Available online at www.sciencedirect.com

ScienceDirect

Biomedical Journal

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

Original Article

Glomerular disease classification and lesion identification by machine learning

Cheng-Kun Yang [a,1], Ching-Yi Lee [a,1], Hsiang-Sheng Wang [b], Shun-Chen Huang [c], Peir-In Liang [d], Jung-Sheng Chen [e], Chang-Fu Kuo [e], Kun-Hua Tu [f], Chao-Yuan Yeh [a], Tai-Di Chen [b,*]

[a] aetherAI, Co., Ltd., Taipei, Taiwan
[b] Department of Anatomic Pathology, Chang Gung Memorial Hospital at Linkou, Taoyuan, Taiwan
[c] Department of Anatomic Pathology, Chang Gung Memorial Hospital at Kaohsiung, Kaohsiung, Taiwan
[d] Department of Pathology, Kaohsiung Medical University Hospital, Kaohsiung, Taiwan
[e] Center for Artificial Intelligence in Medicine, Chang Gung Memorial Hospital at Linkou, Taoyuan, Taiwan
[f] Department of Nephrology, Chang Gung Memorial Hospital at Linkou, Taoyuan, Taiwan

ARTICLE INFO

Article history:
Received 20 October 2020
Accepted 31 August 2021
Available online 8 September 2021

Keywords:
Machine learning
Deep learning
Kidney biopsy
Glomerulonephritis

ABSTRACT

Background: Classification of glomerular diseases and identification of glomerular lesions require careful morphological examination by experienced nephropathologists, which is labor-intensive, time-consuming, and prone to interobserver variability. In this regard, recent advance in machine learning-based image analysis is promising.

Methods: We combined Mask Region-based Convolutional Neural Networks (Mask R–CNN) with an additional classification step to build a glomerulus detection model using human kidney biopsy samples. A Long Short-Term Memory (LSTM) recurrent neural network was applied for glomerular disease classification, and another two-stage model using ResNeXt-101 was constructed for glomerular lesion identification in cases of lupus nephritis.

Results: The detection model showed state-of-the-art performance on variedly stained slides with F1 scores up to 0.944. The disease classification model showed good accuracies up to 0.940 on recognizing different glomerular diseases based on H&E whole slide images. The lesion identification model demonstrated high discriminating power with area under the receiver operating characteristic curve up to 0.947 for various glomerular lesions. Models showed good generalization on external testing datasets.

Conclusion: This study is the first-of-its-kind showing how each step of kidney biopsy interpretation carried out by nephropathologists can be captured and simulated by machine learning models. The models were integrated into a whole slide image viewing and annotating platform to enable nephropathologists to review, correct, and confirm the inference results. Further improvement on model performances and incorporating inputs from immunofluorescence, electron microscopy, and clinical data might realize actual clinical use.

* *Corresponding author*. Department of Anatomic Pathology, Chang Gung Memorial Hospital Linkou Main Branch, 5, Fuxing St., Guishan Dist., Taoyuan 333, Taiwan.
E-mail address: b8902028@msn.com (T.-D. Chen).

Peer review under responsibility of Chang Gung University.

[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.bj.2021.08.011

At a glance commentary

Scientific background on the subject

Pathological examination of kidney biopsy requires careful searching for glomeruli, distinguishing different kinds of glomerulonephritides, and identifying specific glomerular lesions, which is repetitive and time consuming. Advances in computer-aided image analysis and handy digital pathology tools might provide an opportunity to establish a better diagnostic workflow.

What this study adds to the field

The study here described the construction of three machine models capturing the crucial aspects of kidney biopsy interpretation including glomerular detection, glomerulonephritis classification, and multi-label glomerular lesion identification. In addition, we demonstrated how these models can be integrated in pathologists' workflow on a computational pathology platform to assist daily practice.

Introduction

Machine learning-based image analysis has gradually attracted much attention in the field of pathology. In recent years, deep learning techniques such as convolutional neural networks (CNN) have become the de facto standard in computer-aided image analysis [1–3]. Several utilities such as breast cancer lymph node metastasis identification [4], prostate cancer detection [5], and colon cancer outcome prediction have been described [6]. Most machine learning studies in nephrology focus on utilizing clinical parameters and biomarkers [7]. Prediction of developing acute kidney injury after coronary artery intervention, forecasting rapid kidney function decline in patients with diabetes, subtyping transplant glomerulopathy in renal transplant recipients, and risk stratifying IgA nephropathy have been reported [8–11]. On the other hand, research on deep learning-based image analysis is still scarce in the field of medical renal pathology. Kolachalama et al. reported predicting renal survival from pathological images [12]. Zeng et al. utilized deep learning to identify different intrinsic glomerular cell types [13]. Few other studies have focused on different structural compartments such as interstitium, peritubular capillaries, and interstitial fibrosis [14–16], and some hand-crafted feature-based methods or CNN-based methods for glomerulus detection have been reported [17–23].

Pathological examination of kidney biopsy requires careful inspection of glomerular, tubular, interstitial, and vascular compartments in the biopsy. In most of the time, examination of glomeruli is crucial for diagnosis. However, examination of glomeruli does not end at glomerulus detection but needs further interpretations including distinguishing different kinds of glomerulonephritides and identifying glomerular lesions. For example, by combining glomerular histologic features and information from immunofluorescence and clinical laboratory data, nephropathologists first identify a kidney biopsy as a case of lupus nephritis and then search for different kinds of lesions in glomeruli. The searching needs to be repeated on every glomerulus to get a final interpretation. It is tedious, time consuming, and suffers from poor concordance even among expert nephropathologists [24,25]. In this regard, computer-aided image analysis may be helpful in establishing a better diagnostic workflow [26]. However, studies on identifying glomerular lesions using computer-aided image analysis techniques are scarce, and they focused on only one or few pathological features such as global sclerosis, hypercellularity, or glomerular capillary loop thickening [20,27–30]. Furthermore, a model distinguishing different glomerular disease based on light microscopic morphology has not been reported.

In this study, we demonstrated an integrated computational pathology workflow on kidney biopsy whole slide images (WSIs) by a sequence of glomerular detection, glomerulonephritis classification, and glomerular lesion identification models. Previous studies mostly focused on single task such as glomerulus detection or single lesion identification. The current study represents the first-of-its-kind aiming at capturing and integrating multiple tasks what nephropathologists carry out when interpreting kidney biopsies on a computational pathology platform.

Methods

Datasets

Kidney biopsy slides between 2015 and 2019 from Linkou Chang Gung Memorial Hospital (LKCGMH), Kaohsiung Chang Gung Memorial Hospital (KSCGMH), and Kaohsiung Medical University Memorial Hospital (KMUMH) were used. Numbers of case for model building are described separately in the following sections, and a comprehensive list of cases/lesions for model testing is shown in Table 1. Patient characteristics of each model are shown in supplemental Table 1-3. The slides were centrally scanned at LKCGMH by a NanoZoomer S360 Digital slide scanner C13220-01 at 400X magnification, resulting in WSIs with highest resolution (0.23 μm) possible. Human data collection and the use of tissue sections followed protocols approved by the Chang Gung Medical Foundation Institutional Review Board (IRB No.: 201900002B0C501) and Kaohsiung Medical University Memorial Hospital (KMUHIRB-E(I)-20200,193).

Glomerulus detection

1379 kidney biopsy slides from LKCGMH were used, and 15,298, 5649, 5641, and 5679 glomeruli on hematoxylin and eosin (H&E), periodic acid–Schiff (PAS), periodic acid methenamine-silver (PAM), and trichrome stained sections were annotated. Annotations were done by three trained assistants and reviewed by a nephropathologist through a Django-based tool

Table 1 Number of cases/lesions for model testing.

Model							
Detection		Case Number	Number of Glomerulus				
	LKCGMH	20	1585				
	KSCGMH	20	4211				
	KMUMH	20	2837				
	Total	60	8633				
Classification		Glomerular Disease Categories and Number for Testing					
		Total	DM	IGAN	LN	MCD/FSGS	MN
	LKCGMH	50	10	10	10	10	10
	KSCGMH	103	21	15	30	24	13
	KMUMH	105	22	26	22	16	19
	Total	258	53	51	62	50	42
Identification		Lesion Categories and Numbers Used for Testing (from 20 Cases of Each Hospital)					
		GS	CR	EC	HD	NK	SS
	LKCGMH	125	70	339	91	118	57
	KSCGMH	120	12	117	15	17	55
	KMUMH	57	20	300	61	28	27
	Total	302	102	756	167	163	139

DM, diabetic nephropathy; IGAN, IgA nephropathy; LN, lupus nephritis; MCD/FSGS, minimal change disease/focal segmental glomerulosclerosis; MN, membranous nephropathy; GS, global sclerosis; CR, cellular/fibrocellular crescents; EC, endocapillary hypercellularity; HD, hyaline deposits; NK, neutrophils/karyorrhexis; SS, segmental sclerosis.

developed by aetherAI, Co., Ltd., an image AI company specialized in digital pathology. A two-stage glomerulus detection model was proposed. The model detected and segmented glomeruli from WSIs by trained Mask Region Based Convolutional Neural Networks (Mask R−CNN) [31] at the first stage and reduced false positive inferences by a second refining stage. We used ImageNet-pretrained ResNet-101 as the backbone for the Mask R−CNN. The WSIs were too big to be fed into the model. Therefore, we applied bilinear interpolation to downscale the WSIs by 64 times, and the annotated glomeruli were cropped out at a fixed size of 512 × 512 pixels. The cropped image patches were randomly shifted from −128 to +128 pixels for data augmentation. The model was firstly built with H&E images, and transfer learning was adopted to train three additional models for PAS, PAM, and trichrome. For inference, a sliding window method of 512 × 512 pixels with strides of 256 was used to enumerate the WSIs. The second refining stage was trained with false-positive inference output from the first stage Mask R−CNN and ground truth glomeruli using ResNeXt-101 [32]. All inferences output from the Mask R−CNN were fed into the second refining stage to get the final prediction. 20 additional biopsies from each of the three hospitals (8633 glomeruli in total) were randomly chosen for testing.

Glomerular disease classification

Biopsies diagnosed as diabetic nephropathy (104 case), IgA nephropathy (123 case), ISN/RPS class III or IV lupus nephritis (148 case), minimal change disease/focal segmental glomerulosclerosis (202 cases), and membranous glomerulopathy (76 case) at LKCGMH from 2015 to 2019 were used for training. Glomerular images were cropped at a fixed size of 1024 × 1024 pixels by the glomerular detection model mentioned above, and a label was given based on the diagnosis of the case. For example, all glomeruli from a case of lupus nephritis were labeled as so. A two-stage model was proposed. The first stage of the model is a single glomerulus classification task predicting the glomerulus label. A DenseNet-based CNN model [33] was trained with balanced data generated by data augmentation and output probabilities of each of the five disease classes for every single glomerulus. For tuning, we sequentially doubled layer units or deleted layer units for each layer from the input to the output while freezing other layers. Focal loss [34] was applied to force the model focus on the difficult glomeruli while training. The second stage of the model is a 4-layer-stacked bi-directional Long Short-Term Memory (LSTM) model [35] using the 5-class probabilities of each glomerulus output from the first stage CNN for disease classification on case basis. For a case with N glomeruli on the tissue section, a 5 x N matrix array was created by the 5-class probabilities of each of the N glomeruli sorted according to the order of class probability. This process was repeated for five times for each of the five disease classes, and a (5 × 5) x N combined matrix as the input of LSTM was created for each of the N glomeruli from a single case. Model generalization was achieved by randomly combining glomeruli from the same category to create a corresponding pseudo-sample for each training example.

Multiclass and multilabel identification of glomerular lesions

A two-stage glomerular lesion identification model was proposed. For training, 146 class III or IV (±class V) lupus nephritis biopsies diagnosed at LKCGMH were used. 5459 glomerular images were extracted by the glomerular detection model from H&E WSIs and annotated by a nephropathologist. Pure class I, II, and V lupus nephritis cases are not included because they do not have scorable lesions defined by ISN/RPS classification.

Three exclusive labels: unremarkable (the glomerulus is either normal or shows only reactive change), global sclerosis (the glomerulus is totally sclerotic), and abnormal, NOS (the glomerulus has pathological findings, but the findings are not scorable lesions defined in lupus nephritis ISN/RPS classification), or one or more non-exclusive labels (endocapillary hypercellularity, neutrophils/karyorrhexis, fibrinoid necrosis, hyaline deposits, cellular/fibrocellular crescents, segmental sclerosis, and fibrous crescents) were assigned to each glomerulus. The term "exclusive label" used here means the label is stand-alone and does not co-exist with other labels. On the other hand, "non-exclusive" means the label can co-exist with other labels. Fibrinoid necrosis and fibrous crescents were not included given these two lesions were very few in our datasets. For training and testing, glomeruli labeled "unremarkable" and "abnormal, NOS" were grouped together. In the first stage of the model, we used an ImageNet-pretrained ResNeXt-101 which outputs probabilities of three group labels which are Unremarkable/Abnormal, NOS (no scorable lesions), Global sclerosis, and Multi-label (having one or more scorable lesions). If Multi-label is favored according to the output, the image will subsequently be fed into second stage of the model to determine the presence/absence of various non-exclusive labels. The second stage of the model used the same ImageNet-pretrained ResNeXt-101 as the backbone but the last layer was changed to sigmoid activation. Rotation, random shifting, hue adjustment, color augmentation, and random oversampling in minority class were adopted. Early stopping was used to prevent overfitting. 20 cases from each of the three hospitals (2482 glomeruli in total) were randomly chosen for testing.

Statistics and performance analysis

The performance of models was assessed by area under the receiver operating characteristic curve (AUC), accuracy, balanced F score (F1 score), recall, and precision. Comparisons between human and model performance and comparisons between different datasets were assessed by paired or independent t-test.

Results

Glomerulus detection

To facilitate downstream glomerular disease classification and lesion identification, a robust glomerular instance segmentation model was developed. 32,267 glomeruli of 1379 kidney biopsy slides were used for training and validation, and the model was tested on 8633 glomeruli of 60 biopsies obtained from three hospitals [Table 1]. The performance of the model on four different histochemically stained kidney biopsy WSIs from three hospitals is summarized in Table 2. Overall F1 scores, which strike a balance between recall (true positive rate) and precision (positive predictive value), were around 0.9. In general, the model did best on trichrome stained slides (highest F1 score: 0.944). Detection performance on H&E slides from LKCGMH, but not on other stainings, was better over slides from KSCGMH and KMUMH [Fig. 1A]. The numbers of glomerulus correctly inferred by the model were significantly higher than the numbers reported in pathological reports ($p < 0.001$, Fig. 1B). The performance on external testing dataset with lower F1 score was comparable with that on internal testing dataset and did not show further improvement after fine tuning, suggesting good generalization of the model [Fig. 1C]. The number of glomerulus documented on pathological reports, annotated by nephropathologists (representing the ground truth), and correctly inferred by glomerular detection model showed nearly perfect correlation ($R^2 = 0.7888–0.9962$ by Pearson correlation coefficient, $p < 0.0001$; Fig. 2A and B).

Glomerular disease classification

Further lesion identification and interpretation of glomerular diseases depend on primary glomerular disease classification. For example, ISN/RPS class will be assigned to a case of lupus nephritis [36], and Tervaert classification would be applied on cases of diabetic nephropathy [37]. Therefore, a glomerular disease classification model was trained and validated by 653 kidney biopsies and tested on 258 biopsies obtained from three hospitals. Detailed numbers and distribution of glomerular disease categories in the testing dataset were shown in Table 1. A representative image of the glomerulus belonging to the most frequently encountered glomerular diseases and the performance of the classification model were showed in Fig. 3A. The average classification accuracies were 0.864, 0.794, and 0.783 on LKCGMH, KSCGMH, and KMUMH datasets. The best performance on LKCGMH testing dataset was achieved by lupus nephritis (accuracy: 0.940; recall: 0.800; precision: 0.889) and followed by diabetic nephropathy (accuracy: 0.880; recall: 0.900; precision: 0.643). The classification accuracies of lupus nephritis among different datasets are much varied (0.699–0.940) compared to that of diabetic

Table 2 Performance of the glomerular detection model.

	H&E	PAS	PAM	TRI
	F1 Score / Precision / Recall	F1 Score / Precision / Recall	F1 Score / Precision / Recall	F1 Score / Precision / Recall
LKCGMH	0.935 / 0.922 / 0.947	**0.881 / 0.871 / 0.891**	0.887 / 0.870 / 0.904	**0.944 / 0.950 / 0.938**
KSCGMH	0.890 / 0.908 / 0.873	0.889 / 0.934 / 0.848	0.908 / 0.908 / 0.908	0.932 / 0.969 / 0.898
KMUMH	0.927 / 0.901 / 0.955	0.898 / 0.949 / 0.852	0.911 / 0.936 / 0.888	N/A

Dark bold font indicates the case group with the best performance. Grey bold font indicates the case group with the worst performance. N/A: case group not available (trichrome stain is not routinely done in KMUMH).

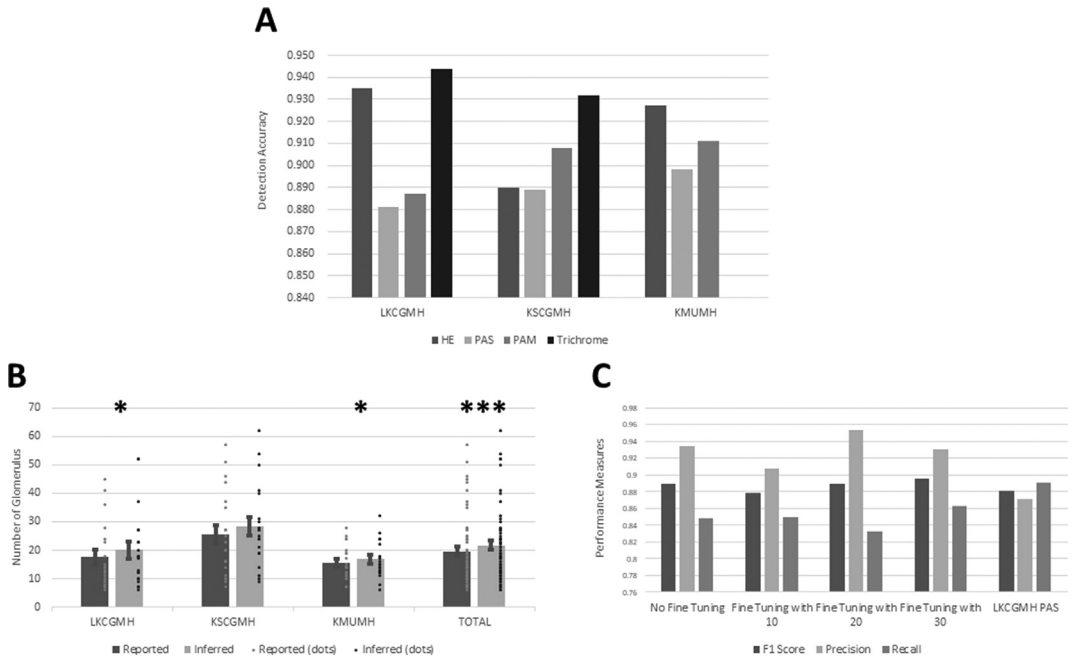


Fig. 1 **Performance of the glomerulus detection model.** (A) Grouped bar chart of the performance of glomerular detection model on variedly stained slides from three different hospitals. (B) Comparison of the numbers of glomerulus documented in pathological reports and the numbers of glomerulus correctly inferred by the glomerular detection model. *$p < 0.05$, ***$p < 0.001$. Presented as grouped bar chart with dot plots (individual numbers) and error bars as mean ± SEM. (C) Performance of the glomerulus detection model on KSCGMH PAS-stained dataset before and after fine tuning. There was no obvious improvement on F1 score, precision, nor recall after fine tuning with up to 30 additional cases. Rightmost bars were the performance of LKCGMH PAS-stained dataset for comparison.

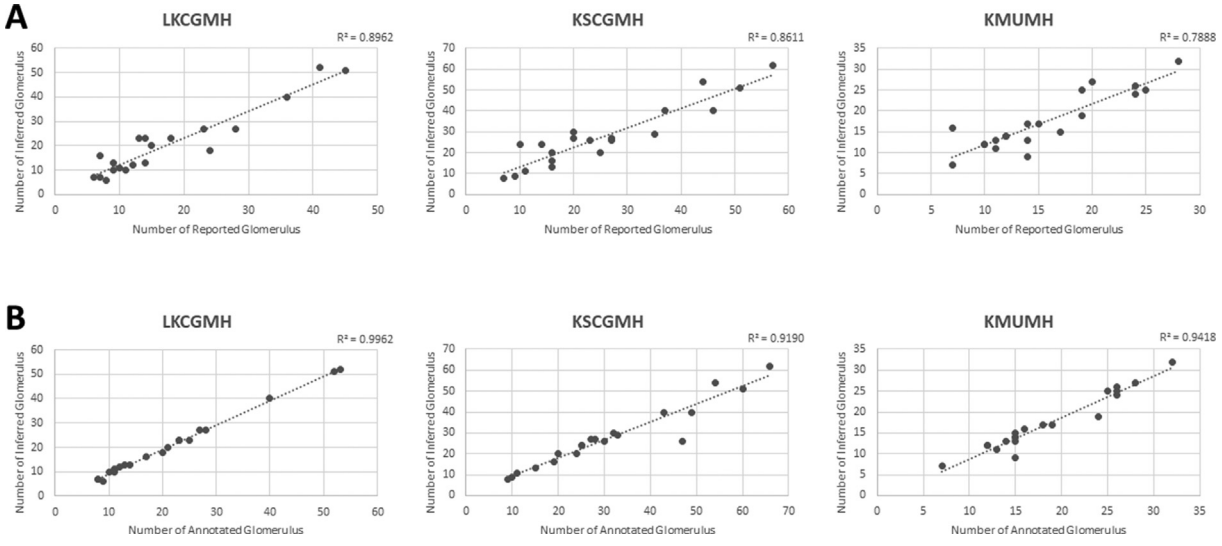
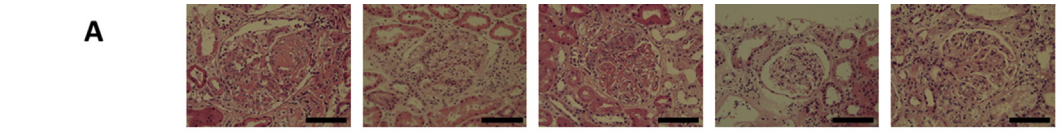


Fig. 2 **Correlation between correctly inferred glomeruli, reported (documented) glomeruli, and annotated (ground truth) glomeruli.** (A) Correlation between correctly inferred glomeruli and reported glomeruli. (B) Correlation between correctly inferred glomeruli and annotated glomeruli.

nephropathy (0.857–0.903) and membranous glomerulopathy (0.838–0.883), which might reflex the heterogeneous nature of lupus nephritis morphology. The confusion matrix demonstrated that lupus nephritides were occasionally misclassified as IgA nephropathy and membranous glomerulopathy [Fig. 3B]. Among classes with poorer performance, minimal change disease was easily misclassified as membranous glomerulopathy, and IgA nephropathy was easily misclassified as diabetic nephropathy, which generally correlate well with their similar glomerular morphologies.



Accuracy of	DM	IGAN	LN	MCD	MN
LKCGMH	0.880	0.820	**0.940**	0.840	0.840
KSCGMH	**0.903**	0.806	0.699	0.680	0.883
KHUMH	**0.857**	0.771	0.705	0.743	0.838

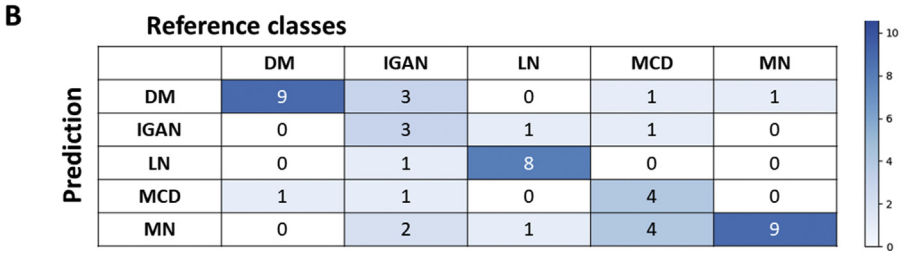


Fig. 3 **Performance of the glomerular disease classification model.** (A) Representative glomerular image of each glomerular disease and the corresponding model performance. Red color indicates the best performance in each testing dataset. Bars = 100 μm (B) Confusion matrix of glomerular disease classification model on LKCGMH testing dataset. Recall of LN was slightly lower than that of DM and MN, but LN had much fewer false positive prediction instances, resulting in highest accuracy (0.940). Note the poorer performances on IGAN and MCD resulted from their morphological mimickers DM and MN, respectively.

Multiclass and multilabel identification of glomerular lesions

Glomerular lesions are the focus and key findings in kidney biopsy interpretation. The glomerular lesion identification model was trained and validated by 5459 glomeruli of 146 kidney biopsies and tested on 60 biopsies obtained from three hospitals. Detailed numbers and distribution of glomerular lesions in the testing dataset were shown in Table 1. The first stage of the glomerular lesion identification model showed high recall (93.6%) for the Multi-label class, which indicates that most of the glomeruli need to be scored were correctly found out [Table 3]. In addition, the glomerular lesion identification model accurately identified globally sclerotic glomeruli with nearly perfect accuracies (0.98–0.99) on all testing datasets. Among scorable lesions, the area under the receiver operating characteristic curve (AUC) of each kind of lesions on testing datasets are from 0.687 to 0.947 [Fig. 4]. Best performance was achieved on cellular/

Table 3 Confusion matrix for the first stage of the glomerular lesion identification model.

Prediction \ Reference classes	Unremarkable / Abnormal, NOS	Global sclerosis	Multi-label	Precision
Unremarkable / Abnormal, NOS	222	1	25	89.5% (222/248)
Global sclerosis	3	117	7	94.4% (117/124)
Multi-label	199	7	427	67.5% (427/633)
Recall	52.4% (222/424)	93.6% (117/125)	93.6% (427/456)	

Classification of global sclerosis showed both best recall and precision. High recall (93.6%) ensured that most of the glomeruli in Multi-label class would go through second stage of the model for scorable lesion identification.

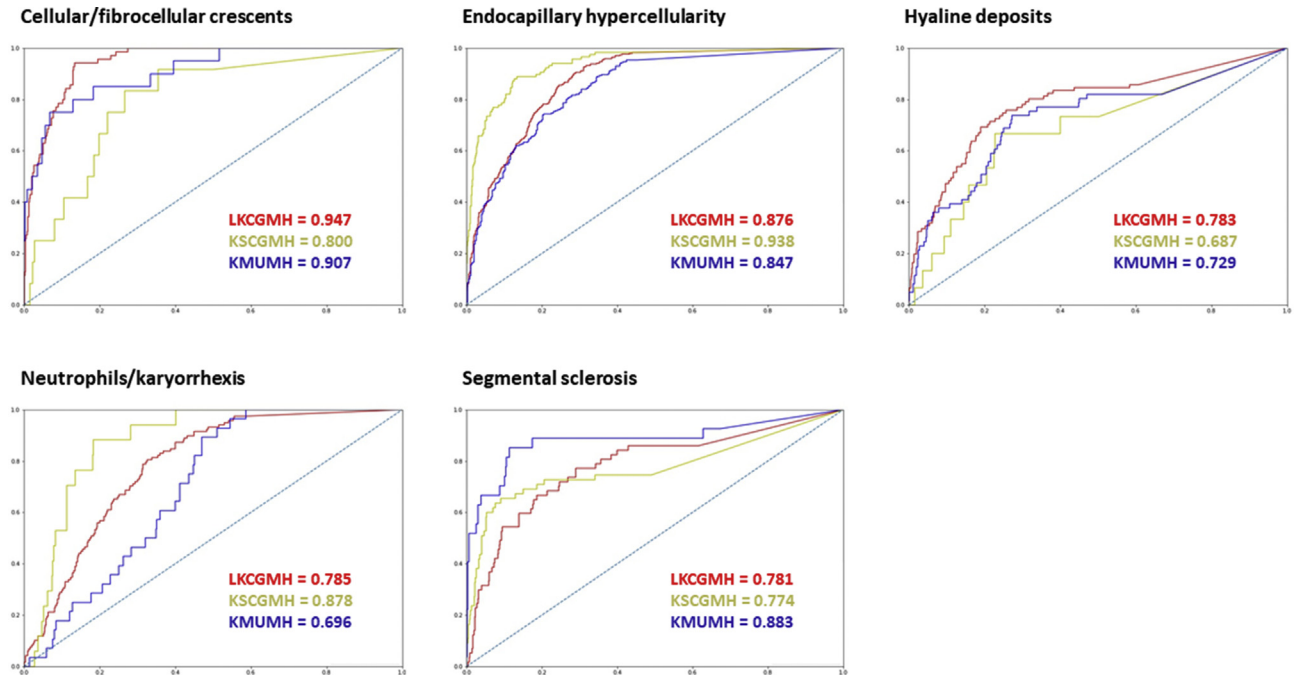


Fig. 4 **Performance of the glomerular lesion identification model.** Area under the receiver operating characteristic curve for cellular/fibrocellular crescents, endocapillary hypercellularity, hyaline deposits, neutrophils/karyorrhexis, and segmental sclerosis of lupus nephritis cases from LKCGMH, KSCGMH, and KMUMH.

fibrocellular crescents (AUC = 0.80–0.95; overall accuracy = 0.95; Table 4), followed by endocapillary hypercellularity (AUC = 0.85–0.94; accuracy = 0.81). The model did not show a clear predilection of better performance on cases from LKCGMH, suggesting good generalization. Localization maps produced by gradient-weighted class activation mapping (Grad-CAM) highlighted the most informative areas for the model to identify glomerular lesions [Fig. 5].

Discussion

Pathologists spend much time on detection, classification, and quantification of histopathological features. Our study showed the prospect of outsourcing these mundane and repetitive work to trained machine learning models, which might effectively reduce pathologists' workload and enable them to focus on more complex tasks such as clinicopathological correlation, gestalt interpretation, and finalizing reports.

A glomerular detection model can significantly reduce the time spending on slides finding glomeruli, and various methods were proposed with different degrees of success [19–21,26]. Hermsen et al. reported F1 score of 0.95 on glomerulus detection [14]. Their study focused on transplant biopsies and used semantic segmentation; therefore, the result cannot be directly compared to ours. The only comparable one by Kawazoe et al. [38] reached best F1 score of 0.876–0.928, but their study did not include external datasets for validation. Our model achieved F1 score of 0.881–0.944 on internal testing datasets and 0.889 to 0.932 on external testing datasets, representing state-of-the-art performance on glomerulus instance segmentation for non-selective human kidney biopsies with diverse histological stainings obtained from clinical settings. A major advantage of our method is that our model, using mask R–CNN, can output an exact mask of each inferred glomerulus instead of a bounding box or a semantic area. This property will be especially useful if focusing on glomeruli is needed in subsequent image analysis, because

Table 4 AUC/accuracy of glomerular lesion detection model.

	Global Sclerosis	Crescents	Endocapillary Hypercellularity	Hyaline Deposits	Neutrophils/karyorrhexis	Segmental Sclerosis
LKCGMH	>0.99/0.99	0.95/0.94	0.88/0.78	0.78/0.80	0.78/0.68	0.78/0.92
KSCGMH	>0.99/0.98	0.80/0.96	0.94/0.92	0.69/0.95	0.88/0.81	0.77/0.92
KMUMH	>0.99/0.99	0.91/0.97	0.85/0.75	0.73/0.88	0.70/0.70	0.88/0.97
Overall	>0.99/0.98	0.92/0.95	0.90/0.81	0.77/0.86	0.79/0.82	0.77/0.93

The glomerular detection model showed nearly perfect performance on detecting global sclerosis. Except for crescents, the model did not show better performance on cases of LKCGMH over the other two datasets.

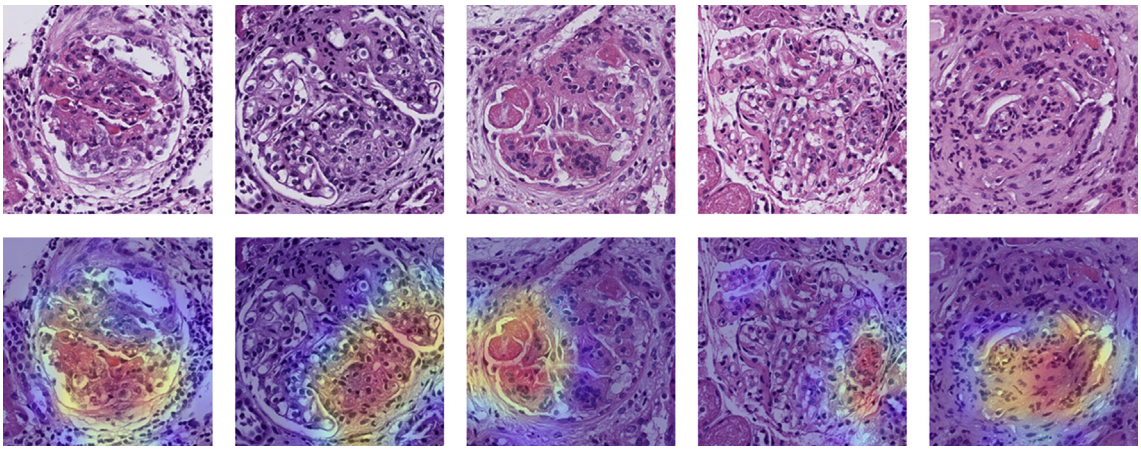


Fig. 5 **Results of gradient-weighted class activation mapping (Grad-CAM) on glomerular lesions inferred by the model.** The upper row is the original H&E images, and the lower row is the H&E images overlayed by the result of Grad-CAM. Note that the important region for classifying a glomerulus correlated well with the location of the particular lesion within the glomerulus (from left to right: cellular/fibrocellular crescents, endocapillary hypercellularity, hyaline deposits, neutrophils/karyorrhexis, segmental sclerosis).

tissue surrounding glomeruli can be easily excluded. A drop of detection accuracy on certain external testing datasets would be within expectation as the model needs to overcome the staining variation among slides processed in different institutions. However, we did not find such decay except for H&E staining. In addition, fine tuning with additional cases did not result better performance. This finding implies a good model generalization, for differences on performance simply reflexes dataset variances rather than a decline of performance due to overfitting. Glomerular feature clustering maps also support the above observation (supplemental Figure). Most importantly, even without fine tuning, the numbers of correctly inferred glomerulus were already significantly higher than the numbers of glomerulus documented in pathological reports, which provides us fair confidence on the representativeness of glomeruli found out by the model.

Our study is the first one to classify different glomerular diseases purely based on tissue section images using modern machine learning techniques. The LSTM model for this task takes a series of input (images of glomeruli) to determine the output (diagnosis). The logic is remarkably like that of nephropathologists on examining glomerular diseases. We recognized that diagnosis of glomerular disease requires a combination of clinical, light microscopic, immunofluorescence, and electron microscopic studies. Therefore, the scope here is to identify cases need to be further scored, e.g., class III or IV lupus nephritis or diabetic nephropathy. Our model successfully classified most common glomerulonephritides encountered in daily practice with good accuracies. The glomerular lesions therefore can be inferred by trained identification model before nephropathologists seeing the WSIs, which is clinically important for smoothing and accelerating diagnostic workflow. Pre-selecting cases which will benefit from other computer-aided image analysis or special handling (e.g., ordering a PLA2R IHC for membranous glomerulopathy) is also an important integration part of computational pathology pipeline. Our study showed within a confined scenario a classification model can fulfill the need. However, how far it can be generalized to hundreds of possible diagnoses on kidney biopsies, of which many are rare, and how to combine the model with immunofluorescence study, electron microscopy, and clinical data to get a better prediction, needs further investigation.

Glomerular lesion identification is the key to all glomerular diseases. We here demonstrated this particularly challenging task can also be tackled by trained machine learning models with high degrees of accuracy. On this topic, only a handful of studies are present, and the lesions analyzed are limited [20,27–30]. In contrast, our study evaluated six major glomerular lesions required for proper classification of lupus nephritis. Our datasets are highly imbalanced, reflexing the nature of glomerular findings in daily practice. More training data might improve the performance of low accuracy categories. Modern machine learning techniques such as few-shot learning or generative networks might help [39,40]. Incorporating quantitative techniques or identification of glomerular intrinsic cell types could be useful to force the models to focus on important morphological clues and might increase model performance [13,28]. Unlike tumor classification and segmentation, in glomerular diseases the “lesions” almost always have a large morphological spectrum, and both the physical and conceptual borders between normal and abnormal are often very subtle and poorly defined, especially in real world cases. Even expert nephropathologists cannot achieve good interobserver and intraobserver reproducibility. Our model was trained on dataset annotated by a single nephropathologist, so it only captured the interpretation and preference of the annotating pathologist. However, it is well known that scoring of lupus nephritis (and other glomerular diseases) exhibits poor interobserver agreement [24]. A collaboration including large number of nephropathologists might be needed to develop a consensus ground truth. At present, nephropathologists

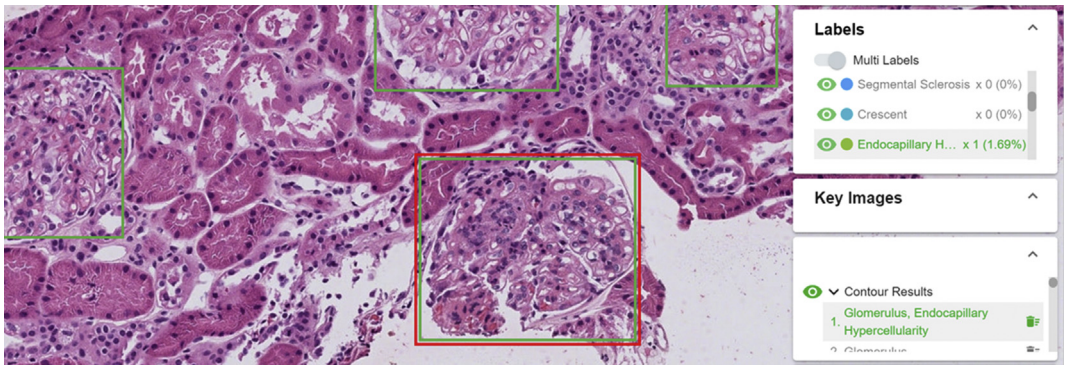


Fig. 6 **Representation of glomerulus detection and lesion inference/annotation in operation.** Glomeruli found by the detection model were boxed in green. Lesions found in the target glomerulus (boxed in red) were showed on the right lower panel, and users can modify annotations using the right upper control panel. A video of how the platform works is provided in supplemental material.

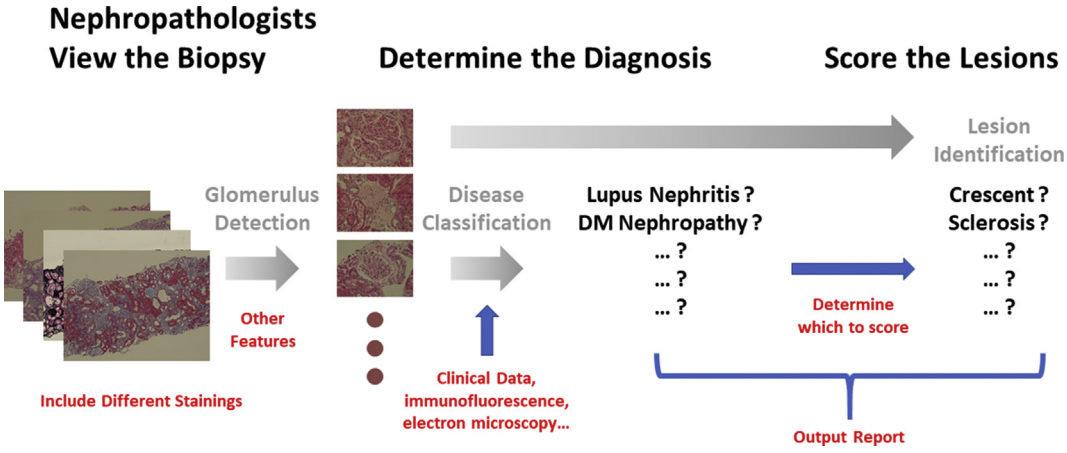


Fig. 7 **The ideal full model composed of all relevant steps imitating nephropathologists' workflow on kidney biopsy interpretation and reporting.** In this paper, three essential steps were carried out by machine learning models (colored in grey). In addition to models described in this paper, comprehensive interpretation of a kidney biopsy will need additional inputs from light microscopy, clinical data, immunofluorescence, and electron microscopy. Items colored in red are steps also required for a complete computer-aided analysis of kidney biopsy interpretation.

can review, correct, or confirm the inference results on a computational pathology platform [Fig. 6], and the feedbacks can also be helpful in further improving the model performance.

Supplementary video related to this article can be found at https://doi.org/10.1016/j.bj.2021.08.011

Our models showed good and sometimes even better performance on certain external testing datasets on various tasks. This finding might be explained by that higher variances of small testing datasets can result better performance just by chance (when cases of external testing datasets are easier to interpret compared to internal testing datasets). We took it as evidence of good representation of our training datasets and good generalization of trained models.

In this study the detection of glomerulus, the determination of glomerular disease, and the identification of glomerular lesion were carried out on a single section and/or slide basis. In real world information from multiple sections/slides must be integrated to a final decision. We acknowledged this weakness and are currently working on glomerulus registration and image overlay. In addition, tubulointerstitial and vascular compartment may provide clues and are occasionally essential to the diagnosis of kidney diseases. They should be incorporated into current workflow stepwise.

Conclusion

To summarize, we showed how the concept and workflow of kidney biopsy interpretation can be simulated by machine learning models. A detection model was built for glomerulus detection, and the glomeruli found by the detection model went through a classification model to determine the glomerular disease. A lesion identification model was then applied to find out glomerular lesions relevant to the disease. The whole process imitates the way nephropathologists interpreting kidney biopsies, which is the first-of-its-kind representation how computational pathology can be implemented into daily practice of medical renal pathology. The

models were integrated into a WSI viewing and annotating platform for nephropathologists to review, correct, and confirm the inference results. Further improvement with more training data, refining algorithms, and integration of multimodal input will enable the models for actual clinical use [Fig. 7].

Funding

None.

Conflicts of interest

C.-Y.Y. is the chairman and Chief Executive Officer and a cofounder of aetherAI. C.-K.Y. and C.-Y.L. are data scientists of aetherAI. Other authors declare no conflicts of interest.

Acknowledgements

The authors appreciate the statistical consultation and acknowledge the support of the Maintenance Project of the Center for Artificial Intelligence in Medicine (Grant CLRPG3H0012, CIRPG3H0012) at Chang Gung Memorial Hospital.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.bj.2021.08.011.

REFERENCES

[1] Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. Neural Inf Process Syst Conf 2012:pp1097–105.
[2] LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015;521:436–44.
[3] Litjens G, Sanchez CI, Timofeeva N, Hermsen M, Nagtegaal I, Kovacs I, et al. Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis. Sci Rep 2016;6:26286.
[4] Ehteshami Bejnordi B, Veta M, Johannes van Diest P, van Ginneken B, 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;318:2199–210.
[5] Campanella G, Hanna MG, Geneslaw L, Miraflor A, Werneck Krauss Silva V, Busam KJ, et al. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. Nat Med 2019;25:1301–9.
[6] Skrede OJ, De Raedt S, Kleppe A, Hveem TS, Liestol K, Maddison J, et al. Deep learning for prediction of colorectal cancer outcome: a discovery and validation study. Lancet 2020;395:350–60.
[7] Nadkarni GN, Chaudhary K, Coca SG. Machine learning in glomerular diseases: promise for precision medicine. Am J Kidney Dis 2019;74:290–2.
[8] Huang C, Murugiah K, Mahajan S, Li S, Dhruva SS, Haimovich JS, et al. Enhancing the prediction of acute kidney injury risk after percutaneous coronary intervention using machine learning techniques: a retrospective cohort study. PLoS Med 2018;27:e1002703.
[9] Nadkarni GN, Fleming F, McCullough JR, Chauhan K, Verghese DA, He JC, et al. Prediction of rapid kidney function decline using machine learning combining blood biomarkers and electronic health record data. bioRvix:587774v1[Preprint]. 2019 [cited 2020 Jul 14]. Available from: https://doi.org/10.1101/587774.
[10] Aubert O, Higgins S, Bouatou Y, Yoo D, Raynaud M, Viglietti D, et al. Archetype Analysis identifies distinct profiles in renal transplant recipients with transplant glomerulopathy associated with allograft survival. J Am Soc Nephrol 2019;30:625–39.
[11] Chen T, Li X, Li Y, Xia E, Qin Y, Liang S, et al. Prediction and risk stratification of kidney outcomes in IgA nephropathy. Am J Kidney Dis 2019;74:300–9.
[12] Kolachalama VB, Singh P, Lin CQ, Mun D, Belghasem ME, Henderson JM, et al. Association of pathological fibrosis with renal survival using deep neural networks. Kidney Int Rep 2018;3:464–75.
[13] Zeng C, Nan Y, Xu F, Lei Q, Li F, Chen T, et al. Identification of glomerular lesions and intrinsic glomerular cell types in kidney diseases via deep learning. J Pathol 2020;252:53–64.
[14] Hermsen M, de Bel T, den Boer M, Steenbergen EJ, Kers J, Florquin S, et al. Deep learning-based histopathologic assessment of kidney tissue. J Am Soc Nephrol 2019;30:1968–79.
[15] Kim YG, Choi G, Go H, Cho Y, Lee H, Lee AR, 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 2019;9:5123.
[16] Ginley G, Jen K, Han SS, Rodrigues L, Jain S, Fogo AB, et al. Automated computational detection of interstitial fibrosis, tubular atrophy, and glomerulosclerosis. J Am Soc Nephrol 2021;32:837–50.
[17] Kato T, Relator R, Ngouv H, Hirohashi Y, Takaki O, Kakimoto T, et al. Segmental HOG: New descriptor for glomerulus detection in kidney microscopy image. BMC Bioinf 2015;16:316.
[18] Ginley B, Tomaszewski JE, Yacoub R, Chen F, Sarder P. Unsupervised labeling of glomerular boundaries using Gabor filters and statistical testing in renal histology. J Med Imag (Bellingham) 2017;4:021102.
[19] Simon O, Yacoub R, Jain S, Tomaszewski JE, Sarder P. Multi-radial LBP features as a tool for rapid glomerular detection and assessment in whole slide histopathology images. Sci Rep 2018;8:2032.
[20] Marsh JN, Matlock MK, Kudose S, Liu TC, Stappenbeck TS, Gaut JP, et al. Deep learning global glomerulosclerosis in transplant kidney frozen sections. IEEE Trans Med Imaging 2018;37:2718–28.
[21] Kannan S, Morgan LA, Liang B, Cheung MG, Lin CQ, Mun D, et al. Segmentation of glomeruli within trichrome images using deep learning. Kidney Int Rep 2019;4:955–62.
[22] Bueno G, Fernandez-Carrobles MM, Gonzalez-Lopez L, Deniz O. Glomerulosclerosis identification in whole slide images using semantic segmentation. Comput Methods Progr Biomed 2020;184:105273.
[23] Gallego J, Pedraza A, Lopez S, Steiner G, Gonzalez L, Laurinavicius A, et al. Glomerulus classification and detection based on convolutional neural networks. J Imag 2018;4:20.
[24] Dasari S, Chakraborty A, Truong L, Mohan C. A systematic review of interpathologist agreement in histologic

classification of lupus nephritis. Kidney Int Rep 2019;4:1420–5.
[25] Furness PN, Taub N. International variation in the interpretation of renal transplant biopsies: report of the CERTPAP Project. Kidney Int 2001;60:1998–2012.
[26] Aeffner F, Zarella MD, Buchbinder N, Bui MM, Goodman MR, Hartman DJ, et al. Introduction to digital image analysis in whole-slide imaging: a white paper from the digital pathology association. J Pathol Inf 2019;10:9.
[27] Barros GO, Navarro B, Duarte A, Dos-Santos WLC, PathoSpotter K. A computational tool for the automatic identification of glomerular lesions in histological images of kidneys. Sci Rep 2017;7:46769.
[28] Ginley B, Lutnick B, Jen KY, Fogo AB, Jain S, Rosenberg A, et al. Computational segmentation and classification of diabetic glomerulosclerosis. J Am Soc Nephrol 2019;30:1953–67.
[29] Chagas P, Souza L, Araujo I, Aldeman N, Duarte A, Angelo M, et al. Classification of glomerular hypercellularity using convolutional features and support vector machine. Artif Intell Med 2020;103:101808.
[30] Hao F, Li M, Liu X, Li X, Yue J, Han W. Classification of glomeruli with membranous nephropathy on renal digital pathological images with deep learning. In: CAIH2020: proceedings of the 2020 conference on artificial intelligence and healthcare. New York: Association for Computing Machinery; 2020. p. 239–43.
[31] He K, Gkioxari G, Dollár P, Girshick R, Mask R-CNN. IEEE Trans Pattern Anal Mach Intell 2020;42:386–97.
[32] Xie S, Girshick R, Dollár P, Tu Z, He K. Aggregated residual transformations for deep neural networks. Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR). 2017. p. 5987–95.
[33] Huang G, Liu Z, Maaten LVD, Weinberger KQ. Densely connected convolutional networks. Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR). 2017. p. 2261–9.
[34] Lin T, Goyal P, Girshick R, He K, Dollár P. Focal loss for dense object detection. IEEE Trans Pattern Anal Mach Intell 2020;42:318–27.
[35] Ullah A, Ahmad J, Muhammad K, Sajjad M, Baik SW. Action recognition in video sequences using deep Bi-directional LSTM with CNN features. IEEE Access 2018;6:1155–66.
[36] Bajema IM, Wilhelmus S, Alpers CE, Bruijn JA, Colvin RB, Cook HT, et al. Revision of the International Society of Nephrology/Renal Pathology Society classification for lupus nephritis: clarification of definitions, and modified National Institutes of Health activity and chronicity indices. Kidney Int 2018;93:789–96.
[37] Tervaert TW, Mooyaart AL, Amann K, Cohen AH, Cook HT, Drachenberg CB, et al. Pathologic classification of diabetic nephropathy. J Am Soc Nephrol 2010;21:556–63.
[38] Kawazoe Y, Shimamoto K, Yamaguchi R, Shintani-Domoto Y, Uozaki H, Fukayama M, et al. Faster R-CNN-based glomerular detection in multistained human whole slide images. J Imag 2018;4:91.
[39] Murali LK, Lutnick B, Ginley B, Tomaszewski JE, Sarder P. Generative modeling for renal microanatomy. Proc SPIE-Int Soc Opt Eng 2020;11320:113200F.
[40] Quellec G, Lamard M, Conze PH, Massin P, Cochener B. Automatic detection of rare pathologies in fundus photographs using few-shot learning. Med Image Anal 2020;61:101660.