MorphSet: Improving Renal Histopathology Case Assessment Through Learned Prognostic Vectors

Pietro Antonio Cicalese[1](✉), Syed Asad Rizvi[1], Victor Wang[1], Sai Patibandla[1], Pengyu Yuan[1], Samira Zare[1], Katharina Moos[2], Ibrahim Batal[3], Marian Clahsen-van Groningen[4], Candice Roufosse[5], Jan Becker[2], Chandra Mohan[1], and Hien Van Nguyen[1]

[1] University of Houston, Houston, TX, USA
pcicalese@uh.edu
[2] Institute of Pathology, University Hospital of Cologne, Cologne, Germany
[3] Columbia University College of Physicians and Surgeons, New York, NY, USA
[4] Department of Pathology, Erasmus MC, Rotterdam, The Netherlands
[5] Department of Medicine, Imperial College, London, UK

Abstract. Computer Aided Diagnosis (CAD) systems for renal histopathology applications aim to understand and replicate nephropathologists' assessments of individual morphological compartments (e.g. glomeruli) to render case-level histological diagnoses. Deep neural networks (DNNs) hold great promise in addressing the poor intra- and interobserver agreement between pathologists. This being said, the generalization ability of DNNs heavily depends on the quality and quantity of training labels. Current "consensus" labeling strategies require multiple pathologists to evaluate every compartment unit over thousands of crops, resulting in enormous annotative costs. Additionally, these techniques fail to address the underlying reproducibility issues we observe across various diagnostic feature assessment tasks. To address both of these limitations, we introduce MorphSet, an end-to-end architecture inspired by Set Transformers which maps the combined encoded representations of Monte Carlo (MC) sampled glomerular compartment crops to produce Whole Slide Image (WSI) predictions on a case basis without the need for expensive fine-grained morphological feature labels. To evaluate performance, we use a kidney transplant Antibody Mediated Rejection (AMR) dataset, and show that we are able to achieve 98.9% case level accuracy, outperforming the consensus label baseline. Finally, we generate a visualization of prediction confidence derived from our MC evaluation experiments, which provides physicians with valuable feedback.

Keywords: Self attention · Antibody Mediated Rejection · Morphology

1 Introduction

Histopathology is on the verge of transforming into a highly quantitative and computational discipline. Within the next decade, Deep Neural Network (DNN) based Computer Aided Diagnosis (CAD) systems are expected to become indispensable for

M. de Bruijne et al. (Eds.): MICCAI 2021, LNCS 12908, pp. 319–328, 2021.
https://doi.org/10.1007/978-3-030-87237-3_31

histopathologists in their daily routine diagnostics, improving reproducibility and accuracy at considerably lower costs and with better transparency than molecular tests. Nephropathology is a subspecialty of histopathology that integrates paraffin histology, immunohistology and transmission electron microscopy observations into a single diagnosis for non-tumor diseases in native and transplant kidneys. In nephropathology, the diagnosis of disease entities like glomerulonephritis (as either present or absent) is highly reproducible. On the other hand, the assistance of such CADs is much more urgently needed for fine-grained prognostic details and disease entities with discriminative, gradual biology, such as Antibody-Mediated Rejection (AMR). For diagnostic simplicity, AMR is diagnosed as present (chronic active, chronic, or active) or absent according to the Banff Classification of histopathological and clinical parameters [11]. To address the limitations observed in existing classification methods for diseases with gradual manifestation, several research groups have developed supervised Convolutional Neural Network (CNN) classification architectures that can approximate the scoring criteria developed for several renal pathologies with variable success. They do not, however, solve the issues underlying pathologists' scores of individual compartment units.

Various research groups have opted to have multiple renal pathologists annotate the same images independently, treating their assessments as votes that can be used to approximate the general concepts underlying the given scoring criteria. This allows the CNN to generalize well to new cases, but it introduces significant annotative cost. In addition, consensus voting schemes aim to reflect the general opinion of the pathologists, but the fact that the minority votes are discarded may be harmful to performance since these discarded votes may still be informative for a disease with a gradual manifestation [12]. While other approaches may attempt to make use of all annotations, these processes still depend on votes generated with scoring criteria that lack reproducibility and undergo frequent revision [10]. We were interested in bypassing the scoring process for individual compartments entirely, leveraging the ability of DNNs to learn prognostic features in order to achieve reliable case-level accuracy. This could bypass the need for standardized nephropathology descriptors, as recently defined by Haas *et al.* [5], which have unknown underlying biology and uncertain reproducibility. Moreover, it would obviate the need for pre-analytical standardization of laboratory procedures as proposed by Barisoni *et al.* [3]. Thus, we introduce MorphSet, an architecture capable of attributing prognostic features to individual morphological compartments through a mechanism inspired by Set Transformers [9]. Our contributions are as follows:

- We propose a novel Monte Carlo (MC) glomeruli sampling method for generating unique subsets of available images for a given case, which we use to produce case-level predictions while improving regularization.
- We introduce two case-level architectures; the first architecture uses convolutional layers to process the input images for a given case, while the second architecture, which we call MorphSet, utilizes a multi-head self attention mechanism to compare embeddings of input images to various learned prognostic vectors.
- We repeat the MC sampling step which allows us to produce an aggregate prediction that is more representative of the input case. We subsequently use these MC predictions to generate model confidence visualizations, which provide meaningful feedback to the pathologist.

Related Works. Much of the recent DNN-based classification work in renal histopathology has been centered around the prediction of pathological findings in individual morphological compartments, particularly in glomeruli due to its diagnostic or prognostic relevance. Uchino *et al.* developed glomerular classifiers for seven pathological findings by fine-tuning an InceptionV3 network and using a majority decision approach to predict consensus labels on glomerular image crops [14]. While promising, we note that several important pathological findings could not be accurately classified, meaning that performance across the full spectrum of renal diseases remains elusive. Possible reasons for this are data scarcity and interobserver disagreement for the compartment unit labels, even with consensus descriptors. Overcoming scarcity problems and correcting for annotative disagreements in available nephropathology datasets will be a critical step for the further digitization of histopathology.

Other techniques for case-level classification of renal diseases have focused on resisting label noise and non-descriptive images in renal datasets, as well as patch-evaluation methods for WSI prediction. Cicalese *et al.* proposed an uncertainty-guided CAD system for kidney-level case prediction that assigned case-level labels to individual morphological compartments in images and allowed the classifier to filter out non-descriptive images in its predictions [4]. Xu *et al.* used a Multiple-Instance Learning (MIL) framework to classify high resolution colon histopathology images, aggregating patch predictions by a voting criterion in which a WSI is predicted positive if it contains a patch that is predicted positive [16]. While promising, the vote aggregating criterion used is not robust, potentially giving single patches disproportionate influence over the final WSI prediction. Hou *et al.* addressed this by training a decision fusion model to aggregate high-resolution patch predictions into WSI-level predictions, outperforming both max-pooling and voting aggregation mechanisms on glioma and Non-Small-Cell Lung Carcinoma WSI cases [6]. This technique, however, relies on the automated extraction of discriminative crops from WSIs through the use of an Expectation Maximum (EM) based CNN method, which risks rejecting patches that are hard to classify but still biologically relevant. We were interested in explicitly learning prognostic features which we could use to produce unique embeddings corresponding to individual discriminative concepts with prognostic relevance. To accomplish this, we took inspiration from the Set Transformer architecture, which computes pixel-wise interactions with respect to a series of learned concepts for an input set [9]. We could then use these learned concepts to map a set of glomerular images to its relevant disease diagnosis, thus circumventing the need for fine-grained glomerular labels.

2 Methodology

2.1 AMR Dataset Generation and Annotation

To evaluate the effectiveness of our method with respect to a fine-grained annotation scheme, we used an Antibody Mediated Rejection (AMR) glomerular crop dataset. We chose this dataset given that we knew the case-level ground truths prior to data processing (i.e. we knew which transplants were positive for AMR through other criteria as donor-specific antibodies or C4d positive on immunohistology). We randomly selected a total of 89 (51 chronic active, chronic, or active AMR and 38 Non-AMR) blood group

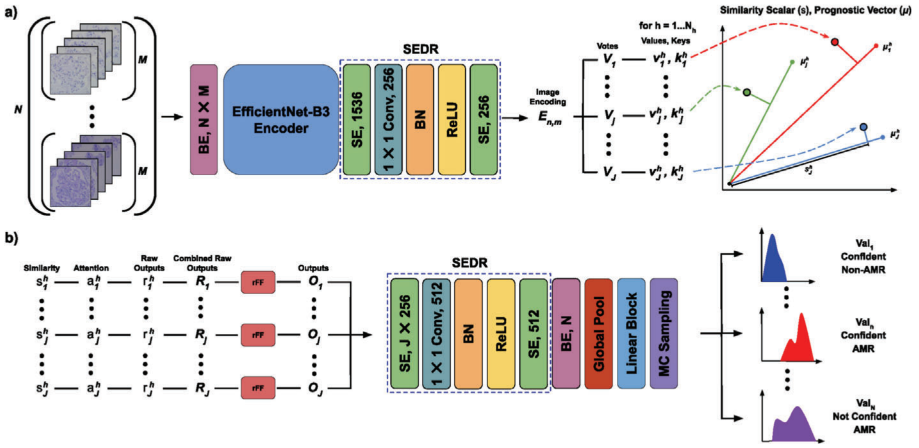

Fig. 1. The overall MorphSet architecture. **a)** We begin by sampling our M images across N cases, and encode each image as an individual batch element. After passing our encoded images through our Squeeze-Excitation Dimensionality Reduction (SEDR) block, we compare each image to a set of J parametrized Prognostic Vectors (PVs). **b)** Once we have computed our similarity scores, we proceed to produce unique embeddings for each $\boldsymbol{\mu}_j$, thus generating image-level assessments for each PV. We then pass our outputs through a second SEDR block, and concatenate the M encoded image embeddings together, yielding a batch size of N. After our global pooling and linear classification blocks, we can then perform Monte Carlo (MC) sampling during the evaluation phase to generate probability density curves, providing valuable feedback to the physician.

ABO- compatible, paraffin embedded kidney transplant biopsies, all of which satisfied the minimum sample criteria ($\geq$7 glomeruli, $\geq$1 artery) [11]. All sections were cut to 2 µm and were Periodic acid-Schiff (PAS) stained in the same pathology lab over a two year time frame. Micrographs were taken from all non-globally sclerosed glomeruli that were at least four levels apart at a resolution of 1024 × 768, yielding a total of 1,655 glomerular crops. Each of these images were then labeled by any combination of three experienced nephropathologists (from a group of four) using the LabelBox platform, with choices being AMR, non-AMR, or inconclusive [1]. In the event of a three way disagreement, the fourth pathologist would break the tie (54 tiebreakers, yielding a total of 5,019 annotations). Each image was then manually segmented by a single experienced pathologist using QuPath, a digital pathology software, to produce fine-grained masks for the biologically relevant glomerular compartment unit which were then used to extract the glomerular information prior to classification [2].

2.2 MorphSet

Given the variable number of glomerular crops that we see for any of our given N cases, we chose to pursue a MC sampling scheme; this allows us to sample M unique images at each iteration of training for a given case n. By doing this, we ensure that the architecture sees a new combination of images for each pass through the network, allowing it to account for the variability we see between glomerular crops and entire cases. In the event that there are fewer than M images available for a given case, we simply use online augmentation on the available set to produce the remaining images. We then pass these images through a CNN encoder to produce our initial feature embeddings. Once this is done, we then pass the outputs of our encoder network through a Squeeze-Excitation Dimensionality Reduction (SEDR) block (see Fig. 1), which consists of Squeeze-Excitation (SE) attention, followed by a linear layer, batch normalization, a ReLU activation function, and another SE attention operation [7].

A key component of our architecture lies in its ability to generate output representations for each image with respect to every encoded Prognostic Vector (PV), which we refer to as the MorphSet operation. We define PVs as learned discriminative feature embeddings that can be used to interpret individual images with respect to a specified number of key concepts (i.e. the number of PVs). Let $\boldsymbol{E}^{m,n} \in \mathbb{R}^{p \times p}$ represent the encoded representation of the m^{th} image from the n^{th} case, with $p \times p$ features. These embeddings are then transformed to yield our votes $\boldsymbol{V}_j^{m,n} \in \mathbb{R}^{p \times p}$, where j represents a particular PV that we wish to learn. We generate our votes using shared learned transformation matrices $\boldsymbol{W}_j \in \mathbb{R}^{p \times p}$, following

$$\boldsymbol{V}_j^{m,n} = \boldsymbol{W}_j \boldsymbol{E}^{m,n} \tag{1}$$

The vote generation process can be interpreted as a preparatory step which we will use to learn to compare each discriminative feature separately, mimicking how pathologists assess tissue morphology with respect to each relevant scoring task.

We chose to use an attention mechanism similar to that described in the Set Transformer architecture to compute the similarity between a set of votes $\boldsymbol{V}_j$ and their respective parametrized PVs μ_j, which were Kaiming initialized [9]. We accomplish this by generating a similarity measure between each vote and its parametrized PV, thus biasing the network to information that is representative of the given discriminative marker. Using this set operation also allows our comparison mechanism to retain linear time complexity $\mathcal{O}(J)$, where J is the number of learned PVs. To describe our comparison mechanism, we adopt the following naming conventions: the number of attention heads is denoted by N_h, while the number of dimensions for the key and value vectors are given by d_k and d_v, respectively. We define multi-head attention as evenly dividing the features in d_k and d_v into N_h pairs of output vectors such that d_k^h and d_v^h represent the key and value vectors of head h. We will omit the image indices m and n for the sake of simplicity while describing this mechanism.

We begin by flattening our vote matrices $\boldsymbol{V}_j$ to then generate N_h different d_k^h and d_v^h dimensional key ($\boldsymbol{k}_j^h$) and value ($\boldsymbol{v}_j^h$) vectors, using a set of learnable transformation matrices $\Lambda = \{\boldsymbol{W}_k^h, \boldsymbol{W}_v^h\}_{h=1}^{N_h}$, where $\boldsymbol{W}_k^h \in \mathbb{R}^{p^2 \times d_k^h}$ and $\boldsymbol{W}_v^h \in \mathbb{R}^{p^2 \times d_v^h}$. To simplify our presentation, we will set $d = d_v^h = d_k^h$ throughout the remainder of the paper. We parametrize our PV as $\boldsymbol{\mu}^h$ and compute each element of our similarity matrix $\boldsymbol{S} \in \mathbb{R}^{N_h \times J}$ following

$$\boldsymbol{s}_j^h = \boldsymbol{k}_j^h \cdot \boldsymbol{\mu}_j^h \tag{2}$$

We can now generate our output using our computed similarity matrix with

$$\boldsymbol{r}_j^h = \boldsymbol{\mu}_j^h + \boldsymbol{a}_j^h \cdot \boldsymbol{v}_j^h \quad \text{where} \quad \boldsymbol{a}^h = \texttt{softmax}(\boldsymbol{s}^h/\sqrt{d}) \tag{3}$$

We scale the similarity vectors $\boldsymbol{s}^h$ by a factor of $1/\sqrt{d}$ to avoid the vanishing gradient problem described in [15], and then softmax the result to generate our final attention coefficients $\boldsymbol{a}^h$. We can think of each PV $\boldsymbol{\mu}_j^h$ as a static memory component, encoding the typical appearance of some discriminative feature, while the dynamic component $\boldsymbol{a}_j^h \cdot \boldsymbol{v}_j^h$ represents the degree to which a given input image deviates from the static concept. We then transform our concatenated outputs following

$$\boldsymbol{R} = \texttt{Norm}[\texttt{concat}(\boldsymbol{r}^1, ..., \boldsymbol{r}^{N_h})\boldsymbol{W}_o] \tag{4}$$

using Batch-Normalization (BN) for our `Norm` computation [8]. Finally, the fully processed output representation is given by

$$\boldsymbol{O} = \texttt{Norm}[\boldsymbol{R} + \texttt{rFF}(\boldsymbol{R})] \tag{5}$$

with `rFF` corresponding to a linear transformation that processes the inputs identically.

We then pass our feature embeddings through another SEDR dimensionality reduction block to facilitate our case-level computations. At this point, we edit the batch size to be of size N, where each stack of M image embeddings correspond to a single batch element n. Once this is done, we perform global pooling and pass our model through three linear layers to generate our case-level predictions. Another advantage of this architecture lies in its MC sampling protocol, which allows us to construct confidence metrics during evaluation. We may chose to sample multiple times in order to construct a probability density curve, which provides the operating physician with valuable feedback about the model's confidence in its assessment.

3 Results and Discussion

Training Settings. Throughout our experiments, we used an EfficientNet-B3 encoder that was pre-trained on ImageNet, given its computational efficiency and high performance [13]. To analyze the benefits of treating each case as an unlabeled set during classification, we also trained an EfficientNet-B3 AMR/Non-AMR glomerulus level classifier using a consensus labeling scheme (pre-trained on ImageNet, all inconclusive glomeruli removed). To analyze the impact of our PV embeddings, we trained a separate model that replaced the MorphSet operation with a simple convolution with the same output dimensionality, yielding our convolutional baseline model. All models were trained using a Binary Cross Entropy loss function with the Adam optimizer for 400 epochs with a learning rate of 1×10^{-4}, a β_1 value of 0.9, β_2 value of 0.999, and L2 coefficient of 0.01. For all of our experiments, we used a five-fold cross validation scheme, and all images were resized to 256×256 before being passed through each model. During training of both the convolutional baseline model and MorphSet, a batch size of three cases was used, with 12 glomerular crops being sampled from each case, yielding 36 input images per batch. To ensure that our comparisons were fair, we trained the EfficientNet-B3 glomerulus level classifier using a batch size of 36.

All images were ImageNet normalized, and were augmented during training using standard online transformations, where each image had a 50% chance of being horizontally flipped, vertically flipped, cropped to 70%–100% of its input size, and rotated between 0–90°, in that order.

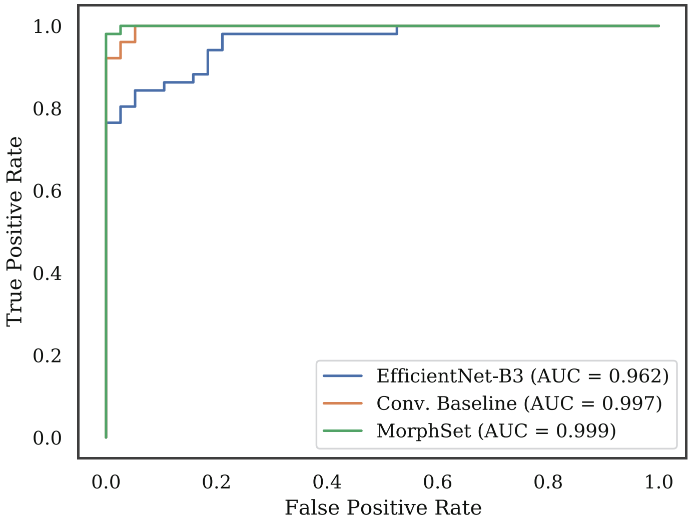

Fig. 2. ROC case-level curves for our three experimental models. The EfficientNet-B3 model's ROC curve was computed using the percentage of glomeruli classified as AMR for each case to avoid introducing a threshold bias. The models trained using the MC sampling scheme yielded higher AUC scores when compared to the EfficientNet-B3 baseline model.

MorphSet Performance. After encoding each of our input images with our EfficientNet-B3 encoder, we then reduce the dimensionality of the output from $[36, 1536, 8, 8]$ to $[36, 256, 8, 8]$ using our SEDR block (as shown in Fig. 1a). We train our architecture using eight parametrized PVs with one attention head ($k = 3$, $s = 2$, $p = 1$), resulting in an output dimensionality of $[32, 2048, 4, 4]$. We then pass the resulting outputs through our second SEDR block, concatenate the image embeddings for each case, and perform global pooling, producing an output dimensionality of $[3, 6144]$ (as shown in Fig. 1b). Our linear block consists of three linear layers, two of which reduce the dimensionality by a factor of two with batch normalization, followed by an output layer. We then perform MC sampling 100 times on each validation set, taking the average of our sigmoid activation outputs to produce our final predictions. To compare our EfficientNet-B3 baseline to both MC architecture's case-level scores, we chose to assign the percentage of glomeruli classified as AMR by the EfficientNet-B3 architecture for each case as it's respective case-level score. We then used an ROC curve to compare the performance between the three models (as shown in Fig. 2). We did this as opposed to fixing some classification threshold for EfficientNet-B3 AMR case level predictions (i.e. $>50\%$ of glomeruli classified as AMR constitutes an AMR case prediction) because pathologists do not generate case level diagnoses by using a hard set threshold on their glomerular assessments. Reporting case level accuracy in this way would therefore not be particularly meaningful from a medical standpoint, whereas reporting our results using an ROC curve allows us to avoid introducing a threshold bias, thus allowing us to compare the models fairly. Our resulting ROC curve highlights the performance improvements that we can attribute to our MC sampling set approach, with AUC increases in both the MorphSet and convolutional baseline models when compared to the glomerular level classifier. The convolutional baseline and MorphSet models achieve a case-level validation accuracy of 97.8% and 98.9%, respectively.

Probability Density Curves. To better understand the differences between MorphSet and the convolutional baseline model, we produced probability density curves using the 100 sigmoid activation outputs generated for each case during the MC sampling step. We found that MorphSet tended to produce higher probability point estimates for each AMR case, while also remaining more confident than the convolutional baseline model, as determined by our standard deviation (STD) computation, illustrated in Fig. 3. This result implies that MorphSet was better suited to learning the glomerular characteristics of AMR, and highlights the potential of our PV approach.

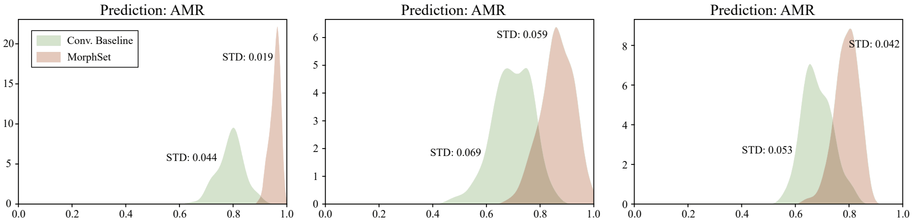

Fig. 3. Density versus probability point estimate curves generated using the convolutional baseline architecture (green) and MorphSet (red) for three AMR examples. We note that higher prediction densities at higher probability point estimate values imply that a model is more confident in its case level AMR prediction. MorphSet tended to produce AMR predictions with higher confidence, suggesting that MorphSet achieved a better understanding of glomerular AMR characteristics. (Color figure online)

4 Conclusion

In this work, we present an MC sampling approach for the case-level assessment of AMR in PAS stained renal histopathology glomerular crops, relieving the need for fine-grained structural annotation. We introduce both a convolutional case-level classifier and MorphSet, which learns unique Prognostic Vectors (PVs) meant to represent the discriminative concepts used by a pathologist when assessing tissue biopsies. We show that both of our proposed models outperform our fine-grained glomerulus classifier without having to remove inconclusive images or rely on using glomerular-level annotations. We also show that MorphSet was more confident in its AMR predictions while producing higher probability point estimates, suggesting it achieved a stronger understanding of the disease characteristics. Future works should aim to investigate the performance of MorphSet with larger datasets and multiple disease cases to improve our understanding of its generalization ability and how well it adapts to an increased number of discriminative concepts. The ability of the architecture to identify discriminative images for cases is another potential area of further study, as is the possibility of scaling up learning sets through reference pathologist cases without the need for fine-grained annotation.

References

1. Labelbox, labelbox, online (2020). https://labelbox.com
2. Bankhead, P., et al.: QuPath: open source software for digital pathology image analysis. bioRxiv (2017). https://doi.org/10.1101/099796, https://www.biorxiv.org/content/early/2017/03/06/099796
3. Barisoni, L., et al.: Digital pathology imaging as a novel platform for standardization and globalization of quantitative nephropathology. Clin. Kidney J. **10**(2), 176–187 (2017). https://doi.org/10.1093/ckj/sfw129
4. Cicalese, P.A., Mobiny, A., Shahmoradi, Z., Yi, X., Mohan, C., Van Nguyen, H.: Kidney level lupus nephritis classification using uncertainty guided Bayesian convolutional neural networks. IEEE J. Biomed. Health Inform. **25**(2), 315–324 (2021). https://doi.org/10.1109/JBHI.2020.3039162
5. Haas, M., et al.: Consensus definitions for glomerular lesions by light and electron microscopy: recommendations from a working group of the Renal Pathology Society. Kidney Int. **98**(5), 1120–1134 (2020). https://doi.org/10.1016/j.kint.2020.08.006
6. Hou, L., Samaras, D., Kurc, T.M., Gao, Y., Davis, J.E., Saltz, J.H.: Patch-based convolutional neural network for whole slide tissue image classification (2016)
7. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7132–7141 (2018)
8. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. CoRR arXiv:1502.03167 (2015)
9. Lee, J., Lee, Y., Kim, J., Kosiorek, A., Choi, S., Teh, Y.W.: Set transformer: a framework for attention-based permutation-invariant neural networks. In: International Conference on Machine Learning, pp. 3744–3753. PMLR (2019)
10. Liapis, G., Singh, H.K., Derebail, V.K., Gasim, A.M.H., Kozlowski, T., Nickeleit, V.: Diagnostic significance of peritubular capillary basement membrane multilaminations in kidney allografts: old concepts revisited. Transplantation **94**(6), 620–629 (2012)
11. Roufosse, C., et al.: A 2018 reference guide to the Banff classification of renal allograft pathology. Transplantation **102**(11), 1795–1814 (2018)
12. Smith, B., et al.: A method to reduce variability in scoring antibody-mediated rejection in renal allografts: implications for clinical trials - a retrospective study. Transpl. Int. **32**(2), 173–183 (2019). https://doi.org/10.1111/tri.13340
13. Tan, M., Le, Q.: EfficientNet: rethinking model scaling for convolutional neural networks. In: Chaudhuri, K., Salakhutdinov, R. (eds.) Proceedings of the 36th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 97, pp. 6105–6114. PMLR, June 2019. http://proceedings.mlr.press/v97/tan19a.html
14. Uchino, E., et al.: Classification of glomerular pathological findings using deep learning and nephrologist-AI collective intelligence approach. Int. J. Med. Inform. **141**, 104231 (2020)

15. Vaswani, A., et al.: Attention is all you need (2017)
16. Xu, Y., Mo, T., Feng, Q., Zhong, P., Lai, M., Chang, E.I.: Deep learning of feature representation with multiple instance learning for medical image analysis. In: 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1626–1630 (2014). https://doi.org/10.1109/ICASSP.2014.6853873