Deep Learning Segmentation of Glomeruli on Kidney Donor Frozen Sections

Richard C. Davis, MD, MSPH,^{1,*} Xiang Li, MS,^{2,*} Yuemei Xu, MD^{1,3}, Zehan Wang, MA⁴, Nao Souma, MD, PhD⁵, Gina Sotolongo, MD¹, Jonathan Bell, BS¹, Matthew Ellis, MD^{5,6}, David Howell, MD, PhD¹, Xiling Shen, PhD⁴, Kyle Lafata, PhD^{2,7,8,**}, Laura Barisoni, MD^{1,5,**}

¹ Department of Pathology, Duke University, Durham, NC, USA

² Department of Electrical and Computer Engineering, Duke University, Durham, NC, USA

³ Department of Pathology, Nanjing Drum Tower Hospital, Nanjing, China

⁴ Department of Biomedical Engineering, Duke University, Durham, NC, USA

⁵ Department of Medicine, Division of Nephrology, Duke University, Durham, NC, USA

⁶ Department of Surgery, Duke University, Durham, NC, USA

⁷ Department of Radiation Oncology, Duke University, Durham, NC, USA

⁸ Department of Radiology, Duke University, Durham, NC, USA

ABSTRACT

Purpose: Recent advances in computational image analysis offer the opportunity to develop automatic quantification of histologic parameters as aid tools for practicing pathologists. This work aims to develop deep learning (DL) models to quantify non-sclerotic and sclerotic glomeruli on frozen sections from donor kidney biopsies.

Approach: A total of 258 whole slide images (WSI) from cadaveric donor kidney biopsies performed at our institution (n=123) and at external institutions (n=135) were used in this study. WSIs from our institution were divided at the patient level into training and validation datasets (Ratio: 0.8:0.2) and external WSIs were used as an independent testing dataset. Non-sclerotic (n=22767) and sclerotic (n=1366) glomeruli were manually annotated by study pathologists on all WSIs. A 9-layer convolutional neural network based on the common U-Net architecture was developed and tested for the segmentation of non-sclerotic and sclerotic glomeruli. DL-derived, manual segmentation and reported glomerular count (standard of care) were compared.

Results: The average Dice Similarity Coefficient testing was 0.90 and 0.83. and the F1, Recall, and Precision scores were 0.93, 0.96, and 0.90, and 0.87, 0.93, and 0.81, for non-sclerotic and sclerotic glomeruli, respectively. DL-derived and manual segmentation derived glomerular counts were comparable, but statistically different from reported glomerular count.

Conclusions: DL segmentation is a feasible and robust approach for automatic quantification of glomeruli. This work represents the first step toward new protocols for the evaluation of donor kidney biopsies.

** These authors contributed equally to the manuscript and are listed in alphabetical order.*

** Correspondence:

Laura Barisoni, MD (Laura.Barisoni@duke.edu)

Kyle J. Lafata, PhD (Kyle.Lafata@duke.edu)

1. Introduction

Renal allograft transplantation has superior long-term survival compared to dialysis¹. However, fewer allografts remain available for transplantation than the number of patients on the transplant waiting list²⁻⁴. To address this problem, the extended criteria donor (ECD) program was introduced in 2002 in the United States, which allowed for transplantation of allografts from cadaveric donors > 60 years of age and those with comorbidities². To increase the utilization of cadaveric kidneys and improve the prediction of their function, the ECD was later substituted by the Kidney Donor Profile Index (KDPI)^{5,6}. While the utility of morphologic parameters to predict allograft function and outcomes is controversial^{4,5,7-13}, histologic analyses of frozen sections of kidney wedge biopsies stained with hematoxylin and eosin (H&E) remain the current practice in North America for determining the suitability of cadaveric kidneys for transplantation.

Semi-quantitative assessment of interstitial fibrosis and inflammation, tubular atrophy and acute tubular injury, arterial intimal fibrosis and arteriolar hyalinosis, presence of intravascular thrombi or glomerular pathology, and percent of globally sclerotic glomeruli are routinely evaluated prior to implantation to prognosticate post-transplant kidney function¹⁴. The overall longevity of the allograft, however, depends also on other post-transplant superimposed events, such as episodes of T cell or antibody mediated rejection, measured by the Banff scores¹⁵, response to immunosuppression, viral infections, and other recipient-specific clinical conditions.

More recently, the i-Box, integrating demographic, clinical and morphologic data, was proven to be a useful tool to better predict allograft long term outcomes¹⁶. Semi-quantitative features, such as interstitial fibrosis, tubular atrophy, arteriosclerosis (i.e. arterial intimal fibrosis), arteriolar hyalinosis, interstitial inflammation, and percent of globally sclerotic glomeruli are routinely evaluated prior to implantation, often by pathologists on-call who do not always have a

kidney domain expertise^{3-5,8,9}, resulting in implantation of suboptimal donor renal allografts or improper discarding of otherwise suitable allografts^{3,5}. High inter- and intra-observer variability further complicate this paradigm^{14,17}.

Since the introduction of commercial whole slide scanners in 1998, digital pathology has become an increasingly important aspect of pathology workflows in both research and clinical practice¹⁸⁻²². Digital pathology enables computational image analysis to be applied to digitized tissue samples. In particular, deep learning (DL), a specific type of machine learning²³, is a useful tool for image representation and image analysis tasks²⁴. DL methods have been implemented for a wide array of digital pathology domains^{25,26}, such as cell detection²⁷ and segmentation²⁸, detection of breast cancer metastases in lymph nodes²⁹, and grading of gliomas³⁰. DL approaches have also been applied to kidney biopsies for the automatic detection of normal structures (e.g., glomeruli, urinary space, tubules, and vessels)²⁸, and abnormal structures (e.g., global sclerosis, interstitial fibrosis and tubular atrophy)^{14,31-36}, using WSIs derived from formalin-fixed and paraffin embedded sections. However, DL studied on frozen sections of kidney remain limited³⁷.

Here, we present a DL approach to automatically detect and segment sclerotic and non-sclerotic glomeruli on frozen sections of donor kidney biopsies. We anticipate that this initial pipeline can be enriched by the DL segmentation and quantification of other relevant histologic parameters, resulting into a robust interactive human-machine protocol for the assessment of donor kidney biopsies with potential scalability in clinical practice.

2. Materials and Methods

2.1 Whole Slide Imaging Dataset

This study was approved by the Duke University Institutional Review Board. A total 211 kidney donors deceased between January 2015 and January 2020 were included in the study, for a total of 268 frozen section H&E-stained slides. Of these, 75 donors had a wedge kidney biopsy performed, frozen, cut, and stained with Hematoxylin & Eosin (H&E) at Duke University Medical Center (DUMC) for a total of 128 frozen section H&E-stained glass slides (Internal cases). Of these 75 donors, 53 had bilateral biopsies and 22 had unilateral biopsies performed. The remaining 136 deceased donors had a wedge kidney biopsy performed, frozen, cut, and stained with H&E in other institutions (External cases) and subsequently reviewed at DUMC, for a total of 140 frozen section H&E-stained slides (Figure 1 - a1).

Whole slide images (WSIs) were acquired at 40X magnification on a Leica AT2 whole-slide scanner located in the Duke University Department of Pathology's BioRepository and Precision Pathology Center for all cases. All WSI were reviewed for image quality assurance, and 10 WSI excluded because they had with severe artifacts, including excessive folding, poor quality of staining, and the presence of bubbles under the coverslip. The final WSI dataset included a total of 258 WSIs (123 Internal and 135 External).

Figure 1: Overall research design. (a) The material preparation process consisted of collection of cases from Duke University medical center and outside institutions, followed by annotation by 3 pathologists, QC by expert renal pathologists and mask generation. (b) The training process includes selection and augmentation of training samples, followed by model optimization for segmenting the glomeruli. (c) Predictions are made based on sequential patches, which are stitched together to recover a WSI prediction. (d) The model performance is further investigated by comparing to the standard of care report.

2.2 Manual Segmentation of Glomeruli

Three primary annotators, 1 post-graduate year one (GS) and 1 post-graduate year two (RD) trainees in the Duke accredited residency program, and 1 internationally trained pathologist (YM), manually segmented non-sclerotic glomeruli and globally sclerotic glomeruli. Segmentation was achieved by manually outlining glomeruli in all 258 WSIs using a publicly available digital pathology tool (QuPath, version 0.1.2)^{8,13}. For non-sclerotic glomeruli,

annotations were made by tracing the Bowman's capsule and through the vascular and tubular pole of the glomerulus, when visible, to maintain a continuous circular annotation outline of the individual glomeruli. As the Bowman's capsule of globally sclerotic glomeruli is generally inconsistent or separated from the tuft by a white space representing a processing/freezing artifact, only the sclerosed tuft was outlined during the annotation process (Figure 2).

All manual segmentations were reviewed by two expert renal pathologists (LB & DH) to assure accuracy of glomerular detection and boundaries. The expert pathologists traced non-sclerotic and sclerotic glomeruli when missed by the primary annotators and retraced the boundaries in those glomeruli where the primary annotator did not follow the segmentation criteria. Matched pairs of shift-invariant WSIs and manual segmentations were considered as DL training samples (Figure 1 - a2).

Figure 2: An illustrating example of a whole slide image with glomerular segmentations. Manual segmentations of non-sclerotic glomeruli (blue circle) and sclerotic glomeruli (red circle – black arrow) on whole slide images are generated using QuPath.

2.3 Deep Learning Implementation and Performance Evaluation

The 75 Internal cases (123 WSIs) were divided at the patient level into training (D1) and validation (D2) datasets with an 0.8:0.2 ratio, respectively, and used to train/fine-tune the DL model. The 135 External cases (135 WSIs) were used as an independent testing dataset (D3).

2.3.1 Network architecture

A DL framework was developed to automatically identify and segment non-sclerotic vs. globally sclerotic glomeruli on frozen section WSIs. A 9-module convolutional neural network (CNN) based on the common U-Net architecture³⁸ with a dilated bottleneck was developed for glomerular segmentation (Figure 3). Our U-Net architecture is a symmetric encoder-decoder fully convolutional network with a 256x256x3 input layer that produces pixel-level segmentation results. Each encoder module contains two convolution blocks consisting of a 3x3 convolutional layer, a batch normalization procedure, and a rectified linear unit (ReLU) activation function. Modules are connected by a 2x2 max-pooling layer for down-sampling.

The down-sampled bottleneck module consists of three dilation operations³⁹ for each convolutional layer, which enlarge the receptive field to capture coarse field-of-view imaging details within the high-level feature maps. The decoder modules recover and up-sample the semantic information generated from the bottleneck module, each of which includes a 2x2 transpose convolution block and a 3x3 convolution block. Long-term skip connections are used for enhancing different scale texture details provided in the encoding layers. Finally, a 1x1 convolutional layer is used to output a 2-layer probability map representing glomerular foreground vs. background.

Figure 3: Deep learning model architecture. An input glomeruli patch (left) of size 256x256x3 is fed into a 9-module U-Net model. For this model, each of the four encoder (i.e., left side of the network) modules are loaded with VGG pre-trained weights, which are then fine-tuned during training. The bottleneck module consists of three dilated convolution layers with different dilation rates. The feature maps at this level are added element-wise at the end of the bottleneck. The four decoder modules (i.e., right side of the network) use transpose convolutions to up-sample the feature maps, and skip-connections incorporate the encoder information. Finally, a 1x1 convolution layer maps the information to a 256x256 binary image of pixel-level predictions of glomerular locations.

2.3.2 Model training and validation

Two independent models were trained in parallel for normal and globally sclerotic glomeruli, respectively. Using the training data, random 256x256 image patches were extracted at 5x magnification from regions of relatively high glomerular density. Matched pairs of shift-invariant image patches and corresponding manual segmentations were used as training examples to train the DL model. To boost model generalization, a data augmentation procedure was applied, including basic operations (e.g., Horizontal Vertical Flip, Crop Resize), morphological

transformations (e.g., Shift Scale Rotate, Elastic Transform), color distortions (e.g., Contrast Brightness, Hue Saturation Value), and other image processing operations (Gaussian Blur, Random Cutout).

Training utilized transfer learning, where the network's first four fully convolutional blocks were initialized as the pre-trained ImageNet⁴⁰ weights of the publicly available VGG16³¹ data. During the training process, a cross entropy loss function was minimized based on Adam optimization³² to learn optimal model hyper-parameterization. Training was run for 200 epochs with a batch size of 16 input patches and an initial learning rate of 0.001. The training implementation achieving the lowest validation loss was computationally locked down and deployed as the final model for testing.

2.3.3 Model testing

Model performance was independently evaluated on the test set data by comparing model output to corresponding manual segmentation results (Figure 1-c). For a given WSI, the model was applied to sequential patches, and the results were concatenated together to generate a biopsy-level prediction at the full 40X field-of-view. Segmentation accuracy was quantified based on the dice similarity coefficient (DSC)⁴¹, which measures the pixel-level overlap between DL-generated segmentation results and manual segmentation results. Additionally, model accuracy, sensitivity, and specificity of sclerotic vs. non-sclerotic glomeruli was quantified based on F1, Precision, and Recall scores⁴².

2.3.4 Transfer learning

The effect of transfer learning and sample size on model performance was evaluated for both non-sclerotic and sclerotic glomeruli segmentation models. Model performance metrics with and without the transferred VGG16 weights were compared and their differences quantified.

2.4 Deep Learning Glomerular Count vs. Standard of Care Pathology Reporting

External cases where the reported glomerular count was available (N=47) were used to compare the DL-derived glomerular count to the current standard of care. The reported glomerular counts for non-sclerotic and globally sclerotic glomeruli performed on the frozen section, along with the sclerotic-non-sclerotic glomerular ratio, were compared to both the corresponding manual segmentation-derived counts and the DL-derived counts (Figure 1-d).

Pair-wise t-tests and Pearson correlation coefficients were used to quantify statistical differences between the (a) historically reported, (b) manual segmentation-derived, and (c) DL-derived glomerular count. The Bonferroni-Holm method⁴³ was implemented to correct p-values for multiple hypotheses testing. A corrected p-value lower than 0.05 was considered statistically significant.

3. Results

3.1 Manual Segmentation of Glomeruli

A total of 21146 non-sclerotic (8897 from the Internal dataset and 12249 from the External dataset) and 1322 sclerotic glomeruli (682 from the Internal dataset and 640 from the External dataset) were manually segmented on 258 images.

3.2 Deep Learning Detection and Segmentation Performance

DL model performance results on the External testing WSIs are summarized in Table 1. The non-sclerotic glomeruli model achieved a DSC of 0.91 (implying high spatial overlap compared to manual segmentation), an F1 score of 0.93 (implying strong overall detection performance), high recall of 0.96 (implying accurate recognition of true positive non-sclerotic glomeruli), and high precision of 0.90 (indicating a high over-prediction of non-sclerotic glomeruli compared to sclerotic glomeruli). Similarly, the sclerotic glomeruli model achieved a DSC of 0.83, an F1 score of 0.87, recall of 0.93, and a precision of 0.81. A WSI final prediction example is visualized in Figure 4.

Table 1: Model performance on the test set. Glomerular detection and segmentation performance for the non-sclerotic glomeruli were >0.9 by all measures, indicating robust performance. The sclerotic performance was slightly less robust but still good with measures >0.8 with recall being 0.91. The slightly worse precision compared to recall means there were more false positives than false negatives. When compared to the performance of the non-sclerotic algorithm, the precision is less robust which is consistent with the smaller amount of data in the sclerotic cohort and the relatively greater variety histologic mimics of sclerotic glomeruli on the WSIs.

	Detection			Segmentation
	Precision	Recall	F1	DSC
Non-sclerotic	0.9	0.96	0.93	0.91
Sclerotic	0.81	0.93	0.87	0.83

Figure 4: An illustrating example of a deep learning prediction on a whole slide image. Results are color coded relative to reference manual annotations to demonstrate the performance of the deep learning algorithm on testing data.

Figure 5 displays examples of false positive and false negative predictions of sclerotic and non-sclerotic models. Sources of model error included two major categories: procedure artifacts (e.g., tissue processing artifacts such as overstaining, folds, air bubbles, and chatter) and glomerular histologic mimics (e.g., fibrosis of the urinary space, dense interstitial fibrosis, red blood cell casts). The global sclerosis model had a relatively high false positive rate, due to the variety of sclerotic textures and the relative lack of global sclerosis training data. Hence, most of the incorrect predictions were due to histology mimics. The non-sclerotic model generally learned well. Extreme procedure artifacts (e.g., distorted glomeruli, tangential cuts with small glomerular profiles, and poor staining) were the major reasons for failed predictions.

Figure 5: False Positive and False negative predictions for globally sclerotic and non-sclerotic glomeruli. (A-C) False positive for the global sclerosis model; **(D-F)** False negative for the global sclerosis model; **(G-I)** False positive for the non-global sclerosis model. A probable non-globally sclerotic glomerulus **(I)** was not manually annotated by the primary annotator nor by the quality control pathologist, but it was detected by the DL model; **(J-L)** false negative for the non-global sclerosis model.

3.3 Effect of Transfer Land Sample Size on Deep Learning Performance

As demonstrated in Figure 6, when training on a relatively small sample size (e.g., number of glomeruli < 600), transfer learning had a significant effect on DL model performance. Despite less data, model performance was improved based on the transfer learning procedure. The effect of transfer learning was less significant when >600 glomeruli were used for training. The non-sclerotic glomeruli model required less training samples (i.e., 1500 non-sclerotic glomeruli) to achieve performance saturation (i.e., DSC = 0.9) compared to the sclerotic glomeruli model. Meanwhile, the sclerotic glomeruli model appeared to not achieve performance saturation, implying that more sclerotic data would likely improve performance.

Figure 6: Effect of transfer learning and sample size on model performance of (left) non-sclerotic glomeruli and (right) sclerotic glomeruli. Transfer learning improved model performance significantly with limited data (i.e. <600 glomeruli) for both **(left)** non-sclerotic and **(right)** sclerotic model. Less glomeruli samples are needed for the non-sclerotic model to reach a performance saturation (i.e., DSC = 0.9 with 1500 glomeruli), compared to sclerotic model.

3.4 Deep Learning Glomerular Count vs. Standard of Care Pathology Reporting

Direct statistical comparisons between (i) historically reported glomerular counts, (ii) manual segmentation glomerular counts, and (iii) DL-model glomerular counts are reported in Table 2. When the glomerular count from the manual segmentation was compared to the glomerular count from the DL model, statistically similar counts were observed for both non-sclerotic ($p=0.837$) and sclerotic glomeruli ($p=0.0950$). This implies that the DL-model is operating at a non-inferior counting performance relative to an expert renal pathologist. When the glomerular counts from the manual and DL segmentation were compared to the historically reported standard-of-care glomerular counts, a statistically significant difference was observed for both non-sclerotic ($p<0.0001$) and sclerotic ($p=0.002$) glomeruli. This implies that both manual counting by a renal pathologist and automatic counting by the DL model are both more accurate than historically reported clinical data, which is otherwise prone to high inter-observer variability. Correlation plots of the three category results among three methods for each data point are shown in Figure 7. In Figure 8, testing WSIs on the correlation plot are displayed with different procedure artifact conditions, which were a major source of model performance deviation.

Table 2: Comparison of glomerular count. For the model-annotation count is not significantly different from the reference segmentations, whereas the standard of care count is different from the reference segmentations. The same is true for the sclerotic model, which is not significantly different from the reference segmentations, whereas the standard of care has a significantly different count from the reference segmentations.

	Model vs Reference Segmentation	Report vs Reference Segmentation
Non-sclerotic	$p=0.837$	$p=3.695e-7$
Sclerotic	$p=0.095$	$p=7.920e-4$

Figure 7: Comparison among glomerular counting modalities. Comparison for percentage of globally sclerosed glomeruli, count of non-sclerotic glomeruli, and count of sclerotic glomeruli among DL Model Prediction, Manual Segmentation, and Pathology Report.

Figure 8: Visualization of whole slide images relative to model performance on testing data. (a,b,c) WSIs contain tissue freezing and folding artifacts (a), glass slide (bubble between the cover slip and the glass slide) and cutting artifacts (b), and tissue folding artifacts (c), which may have contributed to decreased performance. (d) The tissue section is intact and without artifacts, with accuracy for sclerosis from the DL model prediction.

4. Discussion

Digital pathology and state-of-the-art computational techniques are changing the landscape of pathology practice⁴⁴. Unlike in oncologic pathology – where computational image analysis has been extensively developed and is slowly being introduced into clinical practice, drug development, and clinical trials⁴⁵ – native and transplant nephropathology still relies entirely on visual assessment. In the current study, we developed a robust, accurate, and generalizable DL model for automatic segmentation of non-sclerotic and globally sclerotic glomeruli on H&E frozen section WSIs from donor kidney wedge biopsies. Our results demonstrate high predictive

performance on an independent dataset and show that machine-derived glomerular counts are more accurate than historically reported standard-of-care clinical data. These positive findings provide hypothesis generating data and motivate future applications of AI techniques to transplant nephropathology.

While other groups have primarily investigated DL-based glomerular detection and segmentation on WSIs of paraffin-embedded tissue^{46–48}, our study represents the largest application using frozen sections from kidney donor wedge biopsies. Previous work by Marsh et. al. successfully demonstrated elliptical detection³⁷ and quantification⁴⁹ of glomeruli on frozen sections. Although their approach to DL was different than ours and their cross-validated model lacked independent testing, our key findings are comparable and thus both provide promising insight for this emerging technology.

Generalizability of a DL model is critical for future scalability in clinical practice. Even though laboratories in the United States follow general College of American Pathology (CAP)^{20,34} and Clinical Laboratory Improvement Amendments (CLIA) guidelines³⁵, pre-analytical variability is still significant. Computational pathology techniques are highly sensitive to pre-analytical tissue variations⁴⁴, including room and freezing temperature, wedge biopsy tissue thickness, the percentage of water in the biopsy tissue (freezing artifacts), frozen section thickness, the presence of folds (cutting artifacts), and composition of the stain solutions (stain artifacts)³³. Furthermore, as cadaveric donor biopsies are obtained from deceased patients, the time from death to harvesting is often highly variable. Longer time delays from death to implantation often lead to greater autolysis artifacts on the frozen sections. Pre-existing conditions leading to patient death can also degrade tissue presentation, although these generally affect tubules and interstitium more so than glomeruli. Unlike the previously published work (37,49), our model was independently evaluated

on a multi-institutional test dataset. Thus, our results provide a reasonable representation of how well the model generalized to new information not observed during training. Since independent model testing is a fundamental principle of machine learning, this is a key novelty of our research design.

While several studies used multiple pathologists to manually annotate the same WSIs to train the model on a heterogeneous group of annotations, we established a protocol where the annotation of trainees were reviewed by senior pathologists. This protocol was designed to correct the missed glomeruli due to fatigue of the primary annotator and to mimic current clinical practice, where trainees provide the first read and senior pathologists review and correct, so that a robust dataset for reference segmentations for both detection and boundaries could be generated.

Our data have also shown that the DL algorithm on digital images operates with more accuracy than clinical reads using conventional microscopy. Several reasons may account for this increased performance. First, often, pathologists assessing frozen sections donor biopsies not only operate overnight, but also are not subspecialty trained in renal pathology. Second, the spatial-visual memory of any human is limited, resulting in missing some glomeruli while counting, or counting the same glomeruli twice. The manual annotation of all the glomeruli allows for the mapping of the glomeruli on the WSIs, so that the count can be more accurate⁵⁰ and better serve as ground truth.

In this work, we chose to implement a UNET architecture that incorporated several state-of-the-art techniques, including transfer learning, data augmentation, and convolutional dilation⁵¹. The motivation behind this design choice was based on recently published work, where UNET was successfully implemented to segment glomeruli on non-frozen tissue^{52–54}. Since our work is the first of its kind on frozen tissue, we chose a relatively simple model architecture that is

commonly used in diverse biomedical image segmentation applications. We acknowledge, however, that there are newer deep learning architectures, some of which have been applied to glomeruli detection, including the Inception V3 Architecture⁵⁵ and the SegNet architecture⁵⁶. Future work will focus on comparing different model architectures on frozen tissue, including advanced segmentation networks such as SegNet, DeepLab V3⁵⁷, Mask R-CNN⁵⁸ and effective network modules such as Inception V3, Residual blocks⁵⁹, Attention blocks⁶⁰. While such an analysis is out of scope of the current work, characterization of different model architectures is essential to eventually implementing these new technologies in clinical practice.

Nevertheless, we believe our research design is suitable for the current dataset. First, transfer learning was shown to boost model performance by using pre-trained weights obtained from the publicly available ImageNet database as initial parameter conditions⁶¹. This implies that these natural images encode generalized features of quantitative image representation that are applicable to digital pathology tasks. Our results demonstrate that the relative effect of transfer learning is indirectly proportional to sample size. This finding is consistent with machine learning theory^{62,63} and implies that the performance of our sclerotic glomeruli model may asymptotically improve with more data. Second, data augmentation was also shown to increase the generalization of our DL results, suggesting artificial augmentations may capture important characteristics of renal pathology⁶⁴. For example, *shape deformation* was shown to be effective in improving the recognition of various shapes and sizes of glomeruli. In addition, *color jitter* was useful to harmonize the large variation in stain quality, especially between the training data and the testing data. We observed sclerotic glomeruli to be most affected by *color jitter*, while non-sclerotic glomeruli were more sensitive to *texture deformation*. We hypothesize that this is due to sparser image texture in the compact structure of sclerotic glomeruli. Other augmentation methods (e.g.,

Gaussian blur) generally yielded better performance in the presence of freezing artifacts. Finally, our network architecture included a dilated block at the bottleneck. The rationale behind this design choice was to enable a larger receptive field-of-view to aggregate multi scale imaging information. Based on our results, we found that this dilation procedure reduced misclassification errors in the presence of tissue fold artifacts.

While our results are promising, this study has some limitations. First, all images were acquired on the same whole slide scanner. As such, it is currently unclear how scanning variability will affect the quality and performance of our trained DL models. Second, differences in model performance noted between the non-sclerotic and globally sclerotic glomeruli were due to the limited number of globally sclerotic glomeruli in the WSI dataset. We anticipate that model performance will increase based on more examples of sclerotic glomeruli. Third, our study only focused on glomeruli, which is only one aspect of pre-implant pathology. Future work will be dedicated to building additional DL models for other relevant histologic parameters such as interstitial fibrosis, acute and chronic tubular injury, vascular damage and other glomerular pathology⁶⁵. Additionally, our correlation analyses for glomerular counts among the DL models, the QA annotations, and the report of record were only performed on a limited number of cases where the outside reporting was available. Last, while our approach generated a robust reference segmentation dataset, inter-observer variability of segmentation was not addressed.

Digital pathology and automatic image analysis enable solutions that may aid in the clinical transplant nephropathology environment by providing robust and standardized quantitative observations, higher efficiency, centralize interpretation by expert pathologists with overall reduced error rates⁶⁶, and by reducing the known limitations associated with visual examination^{14,17}. Additionally, a digital solution offers a more rapid and efficient allocation of the kidney

overcoming the limitations, expenses, loss of precious time from the transferring of a donor organ and associated frozen section of the wedge biopsy from institution to institution, in search of a recipient.

Disclosures

Conflicts of Interest Statement

The authors have no conflict of interests to disclose.

Funding Statement

This work was supported by the Nephcure foundation and by Duke University institutional funding.

Data Availability Statement

The raw data for this study can be obtained through correspondence with the corresponding authors.

Acknowledgments

This collaborative work between the Department of Pathology, Division of AI and Computational Pathology, the Department of Medicine, Division of Nephrology, and the Woo Center for Big Data and Precision Health at Duke University.

References

1. Wolfe RA, Ashby VB, Milford EL, et al. Comparison of Mortality in All Patients on Dialysis, Patients on Dialysis Awaiting Transplantation, and Recipients of a First Cadaveric Transplant. *N Engl J Med*. 1999;341(23):1725-1730. doi:10.1056/NEJM199912023412303
2. Perico N, Ruggenenti P, Scalapogno M, Remuzzi G. Tackling the Shortage of Donor Kidneys: How to Use the Best that We Have. *Am J Nephrol*. 2003;23(4):245-259. doi:10.1159/000072055
3. Hart A, Smith JM, Skeans MA, et al. *OPTN/SRTR 2018 Annual Data Report: Kidney*; 2020.
4. Angeletti A, Cravedi P. Making Procurement Biopsies Important Again for Kidney Transplant Allocation. *Nephron*. 2019;142(1):34-39. doi:10.1159/000499452
5. Kasiske BL, Stewart DE, Bista BR, et al. The Role of Procurement Biopsies in Acceptance Decisions for Kidneys Retrieved for Transplant. *Clin J Am Soc Nephrol*. 2014;9:562-571. doi:10.2215/CJN.07610713
6. Dahmen M, Becker F, Pavenstädt H, Suwelack B, Schütte-Nütgen K, Reuter S. Validation of the Kidney Donor Profile Index (KDPI) to assess a deceased donor's kidneys' outcome in a European cohort. *Sci Rep*. 2019;9(1):11234. doi:10.1038/s41598-019-47772-7
7. Carpenter D, Husain SA, Brennan C, et al. Procurement Biopsies in the Evaluation of Deceased Donor Kidneys. *Clin J Am Soc Nephrol*. 2018;13:1876-1885. doi:10.2215/CJN.04150418
8. Teixeira AC, Freire De Carvalho CC, Mororó GP, Pereira LDM, Lacerda VS, Esmeraldo RM. Evaluation of Frozen and Paraffin Sections Using the Maryland Aggregate Pathology Index Score in Donor Kidney Biopsy Specimens of a Brazilian Cohort. *Transpl Proc*. 2017;49:2247-2250. doi:10.1016/j.transproceed.2017.11.004
9. Sagasta A, Sanchez-Escuredo A, Oppenheimer F, et al. Pre-implantation analysis of kidney biopsies from expanded criteria donors: testing the accuracy of frozen section technique and the adequacy of their assessment by on-call pathologists. *Transpl Int*. 2016;29:234-240. doi:10.1111/tri.12709
10. El-Husseini A, Sabry A, Zahran A, Shoker A. Can Donor Implantation Renal Biopsy Predict Long-Term Renal Allograft Outcome? *Am J Nephrol*. 2007;27(2):144-151. doi:10.1159/000099944
11. Salvadori M. Histological and clinical evaluation of marginal donor kidneys before transplantation: Which is best? Conflict-of-interest statement. *World J Transplant*. 2019;9(4):62-80. doi:10.5500/wjt.v9.i4.62
12. Goumenos DS, Kalliakmani P, Tsamandas AC, et al. The prognostic value of frozen section preimplantation graft biopsy in the outcome of renal transplantation The prognostic value of frozen section preimplantation graft biopsy in the outcome of renal transplantation Preimplantation biopsy in renal transpl. *Ren Fail*. 2010;32:434-439. doi:10.3109/08860221003658241
13. Munivenkatappa RB, Schweitzer EJ, Papadimitriou JC, et al. The Maryland Aggregate Pathology Index: A Deceased Donor Kidney Biopsy Scoring System for Predicting Graft Failure. *Am J Transplant*. 2008;8(11):2316-2324. doi:10.1111/j.1600-6143.2008.02370.x
14. Liapis H, Gaut JP, Klein C, et al. Banff Histopathological Consensus Criteria for Preimplantation Kidney Biopsies. *Am J Transplant*. 2017;17(1):140-150. doi:10.1111/ajt.13929
15. Roufosse C, Simmonds N, Clahsen-van Groningen M, et al. A 2018 Reference Guide to the Banff

- Classification of Renal Allograft Pathology. *Transplantation*. 2018;102(11):1795-1814. doi:10.1097/TP.0000000000002366
16. Loupy A, Aubert O, Orandi BJ, et al. Prediction system for risk of allograft loss in patients receiving kidney transplants: International derivation and validation study. *BMJ*. 2019;366. doi:10.1136/bmj.l4923
17. Antonieta Azancot M, Moreso F, Salcedo M, et al. The reproducibility and predictive value on outcome of renal biopsies from expanded criteria donors. *Kidney Int*. 2014;85(5):1161-1168. doi:10.1038/ki.2013.461
18. Pantanowitz L, Sharma A, Carter A, Kure T, Sussman A, Saltz J. 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*. 2018;9(1):40. doi:10.4103/jpi.jpi_69_18
19. Retamero JA, Aneiros-Fernandez J, del Moral RG. Complete digital pathology for routine histopathology diagnosis in a multicenter hospital network. *Arch Pathol Lab Med*. 2020;144(2):221-228. doi:10.5858/arpa.2018-0541-OA
20. Pantanowitz L, Sinard JH, Henricks WH, et al. Validating whole slide imaging for diagnostic purposes in Pathology: Guideline from the College of American pathologists Pathology and Laboratory Quality Center. *Arch Pathol Lab Med*. 2013;137(12):1710-1722. doi:10.5858/arpa.2013-0093-CP
21. Brachtel E, Yagi Y. Digital imaging in pathology - current applications and challenges. *J Biophotonics*. 2012;5(4):327-335. doi:10.1002/jbio.201100103
22. Baidoshvili A, Bucur A, van Leeuwen J, van der Laak J, Kluin P, van Diest PJ. Evaluating the benefits of digital pathology implementation: time savings in laboratory logistics. *Histopathology*. 2018;73(5):784-794. doi:10.1111/his.13691
23. Abels E, Pantanowitz L, Aeffner F, et al. Computational pathology definitions, best practices, and recommendations for regulatory guidance: a white paper from the Digital Pathology Association. *J Pathol*. 2019;249(3):286-294. doi:10.1002/path.5331
24. Lecun Y, Bengio Y, Hinton G. Deep learning. *Nature*. 2015;521(7553):436-444. doi:10.1038/nature14539
25. Janowczyk A, Madabhushi A. Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases. *J Pathol Inform*. 2016;7(1). doi:10.4103/2153-3539.186902
26. Madabhushi A, Lee G. Image analysis and machine learning in digital pathology: Challenges and opportunities. *Med Image Anal*. 2016;33:170-175. doi:10.1016/j.media.2016.06.037
27. Litjens G, Kooi T, Bejnordi BE, et al. A survey on deep learning in medical image analysis. *Med Image Anal*. 2017;42:60-88. doi:10.1016/j.media.2017.07.005
28. Wang S, Yang DM, Rong R, Zhan X, Xiao G. Pathology Image Analysis Using Segmentation Deep Learning Algorithms. *Am J Pathol*. 2019;189(9):1686-1698. doi:10.1016/j.ajpath.2019.05.007
29. Bejnordi BE, Veta M, Van Diest PJ, et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA - J Am Med Assoc*. 2017;318(22):2199-2210. doi:10.1001/jama.2017.14585

30. Ertosun MG, Rubin DL. Automated Grading of Gliomas using Deep Learning in Digital Pathology Images: A modular approach with ensemble of convolutional neural networks. *AMIA . Annu Symp proceedings AMIA Symp.* 2015;2015:1899-1908. [/pmc/articles/PMC4765616/?report=abstract](https://pubmed.ncbi.nlm.nih.gov/31899916/). Accessed July 11, 2020.
31. Simonyan K, Zisserman A. Very Deep Convolutional Networks For a Large-Scale Image Recognition. In: *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings.* ; 2015. <http://www.robots.ox.ac.uk/>. Accessed August 8, 2020.
32. Kingma DP, Lei Ba J. Adam: A method for stochastic optimization. In: *3rd International Conference on Learning Representations, ICLR 2015.* ; 2015.
33. Bauer T, Slaw R, McKenney J, Patil D. Validation of whole slide imaging for frozen section diagnosis in surgical pathology. *J Pathol Inform.* 2015;6(1):49. doi:10.4103/2153-3539.163988
34. Evans AJ, Bauer TW, Bui MM, et al. US Food and Drug Administration approval of whole slide imaging for primary diagnosis: A key milestone is reached and new questions are raised. *Arch Pathol Lab Med.* 2018;142(11):1383-1387. doi:10.5858/arpa.2017-0496-CP
35. U.S.C. Title 42-Chapter 6A-THE PUBLIC HEALTH AND WELFARE. <https://www.govinfo.gov/content/pkg/USCODE-2011-title42/pdf/USCODE-2011-title42-chap6A-subchapII-partF-subpart2-sec263a.pdf>. Accessed August 15, 2020.
36. Gaber L, Moore L, Alloway RR, Amiri MH, Vera S, Gaber AO. Glomerulosclerosis As a Determinant of Posttransplant Function of Older Donor Renal Allografts. *Transplantation.* 1995;60(4):334-339.
37. Marsh JN, Matlock MK, Kudose S, et al. Deep Learning Global Glomerulosclerosis in Transplant Kidney Frozen Sections. *IEEE Trans Med Imaging.* 2018;37(12):2718-2728. doi:10.1109/TMI.2018.2851150
38. Ronneberger O, Fischer P, Brox T. *U-Net: Convolutional Networks for Biomedical Image Segmentation.*; 2015. <http://lmb.informatik.uni-freiburg.de/>. Accessed November 19, 2020.
39. Chen L-C, Papandreou G, Member S, Kokkinos I, Murphy K, Yuille AL. *DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs.*; 2017. <http://liangchiehchen.com/projects/>. Accessed November 19, 2020.
40. Russakovsky O, Deng J, Su H, et al. ImageNet Large Scale Visual Recognition Challenge. *Int J Comput Vis.* 2015;115(3):211-252. doi:10.1007/s11263-015-0816-y
41. Zou KH, Warfield SK, Bharatha A, et al. Statistical Validation of Image Segmentation Quality Based on a Spatial Overlap Index. *Acad Radiol.* 2004;11(2):178-189. doi:10.1016/S1076-6332(03)00671-8
42. Sokolova M, Lapalme G. A systematic analysis of performance measures for classification tasks. *Inf Process Manag.* 2009;45(4):427-437. doi:10.1016/j.ipm.2009.03.002
43. Aickin M, Gensler H. Adjusting for multiple testing when reporting research results: The Bonferroni vs Holm methods. *Am J Public Health.* 1996;86(5):726-728. doi:10.2105/AJPH.86.5.726
44. Barisoni L, Lafata KJ, Hewitt SM, Madabhushi A, Balis UGJ. Digital pathology and computational image analysis in nephropathology. *Nat Rev Nephrol.* 2020;1. doi:10.1038/s41581-020-0321-6

45. Bera K, Schalper KA, Rimm DL, Velcheti V, Madabhushi A. Artificial intelligence in digital pathology — new tools for diagnosis and precision oncology. *Nat Rev Clin Oncol*. 2019;16(11):703-715. doi:10.1038/s41571-019-0252-y
46. Bukowy JD, Dayton A, Cloutier D, et al. Region-based convolutional neural nets for localization of glomeruli in trichrome-stained whole kidney sections. *J Am Soc Nephrol*. 2018;29(8):2081-2088. doi:10.1681/ASN.2017111210
47. Simon O, Yacoub R, Jain S, Tomaszewski JE, Sarder P. Multi-radial LBP Features as a Tool for Rapid Glomerular Detection and Assessment in Whole Slide Histopathology Images. *Sci Rep*. 2018;8(1). doi:10.1038/s41598-018-20453-7
48. Kannan S, Morgan LA, Liang B, et al. Segmentation of Glomeruli Within Trichrome Images Using Deep Learning. *Kidney Int Reports*. 2019;4(7):955-962. doi:10.1016/j.ekir.2019.04.008
49. Marsh JN, Liu T-C, Wilson PC, Swamidass SJ, Gaut JP. Development and Validation of a Deep Learning Model to Quantify Glomerulosclerosis in Kidney Biopsy Specimens. *JAMA Netw Open*. 2021;4(1):e2030939. doi:10.1001/jamanetworkopen.2020.30939
50. Rosenberg AZ, Palmer M, Merlino L, et al. The Application of Digital Pathology to Improve Accuracy in Glomerular Enumeration in Renal Biopsies. Tan M-H, ed. *PLoS One*. 2016;11(6):e0156441. doi:10.1371/journal.pone.0156441
51. Rashidi HH, Tran NK, Betts EV, Howell LP, Green R. Artificial Intelligence and Machine Learning in Pathology: The Present Landscape of Supervised Methods. *Acad Pathol*. 2019;6. doi:10.1177/2374289519873088
52. Jayapandian CP, Chen Y, Janowczyk AR, et al. Development and evaluation of deep learning-based segmentation of histologic structures in the kidney cortex with multiple histologic stains. *Kidney Int*. 2021;99(1):86-101. doi:10.1016/j.kint.2020.07.044
53. Gallego J, Swiderska-Chadaj Z, Markiewicz T, Yamashita M, Gabaldon MA, Gertych A. A U-Net based framework to quantify glomerulosclerosis in digitized PAS and H&E stained human tissues. *Comput Med Imaging Graph*. 2021;89:101865. doi:10.1016/j.compmedimag.2021.101865
54. Hermesen M, Bel T, Boer M Den, et al. Deep learning-based histopathologic assessment of kidney tissue. *J Am Soc Nephrol*. 2019;30(10):1968-1979. doi:10.1681/ASN.2019020144
55. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the Inception Architecture for Computer Vision. In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. Vol 2016-Decem. IEEE Computer Society; 2016:2818-2826. doi:10.1109/CVPR.2016.308
56. Li J, Sarma K V., Chung Ho K, Gertych A, Knudsen BS, Arnold CW. A Multi-scale U-Net for Semantic Segmentation of Histological Images from Radical Prostatectomies. *AMIA . Annu Symp proceedings AMIA Symp*. 2017;2017:1140-1148. /pmc/articles/PMC5977596/. Accessed May 14, 2021.
57. Chen LC, Papandreou G, Schroff F, Adam H. Rethinking atrous convolution for semantic image segmentation. *arXiv*. June 2017. <https://arxiv.org/abs/1706.05587v3>. Accessed May 3, 2021.
58. He K, Gkioxari G, Dollár P, Girshick R. Mask R-CNN. *IEEE Trans Pattern Anal Mach Intell*. 2020;42(2):386-397. doi:10.1109/TPAMI.2018.2844175
59. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. Vol 2016-

- December. IEEE Computer Society; 2016:770-778. doi:10.1109/CVPR.2016.90
60. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. In: *Advances in Neural Information Processing Systems*. Vol 2017-December. Neural information processing systems foundation; 2017:5999-6009. <https://arxiv.org/abs/1706.03762v5>. Accessed May 3, 2021.
 61. Yosinski J, Clune J, Bengio Y, Lipson H. *How Transferable Are Features in Deep Neural Networks?*; 2014.
 62. Romero M, Interian Y, Solberg T, Valdes G. Targeted transfer learning to improve performance in small medical physics datasets. *Med Phys*. October 2020:mp.14507. doi:10.1002/mp.14507
 63. Raghu M, Zhang C, Brain G, Kleinberg J, Bengio S. *Transfusion: Understanding Transfer Learning for Medical Imaging.*; 2019.
 64. Tellez D, Litjens G, Bándi P, et al. *Quantifying the Effects of Data Augmentation and Stain Color Normalization in Convolutional Neural Networks for Computational Pathology.*; 2020.
 65. Walker JL, Piedmonte MR, Spirtos NM, et al. Laparoscopy Compared With Laparotomy for Comprehensive Surgical Staging of Uterine Cancer: Gynecologic Oncology Group Study LAP2. *J Clin Oncol*. 2009;27:5331-5336. doi:10.1200/JCO.2009.22.3248
 66. Hanna MG, Reuter VE, Samboy J, et al. Implementation of digital pathology offers clinical and operational increase in efficiency and cost savings. *Arch Pathol Lab Med*. 2019;143(12):1545-1555. doi:10.5858/arpa.2018-0514-OA