International Journal of Medical Informatics 141 (2020) 104231

Contents lists available at ScienceDirect

International Journal of Medical Informatics

journal homepage: www.elsevier.com/locate/ijmedinf

Classification of glomerular pathological findings using deep learning and nephrologist–AI collective intelligence approach

Eiichiro Uchino [a,b,1], Kanata Suzuki [c,1], Noriaki Sato [b,d], Ryosuke Kojima [d], Yoshinori Tamada [a], Shusuke Hiragi [b,e], Hideki Yokoi [b], Nobuhiro Yugami [c], Sachiko Minamiguchi [f], Hironori Haga [f], Motoko Yanagita [b,g,**], Yasushi Okuno [d,h,*]

[a] Department of Medical Intelligent Systems, Graduate School of Medicine, Kyoto University, Kyoto, Japan
[b] Department of Nephrology, Graduate School of Medicine, Kyoto University, Kyoto, Japan
[c] Fujitsu Laboratories LTD., Kawasaki, Japan
[d] Department of Biomedical Data Intelligence, Graduate School of Medicine, Kyoto University, Kyoto, Japan
[e] Division of Medical Informatics and Administration Planning, Kyoto University Hospital, Kyoto, Japan
[f] Department of Diagnostic Pathology, Graduate School of Medicine, Kyoto University, Kyoto, Japan
[g] Institute for the Advanced Study of Human Biology (ASHBi), Kyoto University, Kyoto, Japan
[h] RIKEN, The Drug Development Data Intelligence Platform Group, Yokohama, Japan

ARTICLE INFO

Keywords:
Renal pathology
Artificial intelligence
Deep learning
Collective intelligence

ABSTRACT

Background: Automated classification of glomerular pathological findings is potentially beneficial in establishing an efficient and objective diagnosis in renal pathology. While previous studies have verified the artificial intelligence (AI) models for the classification of global sclerosis and glomerular cell proliferation, there are several other glomerular pathological findings required for diagnosis, and the comprehensive models for the classification of these major findings have not yet been reported. Whether the cooperation between these AI models and clinicians improves diagnostic performance also remains unknown. Here, we developed AI models to classify glomerular images for major findings required for pathological diagnosis and investigated whether those models could improve the diagnostic performance of nephrologists.

Methods: We used a dataset of 283 kidney biopsy cases comprising 15,888 glomerular images that were annotated by a total of 25 nephrologists. AI models to classify seven pathological findings: global sclerosis, segmental sclerosis, endocapillary proliferation, mesangial matrix accumulation, mesangial cell proliferation, crescent, and basement membrane structural changes, were constructed using deep learning by fine-tuning of InceptionV3 convolutional neural network. Subsequently, we compared the agreement to truth labels between majority decision among nephrologists with or without the AI model as a voter.

Results: Our model for global sclerosis showed high performance (area under the curve: periodic acid-Schiff, 0.986; periodic acid methenamine silver, 0.983); the models for the other findings also showed performance close to those of nephrologists. By adding the AI model output to majority decision among nephrologists, out of the 14 constructed models, the results of the majority decision showed improvement in sensitivity for 10 models (four of them were statistically significant) and specificity for eight models (five significant).

Conclusion: Our study showed a proof-of-concept for the classification of multiple glomerular findings in a comprehensive method of deep learning and suggested its potential effectiveness in improving diagnostic accuracy of clinicians.

* Corresponding author at: Department of Biomedical Data Intelligence, Graduate School of Medicine, Kyoto University, 53 Shogoin-Kawahara-cho, Sakyo-ku, Kyoto 606-8507, Japan.
** Corresponding author at: Department of Nephrology, Graduate School of Medicine, Kyoto University, 54 Shogoin-Kawahara-cho, Sakyo-ku, Kyoto 606-8507, Japan.
E-mail addresses: motoy@kuhp.kyoto-u.ac.jp (M. Yanagita), okuno.yasushi.4c@kyoto-u.ac.jp (Y. Okuno).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.ijmedinf.2020.104231
Received 31 December 2019; Received in revised form 30 May 2020; Accepted 6 July 2020
Available online 11 July 2020

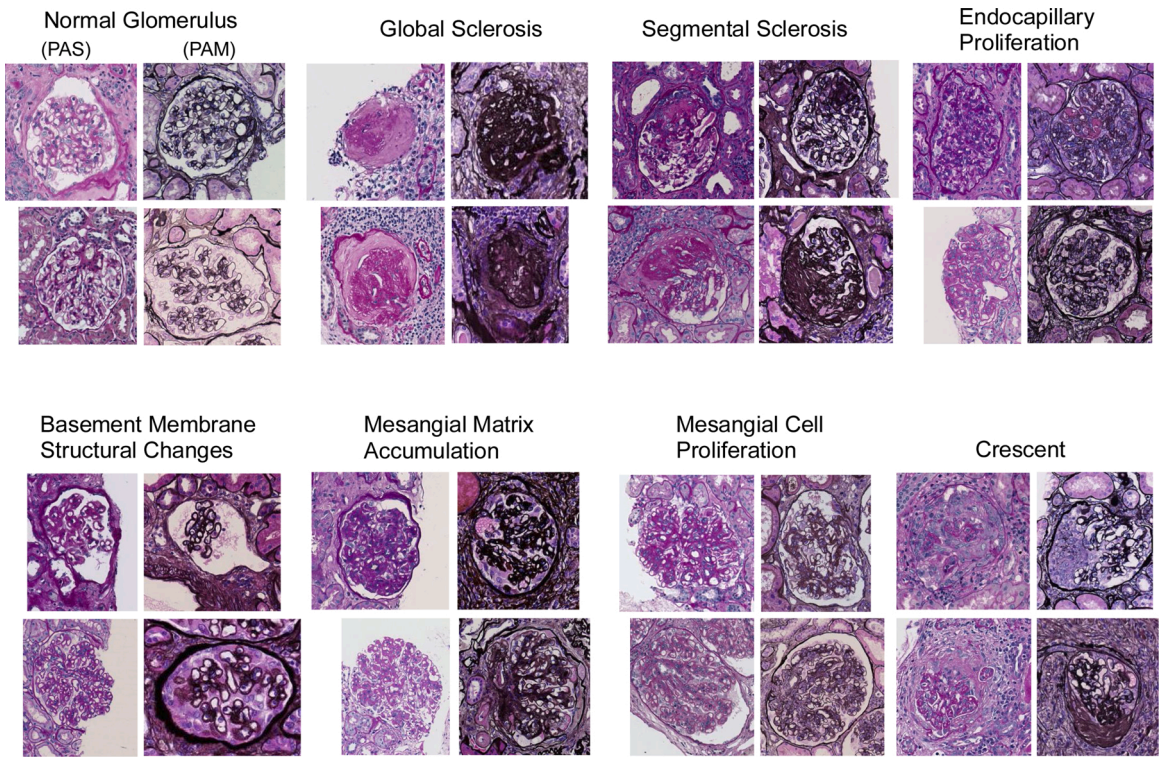


Fig. 1. Examples of glomeruli with each pathological finding.
The images show the representative glomeruli that were annotated as positive for the following findings: global sclerosis, segmental sclerosis, endocapillary proliferation, basement membrane structural changes, mesangial matrix accumulation, mesangial cell proliferation, and crescent formation.
PAS, periodic acid-Schiff; PAM, periodic acid methenamine silver

1. Introduction

Renal pathology is important for the diagnosis and management of patients with kidney disease. The renal survival rate tends to be better with histologic evaluation by renal biopsy than without renal biopsy [1], thus, accurate and robust diagnosis is essential for the proper management of patients with kidney disease. On the other hands, making an accurate diagnosis is a time-consuming process even for experienced pathologists. It has been expected that automated processing to support this procedure will improve the efficiency of renal pathology and contribute to a more objective and standardized diagnosis [2], especially in hospitals, areas, or countries where there are an insufficient number of nephropathologists. A field called digital pathology, which aims to diagnose and quantify disease based on image data obtained by scanning pathological tissue specimens, has rapidly been developed. With the use of current state-of-the-art techniques of deep learning (DL), the artificial intelligence (AI) approach has made a significant progress in medical image analysis of retinal fundus images [3], skin images [4], and pathology mainly on cancer [5]. Currently, the implementation of these technologies in the clinical process and their effect on healthcare workers are of great interest [6].

There are some studies trying to apply DL to renal pathology. While some studies have validated DL models analyzing the structures other than the glomeruli, such as the tubules, blood vessels, and interstitium [7–10], many studies have focused on the glomeruli, which present various histological findings essential for diagnosis. As a first step in the automation of this diagnostic procedure, detection of a glomerulus in a whole slide image (WSI) of renal tissue specimens has been recently attempted in many studies with the use of methods to define various features [11–24] or using convolutional neural networks (CNNs) [25], such as InceptionV3 [26], AlexNet [27], U-Net [28], R-CNN [29,30], or DeepLab V2 ResNet [31].

On the other hands, studies trying to classify pathologic findings from the glomerular images are still very few, and the pathological findings analyzed in these studies are quite limited. Barros et al. [32] constructed a model to classify proliferative lesions. Sheehan et al. [24] quantified mesangial matrix proliferation, numbers of nuclei, and capillary openness. Ginley et al. [31] also quantified nuclei, luminal space, and periodic acid-Schiff-positive component. Kannan et al. [26] and Marsh et al. [33] reported models to distinguish between sclerotic and nonsclerotic glomeruli. The pathological findings analyzed in these studies are quite limited, and do not cover the pathological findings necessary for accurate diagnosis, and there has been no study which enables the comprehensive evaluation of the essential pathological findings necessary for the diagnosis.

In this study, we focused on seven major pathological findings required for pathological diagnosis: global sclerosis, segmental sclerosis, endocapillary proliferation, mesangial matrix accumulation, mesangial cell proliferation, crescent, and basement membrane structural changes, and developed AI models to classify these findings. In addition, we examined whether our AI model can cooperate with nephrologists and improve their diagnostic performance. Although many studies have compared the performance between AI and the specialists [3,4], validation of the effect of the collaboration between AI and clinicians on the diagnostic judgment is also important and clinically relevant. Assuming a situation in which a majority decision of diagnosis is taken among specialists at a case conference, we demonstrated that the diagnostic performance was improved by adding AI model as one of the specialists.

2. Materials and methods

2.1. Data preparation

We used WSIs of 283 renal biopsy cases that were agreed to be used for research at the Kyoto University Hospital between 2012 and 2017. The renal biopsy samples, including the transplanted allografts, were

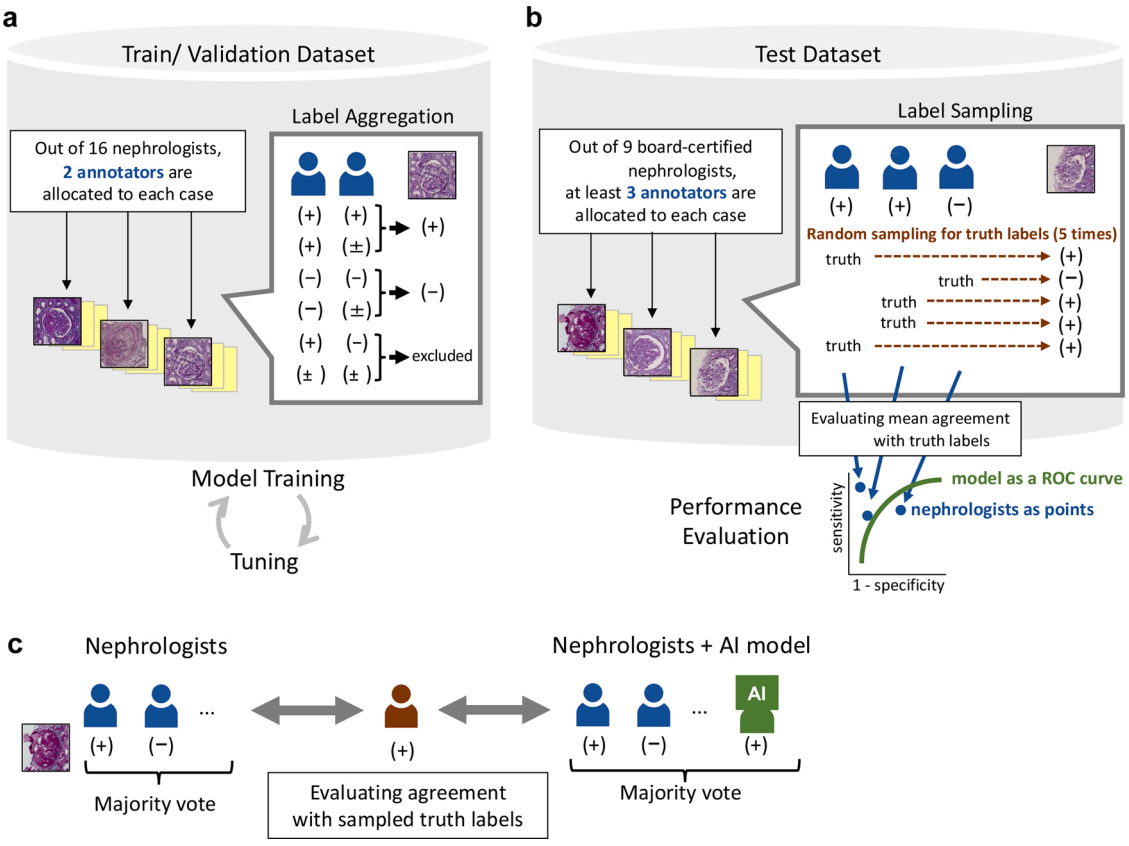


Fig. 2. Dataset construction and framework for training and testing of the models.
(a) Train/validation dataset. Out of a total of 16 nephrologists, two are assigned to each case and annotated glomerular images. For each image, two labels are aggregated to determine a truth label (label aggregation). This dataset is used for model training, tuning hyperparameters, and its validation. (b) Test dataset. Out of a total of nine board-certified nephrologists, at least three are assigned to each case and annotated glomerular images. For each image, a randomly selected label is adopted as a truth label (label sampling). This sampling process is repeated five times, and the average performance of the model is assessed in comparison with the nephrologists. (c) Nephrologist–AI collective intelligence approach by majority decision. It is examined whether adding model results to decision making can improve the performance. We compare the agreement to truth labels between majority decision among nephrologists alone and that with the AI model as a voter.

obtained by needle biopsy. Specimens that were stained by periodic acid-Schiff (PAS) and periodic acid methenamine silver (PAM) were used. Details of the staining and scanning of the slides are provided in the Supplementary Material.

Patients provided written informed consent for the use of the specimens in this research. Moreover, we posted an announcement regarding this research study on our department website and provided information on exclusion from participation in the study. The study protocol was approved by the Ethics Committee on Human Research of the Graduate School of Medicine, Kyoto University (No. R643−2 and G562).

2.2. Annotation of images

Using the ImageJ software [34], two nephrologists annotated and recorded the positions and coordinates of the glomeruli in all the WSIs. Subsequently, the pathological findings in the cropped glomerular images were annotated using an original graphical user interface-based input system (Supplementary Figure S1). The following seven findings in all glomeruli were respectively evaluated as positive (+), undecidable (±), or negative (−): global sclerosis, segmental sclerosis, endocapillary proliferation, basement membrane structural changes, mesangial matrix accumulation, mesangial cell proliferation, and crescent formation (examples in Fig. 1). In this study, basement membrane structural changes were defined as the presence of basement membrane thickening, spike formation, bubbling appearance, or double-contoured basement membrane. Additionally, for all glomeruli, the quality of the sample was evaluated and annotated as "artifact," which represents glomeruli that were not suitable for evaluation of the findings, such as those collapsed by external forces, not in focus, or had dust on them (examples in Supplementary Figure S2).

In Japan, as the number of nephropathologists is still quite small, nephrologists are trained and are practicing renal pathology in most clinical situations. Thus, we asked nephrologists to annotate the datasets. The annotators were blinded to the patient information, clinical information, and diagnosis, because this study aimed at judging the findings based on the image alone.

2.3. Train/validation dataset

We used the images obtained between January 2012 and June 2016 as the train/validation dataset, which was used for training and parameter tuning for the AI models. The train/validation dataset was annotated by a total of 16 nephrologists. Two nephrologists were randomly allocated to each case and independently conducted the annotation (Fig. 2a).

Subsequently, we performed label aggregation process in order to determine a truth label for each image (Fig. 2a). For training the model, an image in the train/validation dataset was defined as positive when the respective labels annotated by the two nephrologists were (+) and (+) or (+) and (±) or as negative when the respective labels were (−) and (−) or (−) and (±). Images that were respectively labeled as (+) and (−) or (±) and (±) were excluded from the dataset. Images with artifact labels were also excluded.

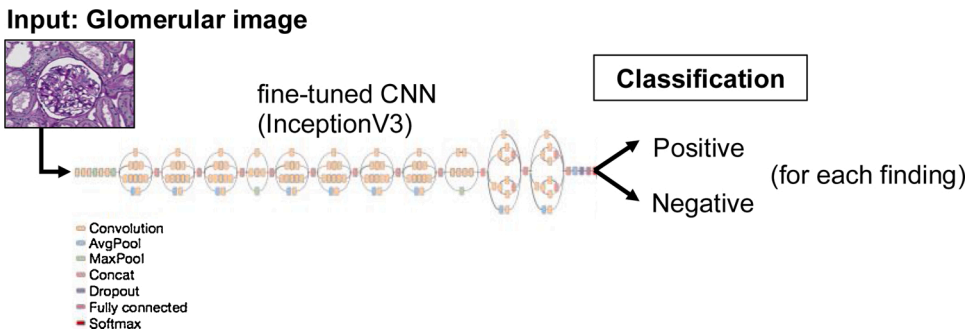


Fig. 3. Model abstract.
Classification of glomeruli by the fine-tuned InceptionV3 CNN in each finding. The model is trained for each of the seven findings of global sclerosis, segmental sclerosis, endocapillary proliferation, basement membrane structural changes, mesangial matrix accumulation, mesangial cell proliferation, and crescent formation. The outputs of the models are the degree of a positive finding when a cropped glomerular image is inputted.
CNN, convolutional neural network

2.4. Development of AI models for glomerular classification

We constructed models to classify a glomerular image as positive or negative for each pathological finding in each staining (Fig. 3). We performed fine-tuning with InceptionV3 [35], which is widely used for the classification problems of other medical images [3,4], with TensorFlow [36] as the backend. The models were trained and tested separately for each pathological finding. The technical details are provided in the Supplementary Material.

2.5. Test dataset

We used the images obtained between July 2016 and June 2017 as the test dataset to evaluate the performances of the models (Fig. 2b). The test dataset was annotated by a total of nine nephrologists, who were different from the evaluators of the train/validation dataset and had been board certified by the Japanese Society of Nephrology. At least three of the nephrologists were assigned to each case, and annotation was performed independently. Specifically for relatively rare remarks, such as crescent formation or basement membrane structural changes, cases with pathological diagnosis containing such remarks were allocated to a maximum of six nephrologists to attain enough numbers of positive labels for evaluating model performance and its comparison to nephrologists.

2.6. Performance evaluation of the models and nephrologists

The performance of the models was evaluated by the test dataset. The glomeruli with artifacts labels were excluded from the dataset in advance. We performed 5-time label sampling processes to determine a truth label for each image and to compare the performance between each nephrologist and the model (Fig. 2b). For each image in the test dataset, an annotated label by a randomly selected nephrologist was adopted as a truth label. The annotated labels by the other nephrologists were used to evaluate their performances. Images that were sampled as truth labels of undecidable (±) were excluded from the dataset within that sampling process. This sampling process was repeated five times, and the average performance of the model was calculated. The mean performance of each nephrologist was also calculated in the five sampling results. Images that were labeled as truth by sampling were excluded from the calculation of the sampled nephrologist's own performance.

The trained models were evaluated by area under the curve (AUC) of receiver operating characteristic (ROC) curve, sensitivity (or recall, true positive rate), and specificity (true negative rate). The performance of each annotator of the test dataset was evaluated by sensitivity and specificity.

2.7. Performance evaluation of the majority decision among nephrologists with the AI models

We examined whether the results of our models improved the sensitivity and specificity of the nephrologists in classifying each finding. We compared the agreement to truth labels between majority decision among nephrologists alone and that with the AI model as a voter (Fig. 2c). The truth labels were sampled by the same method stated above. A nephrologist whose label was chosen as truth was excluded from the voting. In the majority decision, when the number of positive and negative judgments in an image was the same, the result was randomly decided. Outputs of the AI models were determined by the mean cutoff values corresponding to the best Youden's indices. We also compared the performances of each individual nephrologist with and without the AI models. Statistical analyses were described in the Supplementary Material.

Table 1
Baseline case characteristics of the datasets.

	Train/validation dataset (N = 218)		Test dataset (N = 65)	
Sex				
Male	110	(50.5 %)	33	(50.8 %)
Female	108	(49.5 %)	32	(49.2 %)
Age, years				
Mean (SD)	52.1	(18.7)	50.2	(19.7)
Median (IQR)	51	(36–69)	47	(32–68)
Serum creatinine, mg/dL				
Mean (SD)	1.45	(1.25)	1.25	(1.14)
Median (IQR)	1.02	(0.73–1.64)	0.96	(0.65–1.52)
Pathological diagnosis				
IgA nephropathy	35	(16.1 %)	17	(26.2 %)
Lupus nephritis	25	(11.5 %)	7	(10.8 %)
Membranous nephropathy	23	(10.6 %)	3	(4.6 %)
Mesangial proliferative glomerulonephritis	23	(10.6 %)	8	(12.3 %)
Crescentic glomerulonephritis	12	(5.5 %)	1	(1.5 %)
Diabetic nephropathy	12	(5.5 %)	3	(4.6 %)
ANCA-associated crescentic glomerulonephritis	11	(5.0 %)	3	(4.6 %)
Focal segmental glomerulosclerosis	11	(5.0 %)	2	(3.1 %)
Minimal change nephrotic syndrome	11	(5.0 %)	3	(4.6 %)
Sclerosing glomerulonephritis	11	(5.0 %)	2	(3.1 %)
Interstitial nephritis	9	(4.1 %)	3	(4.6 %)
Henoch-Schönlein purpura nephritis	8	(3.7 %)	1	(1.5 %)
Anti-glomerular basement membrane glomerulonephritis	4	(1.8 %)	0	(0%)
Endocapillary proliferative glomerulonephritis	3	(1.4 %)	1	(1.5 %)
Others	20	(9.2 %)	11	(16.9 %)

ANCA, antineutrophil cytoplasmic antibody; SD, standard deviation; IQR, interquartile range.

3. Results

3.1. Patients and annotation of images

The train/validation and test datasets included 218 and 65 cases, respectively. The demographics and pathological diagnoses of these

Table 2
Annotated labels for the train/validation dataset.

	Positive label		Negative label		Excluded	
Global sclerosis						
PAS	834	(18.6 %)	3399	(75.7 %)	256	(5.7 %)
PAM	819	(17.3 %)	3663	(77.2 %)	262	(5.5 %)
Segmental sclerosis						
PAS	25	(0.6 %)	4240	(94.5 %)	224	(5.0 %)
PAM	19	(0.4 %)	4525	(95.4 %)	200	(4.2 %)
Endocapillary proliferation						
PAS	104	(2.3 %)	4018	(89.5 %)	367	(8.2 %)
PAM	100	(2.1 %)	4258	(89.8 %)	386	(8.1 %)
Basement membrane structural changes						
PAS	66	(1.5 %)	4142	(92.3 %)	281	(6.3 %)
PAM	101	(2.1 %)	4211	(88.8 %)	432	(9.1 %)
Mesangial matrix accumulation						
PAS	582	(13.0 %)	2953	(65.8 %)	954	(21.3 %)
PAM	272	(5.7 %)	3567	(75.2 %)	905	(19.1 %)
Mesangial cell proliferation						
PAS	304	(6.8 %)	3426	(76.3 %)	759	(16.9 %)
PAM	59	(1.2 %)	4219	(88.9 %)	466	(9.8 %)
Crescent formation						
PAS	112	(2.5 %)	4159	(92.7 %)	218	(4.9 %)
PAM	124	(2.6 %)	4435	(93.4 %)	185	(3.9 %)

PAS, periodic acid-Schiff; PAM, periodic acid methenamine silver.

Table 3
Annotated labels for the test dataset.

	Positive label		Negative label		Excluded	
Global sclerosis						
PAS	231.8 ± 3.3	(13.5 %)	1459.4 ± 6.2	(84.8 %)	12.8 ± 3.5	(0.7 %)
PAM	250.2 ± 6.2	(13.9 %)	1516.4 ± 6.2	(84.2 %)	10.4 ± 2.3	(0.6 %)
Segmental sclerosis						
PAS	44.8 ± 5.4	(2.6 %)	1651.8 ± 5.6	(96.0 %)	7.4 ± 2.3	(0.4 %)
PAM	35.6 ± 1.8	(2.0 %)	1734.6 ± 3.4	(96.3 %)	6.8 ± 2.3	(0.4 %)
Endocapillary proliferation						
PAS	96.0 ± 4.9	(5.6 %)	1572.2 ± 5.6	(91.4 %)	35.8 ± 4.8	(2.1 %)
PAM	64.2 ± 3.3	(3.6 %)	1672.6 ± 6.8	(92.8 %)	40.2 ± 4.1	(2.2 %)
Basement membrane structural changes						
PAS	40.2 ± 3.8	(2.3 %)	1649.2 ± 6.7	(95.8 %)	14.6 ± 3.9	(0.8 %)
PAM	82.0 ± 2.3	(4.6 %)	1671.8 ± 4.0	(92.8 %)	23.2 ± 4.1	(1.3 %)
Mesangial matrix accumulation						
PAS	847.6 ± 11.5	(49.3 %)	825.2 ± 11.3	(47.9 %)	31.2 ± 1.3	(1.8 %)
PAM	717.0 ± 11.8	(39.8 %)	1033.2 ± 12.7	(57.3 %)	26.8 ± 5.4	(1.5 %)
Mesangial cell proliferation						
PAS	667.4 ± 19.9	(38.8 %)	989.2 ± 19.3	(57.5 %)	47.4 ± 2.3	(2.8 %)
PAM	430.0 ± 12.7	(23.9 %)	1298.8 ± 13.6	(72.1 %)	48.2 ± 2.6	(2.7 %)
Crescent formation						
PAS	44.4 ± 6.9	(2.6 %)	1645.6 ± 9.1	(95.6 %)	14.0 ± 4.6	(0.8 %)
PAM	39.6 ± 3.0	(2.2 %)	1721.4 ± 3.5	(95.9 %)	16.0 ± 1.9	(0.9 %)

Values are expressed as mean ± standard deviation in the five sampling iterations.
PAS, periodic acid-Schiff; PAM, periodic acid methenamine silver.

cases are shown in Table 1. The median numbers of annotated images by one nephrologist were 1625 (532–1698 [minimum, maximum]). In the train/validation dataset, the cropped glomerular images comprised of 5571 images on PAS staining and 5876 images on PAM staining. After removing the images labeled as artifact (examples in Supplementary Figure S2) by at least one annotator, 4489 images on PAS staining and 4744 images on PAM staining were used for model construction. The images of 3.9–21.3 % were excluded from training due to disagreement in the annotators. The numbers of annotated labels in each finding are shown in Table 2. In the test dataset, the cropped glomerular images comprised 2175 images on PAS staining and 2266 images on PAM staining. After removing the images labeled as artifact by at least one annotator, 1704 images on PAS staining and 1777 images on PAM staining were used for performance evaluation. The numbers of annotated labels in each finding are shown in Table 3.

3.2. Performance of AI models for glomerular classification

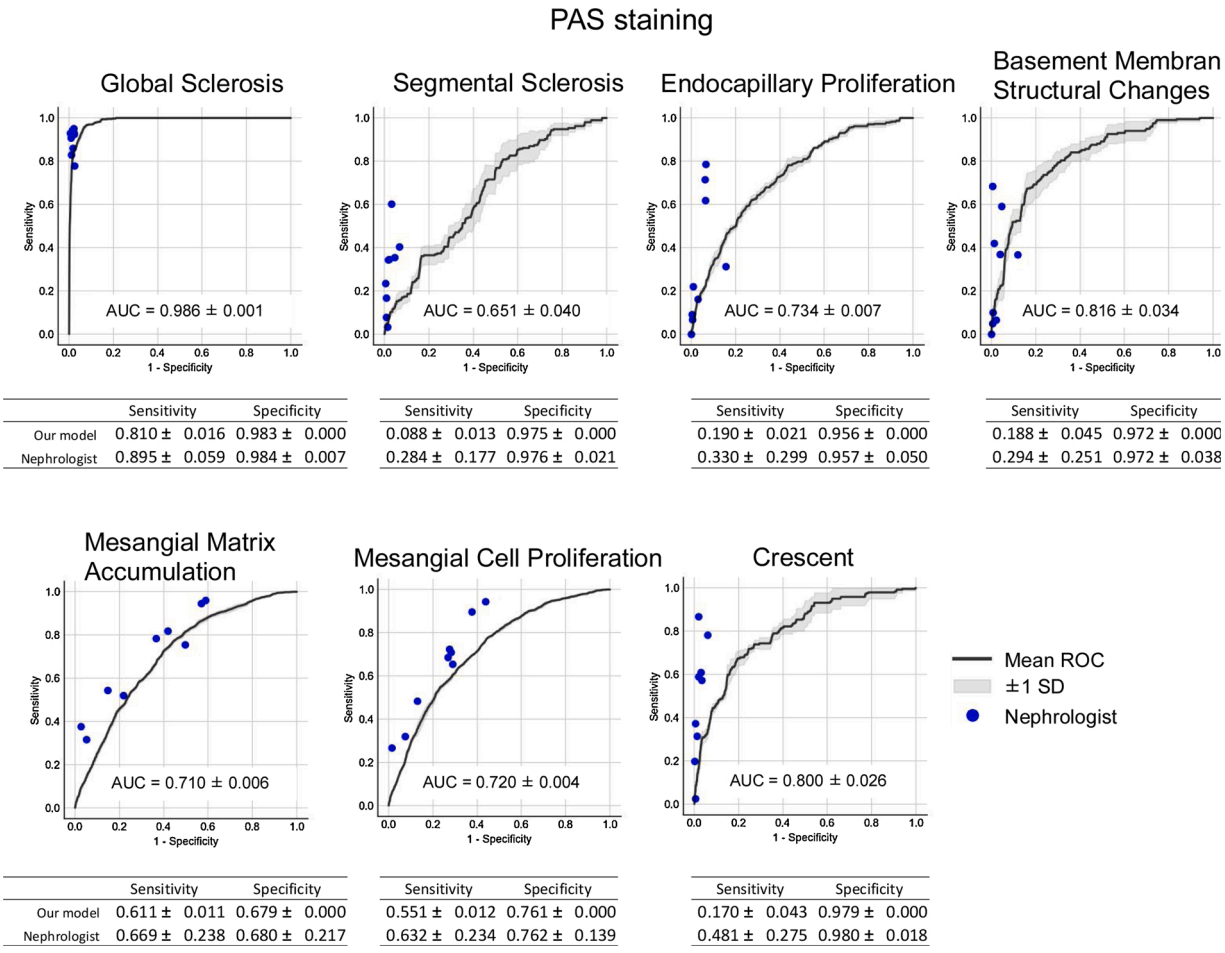
The performances of the models for classification of each pathological finding on PAS and PAM staining are shown in Figs. 4 and 5, respectively. The nephrologists showed high agreement for global sclerosis in both staining, and the models also showed high performance, with an AUC of 0.98. The classification examples of global sclerosis are shown in Supplementary Figure S3. In the other findings, the performance of the model ranged from an AUC of 0.59 to 0.87, with performance variation among nephrologists. For segmental sclerosis, endocapillary proliferation, membrane proliferation, and crescent formation, the nephrologists showed high specificity, but the sensitivity largely varied among them. Therefore, we evaluated the sensitivity of each model output based on a cutoff value that was the closest to the mean specificity of the nephrologists in the test dataset. The sensitivity of each model was lower than the average sensitivity of the nephrologists but exceeded the sensitivity of some nephrologists (Figs. 4 and 5).

3.3. Performance of the majority decision among nephrologists with and without the AI

The performance of the majority decision of with/ without the AI models is shown in Table 4. In the 14 constructed models, the results of the majority decision showed improvement in sensitivity for 10 models (four of them were in levels of p-values < 0.05) and specificity for 8 models (five of them of p-values < 0.05) when the model results were included (Table 4). Out of 1704 PAS and 1777 PAM images in the test dataset, there were 1.7–30.9 % disagreement in majority decision among nephrologists and the AI models substantially made the final decision (Supplementary Table S1). When the AI output was added to each nephrologist's decision, sensitivity was increased but specificity decreased in most of the findings (Supplementary Table S2).

4. Discussion

We constructed AI models to classify several pathological findings of glomerular images. To the best of our knowledge, this is the first study to verify classification models that comprehensively included as many as seven findings essential for renal pathological diagnosis. In the



	Global Sclerosis		Segmental Sclerosis		Endocapillary Proliferation		Basement Membran Structural Changes	
	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity
Our model	0.810 ± 0.016	0.983 ± 0.000	0.088 ± 0.013	0.975 ± 0.000	0.190 ± 0.021	0.956 ± 0.000	0.188 ± 0.045	0.972 ± 0.000
Nephrologist	0.895 ± 0.059	0.984 ± 0.007	0.284 ± 0.177	0.976 ± 0.021	0.330 ± 0.299	0.957 ± 0.050	0.294 ± 0.251	0.972 ± 0.038

	Mesangial Matrix Accumulation		Mesangial Cell Proliferation		Crescent	
	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity
Our model	0.611 ± 0.011	0.679 ± 0.000	0.551 ± 0.012	0.761 ± 0.000	0.170 ± 0.043	0.979 ± 0.000
Nephrologist	0.669 ± 0.238	0.680 ± 0.217	0.632 ± 0.234	0.762 ± 0.139	0.481 ± 0.275	0.980 ± 0.018

Fig. 4. Performance of glomerular classification for seven findings on PAS staining.
The performances of the models and nephrologists in the test dataset are shown as ROC curves and plotted points, respectively. Classification sensitivity and specificity are shown as mean ± SD. The cutoff points of the model outputs are determined by the nearest specificity points of the nephrologists' mean output.
PAS, periodic acid-Schiff; ROC, receiver operating characteristic; SD, standard deviation; AUC, area under the curve

classification of global sclerosis, our model showed high performance, AUC greater than 0.98, and that was also close to the performances of the nephrologists. Marsh et al. [33] reported a model to distinguish between sclerotic and nonsclerotic glomeruli on hematoxylin and eosin-stained sections of a renal graft. Their model, which used cut glomeruli beforehand, achieved a sensitivity of 0.865 and a specificity of 0.962 with the classification of glomerulosclerosis. Although these values are not comparable directly with our study because they vary with the cutoff of the output value from the model, the discriminative ability of our model was almost the same for global sclerosis (Figs. 4 and 5). There has been no previous report comparing experts and AI models in the renal pathology area. Our results showed that the current state-of-the-art AI models was not superior to the mean of experts but showed higher performance than some experts in classifying global sclerosis glomeruli. In the other findings, our AUC values were lower (0.6–0.8) than the results for global sclerosis. As for these findings, comparison of performance is difficult due to the difference of viewpoint from the existing research [24,32]. However, in all the findings, including crescents and membranous lesions that have not been reported for classification by AI, the present study showed performance close to nephrologists by our common method in all the findings. This result suggests the usefulness of DL in classifying many findings in renal pathology. In future studies, evaluation of the other findings, compared to global sclerosis, would need to focus on the specific parts of the glomeruli. For example, compared to an entire glomerulus, the endocapillary area, basement membrane, mesangial area, and extracapillary area are very fine structures. Moreover, there may be segmental findings that correspond to only a portion of the glomeruli. Therefore, performance may be improved by tuning, such as inputting images that are divided finely, rather than as an entire glomerulus. Some previous studies [32] showed that combination of quantification techniques may be important for findings that clinically need to be quantified, such as mesangial proliferation. In automation systems that use machine learning in other fields, performance can be improved by ensemble learning [37], which combines multiple machine learning models. In addition to the present technique, combination with these models or rule-based algorithms may further improve performance.

As an important point in our study, overall performance tended to improve when the output of the models was combined with the majority decision in nephrologists, compared to the majority decision in nephrologists alone. Notably, the current CNN can automatically extract general features from the training data [38], but it is difficult to correctly predict what greatly deviates from the data. In particular, in the pathological images with various phenotypes, compensation for situations, such as unprecedented or complicated results with pathophysiological theory or empirical knowledge, may be necessary. It is important to pay careful attention to how AI models and specialists can cooperate; however, only a few studies reported on the improvement in the prediction with the combined decision of humans and AI [39–41], and the effectiveness of a clinical decision support system that uses the AI technique has not been sufficiently verified [6]. In the research field on clinical decision support, collective intelligence approach has recently attracted

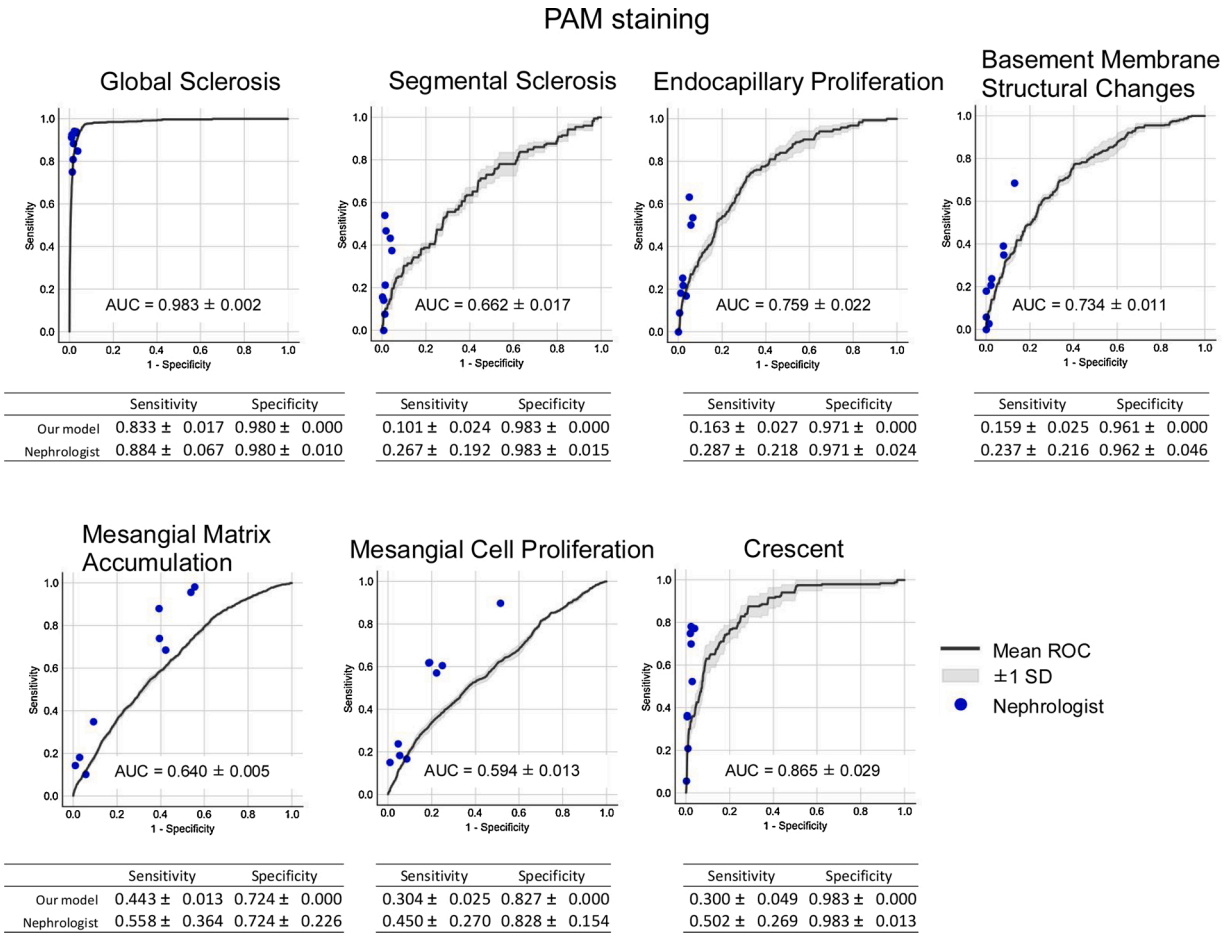


Fig. 5. Performance of glomerular classification for seven findings on PAM staining.
The performances of the models and nephrologists in the test dataset are shown as ROC curves and plotted points, respectively. Classification sensitivity and specificity are shown as mean ± SD. The cutoff points of the model outputs are determined by the nearest specificity points of the nephrologists' mean output.
PAM, periodic acid methenamine silver; ROC, receiver operating characteristic; SD, standard deviation; AUC, area under the curve

Table 4
Classification performance evaluation of the AI models and the majority decision among nephrologists with and without the AI models.

	Sensitivity				Specificity			
	AI model	Nephrologists	Nephrologists + AI model	p-value	AI model	Nephrologists	Nephrologists + AI model	p-value
Global sclerosis								
PAS	0.972 ± 0.005	0.919 ± 0.014	0.955 ± 0.007	0.0019	0.931 ± 0.002	0.984 ± 0.001	0.985 ± 0.002	0.75
PAM	0.967 ± 0.008	0.910 ± 0.024	0.944 ± 0.022	0.048	0.937 ± 0.003	0.984 ± 0.003	0.982 ± 0.002	0.28
Segmental sclerosis								
PAS	0.360 ± 0.049	0.229 ± 0.029	0.199 ± 0.044	0.24	0.831 ± 0.001	0.982 ± 0.002	0.986 ± 0.002	0.0059
PAM	0.101 ± 0.024	0.247 ± 0.035	0.146 ± 0.021	0.0011	0.979 ± 0.001	0.987 ± 0.002	0.995 ± 0.000	0.0013
Endocapillary proliferation								
PAS	0.577 ± 0.020	0.241 ± 0.036	0.331 ± 0.021	0.0021	0.757 ± 0.003	0.962 ± 0.006	0.955 ± 0.003	0.040
PAM	0.406 ± 0.055	0.245 ± 0.042	0.250 ± 0.056	0.88	0.858 ± 0.003	0.978 ± 0.003	0.976 ± 0.004	0.57
Basement membrane structural changes								
PAS	0.751 ± 0.058	0.249 ± 0.021	0.299 ± 0.033	0.024	0.746 ± 0.002	0.986 ± 0.004	0.977 ± 0.002	0.0027
PAM	0.503 ± 0.017	0.183 ± 0.007	0.207 ± 0.020	0.050	0.791 ± 0.001	0.963 ± 0.003	0.964 ± 0.001	0.32
Mesangial matrix accumulation								
PAS	0.574 ± 0.006	0.685 ± 0.014	0.686 ± 0.009	0.88	0.716 ± 0.003	0.653 ± 0.016	0.685 ± 0.006	0.0079
PAM	0.465 ± 0.007	0.570 ± 0.022	0.526 ± 0.011	0.0065	0.704 ± 0.004	0.708 ± 0.012	0.729 ± 0.010	0.021
Mesangial cell proliferation								
PAS	0.581 ± 0.011	0.638 ± 0.015	0.641 ± 0.012	0.78	0.728 ± 0.003	0.755 ± 0.011	0.764 ± 0.010	0.19
PAM	0.378 ± 0.018	0.488 ± 0.008	0.446 ± 0.012	0.00042	0.752 ± 0.006	0.829 ± 0.011	0.863 ± 0.003	0.0013
Crescent formation								
PAS	0.692 ± 0.029	0.375 ± 0.068	0.414 ± 0.053	0.34	0.761 ± 0.002	0.986 ± 0.003	0.984 ± 0.003	0.26
PAM	0.651 ± 0.067	0.441 ± 0.041	0.481 ± 0.055	0.23	0.885 ± 0.002	0.989 ± 0.001	0.989 ± 0.002	0.83

Values are expressed as mean ± standard deviation in five sampling iterations. P-values are evaluated between the majority decision among nephrologists with and without the AI models.
AI, artificial intelligence; PAS, periodic acid-Schiff; PAM, periodic acid methenamine silver.

attention, based on the reported higher diagnostic performance by several specialists than by a single specialist [42,43]. Although the performance of our models alone was not superior to that of the nephrologists, our results suggested that the use of the models for collective intelligence may improve the overall diagnostic performance in the clinical setting; this is a promising approach to improve the accuracy of team-based diagnosis. On the other hand, when an output of the AI models was combined with each individual nephrologist, the overall improvement in performances was not shown (Supplementary Table S2). In this setting, we had to adopt the final decision randomly in cases where their decisions were disagreed, which seems relatively different from the actual clinical situation. In future studies, it is necessary to examine how these models can actually change the decision-making or outcomes in the actual clinical setting.

This study has some limitations. A variety of annotated labels was observed among the annotators. This dataset was thought to reflect the actual variations among nephrologists, because a total of 25 annotators prepared the labels; thus, it is necessary to consider a more robust method to correct the discrepancies between evaluators. The numbers of images should be increased for the findings with the small number of positive labels, such as segmental sclerosis, although fine-tuning method had been used for training with a relatively small dataset. Our dataset was also limited to only PAS and PAM staining. Model construction and verification using a larger dataset will be required in the future. Also, clinical information was not used this study. In renal pathological diagnosis, since a suspicious disease or findings that should not be overlooked vary depending on clinical information, future models may utilize it, for example, to adjust cutoff values for each finding.

In conclusion, we developed a classification model for seven major findings in renal pathology and demonstrated that DL method is effective for classifying these findings. We also showed that the output of the model can improve the diagnostic performance of nephrologists. The use of these models in cooperation with nephrologists, may improve the diagnostic performance for renal pathology. Further study is required to develop models that can be used in the actual clinical setting.

Author statement

All persons who meet authorship criteria are listed as authors, and all authors certify that they have participated sufficiently in the work to take public responsibility for the content, including participation in the concept, design, analysis, writing, or revision of the manuscript. Furthermore, each author certifies that this material or similar material has not been and will not be submitted to or published in any other publication before its appearance in *International Journal of Medical Informatics*.

Transparency document

The Transparency document associated with this article can be found in the online version.

Summary Table

What was already known

- Deep learning models for analyzing renal pathology have been developed to establish an efficient and objective diagnosis in renal pathology, mainly for detecting glomeruli in whole slide images.
- Previous studies for the classification of glomerular pathological findings have focused on a limited number of findings, and there has been no study which enables the comprehensive evaluation of findings necessary for the diagnosis.
- It also remains unknown whether these models can improve diagnostic performance of clinicians.

What this study added to our knowledge

- Deep learning models can classify glomerular images for as many as seven major pathological findings essential for pathological diagnosis.
- Cooperation between nephrologists and these models is a potentially useful method to improve the diagnostic performance for renal pathology.

;1;

Declaration of Competing Interest

E. Uchino and Y. Tamada were given a budget for a joint research project with Fujitsu Ltd. M. Yanagita received research grants from Astellas, Chugai, Daiichi Sankyo, Kyowa Hakko Kirin, Mitsubishi Tanabe Pharma Corporation, MSD, Baxter, Takeda Pharmaceutical, KISSEI PHARMACEUTICAL, Dainippon Sumitomo Pharma, TAISHO TOYAMA PHARM, and Torii. The other authors declare no conflicts of interest.

Acknowledgements

For annotation to the dataset and discussion, we thank the following nephrologists in Department of Nephrology, Graduate School of Medicine, Kyoto University:

Yuki Sato, MD, PhD; Akira Ishii, MD, PhD; Keita P. Mori, MD, PhD; Naohiro Toda, MD, PhD; Keisuke Osaki, MD; Sayaka Sugioka, MD; Shinya Yamamoto, MD; Keiichi Kaneko, MD; Shunsuke Kawamura, MD; Youngna Kang, MD; Takahisa Yoshikawa, MD; Yukiko Kato, MD, PhD; Makiko Kondo, MD; Shigenori Yamamoto, MD; Yuichiro Kitai, MD; Akiko Oguchi, MD; Masahiro Takahashi, MD; Daisuke Takada, MD; Hiroyuki Arai, MD; Mitsuhiro Ichioka, MD; Koji Muro, MD; and Erina Ono, MD. We thank Kei Taneishi for his valuable technical assistance and suggestion.

This research was conducted as a joint research project and used funds from Kyoto University and Fujitsu Ltd. This work was also supported by the JSPS KAKENHI Grant Number JP18H05959. This research was partly conducted under collaborative research program with RIKEN. This research was partly supported by the Japan Agency for Medical Research and Development (AMED) under Grant Number JP18gm5010002 and by World Premier International Research Center Initiative (WPI), MEXT, Japan.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.ijmedinf.2020.104231.

References

[1] K. Iseki, F. Miyasato, H. Uehara, et al., Outcome study of renal biopsy patients in Okinawa, Japan, Kidney Int. 66 (2004) 914–919.
[2] T.J. Fuchs, J.M. Buhmann, Computational pathology: challenges and promises for tissue analysis, Comput. Med. Imaging Graph. 35 (2011) 515–530.
[3] V. Gulshan, L. Peng, M. Coram, et al., Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs, JAMA 316 (2016) 2402–2410.
[4] A. Esteva, B. Kuprel, R.A. Novoa, et al., Dermatologist-level classification of skin cancer with deep neural networks, Nature 542 (2017) 115–118.
[5] A. Janowczyk, A. Madabhushi, Deep learning for digital pathology image analysis: a comprehensive tutorial with selected use cases, J. Pathol. Inform. 7 (2016) 29.
[6] F. Cabitza, R. Rasoini, G. Gensini, Unintended consequences of machine learning in medicine, JAMA 318 (2017) 517–518.
[7] V. Bevilacqua, N. Pietroleonardo, V. Triggiani, et al., An innovative neural network framework to classify blood vessels and tubules based on Haralick features evaluated in histological images of kidney biopsy, Neurocomputing 228 (2017) 143–153.
[8] V.B. Kolachalama, P. Singh, C.Q. Lin, et al., Association of pathological fibrosis with renal survival using deep neural networks, Kidney Int. Rep. 3 (2018) 464–475.

[9] Y.G. Kim, G. Choi, H. Go, et al., A fully automated system using a convolutional neural network to predict renal allograft rejection: extra-validation with giga-pixel immunostained slides, Sci. Rep. 9 (2019) 5123.
[10] M. Hermsen, T. de Bel, M. den Boer, et al., Deep learning–Based histopathologic assessment of kidney tissue, J. Am. Soc. Nephrol. 30 (2019) 1968–1979.
[11] J. Zhang, J. Hu, Glomerulus extraction by optimizing the fitting curve, 2008 International Symposium on Computational Intelligence and Design 2 (2018) 169–172.
[12] J. Ma, J. Zhang, J. Hu, Glomerulus extraction by using genetic algorithm for Edge patching, 2009 IEEE Congress on Evolutionary Computation (2009) 2474–2479.
[13] Y. Hirohashi, R. Relator, T. Kakimoto, et al., Automated quantitative image analysis of glomerular desmin immunostaining as a sensitive injury marker in spontaneously diabetic torii rats, Journal of Biomedical Image Processing. 1 (2014) 20–28.
[14] T. Kato, R. Relator, H. Ngouv, et al., Segmental HOG: new descriptor for glomerulus detection in kidney microscopy image, BMC Bioinformatics 16 (2015) 1–16.
[15] M. Gadermayr, B. Klinkhammer, P. Boor, et al., Do we need large annotated training data for detection applications in biomedical imaging? A case study in renal glomeruli detection, Mach. Learn. Med. Imaging 10019 (2016) 18–26.
[16] R. Marée, S. Dallongeville, J.-C. Olivo-Marin, et al., An approach for detection of glomeruli in multisite digital pathology, 2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI) (2016) 1033–1036.
[17] Y. Zhao, E.F. Black, L. Marini, et al., Automatic glomerulus extraction in whole slide images towards computer aided diagnosis, 2016 IEEE 12th International Conference on E-Science (E-Science) (2016) 165–174.
[18] M. Ishikawa, S. Watanabe, N. Honda, et al., Extraction of glomeruli in whole slide imaging of kidney biopsy specimens, Medical Imaging 2017: Digital Pathology, Proc. of SPIE 10140 (2017), 101400.
[19] O. Simon, R. Yacoub, S. Jain, et al., Multi-radial LBP features as a tool for rapid glomerular detection and assessment in whole slide histopathology images, Sci. Rep. 8 (2018) 2032.
[20] P. Sarder, B. Ginley, J.E. Tomaszewski, Automated renal histopathology: digital extraction and quantification of renal pathology, Medical Imaging 2016: Digital Pathology, Proc. of SPIE 9791 (2016) 97910F.
[21] B. Ginley, J.E. Tomaszewski, P. Sarder, Automatic computational labeling of glomerular textural boundaries, Medical Imaging 2017: Digital Pathology, Proc. of SPIE (2017) 101400G.
[22] M. Gadermayr, D. Eschweiler, A. Jeevanesan, et al., Segmenting renal whole slide images virtually without training data, Comput. Biol. Med. 90 (2017) 88–97.
[23] B. Ginley, J.E. Tomaszewski, R. Yacoub, et al., Unsupervised labeling of glomerular boundaries using Gabor filters and statistical testing in renal histology, J. Med. Imaging Bellingham (Bellingham) 4 (2017), 021102.
[24] S.M. Sheehan, R. Korstanje, Automatic glomerular identification and quantification of histological phenotypes using image analysis and machine learning, Am. J. Physiol. Renal Physiol. 315 (2018) F1644–F1651.
[25] M. Temerinac-Ott, G. Forestier, J. Schmitz, et al., Detection of glomeruli in renal pathology by mutual comparison of multiple staining modalities, Proceedings of the 10th International Symposium on Image and Signal Processing and Analysis (2017) 19–24.
[26] S. Kannan, L.A. Morgan, B. Liang, et al., Segmentation of glomeruli within trichrome images using deep learning, Kidney Int. Rep. 4 (2019) 955–962.
[27] J. Gallego, A. Pedraza, S. Lopez, et al., Glomerulus classification and detection based on convolutional neural networks, J. Imaging 4 (2018) 20.
[28] M. Gadermayr, A.-K. Dombrowski, B.M. Klinkhammer, et al., CNN cascades for segmenting sparse objects in gigapixel whole slide images, Comput. Med. Imaging Graph. 71 (2018) 40–48.
[29] Y. Kawazoe, K. Shimamoto, R. Yamaguchi, et al., Faster R-CNN-Based glomerular detection in multistained human whole slide images, J. Imaging 4 (2018) 91.
[30] J.D. Bukowy, A. Dayton, D. Cloutier, et al., Region-based convolutional neural nets for localization of glomeruli in trichrome-stained whole kidney sections, J. Am. Soc. Nephrol. 29 (2018) 2081–2088.
[31] B. Ginley, B. Lutnick, K.Y. Jen, et al., Computational segmentation and classification of diabetic glomerulosclerosis, J. Am. Soc. Nephrol. 30 (2019) 1953–1967.
[32] G.O. Barros, B. Navarro, A. Duarte, et al., PathoSpotter-K: a computational tool for the automatic identification of glomerular lesions in histological images of kidneys, Sci. Rep. 7 (2017) 46769.
[33] J.N. Marsh, M.K. Matlock, S. Kudose, et al., Deep learning global glomerulosclerosis in transplant kidney frozen sections, IEEE Trans. Med. Imaging 2018 (37) (2018) 2718–2728.
[34] C.A. Schneider, W.S. Rasband, K.W. Eliceiri, NIH Image to ImageJ: 25 years of image analysis, Nat. Methods 9 (2012) 671–675.
[35] C. Szegedy, V. Vanhoucke, S. Ioffe, et al., Rethinking the inception architecture for computer vision, 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016) 2818–2826.
[36] M. Abadi, P. Barham, J. Chen, et al., TensorFlow: a system for large-scale machine learning, Th12 USENIX Symposium on Operating Systems Design and Implementation (2016) 265–283.
[37] L. Breiman, Bagging predictors, Mach. Learn. 24 (1996) 123–140.
[38] G. Hinton, R. Salakhutdinov, Reducing the dimensionality of data with neural networks, Science. 313 (2006) 504–507.
[39] R. Lindsey, A. Daluiski, S. Chopra, et al., Deep neural network improves fracture detection by clinicians, Proc National Acad Sci. 115 (2018) 11591–11596.
[40] P. Lakhani, B. Sundaram, Deep learning at chest radiography: automated classification of pulmonary tuberculosis by using convolutional neural networks, Radiology. 284 (2017) 574–582.
[41] C. Liew, The future of Radiology augmented with Artificial Intelligence: a strategy for success, Eur. J. Radiol. 102 (2018) 152–156.
[42] M. Wolf, J. Krause, P.A. Carney, et al., Collective intelligence meets medical decision-making: the collective outperforms the best radiologist, PLoS One 10 (2015), e0134269.
[43] M.L. Barnett, D. Boddupalli, S. Nundy, et al., Comparative accuracy of diagnosis by collective intelligence of multiple physicians vs individual physicians, JAMA Netw. Open 2 (2019), e190096.

SUPPLEMENTARY MATERIAL

Classification of glomerular pathological findings using deep learning and nephrologist–AI collective intelligence approach

Eiichiro Uchino, Kanata Suzuki, Noriaki Sato, Ryosuke Kojima, Yoshinori Tamada, Shusuke Hiragi, Hideki Yokoi, Nobuhiro Yugami, Sachiko Minamiguchi, Hironori Haga, Motoko Yanagita, Yasushi Okuno

S1. SUPPLEMENTARY METHODS

S1.1. Preparation of whole slide images

All renal biopsy specimens in the study period were manually stained according to the following protocol for clinical practice.

Periodic acid-Schiff (PAS) staining: Formalin-fixed paraffin sections were sliced into 3 μm and stained by a standard procedure using reagents such as 0.5% Orthoperiodic acid, and Schiff reagent (Merck, Darmstadt, Germany), 0.5% sodium pyrosulfite, Meyer's hematoxylin solution.

Periodic acid methenamine silver (PAM) staining: Formalin-fixed paraffin sections were sliced into 2 μm and stained by a standard procedure using reagents such as 0.5% Orthoperiodic acid, 0.5% thiosemicarbazide, mesenamine silver solution, 4% neutral buffered formalin, 0.2% gold chloride aqueous solution, 2% sodium thiosulfate, and Meyer's hematoxylin solution. In the final hematoxylin and eosin staining process, Leica ST5010 AutoStainer XL (Leica Biosystems, Wetzlar, Germany) system was used.

All renal biopsy specimens were scanned with NanoZoomer-2.0HT whole slide imager, digital pathology slide scanner, and the software NDP.scan 3.1.7 (Hamamatsu Photonics, Hamamatsu City, Japan), using ×40 lens (0.23 μm/ pixel). The quality of all the images was checked manually after scanning; if the slides were out of focus, new scans were performed. The image files of the slides were converted from NDPI to JPEG files by a custom Python script, with image shape dimensions ranging from 3072 to 31744 pixels in width and 5440 to 39424 pixels in height.

S1.2. Technical details of the fine-tuned CNN models for glomerular classification

Various parameters (i.e., learning rate 0.01, mini batch 100, and step 4000) were determined by a five-fold cross-validation in the train/validation dataset. Because data augmentation did not improve the performance, in terms of enlargement and rotation of images (data not shown), it was not done in the final model. We used the steepest descent method to optimize the parameters at learning rate of 0.01. The training batch size was 100, and the number of training

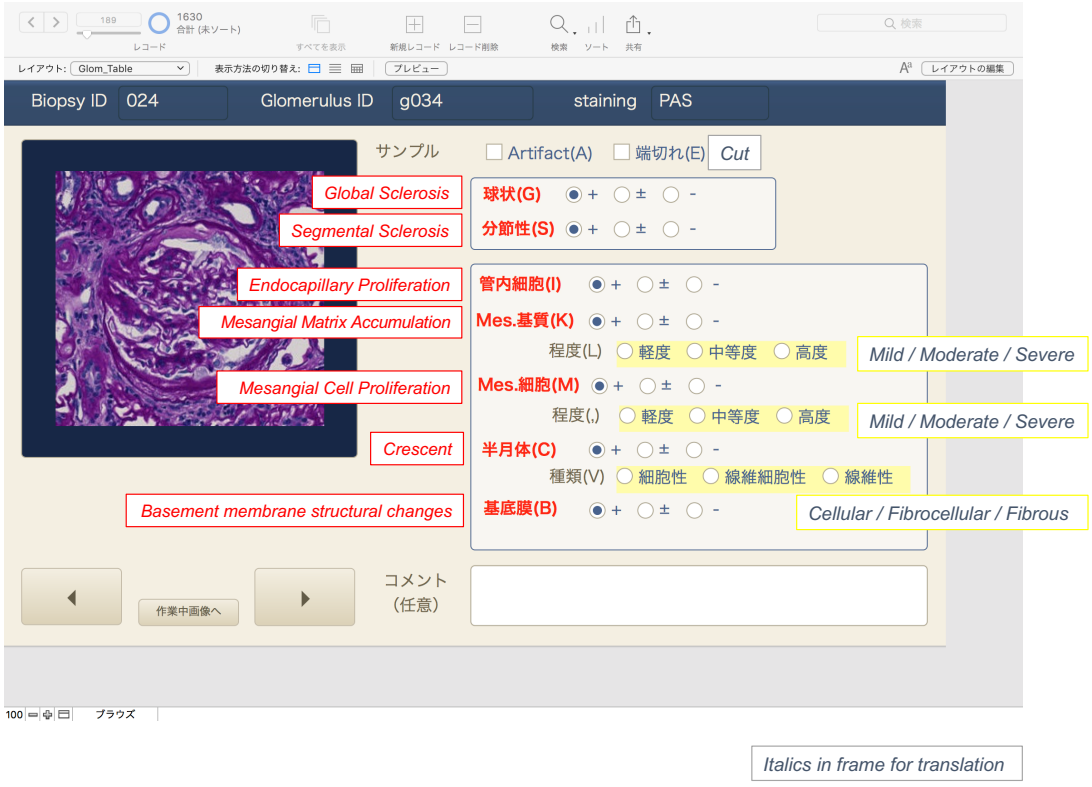
epoch was 4000. For training and testing of the model, a computer that was equipped with a graphics processing unit (GeForce GTX 1080, NVIDIA) was used.

S1.3. Statistics

Results of the sensitivity and specificity by the five-time sampling upon voting among the nephrologists, with and without the models, were evaluated by the Welch's t-test; p values <0.05 were considered statistically significant. The statistical analyses were performed by the R 3.5.1 software.

S2. SUPPLEMENTARY FIGURES AND TABLES

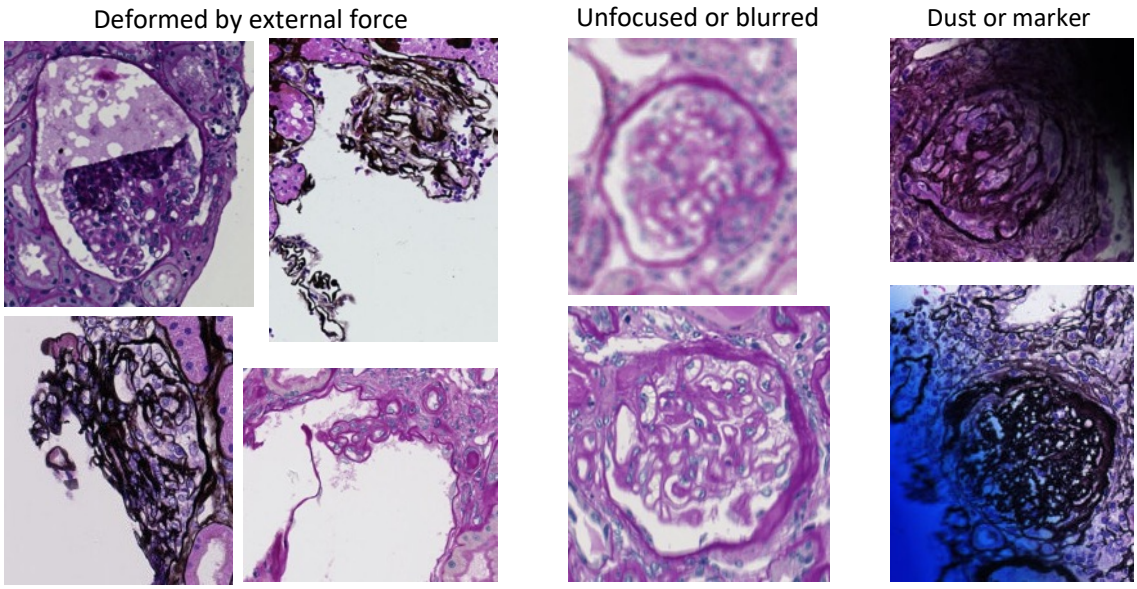
Supplementary Figure S1. Graphical user interface-based annotation system for the pathological findings



An original graphical user interface-based annotation system was developed using the FileMaker Pro 15 Advanced software (FileMaker, Inc.). The system was used in Japanese, but it was shown with an English translation. The following seven findings for all glomeruli were respectively evaluated as positive (+), undecidable (±), or negative (−): global sclerosis, segmental sclerosis, endocapillary proliferation, basement membrane structural changes, mesangial matrix accumulation, mesangial cell proliferation, and crescent formation. Each positive or undecidable mesangial matrix accumulation or cell proliferation was further labeled as severe, moderate, or mild degree. A positive or undecidable crescent formation was labeled as cellular, fibrocellular, or fibrous type of crescent. These labels were not used for constructing the models, because the total number of annotated positive labels for these findings was small. In addition, for all glomeruli, the quality of the sample was evaluated and annotated using the two items “cut” and “artifact.” Glomeruli that were separated by the biopsy needle were labeled

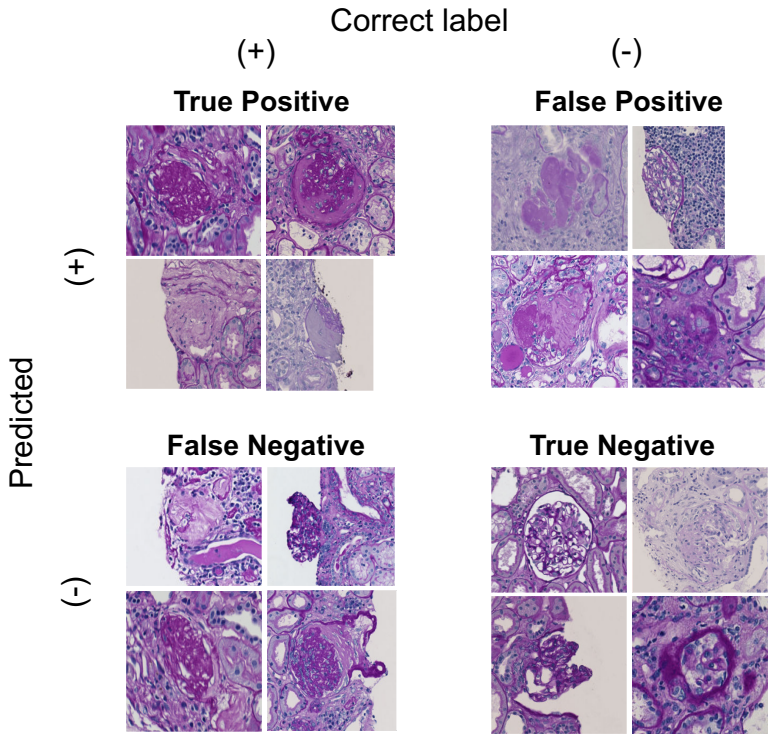
as “cut.” Glomeruli that were not suitable for evaluation of the findings, such as those collapsed by external forces, not in focus, or had dust on them, were labeled as “artifact.”

Supplementary Figure S2. Examples of glomeruli with artifact labels



The images are examples of glomerular images annotated as artifact, such as those deformed or collapsed by external forces, unfocused or blurred, or had dust or colored marker on them. These were excluded from the datasets because they were considered to be less useful for diagnosis in clinical use and inappropriate for evaluation of the findings.

Supplementary Figure S3. Examples of correct and error classification of global sclerosis on PAS staining



The true positive and true negative images are examples of correct classification. The false positive and false negative images represent errors. Notably, these images may be difficult for humans to judge.

PAS, periodic acid-Schiff

Supplementary Table S1. Performance of the AI models for glomeruli with disagreement in majority decision in nephrologists

	Number of disagreements in nephrologists		Sensitivity		Specificity	
			Nephrologists	Nephrologists + AI model	Nephrologists	Nephrologists + AI model
Global sclerosis						
PAS	29.2 ± 1.1	(1.7%)	0.497 ± 0.140	1.000 ± 0.000	0.302 ± 0.113	0.500 ± 0.087
PAM	37.2 ± 5.6	(2.1%)	0.422 ± 0.179	0.902 ± 0.041	0.471 ± 0.066	0.423 ± 0.088
Segmental sclerosis						
PAS	58.0 ± 5.9	(3.4%)	0.463 ± 0.119	0.214 ± 0.080	0.517 ± 0.052	0.697 ± 0.050
PAM	45.2 ± 2.4	(2.5%)	0.475 ± 0.109	0.217 ± 0.058	0.502 ± 0.115	0.912 ± 0.019
Endocapillary proliferation						
PAS	108.4 ± 5.3	(6.4%)	0.499 ± 0.085	0.637 ± 0.045	0.512 ± 0.084	0.481 ± 0.014
PAM	70.4 ± 5.8	(4.0%)	0.557 ± 0.082	0.518 ± 0.050	0.565 ± 0.078	0.634 ± 0.036
Basement membrane structural changes						
PAS	34.2 ± 3.2	(2.0%)	0.500 ± 0.373	1.000 ± 0.000	0.509 ± 0.120	0.578 ± 0.026
PAM	109.4 ± 2.6	(6.2%)	0.590 ± 0.097	0.836 ± 0.077	0.498 ± 0.051	0.549 ± 0.004
Mesangial matrix accumulation						
PAS	476.6 ± 4.4	(28.0%)	0.496 ± 0.037	0.497 ± 0.015	0.494 ± 0.042	0.600 ± 0.016
PAM	549.6 ± 22.3	(30.9%)	0.503 ± 0.056	0.401 ± 0.018	0.497 ± 0.031	0.565 ± 0.005
Mesangial cell proliferation						
PAS	408.8 ± 5.1	(24.0%)	0.484 ± 0.053	0.497 ± 0.011	0.505 ± 0.039	0.544 ± 0.021
PAM	379.8 ± 6.7	(21.4%)	0.507 ± 0.037	0.377 ± 0.025	0.502 ± 0.042	0.679 ± 0.011
Crescent formation						
PAS	40.0 ± 6.3	(2.3%)	0.430 ± 0.236	0.584 ± 0.065	0.501 ± 0.060	0.458 ± 0.054
PAM	30.4 ± 3.0	(1.7%)	0.310 ± 0.095	0.590 ± 0.153	0.470 ± 0.085	0.608 ± 0.053

The numbers of glomeruli in which nephrologists disagreed in majority voting and the AI models helped to the decision in test dataset. Values are expressed as mean ± standard deviation in five sampling iterations. AI, artificial intelligence; PAS, periodic acid-Schiff; PAM, periodic acid methenamine silver

Supplementary Table S2. Classification performance evaluation of each individual nephrologist with and without the AI models

	Sensitivity			Specificity		
	Nephrologist	Nephrologist + AI model	p-value	Nephrologist	Nephrologist + AI model	p-value
Global sclerosis						
PAS	0.874 ± 0.009	0.930 ± 0.005	< 0.0001	0.978 ± 0.003	0.954 ± 0.003	< 0.0001
PAM	0.863 ± 0.019	0.923 ± 0.019	0.0010	0.975 ± 0.005	0.952 ± 0.005	< 0.0001
Segmental sclerosis						
PAS	0.272 ± 0.07	0.344 ± 0.087	0.19	0.972 ± 0.001	0.896 ± 0.003	< 0.0001
PAM	0.261 ± 0.047	0.152 ± 0.041	0.0044	0.980 ± 0.001	0.979 ± 0.001	0.20
Endocapillary proliferation						
PAS	0.288 ± 0.024	0.452 ± 0.030	< 0.0001	0.941 ± 0.003	0.848 ± 0.004	< 0.0001
PAM	0.254 ± 0.039	0.301 ± 0.058	0.18	0.954 ± 0.002	0.903 ± 0.003	< 0.0001
Basement membrane structural changes						
PAS	0.243 ± 0.033	0.528 ± 0.074	< 0.0001	0.966 ± 0.003	0.854 ± 0.004	< 0.0001
PAM	0.229 ± 0.006	0.393 ± 0.031	0.0002	0.953 ± 0.002	0.875 ± 0.004	< 0.0001
Mesangial matrix accumulation						
PAS	0.655 ± 0.007	0.637 ± 0.007	0.0040	0.672 ± 0.004	0.693 ± 0.005	0.00018
PAM	0.547 ± 0.004	0.520 ± 0.014	0.0091	0.716 ± 0.006	0.717 ± 0.007	0.92
Mesangial cell proliferation						
PAS	0.611 ± 0.012	0.615 ± 0.007	0.46	0.746 ± 0.007	0.743 ± 0.009	0.57
PAM	0.434 ± 0.015	0.430 ± 0.019	0.69	0.814 ± 0.002	0.788 ± 0.004	< 0.0001
Crescent formation						
PAS	0.459 ± 0.033	0.613 ± 0.040	0.0002	0.969 ± 0.005	0.860 ± 0.011	< 0.0001
PAM	0.484 ± 0.015	0.626 ± 0.047	0.0016	0.978 ± 0.003	0.929 ± 0.003	< 0.0001

We compared the performances of each nephrologist with and without the AI models for the agreement to the truth labels. For each glomerulus, if a nephrologist and the AI model disagreed, the result was randomly decided. Values are expressed as mean ± standard deviation for the all the combination of annotated glomeruli and nephrologists in five sampling iterations, except for

the labels chosen for truth label. AI, artificial intelligence; PAS, periodic acid-Schiff; PAM, periodic acid methenamine silver