The American Journal of Pathology, Vol. 192, No. 10, October 2022

ELSEVIER

The American Journal of
PATHOLOGY

ajp.amjpathol.org

MACHINE LEARNING, COMPUTATIONAL PATHOLOGY, AND BIOPHYSICAL IMAGING

Convolutional Neural Networks for the Evaluation of Chronic and Inflammatory Lesions in Kidney Transplant Biopsies

Meyke Hermsen,[*] Francesco Ciompi,[*] Adeyemi Adefidipe,[†] Aleksandar Denic,[‡] Amélie Dendooven,[§¶] Byron H. Smith,[‖**] Dominique van Midden,[*] Jan Hinrich Bräsen,[††] Jesper Kers,[†‡‡§§] Mark D. Stegall,[¶¶] Péter Bándi,[*] Tri Nguyen,[‖‖] Zaneta Swiderska-Chadaj,[*,***] Bart Smeets,[*] Luuk B. Hilbrands,[†††] and Jeroen A.W.M. van der Laak[*‡‡‡]

From the Departments of Pathology[*] and Nephrology,[†††] Radboud University Medical Center, Nijmegen, the Netherlands; the Department of Pathology,[†] Amsterdam University Medical Centers, and the Center for Analytical Sciences Amsterdam,[§§] Van 't Hoff Institute for Molecular Sciences, University of Amsterdam, Amsterdam, the Netherlands; the Divisions of Nephrology and Hypertension,[‡] Biomedical Statistics and Informatics,[**] and Transplantation Surgery,[¶¶] and the William J. von Liebig Center for Transplantation and Clinical Regeneration,[‖] Mayo Clinic, Rochester, Minnesota; the Department of Pathology,[§] Ghent University Hospital, Ghent, Belgium; the Faculty of Medicine,[¶] University of Antwerp, Wilrijk, Antwerp, Belgium; the Nephropathology Unit,[††] Institute of Pathology, Hannover Medical School, Hannover, Germany; the Department of Pathology,[‡‡] Leiden University Medical Center, Leiden, the Netherlands; the Department of Pathology,[‖‖] University Medical Center Utrecht, Utrecht, the Netherlands; the Faculty of Electrical Engineering,[***] Warsaw University of Technology, Warsaw, Poland; and the Center for Medical Image Science and Visualization,[‡‡‡] Linköping University, Linköping, Sweden

Accepted for publication
June 22, 2022.

Address correspondence to
Jeroen A. W. M. van der Laak,
Ph.D., Department of Pathology
(824), Radboud University
Medical Center, P.O. Box 9101,
6500 HB Nijmegen, the
Netherlands.
E-mail: jeroen.vanderlaak@radboudumc.nl.

In kidney transplant biopsies, both inflammation and chronic changes are important features that predict long-term graft survival. Quantitative scoring of these features is important for transplant diagnostics and kidney research. However, visual scoring is poorly reproducible and labor intensive. The goal of this study was to investigate the potential of convolutional neural networks (CNNs) to quantify inflammation and chronic features in kidney transplant biopsies. A structure segmentation CNN and a lymphocyte detection CNN were applied on 125 whole-slide image pairs of periodic acid–Schiff– and CD3-stained slides. The CNN results were used to quantify healthy and sclerotic glomeruli, interstitial fibrosis, tubular atrophy, and inflammation within both nonatrophic and atrophic tubuli, and in areas of interstitial fibrosis. The computed tissue features showed high correlation with Banff lesion scores of five pathologists (A.A., A.Dend., J.H.B., J.K., and T.N.). Analyses on a small subset showed a moderate correlation toward higher CD3$^+$ cell density within scarred regions and higher CD3$^+$ cell count inside atrophic tubuli correlated with long-term change of estimated glomerular filtration rate. The presented CNNs are valid tools to yield objective quantitative information on glomeruli number, fibrotic tissue, and inflammation within scarred and non-scarred kidney parenchyma in a reproducible manner. CNNs have the potential to improve kidney transplant diagnostics and will benefit the community as a novel method to generate surrogate end points for large-scale clinical studies. *(Am J Pathol 2022, 192: 1418–1432; https://doi.org/10.1016/j.ajpath.2022.06.009)*

Supported by the ERACoSysMed initiative (SysMIFTA project) as part of the European Union's Horizon 2020 Framework Program offered by ZonMw through grant 9003035004 (M.H.) and by the Dutch Kidney Foundation through grants 17OKG23 (DEEPGRAFT project) (J.K.) and 21OK+012 (DIAGGRAFT project) (M.H. and D.v.M.). The development of the lymphocyte detection network and the automated analysis pipeline was supported by the Dutch Cancer Society grant KUN 2014-7032 (Alpe d'HuZes AQUILA project) (F.C. and Z.S.-C.).

Disclosures: J.A.W.M.v.d.L. is a member of the advisory boards of Philips, the Netherlands, and ContextVision, Sweden, and received research funding from Philips, the Netherlands, ContextVision, Sweden, and Sectra, Sweden, in the last 5 years. He is chief scientific officer and shareholder of Aiosyn BV, the Netherlands. F.C. is advisory board member and shareholder of Aiosyn BV, the Netherlands, and received consulting fees from TRIBVN Healthcare, France.

https://doi.org/10.1016/j.ajpath.2022.06.009

Although much progress has been made toward the prevention of acute kidney transplant rejection, long-term graft loss remains a major issue for donor kidney survival. Scarring of the kidney in the form of interstitial fibrosis and tubular atrophy is the hallmark of progressive transplant failure. Recently, several studies have additionally demonstrated the prognostic value of inflammation and tubulitis in regions with interstitial fibrosis and tubular atrophy (i-IFTA and t-IFTA, respectively).[1−4] Accurate scoring of these chronic, inflammatory parameters is therefore pivotal in strategies to prevent graft loss.

The commonly used scoring system for kidney transplant biopsy assessment is the Banff classification system.[5,6] This system was the first standardized, international classification system for kidney transplant diagnostics and facilitated uniformity in the reporting of renal transplant pathology.[7] It is internationally applied by kidney researchers and physicians, and it is the globally accepted quantification tool for histopathologic transplant evaluation. However, it has increasingly been criticized for its limited reproducibility and its suboptimal patient stratification. Multiple studies show poor to moderate interobserver agreement, specifically for the scoring of fibrotic changes and inflammatory lesions.[8−12] Moreover, the Banff classification system is based on semiquantitative scoring on an ordinal scale, whereas inflammatory and chronic parameters represent a continuous spectrum and should therefore preferably be quantified on a granular, continuous scale.

Quantitative assessment of transplant biopsies may be improved by the application of digital image analysis techniques.[13−15] Specifically, deep learning, the use of data-driven learning systems where multilayered (deep) neural networks are trained to generate output from input, has proven to be a powerful tool for histopathologic tissue assessment.[16−19] The most widely applied neural networks in medical image analysis are convolutional neural networks (CNNs). CNN-based image analysis could benefit biopsy assessment by increasing reproducibility and efficiency. In addition, CNNs can output absolute values, which may provide more insight into the stage of ongoing pathologic processes. A second and important advantage of CNN-based image analysis is the ability to decrease interobserver variability, a major problem in any form of histologic assessment by human observers.

The notable performance of CNNs on medical imaging data has resulted in an increasing number of studies focused on deep learning applications for kidney tissue. These efforts were pioneered by the segmentation and classification of the glomerulus and were expanded toward other applications, such as multiclass segmentation, the segmentation of sclerotic glomeruli and interstitial fibrosis and tubular atrophy (IFTA), and diabetic nephropathy classification.[20−24]

The current study investigated the potential of CNNs as quantification tools for the assessment of chronic and inflammatory lesions, going beyond the current arbitrary semiquantitative thresholds and showing the absolute quantification of tubulointerstitial inflammation as a continuous parameter in areas with and without IFTA. Ideally, CNNs can be used in addition to the Banff classification system to support kidney researchers and physicians in their studies on chronic kidney tissue changes.

For this purpose, two previously developed CNNs aimed at the segmentation of periodic acid−Schiff (PAS)−stained tissue and detection of lymphocytes in immunohistochemistry (IHC) were used.[25,26] The CNNs were retrained and applied on a cohort of PAS- and CD3-stained kidney transplant biopsy slides. Quantifications were performed on the basis of the CNN results. The reliability of the CNN-based quantifications was evaluated by assessing the correlation with the following visually scored components of the Banff classification system: glomerular count, total inflammation (ti), interstitial inflammation (i), tubulitis (t), interstitial fibrosis (ci), tubular atrophy (ct), i-IFTA, and t-IFTA.

Materials and Methods

A visual overview of this study can be found in Figure 1.

Patient Cohort

Tissue Samples

A total of 125 tissue blocks of Bouin-fixed, paraffin-embedded needle-core biopsies obtained for cause (kidney function decline without clear clinical cause) were used from 73 patients who underwent kidney transplantation between 2008 and 2012 in the Radboud University Medical Center (Radboudumc; Nijmegen, the Netherlands). For 39 patients, a single biopsy was included. For the remaining patients, biopsies acquired at two different time points (24 patients), three time points (4 patients), four time points (4 patients), or five time points (2 patients) were used. For the comparison of automatically quantified tissue features and pathologists' visual Banff scoring, multiple biopsies from a single patient were considered as independent samples as they were obtained at different time points and originate from different regions of the kidney. Recipient, donor, and biopsy characteristics based on the original pathology reports are included in Table 1. The local institutional review board waived the need for approval of using Radboudumc tissue blocks in this study (number 2016-2269).

PAS-CD3 Restaining and Whole-Slide Image Preparation

One tissue section (3 μm thick) was cut from every tissue block and placed onto a coated microscope slide. All slides were stained with PAS and digitized using a Pannoramic P1000 whole slide scanner (3DHISTECH, Budapest, Hungary) at a resolution of 0.24 μm per pixel. Subsequently, the slides were washed with an acetone solution to remove the cover film. Potential glue residue was removed by washing the slides in xylene, followed by dehydration in 95% ethanol. The slides were washed in tap water and boiled for epitope retrieval in 10× diluted Tris-borate-

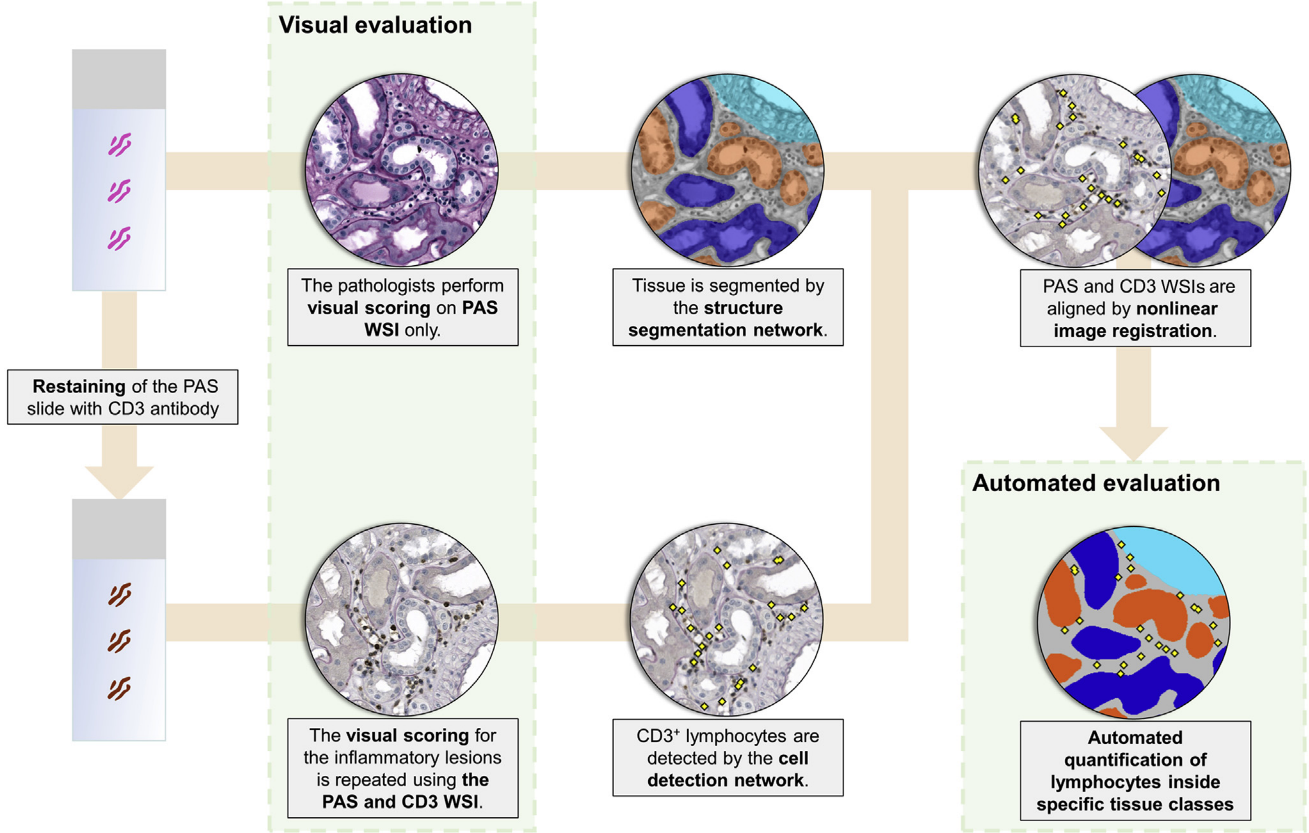
Visual evaluation
The pathologists perform visual scoring on PAS WSI only.
Tissue is segmented by the structure segmentation network.
PAS and CD3 WSIs are aligned by nonlinear image registration.
Restaining of the PAS slide with CD3 antibody
Automated evaluation
The visual scoring for the inflammatory lesions is repeated using the PAS and CD3 WSI.
CD3⁺ lymphocytes are detected by the cell detection network.
Automated quantification of lymphocytes inside specific tissue classes

Figure 1 Overview of the visual and automated evaluation. The periodic acid–Schiff (PAS) slides are digitized, followed by CD3-positivity visualization through immunohistochemistry. The PAS whole-slide images (WSIs) are scored by a panel of pathologists, and a convolutional neural network (CNN) segmented the tissue into relevant tissue classes. After a washout period, the pathologists repeat the visual scoring using the PAS WSI in combination with the digitized CD3 slide. CD3-positive cells are detected by a second CNN in the CD3 WSI. Spatial alignment of PAS and the CD3 WSIs allows for cell quantification inside the segmented structures.

EDTA (VWR Life Sciences, Boxmeer, the Netherlands) buffer. After cooling down, the slides were washed in 3% hydrogen peroxidase solution for endogenous peroxidase blocking, followed by washing in phosphate-buffered saline buffer. After pre-incubation with 1% phosphate-buffered saline/bovine serum albumin, the slides were incubated for 1 hour at room temperature with CD3 antibody solution (1:40; clone SP7; rabbit monoclonal antibody; RM-9107-S; Thermo Scientific, Waltham, MA). After washing in phosphate-buffered saline, the slides were incubated with a horseradish peroxidase (HRP)–conjugated secondary antibody (Poly-HRP-GAMs/Rb IgG; VWRKDPVO999HRP; Immunologic, Duiven, the Netherlands), followed by visualization with 3,3′-diaminobenzidine (Bright-DAB; VWRKBS04; Immunologic) and a counterstaining with hematoxylin. The restained slides were digitized using a Pannoramic P1000 whole slide scanner at a resolution of 0.24 μm per pixel. As a result of this procedure, every tissue sample yielded two whole-slide images (WSIs): one of the PAS-stained slide and one of the restained slide with CD3.

Regions of Interest

The cortical regions were manually annotated using the automated slide analysis platform software version 1.9 (*https://github.com/computationalpathologygroup/ASAP*, last accessed June 13, 2022). The pathologists (A.A., A. Dend., J.H.B., J.K., and T.N.) were asked to perform their analyses within these regions of interest, and the CNN-based quantifications were performed within these same regions. Tissue folds, subcapsular inflammation, and inflammatory infiltrates surrounding large arteries were excluded from the regions of interest.

Visual Pathologists' Assessment of the Patient Cohort Biopsies

Five pathologists (A.A., A.Dend., J.H.B., J.K., and T.N.), specialized in kidney transplant pathology, manually counted the number of glomeruli and scored the following Banff lesion categories on the PAS WSI according to criteria listed in Supplemental Table 1 (based on Banff 2018[5]): ti, i, t, ci, ct, i-IFTA, and t-IFTA. After a washout period of 4 weeks minimum, the pathologists repeated the scoring for the Banff ti, i, t, i-IFTA, and t-IFTA categories, now using the PAS WSI in combination with the CD3 WSI. The interobserver variability was assessed for both scenarios by calculating quadratic weighted Cohen κ coefficients. The visual glomerular counts and Banff ti, i, t, ct, ci, i-IFTA, and

Table 1 Baseline Cohort Characteristics

Characteristics	Values
Recipients (n = 73)	
Age, years	49.1 (19.8 to 70.3)
Female sex	30 (41.1)
Dialysis type	
Hemodialysis	43 (58.9)
Peritoneal dialysis	18 (24.6)
Preemptive	12 (16.4)
Panel-reactive antibodies ≤6	52 (71.2)
Patients with retransplants	11 (15.1)
Graft characteristics (n = 73)	
Donor age, years	57.0 (31.0 to 73.0)
Living	35 (47.9)
Deceased, donation after circulatory death	16 (21.9)
Deceased, donation after brain death	22 (30.1)
HLA-A mismatch	1 (0 to 2)
HLA-B mismatch	1 (0 to 2)
HLA-DR mismatch	1 (0 to 2)
HLA mismatch total	3 (0 to 6)
Cold ischemia time, hours	11.2 (1.8 to 26.5)
Need for dialysis <3 months after transplantation	27 (37.0)
Biopsy characteristics (n = 125)	
Time between transplantation and biopsy, days	40 (3 to 906)
Original diagnosis	
Rejection*	34 (27.2)
Borderline T-cell–mediated rejection	21 (16.8)
Calcineurin inhibitor toxicity	43 (34.4)
Cytomegalovirus	1 (0.8)
Acute tubulus necrosis	16 (12.8)
Recurrence original disease†	2 (1.6)
De novo focal segmental glomerulosclerosis	1 (0.8)
No diagnostic abnormalities	7 (5.6)

Data are given as median (minimum to maximum) or number (percentage).

*Humoral rejection, cellular rejection, or humoral and cellular rejection.

†Membranous glomerulonephritis or membranoproliferative glomerulonephritis.

HLA, human leukocyte antigen.

t-IFTA scores were compared with their equivalent tissue feature, quantified by CNNs (listed in Supplemental Table 2 and described in detail in *Automated Assessment of the Patient Cohort Based on CNN Results*).

Structure Segmentation CNN Development

The authors previously presented a U-net architectural CNN for the multiclass structure segmentation of PAS-stained kidney sections into relevant tissue classes, such as healthy and globally sclerotic glomeruli, interstitium, and proximal, distal, and atrophic tubuli.[25] For the current study, this CNN was improved by including more training data and improved post-processing techniques (see below). There was no overlap between the cases that were used for CNN development and the slides that were used in the formerly described PAS-CD3 patient cohort. A novel method was developed for the segmentation of interstitial fibrosis based on image processing of the multiclass structure segmentation results, further described in *Indirect Segmentation Method for Interstitial Fibrosis and IFTA*.

Ground Truth

For development of the structure segmentation network, the data set (60 WSIs) that was described in the authors' earlier publication on kidney tissue segmentation[25] was complemented with 36 additional PAS-stained transplant biopsies (Radboudumc, n = 19; Mayo Clinic, Rochester, MN, n = 17) and 3 tumor nephrectomy samples (Mayo Clinic), resulting in 99 WSIs. The slides were digitized on a Pannoramic 250 Flash II digital slide scanner (3DHISTECH; Radboudumc) or an Aperio ScanScope XT System scanner (Leica Biosystems, Nussloch, Germany; Mayo Clinic) at a resolution of 0.24 and 0.49 μm/pixel, respectively. The data set was annotated using automated slide analysis platform software, applying the following predefined classes: glomeruli, sclerotic glomeruli, empty Bowman capsules, proximal tubuli, distal tubuli, atrophic tubuli, capsule, arteries/arterioles, interstitium, and border (being the basement membranes of the tubuli). All annotations were checked and corrected where necessary by a pathologist (J.K.). The WSIs were randomly divided into training (n = 63), validation (n = 16), and test (n = 20) sets. The total number of annotations per tissue class is listed in Supplemental Table 3. Mayo Clinic tissue samples were scanned with institutional review board approval (numbers 17-002391 and 10-004644), and digital image file transfer was approved under institutional review board number 18-005592.

Network Design

A U-net architecture was used as the structure segmentation network design.[27] The network was trained for 95 epochs at 512 iterations per epoch with a batch size of eight patches (412 × 412 pixels at a resolution of 1.0 μm/pixel). Adam was used as weight optimization algorithm and categorical cross entropy as loss function.[28] Spatial and color augmentation techniques were applied to increase the network's robustness for variations in tissue morphology, staining intensity, and image quality. Before inference of the structure segmentation network, a tissue-background segmentation network was applied, separating tissue from background and removing dust particles and tissue artifacts.[29]

Post-Processing

Post-processing was used to optimize the structure segmentation results, applying the following steps at a pixel spacing of 1.0 μm/pixel: i) pixels classified as empty glomeruli positioned at the edge of the biopsy were removed; ii) pixels classified as border or interstitium were temporarily set to 0, grouping pixels of all the other classes into discrete objects; iii) holes (ie, value 0 regions) with an area <150 pixels inside objects were filled with their dominant surrounding object label; iv) objects with an area <300 pixels were considered noise and set to the interstitium class; v) objects that consisted of more than one tubule class were assigned to the predominant tubule class, and objects that consisted of more than one glomerulus class were assigned to the predominant glomerulus class; vi) regions <50 pixels inside objects were assigned to their dominant surroundings; vii) objects classified as glomeruli, having an area <2500 pixels, were set to the interstitium class; and viii) pixels classified as border were labeled as interstitium, and all interstitium pixels were subsequently placed back unless they were filled during step 3. The decision to use a minimum area of 2500 pixels for glomeruli was based on the knowledge that the diameter of a complete glomerulus ranges from approximately 100 to 200 μm, depending on the level of sectioning. This corresponds to a minimum area of 7854 pixels [based on the following formula: area = (diameter^2 $*$ π)/4]. By using 2500 pixels as a minimum area, corresponding to a diameter of approximately 56 μm, we avoided the risk of excluding complete glomeruli.

Structure Segmentation Performance

The segmentation performance of the network was assessed by calculating the CNN's precision, recall, and Dice score on pixel level on the test set, where:

$$\text{precision} = \frac{\text{true positive detections (TP)}}{\text{true positive detections (TP)} + \text{false positive detections (FP)}} \tag{1}$$

$$\text{recall} = \frac{\text{true positive detections (TP)}}{\text{true positive detections (TP)} + \text{false negative detections (FN)}} \tag{2}$$

$$\text{F1} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \tag{3}$$

The test set that was used to assess the performance metrics of the structure segmentation CNN was composed of PAS-stained slides from the Mayo Clinic and Radboudumc. Because the material from the current patient cohort contains PAS-stained biopsies from Radboudumc, it can be assumed that the performance on the test set will correspond with that on this patient cohort. Therefore, the performance metrics of the structure segmentation CNN were not additionally calculated for the patient cohort.

Indirect Segmentation Method for Interstitial Fibrosis and IFTA

The structure segmentation CNN was subsequently applied to the 125 PAS WSIs from the patient cohort (see *Materials and Methods*; *Patient Cohort*; *Tissue Samples*). Interstitial fibrosis regions were derived from the structure segmentation masks by computing distance maps for interstitial pixels with respect to atrophic tubuli and to all other structures. Pixels were assigned to the interstitial fibrosis class if they were closer to atrophic tubuli than to any other structure, under the biological assumption that interstitial fibrosis and tubular atrophy develop in tandem. This allowed for the quantification of interstitial fibrosis alone and IFTA. Because the CNN was not directly trained on interstitial fibrosis and IFTA, Dice score, precision, and recall could not be calculated for these classes. Instead, three human observers (A.Deni., D.v.M., and J.K.) visually estimated the percentage interstitial fibrosis and IFTA on 20 cases from the patient cohort. Similar to the automated scoring method, the visual score was a continuous score, ranging from 0% to 100%, and was not limited to categories. To assess the soundness of our automatic interstitial fibrosis/IFTA scoring method, the intraclass correlation coefficient (ICC) was calculated for the percentages given by the human observers and the percentages based on CNN results.

Lymphocyte Detection CNN

A recently developed lymphocyte detection CNN was adapted and used for the detection of lymphocytes in this study. This network was developed in a previous study,[26] in which four network architectures were trained with 171,166 manually annotated CD3$^+$ and CD8$^+$ lymphocytes: a fully convolutional network, a U-net, a you only look at lymphocytes once network, and a locally sensitive method network. The networks were evaluated for their detection performance of lymphocytes within normal tissue, artifacts, and immune cell clusters, using IHC-stained sections originating from nine medical centers. The best performing network for all the tasks was used in the current study (U-net). Because this network was trained on conventional IHC, it was retrained for the current study using 6237 lymphocyte annotations (15 WSIs) in restained kidney slides (PAS-CD3) in addition to the original training data. This retrained network was subsequently used for the cell detections in this study.

Automated Assessment of the Patient Cohort Based on CNN Results

The PAS WSI and the CD3 WSI were analyzed using the structure segmentation network and the lymphocyte detection network, respectively. This resulted in three masks per case: a structure segmentation mask, dividing the tissue in capsule, interstitium, arteries and arterioles, glomeruli, sclerotic glomeruli, proximal tubuli, distal tubuli, and atrophic tubuli; an interstitial fibrosis mask (based on the structure segmentation results, as described); and a lymphocyte detection mask, marking all CD3$^+$ cells in the tissue.

Image Registration

The PAS WSI and the CD3 WSI pairs display the same biopsies and are therefore roughly aligned. Nevertheless, tissue deformations may occur during IHC staining, and the rescanning of the slides causes a slight alteration of the tissue's coordinates in the image. This was corrected by nonlinear image registration, using the noncommercially available software HistokatFusion (version 2019; Fraunhofer MEVIS lab, Bremen, Germany). The software offers a three-step registration pipeline, consisting of a manual or automated pre-alignment, a parametric registration computed on coarse resolution images, and an accurate nonlinear registration.[30] This allowed for an accurate spatial translation of tissue features between slides and corresponding masks.

Automatically Quantified Tissue Features

On the basis of the registered results of the structure segmentation CNN and the lymphocyte detection CNN, the following features were calculated: the number of nonsclerotic glomeruli and globally sclerotic glomeruli; the highest CD3$^+$ cell count inside proximal tubuli or distal tubuli; the highest CD3$^+$ cell count inside atrophic tubuli; the CD3$^+$ cell density inside the total cortical area; the CD3$^+$ cell density inside the cortical area, excluding interstitial fibrosis; and the CD3$^+$ cell density inside regions of interstitial fibrosis.

Correlation between Automated Feature Quantification and Visual Banff Lesion Scoring

To assess the correlation of glomerular counting performed by pathologists with automated glomerular quantification, the average ICC of the pathologists and the average ICC of the pathologists and the CNN are reported.

Spearman correlation coefficients of interstitial fibrosis (pixel percentage), tubular atrophy (object percentage), (total) inflammation (cells/mm^2), inflammation in fibrotic regions (cells/mm^2), tubulitis (highest cell count), and tubulitis in atrophic tubuli (highest cell count) were calculated with the average ci, ct, ti, i, i-IFTA, t, and t-IFTA score of the pathologists, respectively (Supplemental Table 2).

Correlation between Automated and Visual Scoring of Chronic Lesions and the Course of Kidney Function

In contrast to the ordinal scoring by human observers, the deep learning−based results are reported as a continuum. It should be investigated whether these continuous values hold more prognostic information than the current lesion scoring system. As an illustration for such a validation study, we assessed the correlation between manually and automatically scored chronic lesions and long-term change in kidney function. More extensive validation should be performed on a larger data set, specifically designed for this purpose. The Δ estimated glomerular filtration rate (ΔeGFR) was defined as the difference between eGFR measured at 1 week before the biopsy procedure (according to the Modification of Diet in Renal Disease formula) and the eGFR measured at 2 years after the biopsy procedure. These data were available for 46 cases. One biopsy sample per patient was used for these analyses. When biopsy samples from multiple time points were included from a single patient in the patient cohort, only the last sample was included ($n = 39$). Cases were only included if no clinical event occurred (defined as the need for a biopsy for cause) between the biopsy procedure and eGFR measurement 2 years after the biopsy procedure ($n = 29$). Subsequently, 11 cases were excluded, where the biopsy for cause was obtained <60 days after transplantation to avoid that early transplantation-related lesions, such as acute tubular necrosis, would distort the analyses. This resulted in 18 eligible cases for the correlation assessment. Spearman correlation was calculated to assess the relationship between ΔeGFR and visually scored i-IFTA, t-IFTA, ci, and ct scores. The Spearman correlation was also calculated between ΔeGFR and automatically quantified CD3$^+$ cell density inside fibrotic regions, CD3$^+$ cell counts per atrophic tubuli, area percentage of interstitial fibrosis, and percentage of atrophic tubuli.

Results

Visual Banff Lesion Scoring of the Patient Cohort by Pathologists

The patient cohort consisted of 125 WSI pairs of PAS-stained slides and CD3 restained slides from 73 patients (Table 1). A panel of five kidney pathologists (A.A., A.Dend., J.H.B., J.K., and T.N.) visually assessed the PAS WSI for the number of glomeruli, and scored i, t, ti, ci, ct, i-IFTA, and t-IFTA Banff lesions according to the criteria listed in Supplemental Table 1 (based on Banff 2018[5]). After a washout period of 4 weeks minimum, the pathologists repeated the scoring for the Banff i, t, ti, i-IFTA, and t-IFTA categories, using the PAS WSI in combination with the CD3 WSI. The pathologists achieved moderate to good agreement using the PAS WSI (Supplemental Figure 1). Including the CD3 WSI caused a larger spread in Cohen κ values with a lower median interobserver agreement (Supplemental Figure 1). The pathologists' scores based on only PAS were therefore considered to be most reliable to compare with CNN data.

Structure Segmentation Performance of the CNN

The structure segmentation performance was assessed on an unseen test set of 20 WSIs from PAS-stained slides. Table 2 displays the precision, recall, and Dice score per class and the weighted average of the classes combined. The highest Dice score was observed for the glomeruli class, followed by interstitium, proximal tubuli, distal tubuli, sclerotic glomeruli, arteries and arterioles, capsule, atrophic tubuli, and empty Bowman capsules. The confusion matrix, providing insight into how the various predicted classes correspond to the true classes, is depicted in Supplemental Figure 2.

Validation of the Indirect Interstitial Fibrosis and IFTA Segmentation Method with Visually Estimated Percentages

The correlation of automatically generated interstitial fibrosis and IFTA percentages with percentages provided by human observers was assessed to validate the indirect segmentation method of fibrotic regions. The average ICC of three human observers for scoring interstitial fibrosis was 0.655, and the average agreement of the observers and the CNN was 0.667. For the scoring of interstitial fibrosis and tubular atrophy, these values were 0.866 and 0.793, respectively. Visual assessment of the indirect segmentation results of interstitial fibrosis supported these positive findings (Figure 2). This validation confirmed the rationale of the indirect interstitial fibrosis and IFTA segmentation strategy and justified the use of this method to define fibrotic tissue regions in the entire patient cohort. These regions were used to automatically include and exclude interstitial fibrotic regions in CD3$^+$ cell density calculations and to quantify interstitial fibrosis.

Table 2 Performance of the Structure Segmentation Network

Tissue class	Precision	Recall	Dice
Glomeruli	0.96	0.94	0.95
Sclerotic glomeruli	0.78	0.90	0.84
Empty Bowman capsule	0.38	0.58	0.45
Proximal tubuli	0.96	0.88	0.92
Distal tubuli	0.85	0.86	0.86
Atrophic tubuli	0.44	0.63	0.52
Arteries/arterioles	0.60	0.93	0.73
Interstitium	0.91	0.89	0.90
Capsule	0.53	0.90	0.66
Weighted average	—	—	0.88

—, Not calculated.

Segmentations and Cell Detections of the Patient Cohort by the CNNs

An example of a fully automatically assessed PAS-CD3 image pair is depicted in Figure 3. Figures 4 and 5 depict examples of successful and unsuccessful segmentations of

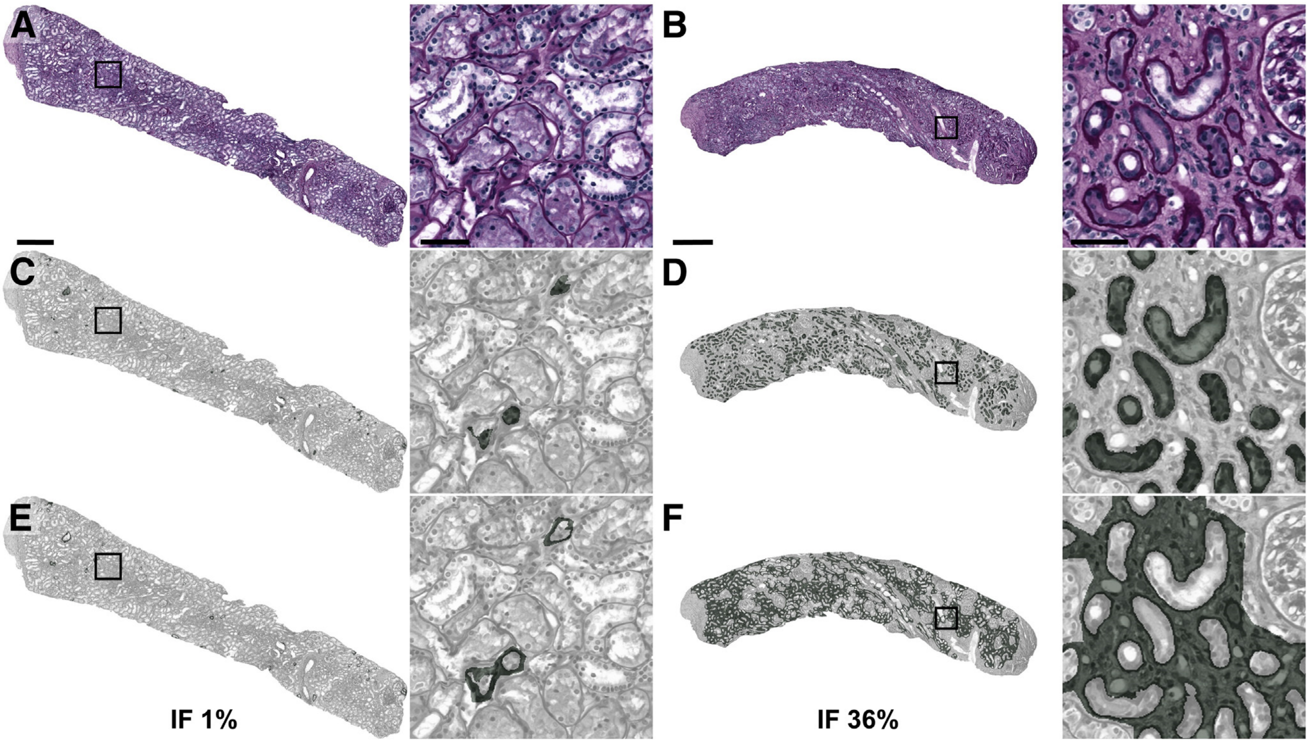

Figure 2 Automated indirect interstitial fibrosis segmentation. **A**, **C**, and **E:** A nonfibrotic biopsy. **B**, **D**, and **F:** Visualization of a fibrotic biopsy. **A**–**F:** The **boxed areas** on the low-resolution images represent the areas depicted in the high-resolution images. **C** and **D:** The segmentation of atrophic tubuli by the structure segmentation convolutional neural network is visualized in green. **E** and **F:** Using image processing, pixels in closer proximity to atrophic tubuli than to any other structures (excluding interstitium) were assigned to the interstitial fibrosis class (green). The interstitial fibrosis (IF) percentage based on the cortical area in this figure is 1% for the nonfibrotic biopsy and 36% for the fibrotic biopsy. Scale bars: 500 μm (**A** and **B**, low resolution); 50 μm (**A** and **B**, high resolution).

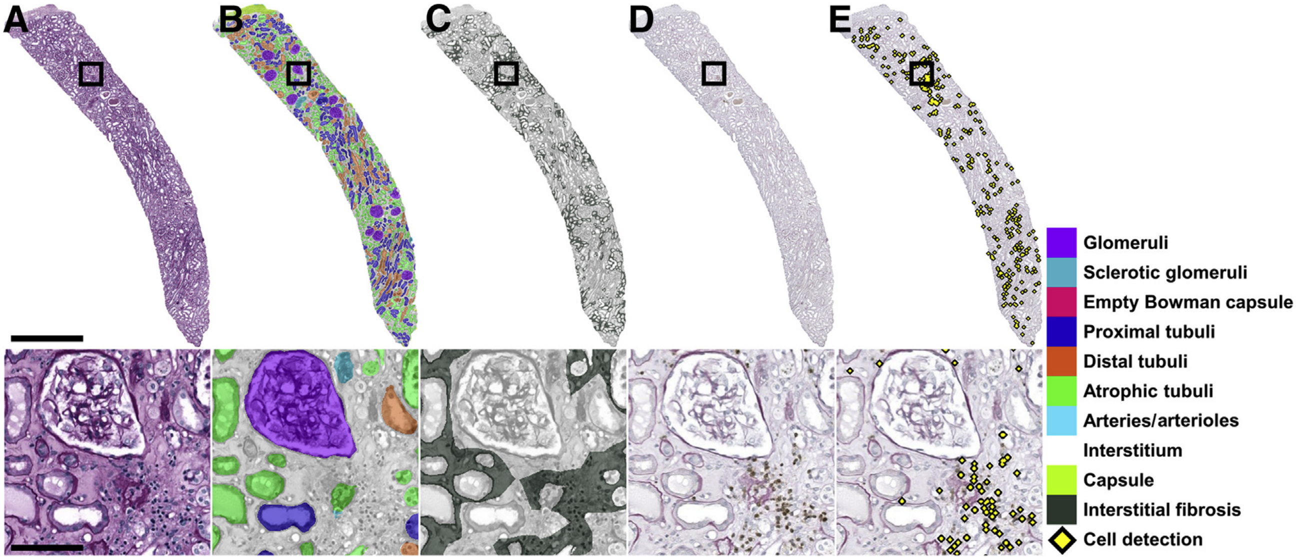

Figure 3 Periodic acid—Schiff (PAS)—CD3 image pair with segmentations and cell detections. Representative image of a restained and automatically analyzed biopsy. This image depicts a PAS whole-slide image (**A**), structure segmentation mask (**B**), interstitial fibrosis mask (**C**), and the corresponding restained CD3 whole-slide image (**D**) with the cell detection mask (**E**). Scale bars: 1 mm (low resolution); 100 μm (high resolution).

glomeruli, sclerotic glomeruli, proximal tubuli, distal tubuli, and atrophic tubuli. Image registration allowed for successful alignment of structure segmentation and cell detection results.

Agreement between Automated Feature Quantification and Visual Banff Lesion Scoring

The results of the structure segmentation CNN and the lymphocyte detection CNN were used to quantify numerous tissue features from the patient cohort. ICCs and Spearman correlations were calculated between these features and the average Banff lesion scoring of five kidney pathologists (A.A., A.Dend., J.H.B., J.K., and T.N.). The mean ICC of the CNN and the panel of pathologists for glomerular counting was 0.941. As supported by Figure 4, visual assessment of the segmentation result showed highly accurate segmentations with occasional false-positive segmentations of sclerotic glomeruli. Limiting the automated glomerular count to nonsclerotic glomeruli led to a mean ICC of the CNN and the pathologists of 0.972 (Table 3).

Next, the CNN assessment of interstitial fibrosis (pixel percentage), tubular atrophy (object percentage), inflammation in the total tubulointerstitium (cells/mm^2), inflammation in nonfibrotic regions (cells/mm^2), inflammation in fibrotic regions (cells/mm^2), tubulitis (highest cell count), and tubulitis in atrophic tubuli (highest cell count) was compared with the average score of pathologists for the following Banff categories: ci, ct, ti, i, i-IFTA, t, and t-IFTA (Table 4 and Figure 6). The highest correlation was reported for automatically assessed CD3$^+$ cell density in the total cortical area with the mean ti score of the pathologists, followed by the CD3$^+$ cell density in non-scarred cortical regions and the mean i score of the pathologists. Good correlations were reported for automatic and visual assessment of interstitial fibrosis and tubular atrophy, as well as for CD3$^+$ cell density in scarred cortical regions and the mean i-IFTA score of the pathologists. The lowest correlations are reported for the highest CD3$^+$ cell count in non-atrophic tubuli and the mean t score of the pathologists, and the highest CD3$^+$ cell count in atrophic tubuli and the mean t-IFTA score of the pathologists.

Correlation between Chronic Tissue Scores and the Course of Kidney Function

The correlation of ci, ct, i-IFTA, and t-IFTA with the long-term course of kidney function was evaluated for the CNN-based quantification method and the visually assessed Banff scores. On average, an improvement of eGFR was found over time in this subset of the patient cohort. Nevertheless, moderate inverse correlations were found between the ΔeGFR and the average i-IFTA score of the pathologists (ICC = −0.567; P = 0.014) (Supplemental Figure 3A), and ΔeGFR and automatically assessed cell density inside interstitial fibrotic regions of the cortex (ICC = −0.515; P = 0.029) (Supplemental Figure 3B). The highest CD3$^+$ cell count inside atrophic tubuli segmented by the structure segmentation CNN also inversely correlated with ΔeGFR (ICC = −0.782; P < 0.001) (Supplemental Figure 4B). A weaker inverse correlation was found between the average t-IFTA score of the pathologists and ΔeGFR compared with the correlation with the automated method (ICC = −0.568; P = 0.014) (Supplemental Figure 4A). The visual ci and ct Banff score and the automatically assessed interstitial fibrosis area percentage and tubular atrophy percentage did not correlate with ΔeGFR.

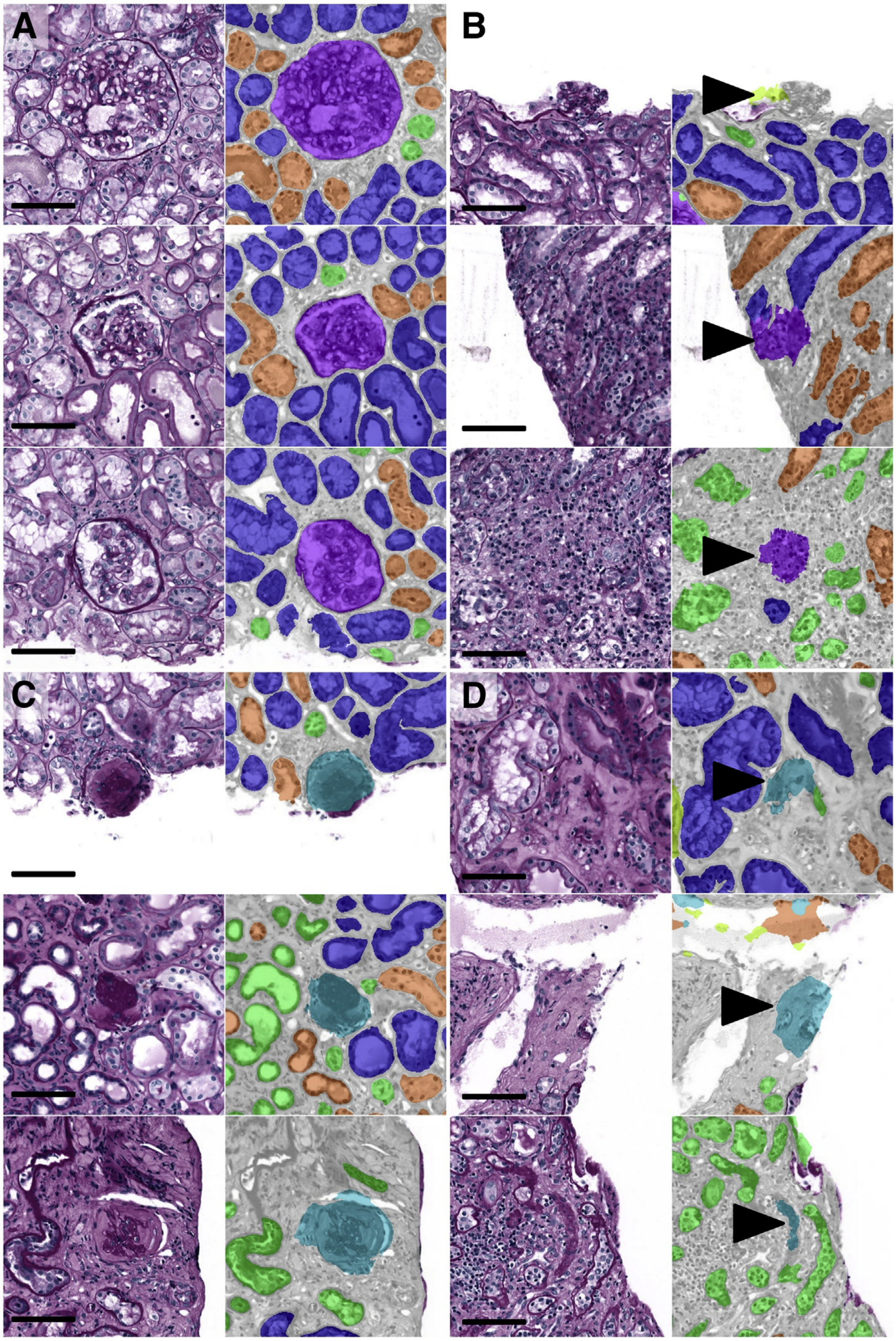

Figure 4 Examples of glomeruli and sclerotic glomeruli segmentations. **A:** Three representative correct glomerulus segmentations are depicted. **B:** An example of a missed glomerulus at the biopsy edge and two examples of false-positive glomeruli segmentations in inflammatory regions (**black arrowheads**). **C:** Three correctly segmented sclerotic glomeruli are depicted. **D:** Two examples depicted of false-positive sclerotic glomerulus segmentations inside fibrotic regions and one example where (possibly) a residue of an atrophic tubule is wrongly segmented as sclerotic glomerulus (**black arrowheads**). Scale bars = 100 μm (**A**–**D**).

Discussion

In this study, deep learning was used to quantify both inflammation and chronic lesions in kidney transplant biopsies. Two CNNs were applied: a structure segmentation CNN for PAS-stained kidney tissue and a lymphocyte detection CNN for IHC-stained slides. Nonlinear image registration allowed quantitative inflammation assessment in specific regions of kidney biopsies, based on individual $CD3^+$ cell detections. The reliability of the CNN-based

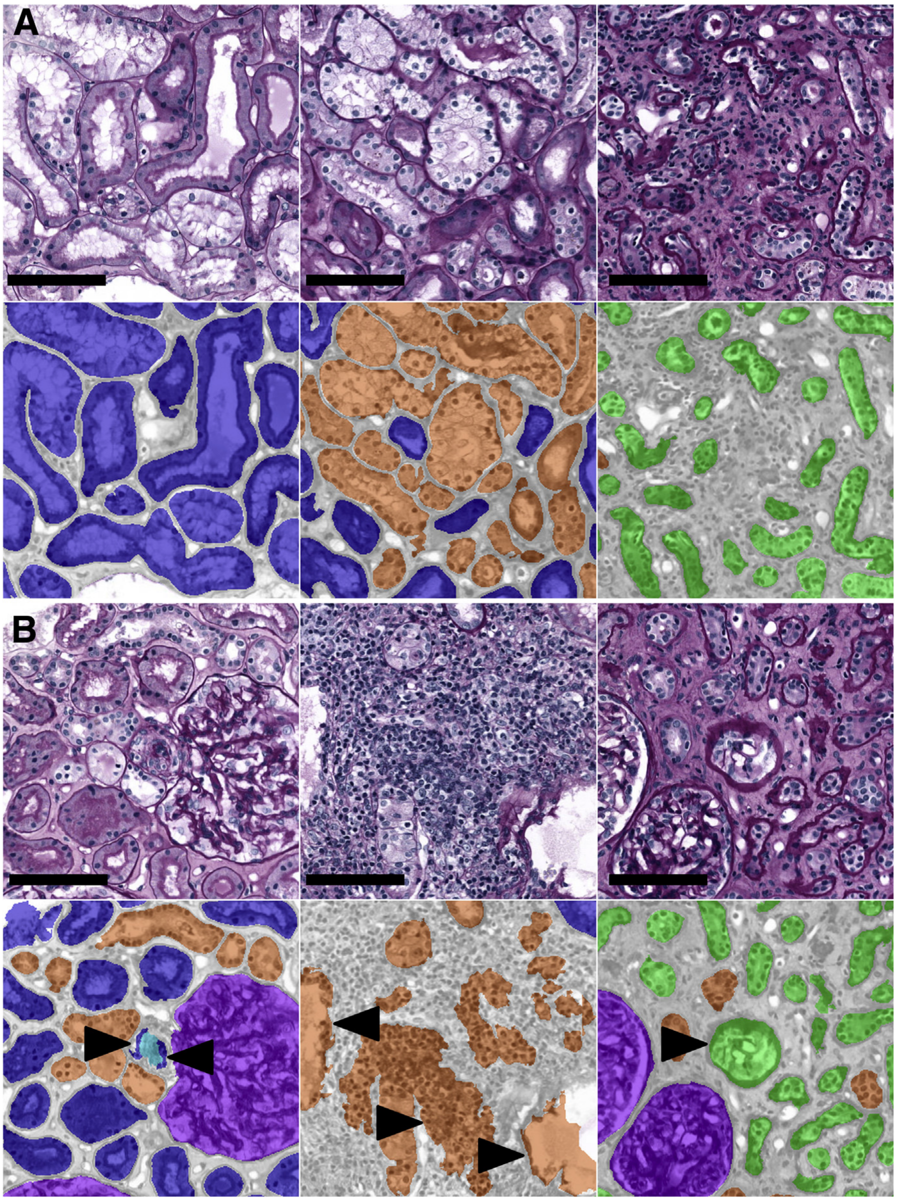

Figure 5 Examples of tubuli segmentations. **A:** Good segmentations of proximal tubuli (**left panels**), distal tubuli (**middle panels**), and atrophic tubuli (**right panels**), with the **top row** displaying the periodic acid—Schiff images and the **bottom row** displaying network segmentations. **B: Left panels:** Two false proximal tubule segmentations inside an arteriole (**black arrowheads**). **Middle panels:** How an inflammatory interstitial region is (partially) incorrectly segmented as distal tubule (**black arrowheads**). **Right panels:** A more superficially cut glomerulus incorrectly segmented as an atrophic tubule (**black arrowhead**). Scale bars = 100 μm (**A** and **B**).

quantifications was assessed by evaluating the correlation of the automatically quantified tissue features with pathologists' Banff lesion scoring. Automatically quantified interstitial fibrosis and tubular atrophy correlated well with Banff ci and ct scoring. Total cortical inflammation, inflammation in non-scarred cortical regions, and inflammation in areas with interstitial fibrosis correlated with Banff ti, i, and i-IFTA scoring, respectively. In addition, glomerular counts based on CNN results correlated highly with visual glomerular counts. A correlation was found between higher inflammatory cell density inside areas of interstitial fibrosis and long-term decline in eGFR. Lower kidney function also correlated with higher inflammatory cell count inside atrophic tubuli. This was in agreement with the correlations that were found for visual Banff i-IFTA and t-IFTA scoring with long-term changes in eGFR.

The literature on kidney tissue segmentation using deep learning has expanded drastically over the past few years.[31–34] Many of the models described in the literature were trained in a binary manner (ie, glomeruli versus

Table 3 ICCs among the Pathologists and for the Pathologists and the CNN for Glomerular Counts

Group	ICC
Mean pathologists	0.977
Mean pathologists: CNN (NSG)	0.972
Mean pathologists: CNN (NSG + GSG)	0.941

CNN, convolutional neural network; GSG, globally sclerotic glomeruli object segmentations; ICC, intraclass correlation coefficient; NSG, non-sclerotic glomeruli object segmentations.

non-glomeruli or tubuli versus nontubuli). The current study demonstrates a segmentation performance for healthy and globally sclerotic glomeruli comparable to that reported in literature, despite the challenge of nonbinary segmentation.[35–37] Also, glomerular quantifications based on our CNN results correlated highly with glomerular counts performed by five pathologists (A.A., A.Dend., J.H.B., J.K., and T.N.).

In a study by Jayapandian et al,[38] multiple networks were presented for segmenting glomerular, vascular, and tubular structures. The authors are one of the few to report separate segmentation performance of proximal and distal tubular segments, with impressive results. Unfortunately, atrophic tubuli were not included in this study.[38] Bouteldja et al[36] demonstrated a multiclass segmentation network for PAS-stained kidney tissue, showing excellent segmentation performances. However, healthy and atrophic tubuli were combined in their evaluation. The current study presents the only multiclass structure segmentation CNN that is developed for the segmentation and classification of the interstitium, healthy and sclerotic glomeruli, and proximal, distal, and atrophic tubuli. Such discrimination (especially that between healthy and atrophic/sclerotic structures) is crucial for developing an assay that yields clinically relevant and actionable data.

Interstitial fibrosis and tubular atrophy have been shown to correlate with chronic kidney disease and chronic rejection in kidney transplants. The quantification of fibrosis has been the subject of several studies.[39–42] Artificial neural networks have been developed for the assessment of fibrosis in trichrome-stained kidney slides[43,44] and recently the first neural network for sclerotic glomeruli and IFTA segmentation in PAS-stained slides was presented, showing good agreement with manual annotations in deceased-donor tissue.[24] In the current study, a novel approach was presented for the segmentation of interstitial fibrosis by generating an interstitial fibrosis mask based on atrophic tubuli segmentations resulting from the structure segmentation CNN. The segmentation of pixels in closer proximity to atrophic tubuli than to other structures resulted in a convincing definition of interstitial fibrotic regions. The correlation of the manual scoring of interstitial fibrosis percentage by three human observers was similar to the correlation between manual scoring and the automated method. In addition, the automated quantification of interstitial fibrosis showed high correlations with the average Banff ci lesion scores of five kidney pathologists. These results convincingly show that the presented CNN can be used as a valid quantification tool for interstitial fibrosis in kidney tissue.

Although the segmentation performance of atrophic tubuli has significantly improved since earlier studies, the Dice coefficient is still relatively low compared with that of some of the other classes. The confusion matrix in Supplemental Figure S2 shows that this can largely be attributed to mix-ups with distal tubuli and interstitium. It can be doubted whether the confusion with distal tubuli can be entirely prevented as the transition from a healthy tubule to an atrophic tubule is a continuous process. However, the false-positive atrophic tubuli segmentations inside (inflamed) interstitium possibly result from a relatively low number of inflamed interstitial regions in the training set. This can be improved in future work, by expanding training data sets.

Over the past two decades, studies have demonstrated the detrimental effect of inflammation within areas of interstitial

Table 4 Spearman Correlation Coefficients for the Computed Tissue Features with the Average Banff Scores of the Pathologists

Computed feature	Banff lesion	Spearman ρ	*P* value
(Interstitial fibrosis pixels/total cortical area pixels) × 100%	Interstitial fibrosis (ci)	0.785	<0.001
(Atrophic tubuli objects/total tubuli objects) × 100%	Tubular atrophy (ct)	0.773	<0.001
$CD3^+$ cell density in total cortical area, cells/mm^2	Total inflammation (ti)	0.838	<0.001
$CD3^+$ cell density in cortical area—interstitial fibrosis, cells/mm^2	Inflammation (i)	0.806	<0.001
Highest $CD3^+$ cell count inside nonatrophic tubuli object segmentations	Tubulitis (t)	0.551	<0.001
$CD3^+$ cell density in interstitial fibrosis area, cells/mm^2	Inflammation in regions with interstitial fibrosis and tubular atrophy (i-IFTA)	0.706	<0.001
Highest $CD3^+$ cell count inside atrophic tubuli object segmentations	Tubulitis in regions with interstitial fibrosis and tubular atrophy (t-IFTA)	0.632	<0.001

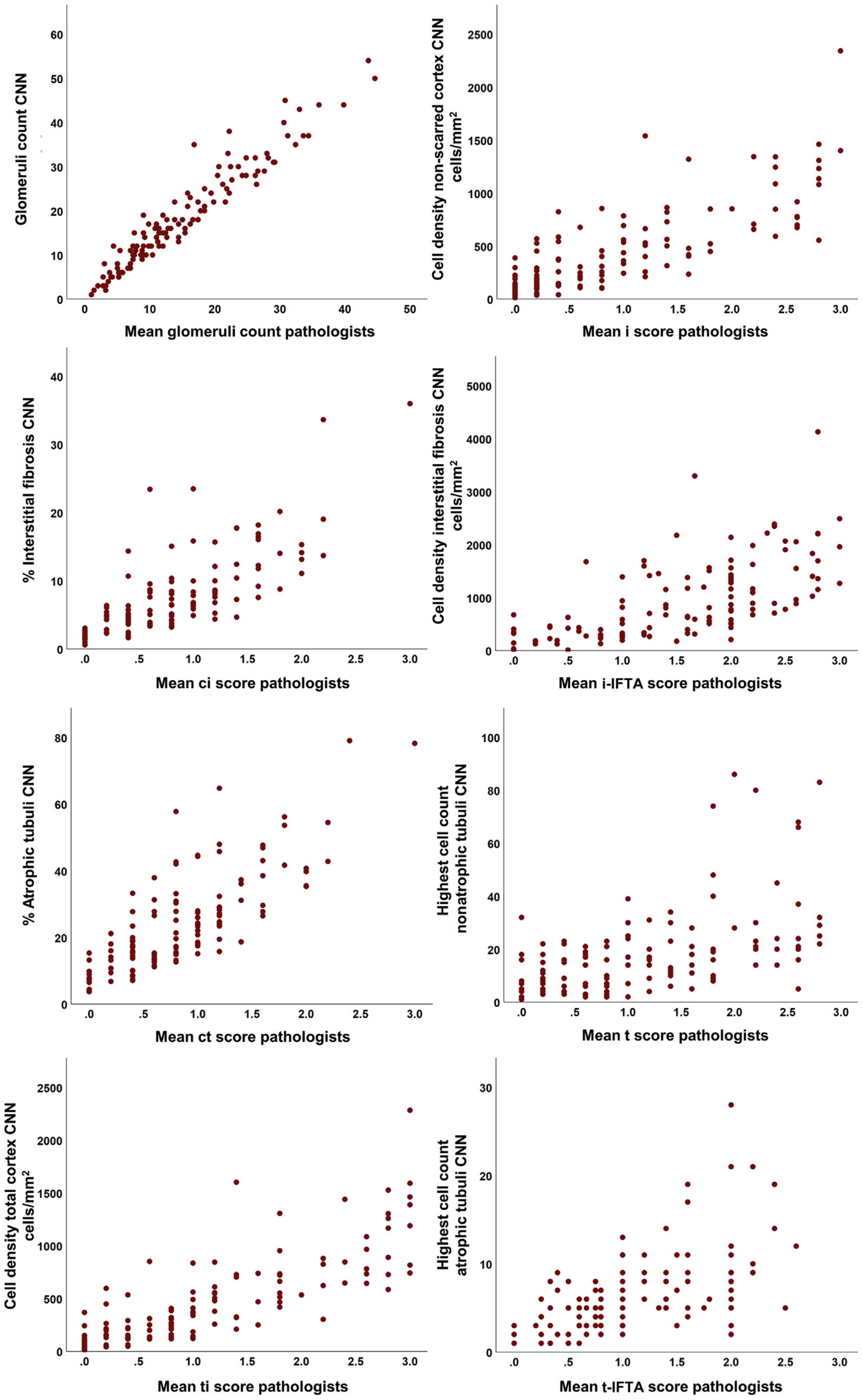

Figure 6 Scatterplots automated and visual assessment of the patient cohort. The computed tissue features by the convolutional neural network (CNN) are depicted on the *y* axes. The average Banff lesion score of five kidney pathologists is depicted on the *x* axes. Ci, interstitial fibrosis; ct, tubular atrophy; i, interstitial inflammation; i-IFTA, inflammation in regions with interstitial fibrosis and tubular atrophy; t, tubulitis; ti, total inflammation; t-IFTA, tubulitis in regions with interstitial fibrosis and tubular atrophy.

fibrosis and tubular atrophy on kidney transplant outcome.[1–4,45–47] As a result, inflammatory fibrosis (i-IFTA) was introduced to the Banff lesion scoring system in 2015.[6] Accurate scoring of i-IFTA requires the visual exclusion of non-scarred parenchyma, followed by an estimation of inflammatory burden inside the scarred region. This makes i-IFTA hard to score, also considering the novelty of the category. The low interobserver agreement for the scoring of i-IFTA in the current study (with and without IHC available) emphasizes the necessity of a supporting scoring tool as presented in this study. Yi et al[48] recently presented the so-called composite damage score, composed of abnormal interstitium areas and tubuli density and areas of mononuclear leukocyte infiltration. Although the authors did not directly compare composite damage score with i-IFTA, it was shown to be predictive for late eGFR decline and patient survival and will possibly approximate this Banff category.[48]

Instead of presenting an entirely new scoring system, the aim of this study was to stay close to the commonly used definitions while increasing the scoring granularity, accuracy, and reproducibility. To do so, the automatically generated segmentations and cell detections were combined using an award-winning image registration technique.[49] This allowed us to calculate CD3$^+$ cell density within scarred and non-scarred parenchyma and perform absolute CD3$^+$ cell counts in healthy and atrophic tubuli, enabling comparison to ti, i, t, i-IFTA, and t-IFTA scores. Automatically quantified cell densities in the complete cortical area were highly correlated to the average ti scores of the pathologists. Excluding scarred regions from the analysis allowed for the calculation of an equivalent for the Banff i score, which showed a high correlation with visual scoring as well. The Banff ti and i scores and their computational equivalents require minimal segmentation of the tissue in specific compartments. This may explain why the highest interobserver agreements and the highest correlations between automated and visual assessment were found for these categories. Lower correlations were found for cell densities inside regions of interstitial fibrosis with visual i-IFTA scores. This was possibly partially due to the low interobserver agreement among pathologists. In addition, false-positive tubuli detections were observed in inflamed interstitial regions. Therefore, the automatically generated interstitial fibrosis mask will not reach these regions, causing an underestimation of i-IFTA. In return, these false-positive segmentations can lead to an overestimation of (atrophic) tubulitis. This can be improved by including more inflamed interstitial regions during development of the structure segmentation network.

Finally, the correlation between the change in kidney function and automatically and visually scored ci, ct, i-IFTA, and t-IFTA was assessed as a proof of principle. Higher serum creatinine levels at time of the biopsy could cause an artifact when looking at ΔeGFR. To avoid this artifact, we used the eGFR measured 1 week before the biopsy for cause as a baseline. In reality, the serum creatinine levels appeared to be close on both time points [mean eGFR at minus 1 week: 29.53 mL/minute per 1.73 m^2 (SD, 8.70 mL/minute per 1.73 m^2); mean eGFR at time of biopsy: 28.26 mL/minute per 1.73 m^2 (SD, 7.99 mL/minute per 1.73 m^2)]. This causes most patients to show an improvement of eGFR over time. Nonetheless, a significant, inverse, correlation was found between the inflammatory burden inside areas of interstitial fibrosis and the subsequent course of kidney function. This held for the automated quantifications by the CNNs and the visual lesion scoring by pathologists. This shows that the presented method can support uniform assessment of inflammatory burden inside fibrotic and nonfibrotic kidney tissue.

There were some limitations in this study. First, the method presented in this article relies on the restaining of PAS-stained slides with IHC, followed by image registration. Most clinical centers will not include these methods in their routine transplant diagnostics procedure. Therefore, future studies shall be targeted at the development of an inflammatory cell detection network for PAS-stained sections, targeted at macrophages, B lymphocytes, and T lymphocytes. Second, our automated method does not correct for tangential sectioning. Third, the data show a trend toward an inverse correlation between visual and automated scores of inflammation inside areas of interstitial fibrosis and tubular atrophy and the course of kidney function. However, the number of patients eligible for these analyses was too small to draw strong conclusions from these results. The predictive potential of automated quantification of specific tissue features should be assessed in a larger cohort that was designed for this purpose. Finally, cortical regions were manually annotated as regions of interest for visual and automated assessment. A cortex segmentation CNN is required for fully automated assessment and will therefore be developed in future work.

Although this study supports a positive view toward the inclusion of CNN-based quantifications in routine transplant diagnostics, the true, short-term, clinical value of this study can be found in the application of CNNs for prospective kidney (transplantation) research. The results demonstrate that the presented CNNs produce reliable quantifications of (inflammatory) fibrotic regions that could be used to monitor pathologic processes in detail over time in a uniform manner. In particular, the CNN-based results can be used as surrogate end points in large-scale clinical studies, relieving pathologists from tedious scoring tasks. Predictive models often require histologic revisions of large cohorts, where uniform assessment is challenged by variation between countries, laboratories, and observers. The presented CNNs can be used to compute tissue features in a reproducible manner, which can subsequently function as input for a clinical prediction model. The continuous output of the CNNs can be used to reevaluate the thresholds of the Banff categories, which might result in a different patient grouping and a better prognostic system.

In conclusion, two CNNs were developed, applied, and combined for the segmentation of kidney tissue and the detection of $CD3^+$ inflammatory cells. Good correlations were found for the automated quantification of glomeruli, interstitial fibrosis, and (total) inflammation with the manual scoring of their equivalent Banff lesion categories. The segmentation performance of (atrophic) tubuli should be improved to achieve better correlation with visual scoring of (atrophic) tubulitis and i-IFTA. Analyses on a small subset indicate an inverse correlation between long-term changes in eGFR and inflammation within scarred regions, based on both automated and visual assessment. Further validations are necessary to continuously assess the prospects of deep learning in kidney transplant pathology.

Acknowledgments

We thank Milly van de Warenburg for help with the generation of additional annotations for the structure segmentation network; colleagues from the routine histology laboratory for cutting the slides and staining the patient cohort with periodic acid−Schiff (PAS); Maud Wekking for help with restaining the PAS slides with CD3 immunohistochemistry; and Thomas de Bel for advice on neural network training. The web-based HistokatFusion software enabled easy image registration, thanks to Fraunhofer MEVIS' Nick Weiss and Johannes Lotz.

Supplementary Data

Supplemental material for this article can be found at *http://doi.org/10.1016/j.ajpath.2022.06.009*.

References

1. Mengel M, Reeve J, Bunnag S, Einecke G, Jhangri GS, Sis B, Famulski K, Guembes-Hidalgo L, Halloran PF: Scoring total inflammation is superior to the current Banff inflammation score in predicting outcome and the degree of molecular disturbance in renal allografts. Am J Transpl 2009, 9:1859−1867
2. Lefaucheur C, Gosset C, Rabant M, Viglietti D, Verine J, Aubert O, Louis K, Glotz D, Legendre C, Duong Van Huyen J, Loupy A: T cell-mediated rejection is a major determinant of inflammation in scarred areas in kidney allografts. Am J Transpl 2018, 18:377−390
3. Nankivell BJ, Shingde M, Keung KL, Fung CL-S, Borrows RJ, O'Connell PJ, Chapman JR: The causes, significance and consequences of inflammatory fibrosis in kidney transplantation: the Banff i-IFTA lesion. Am J Transpl 2018, 18:364−376
4. Mannon RB, Matas AJ, Grande J, Leduc R, Connett J, Kasiske B, Cecka JM, Gaston RS, Cosio F, Gourishankar S, Halloran PF, Hunsicker L, Rush D; DeKAF Investigators: Inflammation in areas of tubular atrophy in kidney allograft biopsies: a potent predictor of allograft failure. Am J Transpl 2010, 10:2066−2073
5. Roufosse C, Simmonds N, Clahsen-van Groningen M, Haas M, Henriksen KJ, Horsfield C, Loupy A, Mengel M, Perkowska-Ptasinska A, Rabant M, Racusen LC, Solez K, Becker JU: A 2018 reference guide to the Banff classification of renal allograft pathology. Transplantation 2018, 102:1795−1814
6. Loupy A, Haas M, Solez K, Racusen L, Glotz D, Seron D, et al: The Banff 2015 kidney meeting report: current challenges in rejection classification and prospects for adopting molecular pathology. Am J Transpl 2017, 17:28−41
7. Solez K, Axelsen RA, Benediktsson H, Burdick JF, Cohen AH, Colvin RB, Croker BP, Droz D, Dunnill MS, Halloran PF, Häyry P, Jennette JC, Keown PA, Marcussen N, Mihatsch MJ, Morozumi K, Myers BD, Nast CC, Olsen S, Racusen LC, Ramos E, Rosen S, Sachs DH, Salomon DR, Sanfilippo F, Verani R, von Willebrand E, Yamaguchi Y: International standardization of criteria for the histologic diagnosis of renal allograft rejection: the Banff working classification of kidney transplant pathology. Kidney Int 1993, 44:411−422
8. Veronese FV, Manfro RC, Roman FR, Edelweiss MI, Rush DN, Dancea S, Goldberg J, Gonçalves LF: Reproducibility of the Banff classification in subclinical kidney transplant rejection. Clin Transpl 2005, 19:518−521
9. Furness PN, Taub N; Convergance of European Renal Transplant Pathology Assessment Procedures (CERTPAP): International variation in the interpretation of renal transplant biopsies: report of the CERTPAP Project. Kidney Int 2001, 60:1998−2012
10. Marcussen N, Olsen TS, Benediktsson H, Racusen L, Solez K: Reproducibility of the Banff classification of renal allograft pathology: inter- and intraobserver variation. Transplantation 1995, 60: 1083−1089
11. Furness PN, Taub N, Assmann KJM, Banfi G, Cosyns J-P, Dorman AM, Hill CM, Kapper SK, Waldherr R, Laurinavicius A, Marcussen N, Martins AP, Nogueira M, Regele H, Seron D, Carrera M, Sund S, Taskinen EI, Paavonen T, Tihomirova T, Rosenthal R: International variation in histologic grading is large, and persistent feedback does not improve reproducibility. Am J Surg Pathol 2003, 27:805−810
12. Schinstock CA, Sapir-Pichhadze R, Naesens M, Batal I, Bagnasco S, Bow L, Campbell P, Clahsen-van Groningen MC, Cooper M, Cozzi E, Dadhania D, Diekmann F, Budde K, Lowe F, Orandi BJ, Rowshani AT, Cornell L, Kraus E: Banff survey on antibody-mediated rejection clinical practices in kidney transplantation: diagnostic misinterpretation has potential therapeutic implications. Am J Transpl 2019, 1:123−131
13. Sicard A, Meas-Yedid V, Rabeyrin M, Koenig A, Ducreux S, Dijoud F, Hervieu V, Badet L, Morelon E, Olivo-Martin JC, Dubois V, Thaunat O: Computer-assisted topological analysis of renal allograft inflammation adds to risk evaluation at diagnosis of humoral rejection. Kidney Int 2017, 92:214−226
14. Servais A, Meas-Yedid V, Noël LH, Martinez F, Panterne C, Kreis H, Zuber J, Timsit MO, Legendre Ch, Olivo-Martin JC, Thervet E: Interstitial fibrosis evolution on early sequential screening renal allograft biopsies using quantitative image analysis. Am J Transpl 2011, 11:1456−1463
15. Vuiblet V, Fere M, Gobinet C, Birembaut P, Piot O, Rieu P: Renal graft fibrosis and inflammation quantification by an automated Fourier-transform infrared imaging technique. J Am Soc Nephrol 2016, 27:2382−2391
16. Bándi P, Geessink O, Manson Q, Dijk M van, Balkenhol M, Hermsen M, et al: From detection of individual metastases to classification of lymph node status at the patient level: the CAMELYON17 challenge. IEEE Trans Med Imaging 2018, 38:550−560
17. Bulten W, Pinckaers H, Boven H van, Vink R, Bel T de, Ginneken B van, van der Laak J, Hulsbergen-van de Kaa C, Litjens G: Automated deep-learning system for Gleason grading of prostate cancer using biopsies: a diagnostic study. Lancet Oncol 2020, 21:233−241
18. Ehteshami Bejnordi B, Veta M, Diest PJ van, Ginneken B van, Karssemeijer N, Litjens G, et al: Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. JAMA 2017, 318:2199−2210
19. Litjens G, Kooi T, Ehteshami Bejnordi B, Setio AAA, Ciompi F, Ghafoorian M, van er Laak JAWM, van Ginneken B, Sánchez CI: A

survey on deep learning in medical image analysis. Med Image Anal 2017, 42:60−88

20. Bukowy JD, Dayton A, Cloutier D, Manis AD, Staruschenko A, Lombard JH, Solberg Woods LC, Beard DA, Cowley AW: Region-based convolutional neural nets for localization of glomeruli in trichrome-stained whole kidney sections. J Am Soc Nephrol 2018, 29:2081−2088
21. Kannan S, Morgan LA, Liang B, Cheung MG, Lin CQ, Mun D, Nader RG, Belghasem ME, Henderson JM, Francis JM, Chitalia VC, Kolachalama VB: Segmentation of glomeruli within trichrome images using deep learning. Kidney Int Rep 2019, 4:955−962
22. Ginley B, Lutnick B, Jen K-Y, Fogo AB, Jain S, Rosenberg A, Walavalkar V, Wilding G, Tomaszewski JE, Yacoub R, Rossi GM, Sarder P: Computational segmentation and classification of diabetic glomerulosclerosis. J Am Soc Nephrol 2019, 30:1953−1967
23. Gadermayr M, Dombrowski A-K, Klinkhammer BM, Boor P, Merhof D: CNN cascades for segmenting sparse objects in gigapixel whole slide images. Comput Med Imaging Graph 2019, 71:40−48
24. Ginley B, Jen K-Y, Han SS, Rodrigues L, Jain S, Fogo AB, Zuckerman J, Walavalkar V, Miecznikowski JC, Wen Y, Yen F, Yun D, Moon KC, Rosenberg A, Parikh C, Sarder P: Automated computational detection of interstitial fibrosis, tubular atrophy, and glomerulosclerosis. J Am Soc Nephrol 2021, 32:837−850
25. Hermsen M, Bel T de, Boer M den, Steenbergen EJ, Kers J, Florquin S, JJTH Roelofs, Stegall MD, Alexander MP, Smith BH, Smeets B, Hilbrands LB, van der Laak JAWM: Deep learning-based histopathologic assessment of kidney tissue. J Am Soc Nephrol 2019, 30:1968−1979
26. Swiderska-Chadaj Z, Pinckaers H, Rijthoven M van, Balkenhol M, Melnikova M, Geessink O, Manson Q, Sherman M, Polonia A, Parry J, Abubakar M, Litjens G, van der Laak J, Ciompi F: Learning to detect lymphocytes in immunohistochemistry with deep learning. Med Image Anal 2019, 58:101547
27. Ronneberger O, Fischer P, Brox T: U-Net: convolutional networks for biomedical image segmentation. Edited by Navab N, Hornegger J, Wells W, Frangi A. Medical Image Computing and Computer-Assisted Intervention−MICCAI 2015. Lecture Notes in Computer Science, Vol 9351, p. 234−241.
28. Kingma D, Ba J: Adam: a method for stochastic optimization. ArXiv 2014. doi: 10.48550/arXiv.1412.6980
29. Bándi P, Balkenhol M, Ginneken B van, Laak J van der, Litjens G: Resolution-agnostic tissue segmentation in whole-slide histopathology images with convolutional neural networks. PeerJ 2019, 7:e8242
30. Lotz J, Weiss N, Heldmann S: Robust, fast and accurate: a 3-step method for automatic histological image registration. ArXiv 2019. doi: 10.48550/arXiv.1903.12063
31. Becker JU, Mayerich D, Padmanabhan M, Barratt J, Ernst A, Boor P, Cicalese PA, Mohan C, Nguyen HV, Roysam B: Artificial intelligence and machine learning in nephropathology. Kidney Int 2020, 98:65−75
32. Barisoni L, Lafata KJ, Hewitt SM, Madabhushi A, Balis UGJ: Digital pathology and computational image analysis in nephropathology. Nat Rev Nephrol 2020, 16:669−685
33. Santo BA, Rosenberg AZ, Sarder P: Artificial intelligence driven next-generation renal histomorphometry. Curr Opin Nephrol Hypertens 2020, 29:265−272
34. Niel O, Bastard P: Artificial intelligence in nephrology: core concepts, clinical applications, and perspectives. Am J Kidney Dis 2019, 74:803−810
35. Pedraza A, Gallego J, Lopez S, Gonzalez L, Laurinavicius A, Bueno G: Glomerulus classification with convolutional neural networks. Edited by Valdés Hernández M, González-Castro V: Medical Image Understanding and Analysis. MIUA 2017. Communications in Computer and Information Science, Vol 723; 2017. pp. 839−849
36. Bouteldja N, Klinkhammer BM, Bülow RD, Droste P, Otten SW, Freifrau von Stillfried S, Moellmann J, Sheehan SM, Korstanje R, Menzel S, Bankhead P, Mietsch M, Drummer C, Lehrke M, Kramann R, Floege J, Boor P, Merhof D: Deep learning-based segmentation and quantification in experimental kidney histopathology. J Am Soc Nephrol 2021, 32:52−68
37. Bueno G, Fernandez-Carrobles MM, Gonzalez-Lopez L, Deniz O: Glomerulosclerosis identification in whole slide images using semantic segmentation. Comput Methods Programs Biomed 2020, 184:105273
38. Jayapandian CP, Chen Y, Janowczyk AR, Palmer MB, Cassol CA, Sekulic M, Hodgkin JB, Zee J, Hewitt SH, O'Toole J, Toro P, Sedor JR, Barisoni L, Madabushi A: Nephrotic Syndrome Study Network (NEPTUNE): development and evaluation of deep learning-based segmentation of histologic structures in the kidney cortex with multiple histologic stains. Kidney Int 2021, 99:86−101
39. Grimm PC, Nickerson P, Gough J, McKenna R, Stern E, Jeffery J, Rush DN: Computerized image analysis of Sirius Red-stained renal allograft biopsies as a surrogate marker to predict long-term allograft function. J Am Soc Nephrol 2003, 14:1662−1668
40. Farris AB, Alpers CE: What is the best way to measure renal fibrosis? a pathologist's perspective. Kidney Int Sup 2014, 24:9−15
41. Farris AB, Colvin RB: Renal interstitial fibrosis: mechanisms and evaluation. Curr Opin Nephrol Hypertens 2012, 21:289−300
42. Farris AB, Adams CD, Brousaides N, Della Pelle PA, Collins AB, Moradi E, Smith RN, Grimm PC, Colvin RB: Morphometric and visual evaluation of fibrosis in renal biopsies. J Am Soc Nephrol 2011, 22:176−186
43. Zheng Y, Cassol CA, Jung S, Veerapaneni D, Chitalia VC, Ren K, Bellur SS, Boor P, Barisoni LM, Waikar SS, Betke M, Kolachalama VB: Deep learning-driven quantification of interstitial fibrosis in digitized kidney biopsies. Am J Pathol 2021, 8: 1442−1453
44. Kolachalama VB, Singh P, Lin CQ, Mun D, Belghasem ME, Henderson JM, Francis JM, Salant DJ, Chitalia VC: Association of pathological fibrosis with renal survival using deep neural networks. Kidney Int Rep 2018, 3:464−475
45. Nankivell BJ, Borrows RJ, Fung CLS, O'Connell PJ, Chapman JR, Allen RDM: Delta analysis of posttransplantation tubulointerstitial damage. Transplantation 2004, 78:434−441
46. Sellarés J, de Freitas DG, Mengel M, Sis B, Hidalgo LG, Matas AJ, Kaplan B, Halloran PF: Inflammation lesions in kidney transplant biopsies: association with survival is due to the underlying diseases. Am J Transpl 2011, 11:489−499
47. Mengel M, Gwinner W, Schwarz A, Bajeski R, Franz I, Bröcker V, Becker T, Neipp M, Klempnauer J, Haller H, Kreipe H: Infiltrates in protocol biopsies from renal allografts. Am J Transpl 2007, 7: 356−365
48. Yi Z, Salem F, Menon MC, Keung K, Xi C, Hultin S, Al Rasheed MRH, Li L, Su F, Sun Z, Wei C, Huang W, Fredericks S, Lin Q, Banu K, Wong G, Rogers NM, Farouk S, Cravedi P, Shingde M, Smith RN, Rosales IA, O'Connell PJ, Colvin RB, Murphy B, Zhang W: Deep learning identified pathological abnormalities predictive of graft loss in kidney transplant biopsies. Kidney Int 2021, 101:288−298
49. Borovec J, Kybic J, Arganda-Carreras I, Sorokin DV, Bueno G, Khvostikov AV, Bakas S, Chang EIC, Heldmann S, Kartasalo K, Latonen L, Lots J, Noga M, Pati S, Punithakumar K, Ruusuvuori P, Skalski A, Tahmasebi N, Valkonen M, Venet L, Wang Y, Weiss N, Wodzinski M, Xiang Y, Xu Y, Yan Y, Yushkevich P, Zhao S, Munoz-Barrutia A: ANHIR: automatic non-rigid histological image registration challenge. IEEE Trans Med Imaging 2020, 39: 3042−3052