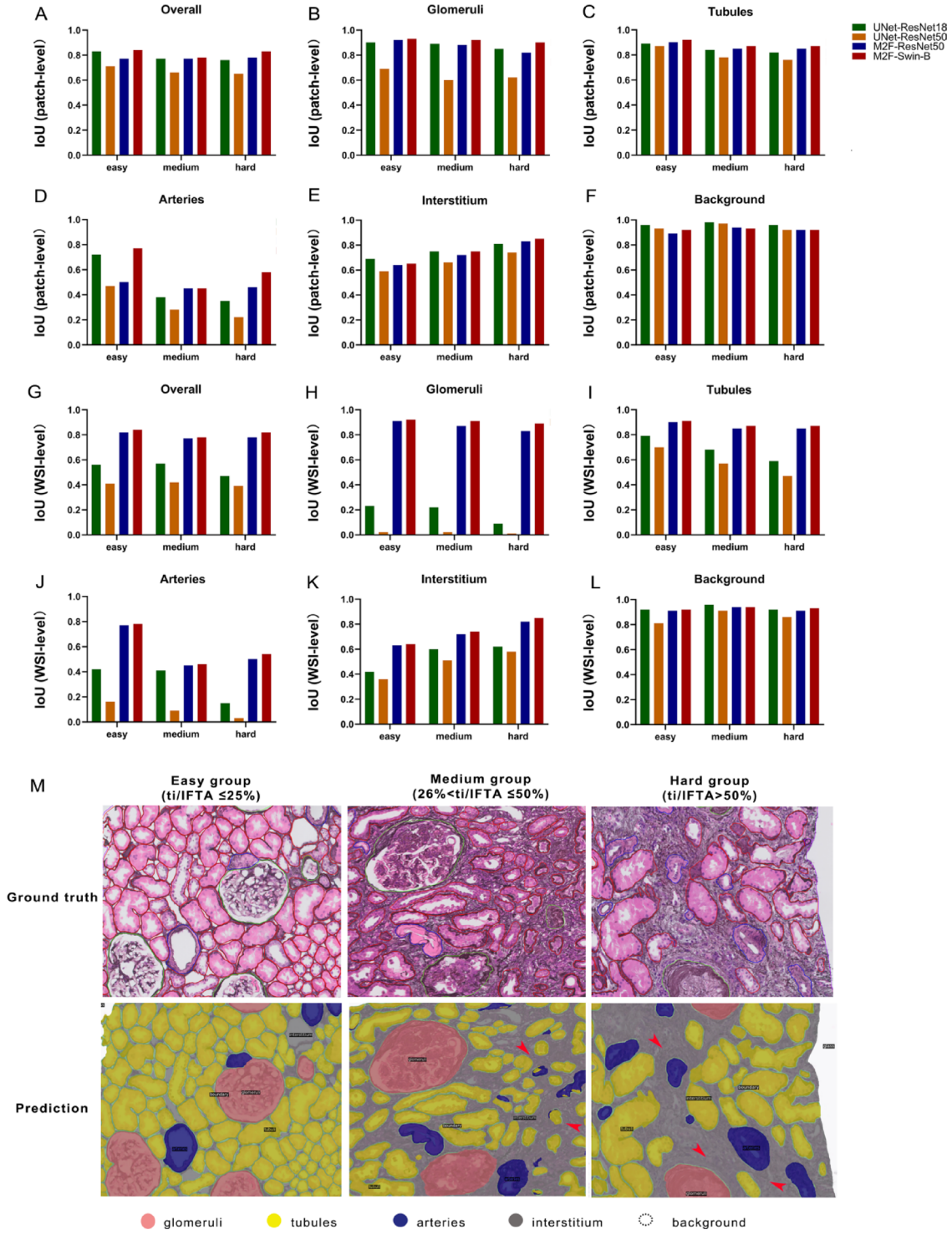


Fig. 3. The performance of different models on the subgrouped external dataset. (**A–F**) IoU of overall class, glomeruli, tubules, arteries, interstitium, and background in the subgrouped external dataset (easy group, medium group, hard group) for different models with the patch-level split. (**G–L**) IoU of overall class, glomeruli, tubules, arteries, interstitium, and background in the subgrouped external dataset (easy group, medium group, hard group) for different models with the WSI-level split. (**M**) The representative images of Jones-stained patches with ground truth annotations and the prediction masks generated by M2F-Swin-B in the subgrouped dataset. The red arrows indicate the challenge areas, which are characterized by significant interstitial inflammation, tubular atrophy, fibrosis, and vascular alterations such as endothelialitis or arteriolar hyalinosis. IoU: Intersection of Union; M2F: Mask2Former.

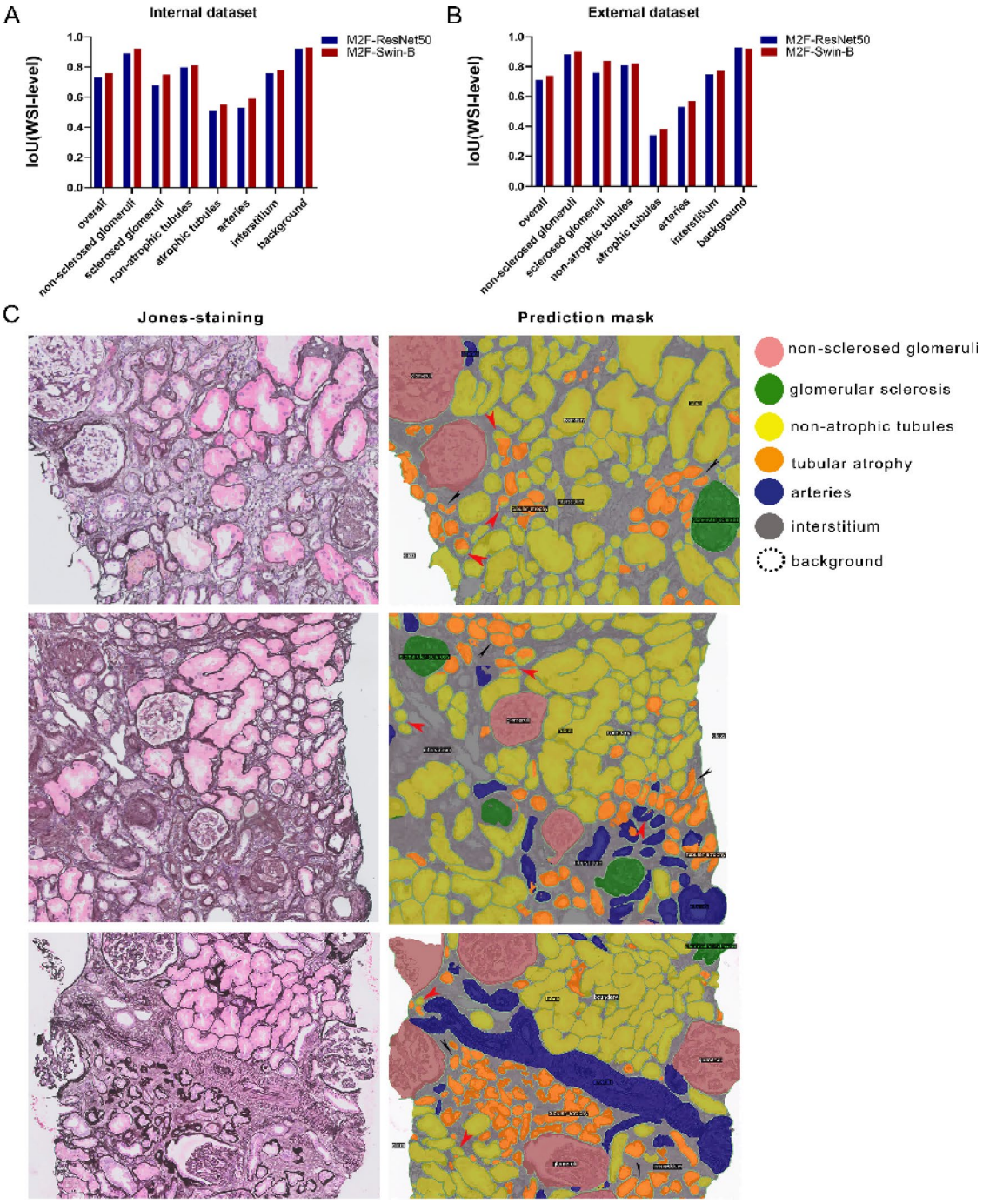


Fig. 4. The performance of Mask2Former on segmenting the normal and abnormal kidney anatomical compartments. (**A–B**) IoU of overall class, non-sclerosed glomeruli, sclerosed glomeruli, non-atrophic tubules, atrophic tubules, arteries, interstitium, background in two different Mask2Former models in the internal dataset (**A**) and external dataset (**B**) with the WSI-level split. (**C**) The representative images of Jones-stained patch images and the prediction mask images generated by M2F-Swin-B model. The red arrows indicate the failure cases of tubular atrophy predictions and the black arrows indicate the success cases of tubular atrophy predictions. IoU: Intersection of Union; M2F: Mask2Former.

Histological parameter	Macro *p* value	Glomeruli*p* value	Tubules *p* value	Vessels *p* value	Interstitium *p* value
Number of glomeruli	0.94 (+)	**0.01** (−)	0.17 (+)	0.93 (−)	0.10 (+)
Percentage of sclerosed glomeruli	0.73 (+)	0.53 (−)	**0.01** (+)	0.25 (+)	**0.07** (−)
Acute tubular injury	**0.002** (+)	0.12 (+)	0.36 (+)	**0.02** (+)	0.51 (+)
Glomerulitis (g-score)	0.23 (−)	0.48 (−)	0.77 (−)	0.21 (−)	0.64 (+)
Transplant glomerulopathy (cg-score)	0.57 (−)	0.82 (+)	0.66 (−)	0.50 (−)	0.52 (−)
Mesangial proliferation (mm-score)	0.94 (−)	0.40 (−)	0.67 (−)	0.68 (+)	0.50 (−)
Interstitial inflammation (i-score)	0.99 (+)	0.25 (+)	0.78 (−)	0.35 (+)	**0.03** (−)
Tubulitis (t-score)	0.17 (−)	0.92 (+)	0.88 (+)	0.58 (−)	**0.07** (−)
Total inflammation (ti-score)	0.84 (−)	0.60 (+)	0.18 (+)	0.59 (+)	**0.007** (−)
Interstitial fibrosis/tubular atrophy (IFTA-score)	0.16 (+)	0.76 (+)	0.27 (+)	0.20 (+)	0.49 (−)
Interstitial inflammation in areas of IFTA (i-IFTA-score)	0.91 (+)	0.32 (−)	0.51 (+)	0.27 (+)	0.28 (−)
Tubulitis in areas if IFTA (t-IFTA-score)	0.27 (−)	0.33 (−)	0.82 (−)	0.93 (+)	**0.05** (−)
Peritubular capillaritis (ptc-score)	0.32 (+)	0.99 (+)	0.51 (−)	0.70 (+)	0.24 (+)
Endothelialitis (v-score)	0.35 (+)	0.71 (−)	**0.008** (+)	0.61 (−)	0.35 (+)
Vascular intima thickening (cv-score)	0.28 (+)	0.45 (+)	**0.07** (+)	0.80 (+)	0.77 (−)
Arteriolar hyalinosis (ah-score)	0.36 (+)	0.96 (−)	0.19 (+)	**0.05** (+)	**0.07** (−)

Table 2. Histological factors contributing to a lower M2F-SwinB segmentation performance. A linear regression analysis on (1-IoU) values calculated for each of the segmented compartments (macro-level, glomeruli, tubules, vessels, interstitium) across the 24 test cases was performed. The presence of acute tubular injury was the only factor significantly correlated with lower IoU across all four compartments. *P* values in bold represent correlations trends with $P < 0.10$. (+) indicates a positive association and (−) indicates a negative correlation.

Data availability

The source code and model weights are publicly accessible on GitHub (https://github.com/amspath). Annotated whole slide images are available upon request at the corresponding author after setting up a data transfer agreement with each of the university medical centers.

Received: 19 March 2025; Accepted: 1 October 2025
Published online: 06 November 2025

References

1. Furness, P. N. et al. International variation in histologic grading is large, and persistent feedback does not improve reproducibility. *Am. J. Surg. Pathol.* **27**, 805–810. https://doi.org/10.1097/00000478-200306000-00012 (2003).
2. Haas, M. et al. Consensus definitions for glomerular lesions by light and electron microscopy: Recommendations from a working group of the renal pathology society. *Kidney Int.* **98**, 1120–1134. https://doi.org/10.1016/j.kint.2020.08.006 (2020).
3. Pantanowitz, L. et al. Twenty years of digital pathology: An overview of the road travelled, what is on the horizon, and the emergence of vendor-neutral archives. *J. Pathol. Inform.* **9**, 40. https://doi.org/10.4103/jpi.jpi_69_18 (2018).
4. Lee, J. G. et al. Deep learning in medical imaging: General overview. *Korean J. Radiol.* **18**, 570–584. https://doi.org/10.3348/kjr.2017.18.4.570 (2017).
5. Marée, R., Dallongeville, S., Olivo-Marin, J. C. & Meas-Yedid, V. In *IEEE 13th International Symposium on Biomedical Imaging (ISBI)* 1033–1036 (2016).
6. Temerinac-Ott, M. et al. In *Proceedings of the 10th International Symposium on Image and Signal Processing and Analysis*. 19–24.
7. Simon, O., Yacoub, R., Jain, S., Tomaszewski, J. E. & Sarder, P. Multi-radial LBP features as a tool for rapid glomerular detection and assessment in whole slide histopathology images. *Sci. Rep.* **8**, 2032. https://doi.org/10.1038/s41598-018-20453-7 (2018).
8. Bukowy, J. D. et al. Region-based convolutional neural nets for localization of Glomeruli in trichrome-stained whole kidney sections. *J. Am. Soc. Nephrol.* **29**, 2081–2088. https://doi.org/10.1681/ASN.2017111210 (2018).
9. Kato, T. et al. Segmental HOG: New descriptor for glomerulus detection in kidney microscopy image. *BMC Bioinform.* **16**, 316. https://doi.org/10.1186/s12859-015-0739-1 (2015).
10. Sarder, P., Ginley, B. G. & Tomaszewski, J. E. In *SPIE Medical Imaging*.
11. Ginley, B. et al. Computational segmentation and classification of diabetic glomerulosclerosis. *J. Am. Soc. Nephrol.* **30**, 1953–1967. https://doi.org/10.1681/ASN.2018121259 (2019).
12. Hermsen, M. et al. Deep learning-based histopathologic assessment of kidney tissue. *J. Am. Soc. Nephrol.* **30**, 1968–1979. https://doi.org/10.1681/ASN.2019020144 (2019).
13. Bueno, G., Fernandez-Carrobles, M. M., Gonzalez-Lopez, L. & Deniz, O. Glomerulosclerosis identification in whole slide images using semantic segmentation. *Comput. Methods Programs Biomed.* **184**, 105273. https://doi.org/10.1016/j.cmpb.2019.105273 (2020).
14. Zeng, C. et al. Identification of glomerular lesions and intrinsic glomerular cell types in kidney diseases via deep learning. *J. Pathol.* **252**, 53–64. https://doi.org/10.1002/path.5491 (2020).
15. Gallego, J. et al. A U-Net based framework to quantify glomerulosclerosis in digitized PAS and H&E stained human tissues. *Comput. Med. Imaging Graph* **89**, 101865. https://doi.org/10.1016/j.compmedimag.2021.101865 (2021).
16. Nguyen, T. T. U. et al. Deep-learning model for evaluating histopathology of acute renal tubular injury. *Sci. Rep.* **14**, 9010. https://doi.org/10.1038/s41598-024-58506-9 (2024).
17. Jiang, L. et al. A deep learning-based approach for Glomeruli instance segmentation from multistained renal biopsy pathologic images. *Am. J. Pathol.* **191**, 1431–1441. https://doi.org/10.1016/j.ajpath.2021.05.004 (2021).
18. Vaswani, A. et al. In *Neural Information Processing Systems*.

19. Cheng, B., Misra, I., Schwing, A. G., Kirillov, A. & Girdhar, R. Masked-attention mask transformer for universal image segmentation 2022. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 1280–1289 (2021).
20. Jayapandian, C. P. et al. Development and evaluation of deep learning-based segmentation of histologic structures in the kidney cortex with multiple histologic stains. *Kidney Int.* **99**, 86–101. https://doi.org/10.1016/j.kint.2020.07.044 (2021).
21. Holscher, D. L. et al. Next-generation morphometry for pathomics-data mining in histopathology. *Nat. Commun.* **14**, 470. https://doi.org/10.1038/s41467-023-36173-0 (2023).
22. Altini, N. et al. Performance and limitations of a supervised deep learning approach for the histopathological Oxford Classification of glomeruli with IgA nephropathy. *Comput Methods Programs Biomed.* **242**, 107814. https://doi.org/10.1016/j.cmpb.2023.107814 (2023).
23. Hussain, T. & Shouno, H. Explainable deep learning approach for multi-class brain magnetic resonance imaging tumor classification and localization using gradient-weighted class activation mapping. *Information* **14**, 642 (2023).
24. Hussain, T., Shouno, H., Mohammed, M. A., Marhoon, H. A. & Alam, T. DCSSGA-UNet: Biomedical image segmentation with DenseNet channel spatial and semantic guidance attention. *Knowl. Based Syst.* **314**, 113233 (2025).
25. Hussain, T. & Shouno, H. MAGRes-UNet: Improved medical image segmentation through a deep learning paradigm of multi-attention gated residual U-Net. *IEEE Access* **12**, 40290–40310 (2024).
26. Alam, T. et al. An integrated approach using YOLOv8 and ResNet, SeResNet & vision Transformer (ViT) algorithms based on ROI fracture prediction in X-ray images of the elbow. *Curr. Med. Imaging* **20**, e15734056309890 (2024).
27. Loupy, A., Mengel, M. & Haas, M. Thirty years of the International Banff Classification for allograft pathology: The past, present, and future of kidney transplant diagnostics. *Kidney Int.* **101**, 678–691. https://doi.org/10.1016/j.kint.2021.11.028 (2022).
28. Okada, T. et al. Histological predictors for renal prognosis in diabetic nephropathy in diabetes mellitus type 2 patients with overt proteinuria. *Nephrology (Carlton)* **17**, 68–75. https://doi.org/10.1111/j.1440-1797.2011.01525.x (2012).
29. Tervaert, T. W. et al. Pathologic classification of diabetic nephropathy. *J. Am. Soc. Nephrol.* **21**, 556–563. https://doi.org/10.1681/ASN.2010010010 (2010).
30. Weening, J. J. et al. The classification of glomerulonephritis in systemic lupus erythematosus revisited. *Kidney Int.* **65**, 521–530. https://doi.org/10.1111/j.1523-1755.2004.00443.x (2004).
31. Working Group of the International Ig, A. N. N. *et al.* The Oxford classification of IgA nephropathy: Pathology definitions, correlations, and reproducibility. *Kidney Int.* **76** 546–556, https://doi.org/10.1038/ki.2009.168 (2009).
32. Sethi, S. et al. A proposal for standardized grading of chronic changes in native kidney biopsy specimens. *Kidney Int.* **91**, 787–789. https://doi.org/10.1016/j.kint.2017.01.002 (2017).
33. Srivastava, A. et al. The prognostic value of histopathologic lesions in native kidney biopsy specimens: Results from the Boston kidney biopsy cohort study. *J. Am. Soc. Nephrol.* **29**, 2213–2224. https://doi.org/10.1681/ASN.2017121260 (2018).
34. Seron, D. & Moreso, F. Protocol biopsies in renal transplantation: prognostic value of structural monitoring. *Kidney Int.* **72**, 690–697. https://doi.org/10.1038/sj.ki.5002396 (2007).
35. Cosio, F. G., El Ters, M., Cornell, L. D., Schinstock, C. A. & Stegall, M. D. Changing kidney allograft histology early posttransplant: Prognostic implications of 1-Year protocol biopsies. *Am. J. Transplant.* **16**, 194–203. https://doi.org/10.1111/ajt.13423 (2016).
36. Myllymaki, J., Saha, H., Mustonen, J., Helin, H. & Pasternack, A. IgM nephropathy: Clinical picture and long-term prognosis. *Am. J. Kidney Dis.* **41**, 343–350. https://doi.org/10.1053/ajkd.2003.50042 (2003).
37. Farris, A. B. et al. Banff fibrosis study: Multicenter visual assessment and computerized analysis of interstitial fibrosis in kidney biopsies. *Am. J. Transplant.* **14**, 897–907. https://doi.org/10.1111/ajt.12641 (2014).
38. Snoeijs, M. G. et al. Histological assessment of pre-transplant kidney biopsies is reproducible and representative. *Histopathology* **56**, 198–202. https://doi.org/10.1111/j.1365-2559.2009.03469.x (2010).
39. Grootscholten, C. et al. Interobserver agreement of scoring of histopathological characteristics and classification of lupus nephritis. *Nephrol. Dial. Transplant.* **23**, 223–230. https://doi.org/10.1093/ndt/gfm555 (2008).
40. Gough, J. et al. Reproducibility of the Banff schema in reporting protocol biopsies of stable renal allografts. *Nephrol. Dial. Transplant.* **17**, 1081–1084. https://doi.org/10.1093/ndt/17.6.1081 (2002).
41. Lutnick, B. et al. An integrated iterative annotation technique for easing neural network training in medical image analysis. *Nat. Mach. Intell.d* **1**, 112–119. https://doi.org/10.1038/s42256-019-0018-3 (2019).
42. Kirillov, A. et al. Segment anything 2023. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 3992–4003 (2023).

Acknowledgements

This work was supported by grants from Chongqing Science and health joint medical research project (general program, 2024MSXM017) and the New Chongqing Talent Attraction Program Research Project (CSTB2024Y-CJH-KYXM0062).

Author contributions

JH and JK designed this study. JK provided critical intellectual input into the manuscript with revisions. JH, JLong and JLi annotated the WSIs, and JK reviewed and corrected the annotations. SF, MN, PK, TQN, SM, AP-JdV, OJdB collected renal WSIs and performed histological scores according to the Banff lesion scoring system. JH, PAV and ZX trained and validated the DL-models. PAV updated the annoation software-slidescape, JH, PAV and ZX interpreted the data and wrote the first draft of the manuscript. FJV performed the data visualization. All authors contributed to data sorting, and read and approved the final version to be published.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-22814-5.

Correspondence and requests for materials should be addressed to J.K.