Glomerular Classification Using Convolutional Neural Networks Based on Defined Annotation Criteria and Concordance Evaluation Among Clinicians

Ryohei Yamaguchi1, Yoshimasa Kawazoe1, Kiminori Shimamoto1, Emiko Shinohara1, Tatsuo Tsukamoto2, Yukako Shintani-Domoto3, Hajime Nagasu4, Hiroshi Uozaki5, Tetsuo Ushiku3, Masaomi Nangaku6, Naoki Kashihara4, Akira Shimizu7, Michio Nagata8 and Kazuhiko Ohe9

1Artificial Intelligence in Healthcare, Graduate School of Medicine, Faculty of Medicine, The University of Tokyo, Tokyo, Japan; 2Department of Nephrology and Dialysis, Tazuke Kofukai Medical Research Institute, Kitano Hospital, Osaka, Japan; 3Department of Pathology, Graduate School of Medicine, The University of Tokyo, Tokyo, Japan; 4Department of Nephrology and Hypertension, Kawasaki Medical School, Okayama, Japan; 5Department of Pathology, Teikyo University School of Medicine, Tokyo, Japan; 6Division of Nephrology and Endocrinology, The University of Tokyo Graduate School of Medicine, Tokyo, Japan; 7Department of Analytic Human Pathology, Nippon Medical School, Tokyo, Japan; 8Kidney and Vascular Pathology, Faculty of Medicine, University of Tsukuba, Ibaraki, Japan; and 9Department of Biomedical Informatics, Graduate School of Medicine, University of Tokyo, Tokyo, Japan

Introduction: Diagnosing renal pathologies is important for performing treatments. However, classifying every glomerulus is difficult for clinicians; thus, a support system, such as a computer, is required. This paper describes the automatic classification of glomerular images using a convolutional neural network (CNN).

Method: To generate appropriate labeled data, annotation criteria including 12 features (e.g., “fibrous crescent”) were defined. The concordance among 5 clinicians was evaluated for 100 images using the kappa (κ) coefficient for each feature. Using the annotation criteria, 1 clinician annotated 10,102 images. We trained the CNNs to classify the features with an average κ = 0.4 and evaluated their performance using the receiver operating characteristic–area under the curve (ROC–AUC). An error analysis was conducted and the gradient-weighted class activation mapping (Grad-CAM) was also applied; it expresses the CNN’s focusing point with a heat map when the CNN classifies the glomerular image for a feature.

Results: The average κ coefficient of the features ranged from 0.28 to 0.50. The ROC–AUC of the CNNs for test data varied from 0.65 to 0.98. Among the features, “capillary collapse” and “fibrous crescent” had high ROC–AUC values of 0.98 and 0.91, respectively. The error analysis and the Grad-CAM visually showed that the CNN could not distinguish between 2 different features that had similar visual structures or that occurred simultaneously.

Conclusion: The differences in the texture or frequency of the co-occurrence between the different features affected the CNN performance; thus, to improve the classification accuracy, methods such as segmentation are required.

Kidney Int Rep (2021) 6, 716–726; https://doi.org/10.1016/j.ekir.2020.11.037

KEYWORDS: artificial intelligence; convolutional neural network; deep learning; glomerular image; renal pathology

© 2020 International Society of Nephrology. Published by Elsevier Inc. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).
To automatically classify the glomerular images using CNN, the ground truth for features is essential for all the images. However, in the routine diagnosis of kidney disease, scores are not always recorded for all of the features for all glomeruli, and they are often recorded only for representative glomeruli. Therefore, the pathological reports recorded in daily medical care often do not have complete ground truth for all glomeruli. Furthermore, the existing annotation criterion for a specific disease, such as the Oxford classification, or the annotation criterion for only particular diseases, such as the multicenter Nephrotic Syndrome Study Network (NEPTUNE) scoring system, cannot be applied to various diseases in their current format. Therefore, annotation criteria that can be applied to various kidney diseases should be newly developed. First, to create the ground truth of glomerular images, this study established the annotation criteria for glomerular images and evaluated the degree of concordance among clinicians based on the criteria. Next, CNNs that classified the various features of the glomerular images were developed, and their performance was evaluated. Finally, the concordance between the CNN score and the clinicians’ score was assessed, and an error analysis was performed to determine computer-related issues for the renal pathology images.

### MATERIALS AND METHODS

#### Dataset Collection

This study collected 293 PAS-stained WSIs from 3 hospitals: 99 WSIs from the University of Tokyo Hospital (UTH) from 2010 to 2017; 88 WSIs from the Tazuke Kofukai Medical Research Institute, Kitano Hospital (KH), from 2014 to 2017; and 106 WSIs from the University of Tsukuba Hospital (UTSH) from 2009 to 2016. From the 293 WSIs, the glomerular image was manually extracted in a rectangular shape so that 1 major glomerulus was included, which yielded 10,202 images. Supplementary Table S1 lists the major diagnoses of the 293 WSIs for the 10,202 images. The extracted glomerular images were converted into the png format using Openslide and were saved as a single image file.

#### Development of Annotation Criteria

Five co-authored clinicians (1 physician and 4 pathologists) developed the annotation criteria for various kidney diseases based on the Delphi method, because we were to guarantee the quality of the annotation criteria and annotated labels; Krause et al. showed that the quality of the annotated label affected the validity of its performance in the machine learning model for diabetic retinopathy. Consequently, the annotation criteria consisted of the following 3 elements: (i) “feature”: finding the name (e.g., “mesangial hypercellularity”); (ii) “score”: the possible value of the feature (e.g., “normal,” “mild,” “moderate,” “severe”); and (iii) “regulation”: definitions for determining the score (e.g., “count the number of mesangial cells ...”). The annotation criteria also included 12 features, which are important to diagnose the PAS-stained images. The 12 features are as follows: (i) capillary collapse (CC); (ii)
sclerosis (Scl); (iii) mesangial hypercellularity (MesHyper); (iv) increased mesangial matrix (IMM); (v) mesangiolysis (MLysis); (vi) endocapillary proliferation (EP); (vii) fibrous crescent (F-Cre); (viii) fibrocellular crescent (Fc-Cre); (ix) cellular crescent (C-Cre); (x) adhesion (Adh); (xi) afferent/efferent arteriolar hyalinosis (AAH); and (xii) increased vasculature around the vascular pole (IVVP). The granularity of the score and the regulation were also discussed and agreed upon by the 5 clinicians. In some images, the score of the particular features could not be obtained; these were classified as “impossible to score.” Supplementary Table S2 lists the annotation criteria that were developed for this study. Supplementary Figure S1 depicts the example images of annotation for each feature. In addition to those criteria, an annotation flowchart was developed to clarify the dependency between the features, which sometimes made the annotation complicated. Figure 1a and b show the annotation flowcharts.

Concordance Evaluation of Annotated Features Between Clinicians

Out of 10,202 images, the first author selected 100 images for concordance evaluation. These 100 images included relatively “easy-to-score” images but also “hard-to-score” images in which different clinicians might assign different scores. Five clinicians annotated all 12 features for the 100 images except for 2 images in which 2 clinicians considered them to be ineligible to be scored. Supplementary Figure S2 shows the 98 images used for concordance evaluation. After the 5 clinicians finished annotating the images, the Cohen kappa (K) coefficients of the 5 clinicians were calculated.

Development and Evaluation of CNN

One physician annotated all 12 features that were included in the developed annotation criteria for the remaining 10,102 glomerular images. The breakdown list of the 10,102 images for the features and scores is presented in Table 1.

The CNNs were developed for the features for which the K coefficients were ≥0.40 to classify the features as positive or negative. The binary class breakdown list of the 10,102 images of the 5 features for the CNN is presented in Table 2. The strength of the correlation between the features was calculated using the phi coefficient (Table 3). There was a strong correlation between CC and F-Cre (phi coefficient = 0.67).

The performance of the CNN was evaluated on the validation, test, and concordance data, respectively, over the 4-fold cross-validation (Figure 2). The performance was measured using the average ROC-AUC and F1 score via 4-fold stratified cross-validation for the 10,102 images, and also for the 98 images used for the concordance evaluation.

This study used a 50-layer residual network (ResNet50) as the structure of the CNN, which demonstrates the high performance in general image recognition and is unlikely to fall into poor local minima because of its structure; hence, it would achieve high classification performance. This investigation also adopted transfer learning, which uses the parameters that were trained by the ImageNet dataset. The weighted softmax cross entropy loss was used as a loss function because the dataset for this study was unbalanced between positive and negative. For the weighted softmax cross entropy, the reciprocal of the ratio of the number of images in the class to the total number of images was set for each class. As the optimizing method, adaptive moment estimation (Adam), which has a fast convergence speed and can easily obtain good accuracy without parameter tuning, was used with an empirically determined hyperparameter (α = 1.0 × 10⁻⁷). All original images were resized to 224 x 224 pixels for the CNN input.

At the time of CNN training, data augmentation was applied to randomly rotate (0°, 90°, 180°, and 270°) and invert the images. The losses for the training and validation data were calculated for each epoch. The training was stopped at the point at which the loss of the validation data did not decrease for at least 50 continuous epochs.

Python3 was used as the programming language, and Chainer (version 4.0.0) was used as the deep learning library. The calculations were performed using NVIDIA DGX-1 (OS: Ubuntu 16.04, main memory: 515 GB, CPU: Intel Xeon CPU E5-2698 v4, GPU: 8 Tesla V100-SXM).

Concordance Evaluation Between the CNN Score and Clinicians’ Score

For the features for which the ROC–AUC for the test data exceeded 0.9, this study compared the CNN’s predictive score with the 5 clinicians’ scores for the concordance data and performed error analysis. Moreover, to perform the error analysis visually and to help understand the CNN’s properties, gradient-weighted class activation mapping (Grad-CAM) was applied for the concordance data; it is a technique that expresses the CNN’s focusing point using a heat map calculated by summing up the intermediate gradient of the CNN for each channel.

RESULTS

Concordance Between Clinicians

Table 4 lists the results of the Cohen K coefficient of the 5 clinicians for the 98 images. The average K coefficient
Start the evaluation of a glomerulus.

First, check if the capillaries are obsolete.

Can any capillary be seen in an image?

Yes

Define the score of “CC” as “Impossible to score.”

No

Create score for “CC.”

Is the score of “CC” “Global?”

Yes

Define the score of “Scl” as “Impossible to score.”

No

Create score for “Scl.”

Is the score of “Scl” “Global?”

Yes

Define the score of “MesHyper,” “IMM,” “MLysis,” “Adh,” and “EP” as “Impossible to score.”

No

Focus on the mesangial area.

Is there any non-sclerotic mesangial area left?

Yes

Focus on the capillaries.

Is any capillary patent?

Yes

Define the score of “EP” as “Impossible to score.”

No

Create score for “EP.”

Evaluate the extracapillary lesion.

Figure 1. (a, b) Flowcharts of the annotation for an image of a glomerulus. These 2 flowcharts represent the order followed to create a feature score. AAH, afferent/efferent arteriolar hyalinosis; Adh, adhesion; CC, capillary collapse; C-Cre, cellular crescent; EP, endocapillary proliferation; F-Cre, fibrous crescent; Fc-Cre, fibrocellular crescent; IMM, increased mesangial matrix; IVVP, increased vasculature around the vascular pole; MesHyper, mesangial hypercellularity; MLysis, mesangiolysis; Scl, sclerosis.
Evaluate the extracapillary lesion.

Create score for “F-Cre,” “Fc-Cre,” and “C-Cre.”

Focus on Bowman capsule.

Can Bowman capsule be seen in the image?

Yes

Define score for “Adh” as “Impossible to score.”

No

Create score for “Adh.”

Focus on the vascular pole.

Can a vascular pole can be seen in the image?

Yes

Define score of “IVVP” as “Impossible to score.”

No

Create score for “IVVP.”

Focus on an afferent or efferent arteriole.

Can an afferent or efferent arteriole be seen in the image?

Yes

Define score of “AAH” as “Impossible to score.”

No

Create score for “AAH.”

Finish evaluation of the glomerulus.

Instruction: Instruct clinicians to focus on certain areas of images.

Branch: Go to a different node according to this node’s assessment by clinicians.

Manual Annotation: Clinicians create score of the feature manually.

Automatic Scoring: The scores of the feature are automatically determined.

Figure 1. Continued
There was 1 completely false-negative (cFN) image that the CNN predicted to be negative, and no clinician scored them as positive. In addition, there were no completely false-positive (cFP) images that the CNN predicted to be positive, and all 5 clinicians scored them positively. Table 7 demonstrates that for the F-Cre, there are no cTP images but there are 20 cFP images, 37 cTN images, and 1 cFN image.

### Visualization of the CNN’s Focusing Point

We show the 3 images of Grad-CAM to understand the properties of the CNN qualitatively. Figure 3 depicts a cTP image for CC where the CNN and all 5 clinicians scored it as positive. In this figure, the focus of the CNN was on the collapsed capillary, which is an actual pathological lesion. On the other hand, Figures 4 and 5 show the cFP images for the F-Cre where the CNN scored it as positive but none of the 5 clinicians scored it as positive. In Figure 4, the focus of the CNN was the extracapillary lesion, which is an actual pathological lesion of the F-Cre; however, the CNN misclassified it as F-Cre. In Figure 5, the focus of the CNN was not the extracapillary lesion but the collapsed capillary lesion.

### DISCUSSION

#### Clinicians’ Concordance

As listed in Table 4, the K coefficient ranged from 0.28 to 0.50, which indicates a fair-to-moderate concordance. The study about concordance in the Oxford classification for IgA nephropathy measured the concordance by using intraclass correlation coefficients (ICCs), and their ICCs ranged from 0.03 to 0.90. The

---

### Table 1. Breakdown of the annotated images

| Feature | Score | CC | Scl | Adh | C-Cre | F-Cre | Fc-Cre | C-Cre | Adh | IVVP | AAH | MLysis | IMM | IVVP |
|---------|-------|----|-----|-----|-------|-------|-------|-------|-----|-----|-----|-------|-----|------|
|         |       | 187| 1022| 82  | 1093  | 448   | 861   | 34    | 1213| 112 | 112 | 5411 | 6826| 9958 |
| CC      | –     | –  | –   | –   | –     | –     | –     | –     | –   | –   | –   | –     | –   | –    |
| Scl     | –     | –  | –   | –   | –     | –     | –     | –     | –   | –   | –   | –     | –   | –    |
| Adh     | –     | –  | –   | –   | –     | –     | –     | –     | –   | –   | –   | –     | –   | –    |
| C-Cre   | –     | –  | –   | –   | –     | –     | –     | –     | –   | –   | –   | –     | –   | –    |
| F-Cre   | –     | –  | –   | –   | –     | –     | –     | –     | –   | –   | –   | –     | –   | –    |
| Fc-Cre  | –     | –  | –   | –   | –     | –     | –     | –     | –   | –   | –   | –     | –   | –    |
| C-Cre   | –     | –  | –   | –   | –     | –     | –     | –     | –   | –   | –   | –     | –   | –    |
| Adh     | –     | –  | –   | –   | –     | –     | –     | –     | –   | –   | –   | –     | –   | –    |
| IVVP    | –     | –  | –   | –   | –     | –     | –     | –     | –   | –   | –   | –     | –   | –    |
| AAH     | –     | –  | –   | –   | –     | –     | –     | –     | –   | –   | –   | –     | –   | –    |
| MLysis  | –     | –  | –   | –   | –     | –     | –     | –     | –   | –   | –   | –     | –   | –    |
| IMM     | –     | –  | –   | –   | –     | –     | –     | –     | –   | –   | –   | –     | –   | –    |
| IVVP    | –     | –  | –   | –   | –     | –     | –     | –     | –   | –   | –   | –     | –   | –    |

For each of the 12 features that were made, the distribution of the scores for all of the 10,102 images are listed. Each number represents the number of images scored for each score. AAH, afferent/efferent arteriolar hyalinosis; Adh, adhesion; C-Cre, cellular crescent; CC, capillary collapse; EP, endocapillary proliferation; F-Cre, fibrous crescent; Fc-Cre, fibrocellular crescent; IMM, increased mesangial matrix; IVVP, increased vasculature around the vascular pole; MesHyper, mesangial hypercellularity; MLysis, mesangiolysis; Scl, sclerosis.

---

### Table 2. Breakdown of data used for the convolutional neural network (CNN)

| Score Feature | Negative | Positive | Ratio of negative to positive |
|---------------|----------|----------|------------------------------|
| CC            | None, Impossible to score 0.04 0.04 1.00 0.03 0.03 0.00 0.03 |
| IMM           | –        | –        | –                            |
| MLysis        | –        | –        | –                            |
| F-Cre         | –        | –        | –                            |
| IVVP          | –        | –        | –                            |

Each number shows the number of images contained in the score. CC, capillary collapse; F-Cre, fibrous crescent; IMM, increased mesangial matrix; IVVP, increased vasculature around the vascular pole; MLysis, mesangiolysis.

---

### Table 3. Phi coefficient among the 5 features

| Feature | Phi coefficient |
|---------|-----------------|
| CC      | 1.00            |
| IMM     | –0.50           |
| MLysis  | –0.04           |
| F-Cre   | 0.67            |
| IVVP    | –0.03           |

This table shows the phi coefficient between the 2 features among the 5 features. Each number shows the phi coefficient between the 2 features. The phi coefficient ranges from –1 to 1, where 1 represents a perfect positive correlation, –1 represents a perfect negative correlation, and 0 represents no correlation at all. CC, capillary collapse; F-Cre, fibrous crescent; IMM, increased mesangial matrix; IVVP, increased vasculature around the vascular pole; MLysis, mesangiolysis.
same variety of ICCs were also reported in the Japanese cohort study using the Oxford classification for IgA nephropathy where the ICCs ranged from 0.26 to 0.89.25 Because the ICCs are analogous with the K coefficient,26 the discordance was also reported in those studies.

As for the feature of MesHyper, the K coefficient is 0.35, whereas the K coefficient for the mesangial hypercellularity reported in the NEPTUNE scoring system16 was 0.54 to 0.64. Even though a direct comparison is difficult because of differences in the 2 study designs, the difference in the number of classes (5 for this study and 2 for NEPTUNE) might have led to the difference in the K score.

One possible reason that may have caused discordance is that single-stained WSIs were used for scoring the images. In some cases, clinicians see the WSIs for several stains, and they try to assign a score for an image. Combining other stained WSIs, such as PAM or Azan, might improve the concordance between clinicians.

CNN Performance
As shown in Table 5, the ROC–AUC for the CNN for the test data exceeded 0.9 for the CC and the F-Cre. The 2 features may have demonstrated a high performance due to the positive-to-negative ratio (described in Table 2), and the width of the pathological lesions were not so narrow. In contrast, features such as IVVP with a

---

Table 4. Results of Cohen \( \kappa \) coefficient by the 5 clinicians for the concordance images

|             | A/B | A/C | A/D | A/E | B/C | B/D | B/E | C/D | C/E | D/E | Average |
|-------------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|---------|
| CC          | 0.33| 0.30| 0.50| 0.35| 0.51| 0.61| 0.40| 0.44| 0.35| 0.41| 0.41    |
| Scl         | 0.18| 0.36| 0.46| 0.44| 0.46| 0.31| 0.26| 0.52| 0.38| 0.44| 0.38    |
| MesHyper    | 0.44| 0.37| 0.35| 0.37| 0.42| 0.27| 0.28| 0.29| 0.36| 0.35| 0.36    |
| IMM         | 0.45| 0.52| 0.62| 0.45| 0.43| 0.42| 0.45| 0.53| 0.45| 0.42| 0.47    |
| MLysis      | 0.54| 0.48| 0.68| 0.53| 0.25| 0.39| 0.41| 0.45| 0.56| 0.56| 0.49    |
| EP          | 0.19| 0.38| 0.54| 0.35| 0.18| 0.30| 0.41| 0.34| 0.35| 0.51| 0.36    |
| F-Cre       | 0.59| 0.50| 0.20| 0.64| 0.72| 0.04| 0.59| 0.02| 0.50| 0.28| 0.41    |
| Fc-Cre      | 0.34| 0.20| 0.30| 0.23| 0.72| 0.39| 0.56| 0.12| 0.44| 0.33| 0.36    |
| C-Cre       | 0.47| 0.57| 0.24| 0.33| 0.44| 0.26| 0.45| 0.22| 0.53| 0.40| 0.39    |
| Adh         | 0.16| 0.58| 0.26| 0.37| 0.18| 0.29| 0.23| 0.08| 0.23| 0.33| 0.28    |
| IVVP        | 0.40| 0.40| 0.45| 0.56| 0.41| 0.55| 0.56| 0.46| 0.63| 0.50| 0.50    |
| AAH         | 0.25| 0.16| 0.41| 0.47| 0.35| 0.46| 0.37| 0.39| 0.26| 0.60| 0.37    |

The rows indicate the features, and the column indicate the pairs of clinicians (e.g., A/B implies clinicians A and B) and their average value. Each number shows the Cohen \( \kappa \) coefficient. The Cohen \( \kappa \) coefficient ranges from –1 to 1, where 1 represents perfect concordance, 0 represents perfect discordance, and 0 represents completely no concordance. Adh, adhesion; AAH, afferent/efferent arteriolar hyalinosis; C-Cre, cellular crescent; CC, capillary collapse; EP, endocapillary proliferation; F-Cre, fibrous crescent; Fc-Cre, fibrocellular crescent; IMM, increased mesangial matrix; IVVP, increased vasculature around the vascular pole; MesHyper, mesangial hypercellularity; MLysis, mesangiolysis; Scl, sclerosis.
low performance demonstrated a high ratio of negative to positive, and a small pathological lesion may have caused the low performance. To improve the accuracy, the data should be collected to improve the class balance with the resampling methods.

As displayed in Table 5, there is a difference in the performance between the concordance data and the test data for some features. This seems to be due to the sampling method. Specifically, the test data were selected to have a good representation of the whole dataset by the stratified folding method; however, the concordance data were intentionally selected to contain “hard-to-score” images.

Concordance Between the CNN and Clinicians
As shown in Tables 6 and 7, for the CC, there was no cFP case and only 1 cFN case out of 98 images. Meanwhile, for the F-Cre, there were 20 cFP cases and 1 cFN case. F-Cre has more cFP images compared with CC. The reason why there are so many cFP images for the F-Cre is discussed in the next section.

Visualization of CNN’s Focusing Point
A visualization of the cTP case for the CC is illustrated in Figure 3, and the cFP cases for F-Cre are presented in Figures 4 and 5. In Figure 4, the CNN cannot distinguish between the 2 different features, which are F-Cre and Fc-Cre. It is hypothesized that F-Cre and Fc-Cre have a similar crescentic structure but different textures. On the other hand, in Figure 5, the CNN cannot distinguish between CC and F-Cre, which tended to occur simultaneously in this dataset (as depicted in Table 3). This is because during the training process, the CNN merely focuses on the frequently occurring lesion, which is not always a true pathological lesion. This phenomenon has been reported as an inherent bias derived from the dataset. By correcting the balance of co-occurrence between the 2 different features, this may enable the CNN to focus on the appropriate pathological lesion and to make the correct classification.

Except for the CC and F-Cre, there are several other strong correlations between the features of the renal pathology in terms of pathogenesis. For example, in most cases of IgA nephropathy, endocapillary proliferation tends to accompany mesangial hypercellularity. To avoid the inherent bias caused by this correlation, it is suggested that instead of classifying the features directly, the glomerular image should be divided into different lesions according to its structures, such as the capillary lesion, the mesangial lesion, and the extracapillary lesion, and semantic segmentation can be applied to extract the lesions from the glomerular image.

In conclusion, this study established the annotation criteria and an annotation flowchart to automatically classify the glomerular images using CNNs. The annotation criteria included 12 features and the degree of

### Table 5. Overall results of the convolutional neural network (CNN) performance

| Feature | Validation data | Test data | Concordance data |
|---------|-----------------|-----------|-----------------|
|         | ROC-AUC | F1 score | ROC-AUC | F1 score | ROC-AUC | F1 score |
| CC      | 0.97    | 0.79    | 0.98    | 0.79    | 0.84    | 0.59    |
| IMM     | 0.82    | 0.62    | 0.79    | 0.79    | 0.76    | 0.72    |
| Mlysis  | 0.87    | 0.12    | 0.76    | 0.08    | 0.37    | 0.04    |
| F-Cre   | 0.90    | 0.61    | 0.91    | 0.63    | 0.69    | 0.53    |
| IVVP    | 0.62    | 0.04    | 0.65    | 0.05    | 0.47    | 0.04    |

Each number shows the average receiver operating characteristic–area under the curve (ROC-AUC) or the F1 score over the 4 folds by each feature. CC, capillary collapse; F-Cre, fibrous crescent; IMM, increased mesangial matrix; IVVP, increased vasculature around the vascular pole; Mlysis, mesangiolysis.

### Table 6. Comparison between the convolutional neural network (CNN) predictive score and the clinicians’ feature for the capillary collapse (CC)

| No. of clinicians out of 5 who scored positive |
|-----------------------------------------------|
| 0 | 1 | 2 | 3 | 4 | 5 |

| CNN’s predictive score | 0.75–1.00 | 0.50–0.75 | 0.25–0.50 | 0.00–0.25 |
|------------------------|-----------|-----------|-----------|-----------|
| score over 4 folds     | 0.00      | 0.25      | 0.50      | 0.75      |
| 0                      | 2         | 2         | 2         | 2         |
| 1                      | 5         | 5         | 5         | 5         |
| 2                      | 0         | 0         | 0         | 0         |
| 3                      | 1         | 1         | 1         | 1         |
| 4                      | 0         | 0         | 0         | 0         |
| 5                      | 1         | 1         | 1         | 1         |

This confusion matrix shows the comparison between the CNN’s predictive score and the clinicians’ feature for the CC. Rows show the CNN’s predictive score for an image, calculated as an average of the CNN’s softmax probability over the 4 folds. Columns show the number of clinicians out of 5 who scored it as positive. Each number represents the number of images.

### Table 7. Comparison between the convolutional neural network (CNN) predictive score and the clinicians’ feature for the fibrous crescent (F-Cre)

| No. of clinicians out of 5 who scored positive |
|-----------------------------------------------|
| 0 | 1 | 2 | 3 | 4 | 5 |

| CNN’s predictive score | 0.75–1.00 | 0.50–0.75 | 0.25–0.50 | 0.00–0.25 |
|------------------------|-----------|-----------|-----------|-----------|
| score over 4 folds     | 0.00      | 0.25      | 0.50      | 0.75      |
| 0                      | 4         | 2         | 2         | 2         |
| 1                      | 0         | 0         | 0         | 0         |
| 2                      | 3         | 1         | 1         | 1         |
| 3                      | 0         | 0         | 0         | 0         |
| 4                      | 1         | 1         | 1         | 1         |
| 5                      | 0         | 0         | 0         | 0         |

This confusion matrix shows the comparison between the CNN’s predictive score and the clinicians’ feature for the F-Cre. Rows show the CNN’s predictive score for an image, which was calculated as an average of the CNN’s softmax probability over the 4 folds. Columns show the number of clinicians out of 5 who scored it as positive. Each number shows the number of images.

This confusion matrix shows the comparison between the CNN’s predictive score and the clinicians’ feature for the F-Cre. Rows show the CNN’s predictive score for an image, which was calculated as an average of the CNN’s softmax probability over the 4 folds. Columns show the number of clinicians out of 5 who scored it as positive. Each number represents the number of images.

This confusion matrix shows the comparison between the CNN’s predictive score and the clinicians’ feature for the F-Cre. Rows show the CNN’s predictive score for an image, which was calculated as an average of the CNN’s softmax probability over the 4 folds. Columns show the number of clinicians out of 5 who scored it as positive. Each number represents the number of images.
Figure 3. Example of the true positive score of the convolutional neural network (CNN) for the capillary collapse (CC). (Left) One of the periodic acid–Schiff (PAS)–stained images from the concordance data. The CNN diagnosed this image as “capillary collapse is positive,” and all 5 clinicians scored it as “capillary collapse is positive.” (Center) The blue lines depict the collapsed capillary, and the yellow lines show the fibrous crescents. (Right) Visualization of the CNN’s focusing points using the Gradient-weighted Class Activation Mapping when the CNN diagnosed the capillary collapse as positive or negative. Red lesions are the CNN’s focusing points. This shows that the CNN was focused correctly on the capillary area.

Figure 4. Example of the false-positive case of the convolutional neural network (CNN) for the fibrous crescent (F-Cre). (Left) One of the periodic acid–Schiff (PAS)–stained images from the concordance data. The CNN diagnosed this image as “F-Cre is positive,” but none of the 5 clinicians scored it as “F-Cre is positive,” whereas 4 of the 5 clinicians scored it as “Fibrocellular Crescent (Fc-Cre) is positive.” (Center) The yellow line shows the area of Fc-Cre. (Right) Visualization of the CNN’s focusing points using the gradient-weighted class activation mapping when the CNN diagnosed the F-Cre as positive or negative. The red lesions are the CNN’s focusing points. This shows that the CNN focused correctly on the extracapillary area but “misdiagnosed” it as “F-Cre is positive.”

Figure 5. Another example of the false-positive case of the convolutional neural network (CNN) for the fibrous crescent (F-Cre). (Left) One of the periodic acid–Schiff (PAS)–stained images from the concordance data. The CNN diagnosed this image as “F-Cre is positive” but none of the 5 clinicians scored it as “F-Cre is positive.” (Center) The blue line shows the area where the capillary was collapsed and obsoleted, and the yellow line depicts the fibrocellular crescent area. (Right) Visualization of the CNN’s focusing points using the gradient-weighted class activation mapping when the CNN diagnosed the F-Cre as positive or negative. The red lesions are the CNN’s focusing points. This reveals that the CNN accidentally focused on the capillary area although it needs to focus on the extracapillary lesion.
concordance. This was indicated by the average K coefficient among the 5 clinicians for each feature that ranged from 0.28 to 0.50. One clinician annotated 10,102 images and the performance of the CNN for the 5 features that were evaluated. The ROC–AUC values for the test data ranged from 0.65 to 0.98. In particular, CC and F-Cre demonstrated high ROC–AUC values of 0.98 and 0.91, respectively. The error analysis indicated that the CNN cannot distinguish features that have a similar visual structure or are likely to occur simultaneously. To address this problem, labeling every pixel with the structure class such as mesangial lesions or capillary lesions and applying semantic segmentation might improve the performance of the glomerular classification.

One of the limitations of this study is that the annotation criteria should be validated by other clinicians, and more images for concordance evaluation are desirable. Another limitation is that the images for the CNN development were annotated only by 1 clinician. Thus, there should be unintended bias in the dataset, and the results of the CNN performance might not be extrapolated to the other datasets.

This study has 3 important implications for future research on automatic glomerular classification. The first implication is that the discordance among the clinicians is similar to what has been reported in the literature. The second implication is that the width of the pathological lesion seems to affect the performance of the CNN. The third implication is that providing the predictive score and visualizing the focus of the CNN might improve the performance of the glomerular classification.

When diagnosing kidney diseases, scoring every glomerulus is important but time-consuming; hence, scoring requires the support of a computer. This study demonstrates the performance of a convolutional neural network when classifying the glomerular features. This study visually demonstrated that there are cases where the convolutional neural network cannot distinguish 2 different features that had a similar visual structure, or they occurred simultaneously. This result suggests the need to divide the glomerular image into different lesions according to its structures for computers to perform a proper assessment.

All the authors declared no competing interests.

This research was supported by the Health Labour Sciences Research Grants (grant number 28030401), Japan, and the Japan Science and Technology Agency, Promoting Individual Research to Nurture the Seeds of Future Innovation and Organizing Unique, Innovative Network (grant number JPMJPR1654), Japan. The Department of Artificial Intelligence in Healthcare, Graduate School of Medicine, the University of Tokyo is an endowment department that is supported with an unrestricted grant from I&H Co., Ltd. and EM SYSTEMS company; however, these sponsors had no control over the interpretation, writing, or publication of this work.

**AUTHOR CONTRIBUTIONS**

RY and YK designed the experiments. RY, YS-D, HU, AS, and MN established the annotation criteria and flowchart and performed the concordance experiment. RY conducted the deep learning and analyzed the data. RY, YK, KS, ES, and KO wrote the article. TT, HN, HU, TU, MN, NK, AS, and MN collected the dataset. KO obtained the funding for this research.

**DATA ACCESS**

The dataset used in this study was collected by UTH, KH, and UTSH. The experiments and the data collection were approved by the institutional review board of each hospital (UTH:11455, KH:P17-05-004, UTSH:H29-160). The datasets are not available to the public, and restrictions apply to their use. The code is available at https://github.com/ryama-tokyo/glomerular_classification.

**DISCLOSURE**

All the authors declared no competing interests.

**ACKNOWLEDGMENTS**

This research was supported by the Health Labour Sciences Research Grants (grant number 28030401), Japan, and the Japan Science and Technology Agency, Promoting Individual Research to Nurture the Seeds of Future Innovation and Organizing Unique, Innovative Network (grant number JPMJPR1654), Japan. The Department of Artificial Intelligence in Healthcare, Graduate School of Medicine, the University of Tokyo is an endowment department that is supported with an unrestricted grant from I&H Co., Ltd. and EM SYSTEMS company; however, these sponsors had no control over the interpretation, writing, or publication of this work.

**REFERENCES**

1. Pfister M, Jakob S, Frey FJ, et al. Judgment analysis in clinical nephrology. Am J Kidney Dis. 1999;34:569–575.
2. Kitterer D, Gürzing K, Segerer S, et al. Diagnostic impact of percutaneous renal biopsy. Clin Nephrol. 2015;84:311–322.
3. He L, Long LR, Antani S, et al. Histology image analysis for carcinoma detection and grading. Comput Methods Programs Biomed. 2012;107:538–556.
4. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521:436–444.
5. Fukushima K, Miyake S. Neocognitron: a new algorithm for pattern recognition tolerant of deformations and shifts in position. Pattern Recognit. 1982;15:455–469.

**AUTHOR CONTRIBUTIONS**

RY and YK designed the experiments. RY, YS-D, HU, AS, and MN established the annotation criteria and flowchart and performed the concordance experiment. RY conducted the deep learning and analyzed the data. RY, YK, KS, ES, and KO wrote the article. TT, HN, HU, TU, MN, NK, AS, and MN collected the dataset. KO obtained the funding for this research.

**DATA ACCESS**

The dataset used in this study was collected by UTH, KH, and UTSH. The experiments and the data collection were approved by the institutional review board of each hospital (UTH:11455, KH:P17-05-004, UTSH:H29-160). The datasets are not available to the public, and restrictions apply to their use. The code is available at https://github.com/ryama-tokyo/glomerular_classification.

**TRANSLATIONAL STATEMENT**

When diagnosing kidney diseases, scoring every glomerulus is important but time-consuming; hence, scoring requires the support of a computer. This study demonstrates the performance of a convolutional neural network when classifying the glomerular features. This study visually demonstrated that there are cases where the convolutional neural network cannot distinguish 2 different features that had a similar visual structure, or they occurred simultaneously. This result suggests the need to divide the glomerular image into different lesions according to its structures for computers to perform a proper assessment.

**SUPPLEMENTARY MATERIAL**

Supplementary File (PDF)

Table S1. Annotation criteria for the 12 features that are important for diagnostics

Table S2. Breakdown list of the whole slide images (WSIs)

Figure S1. Annotation example images for the 12 features.

Figure S2. Images used for concordance evaluation.
6. Saha M, Chakraborty C, Racoceanu D. Efficient deep learning model for mitosis detection using breast histopathology images. Comput Med Imaging Graph. 2018;64:29–40.

7. Gallego J, Pedraza A, Lopez S, et al. Glomerulus classification and detection based on convolutional neural networks. J Imaging. 2018;4:20.

8. Kawazoe Y, Shimamoto K, Yamaguchi R, et al. Faster R-CNN-based glomerular detection in multistained human whole slide images. J Imaging. 2018;4:91.

9. Bukowy JD, Dayton A, Cloutier D, et al. Region-based convolutional neural nets for localization of glomeruli in trichrome-stained whole kidney sections. J Am Soc Nephrol. 2018;29:2081–2088.

10. Hermesen M, Bel T, den Boer M, et al. Deep learning-based histopathologic assessment of kidney tissue. J Am Soc Nephrol. 2019;30:1968–1979.

11. Barros GO, Navarro B, Duarte A, et al. PathoSpotter-K: a computational tool for the automatic identification of glomerular lesions in histological images of kidneys. Sci Rep. 2017;7:1–8.

12. Kannan S, Morgan LA, Liang B, et al. Segmentation of glomeruli within trichrome images using deep learning. Kidney Int Reports. 2019;4:955–962.

13. Marsh JN, Matlock MK, KudoSatore S, et al. Deep learning global glomerulosclerosis in transplant kidney frozen sections. IEEE Trans Med Imaging. 2018;37:2718–2728.

14. Chagas P, Souza L, Araújo I, et al. Classification of glomerular hypercellularity using convolutional features and support vector machine. Artif Intell Med. 2020;103:101808.

15. Roberts ISD, Cook HT, Troyanov S, et al. The Oxford classification of IgA nephropathy: pathology definitions, correlations, and reproducibility. Kidney Int. 2009;76:546–556.

16. Barisoni L, Troost JP, Nast C, et al. Reproducibility of the NEPTUNE descriptor-based scoring system on whole-slide images and histologic and ultrastructural digital images. Mod Pathol. 2016;29:671–684.