Classification of Glomerulonephritis with CNN and Self-Attention Networks in Individual Glomeruli in Nephropathology

Gloria Bueno*
VISILAB - Engineering School
Universidad de Castilla-La Mancha
Ciudad Real, Spain

Anibal Pedraza
VISILAB - Engineering School
Universidad de Castilla-La Mancha
Ciudad Real, Spain

Israel Mateos-Aparicio-Ruiz
VISILAB - Engineering School
Universidad de Castilla-La Mancha
Ciudad Real, Spain

Hien Van Nguyen
Dept. Electrical & Computer Eng.
University of Houston
Houston, USA

Nicola Altini
Dept. Informatica Industriale
Politecnico di Bari
Bari, Italy

Huy Q. Vo
Dept. Electrical & Computer Eng.
University of Houston
Houston, USA

Dejan Dobi
Dept. of Pathology
Fac. of Med., Semmelweis University
Budapest, Hungary

Jean-Baptiste Gibier
Dept. of Pathology
Lille University Hospital
Lille, France

Alessandro Del Gobbo
Dept. of Pathology
Ospedale Maggiore Policlinico
Milan, Italy

Lucia González
Dept. of Pathology
Hospital General Univ. Ciudad Real
Ciudad Real, Spain

Loreto Gesualdo
Dept. of Emerg. & Organ Transpl.
University of Bari Aldo Moro
Bari, Italy

Francesco Pesce
Division of Renal Medicine
Fatebenefratelli IT - Gemelli Isola
Rome, Italy

Michele Rossini
Dept. of Emerg. & Organ Transpl.
University of Bari Aldo Moro
Bari, Italy

Avi Rosenberg
Dept. of Pathology
Johns Hopkins University
Baltimore, USA

Jan Ulrich Becker
Institute of Pathology
University Hospital of Cologne
Cologne, Germany

***Abstract*—Despite the renal biopsy being the gold standard for diagnosing glomerulonephritis, this practice remains inaccessible for many patients worldwide. Nephropathologists typically combine microscopy, immunohistology, transmission electron microscopy, clinical information, and genetic studies for diagnosis. However, variability in nephropathology evaluation has hindered its integration with emerging technologies and personalized medicine. This study proposes the use of deep learning to extract significant features to distinguish glomerulonephritis from PAS sections without other modalities. To test this hypothesis, various AI methods were tested for classifying 12 common glomerulonephritis diagnoses. Finally, a sequential classification was implemented, initially characterizing sclerosed and non-sclerosed glomeruli using Swin-Transformers, followed by classifying the non-sclerosed glomeruli into 12 types of glomerulonephritis using ConvNeXt. The first step achieved an average Balanced Accuracy of 97% and an AUC of 0.96. In the second step, a Balanced Accuracy considering up to the top3 of 79.5% and an avarage AUCs of 0.76 were achieved. This study establishes a baseline for this challenging classification task, demonstrating promising results even on single PAS glomerular crops.**

***Index Terms*—Classification of Glomerulonephritis, Deep Learning in Medical Diagnostics, Machine Learning in Nephropathology, Self-Attention Architectures, Digital Pathology**

The authors acknowledge funding support from the HANS project (Ref. PID2021-127567NB-I00), funded by the Spanish Ministry of Science, Innovation, and Universities, as well as by the European Union NextGenerationEU/PRTR. Additional support was provided by the MICROSCOM project (GA: 101128153), funded by the European Commission.

*Corresponding author: gloria.bueno@uclm.es

I. Introduction

THE digitalization of pathology departments has increased the use of whole slide images (WSIs), improving teleconsultation, education, and image archiving. Progress in machine learning (ML) and hardware has enabled the integration of artificial intelligence (AI), developing tools to assist pathologists in image analysis. Within digital pathology, nephropathology stands out as a subspecialty that deals with low-incidence diseases. Nephropathology has also been impacted by advancements in digital pathology and AI, as well as the work of multidisciplinary teams [1].

Glomerular diseases are conditions that directly affect the glomeruli. Patients with these types of diseases present multiple signs and symptoms (e.g., edema, proteinuria, and hematuria) that are present in several of them. These diseases are often characterized by certain common histopathological properties as well. However, even with diagnosis done this way, patients with the same disease show significant variability in their presentation, progression, and response to treatment [2]. Therefore, the classification of glomerular diseases has often been described as complex by experts [3].

Renal biopsy is the gold standard for diagnosing glomerulonephritis, using conventional microscopy techniques, immunohistology, and transmission electron microscopy. However, this practice remains inaccessible for diagnosing many patients worldwide. Nephropathologists combine these observations with clinical information and genetic studies to diagnose and provide prognostic and theranostic information. Despite efforts to standardize biopsy processing and histopathological descriptors, evaluation in nephropathology remains variable, hindering its integration with emerging technologies and personalized medicine.

The hypothesis of this work is that classification methods based on deep learning models are capable of extracting significant features that allow distinguishing glomerulonephritis from PAS sections, without the need of using other modalities. To test this hypothesis and analyze the capabilities of these methods, various AI methods have been implemented for the classification of 12 common diagnoses of glomerulonephritis. These glomerulonephritis and the acronym used in this work are shown in Table I.

One of the most common applications is to classify glomeruli into sclerosed or non-sclerosed. For this purpose, CNNs have been primarily used, as in [4], [5], and [6]. In this latter publication, segmentation and classification of tubular structures are also performed. Regarding other pathologies, Altini *et al.* [7] uses the Oxford classification [8]–[10] for Immunoglobulin A-associated Nephropathy (IgAN), with datasets containing glomeruli from the 12 types of glomerulonephritis used in this study. Other recent works apply AI techniques to locate glomeruli, identify lesions (global and segmental sclerosis, crescents), and quantify glomerular cells on PAS-stained slides with accuracy values around 91% [11]. In Liu's work [12], an AI-driven framework is developed for quantifying glomerular volume and nodular mesangial sclerosis in diabetic kidney disease, achieving 0.899 accuracy for glomerulus classification and an AUROC score of 0.917 for disease prediction. In the work of Zhuang *et al*. [13] the accuracy issues of ML models for diagnosing and predicting the prognosis of IgA nephropathy is addressed by conducting a systematic review of publications up to February 2024, from Embase, Pubmed, Cochrane Library, and Web of Science. Analyzing 47 studies with 51,935 patients, the pooled C-index was found to be 0.902 for 27 diagnostic models and 0.838 for 144 prognostic models. New ML models perform comparably to traditional ones. Future models should aim to improve sensitivity, especially in moderate-risk prediction, and focus on clinical application.

As far as the authors know, no further studies have been conducted to classify glomeruli from renal biopsies into the 12 classes considered here.

II. Materials

A total of 786 renal biopsies with glomerulonephritis (GN) from seven participating institutions were analyzed. Of these, 461 biopsies came from different patients. Each renal biopsy was diagnosed by expert nephropathologists with one of the 12 types of glomerulonephritis listed in I. The biopsies for all diagnostic classes were randomly selected without consideration of any histopathological descriptors or histological subtypes. Expert nephropathologists assigned and retrospectively confirmed the diagnostic classes prior to inclusion, using a standardized diagnostic work-up that included paraffin histology, immunostaining for immunoglobulin heavy and light chains, and transmission electron microscopy. This thorough process, applied to all cases, ensured a robust and reliable ground truth for the entire dataset.

TABLE I: GN pathologies present in the dataset

Name	Label
Anti-glomerular Basement Membrane antibody GN	AMBGN
Anti-Neutrophil Cytoplasmic Antibody-associated GN	ANCA
C3-GlomeruloNephritis	C3-GN
Cryoglobulinemic GN	CryoGN
Dense Deposit Disease	DDD
Fibrillary GN	Fibrilar
Infection-associated GN	IAGN
Immunoglobulin A-associated GN	IgAGN
Membranous nephopathy	Membranous
idiopathic MembranoProliferative GN	MPGN
Proliferative GN Monoclonal Immunoglobulin Deposits	PGNMID
Systemic Lupus Erythematodes-associated GN class IV	SLEGN-IV

The biopsies used in this work come from three different datasets: the 1st, consisting of 725 renal biopsies from 400 different patients, (dataset A); the 2nd comprising 50 renal biopsies from different patients, where only the sclerotic glomeruli were used (dataset B); and the 3rd (dataset C), used as a hold-out for testing, which consists of 61 renal biopsies from different patients. Datasets A and C contain the 12 types of GN.

The renal biopsies were digitized (WSIs) using various scanners, including the Hamamatsu NanoZoomer, the Axioscan (Carl Zeiss), the Pannoramic Midi Slide Scanner (3DHISTECH), and the Aperio CS2 (Leica). All of the WSIs use Periodic Acid-Schiff (PAS) staining. The glomeruli in the images were automatically detected using the segmentation model described in *[Ref. hidden for double-blind review]*, implemented by the co-authors, and reviewed by expert pathologists. For each glomerulus, a rectangular crop was extracted, leaving a margin of 300 pixels on each side. In total, 12,969 PAS-stained glomerular crops were obtained. Dataset A, used for training and validation, comprises 10,128 crops. Dataset B, used for training and validation of sclerotic glomeruli, comprises 1,170 crops. Dataset C, used for testing, comprises 1,671 crops. Each crop in datasets A and C retains its original

diagnostic label corresponding to one of the 12 classes. Dataset A also contains sclerotic glomeruli (see Figure 1 for the class distribution of glomeruli crops across the entire datasets).

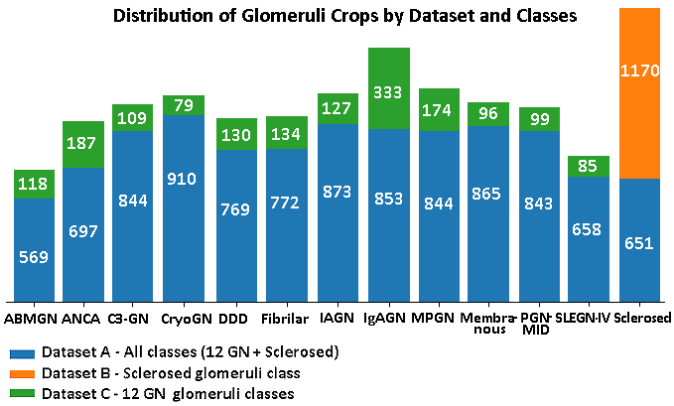


Fig. 1: Distribution of glomeruli crops per class and datasets

Figure 2 shows examples of glomeruli for each of the classes. It can be observed the variability in glomeruli, such as differences in texture and colour, even among glomeruli of the same class. The variety of colors present in the images even when using the same stain, as well as the subtle differences at first glance between the different classes is also observed.

III. Methods

Several advanced classification networks were tested, these include: FocalNet, ConvNeXt, Swin-Transformers, and EfficientNet. These networks aim to optimize neural network architectures to improve efficiency and performance by using innovations in attention mechanisms and convolutions.

Due to the variability and the challenge of visually determining if individual glomeruli belong to the same slide, our objective was to incorporate more variability and broaden the dataset. Consequently, a partition was established at the glomerulus level, allocating 75% of the glomeruli crops for training, 15% for validation, and 10% for testing. The initial results from all networks were highly promising, with an average sensitivity of 0.98, specificity of 0.99 and a balanced accuracy of 0.99. However, when the dataset was partitioned at the slide level, to prevent information leakage across biopsies, the performance decreased drastically. It resulted in an average sensitivity of 0.2, an average specificity of 0.93, and a balanced accuracy of 0.63 with a standard deviation of 0.1 across all methods and classes.

Several tests were done with colour standardization and color transfer for data augmentation but the results only increased around 2%. Therefore, it was concluded that texture and the contextual relationships of the tissue plays a fundamental role in the diagnosis of these glomeruli crops.

Furthermore, activation maps were visualized using various techniques such as Grad-CAM++, ScoreCAM, FullGrad, LayerCAM, and XGradCAM to highlight areas of the images that significantly contribute to predicting specific classes. From the conducted tests, all methods show similar results as depicted

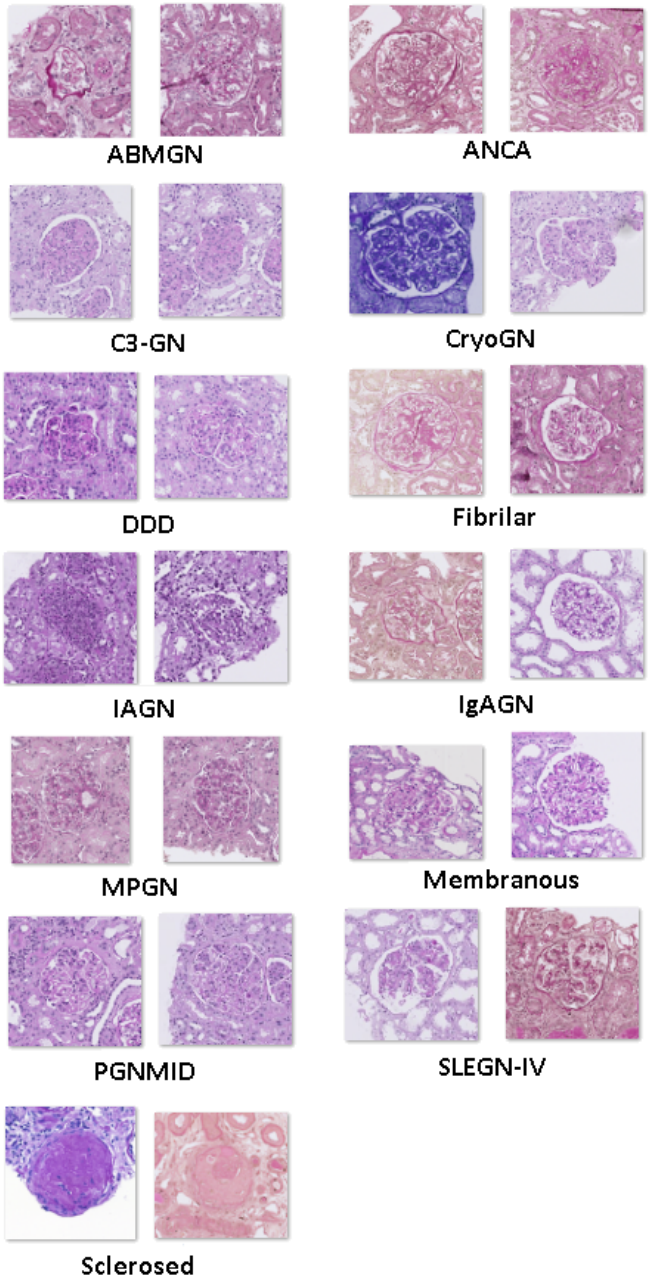


Fig. 2: Example images of each class of glomeruli. Two glomeruli from each class are shown

in Figure 3.This figure shows two glomeruli, one of type DDD and one of type IAGN, along with activation maps from the six methods. No specific area was identified by the network as relevant or contributing to the prediction. Figure 4 shows activation maps using Grad-CAM++ for different glomeruli types: panel a) shows a Cryoglobulinemic-type glomerulus misclassified as ANCA, while panel b) shows a correctly classified DDD-type glomerulus.

By examining Figure 2, it can be observed that sclerosed glomeruli have unique textures distinct from those with GN. Therefore, a sequential process was designed and imple-

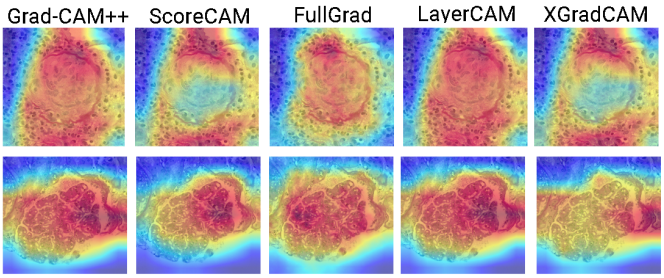
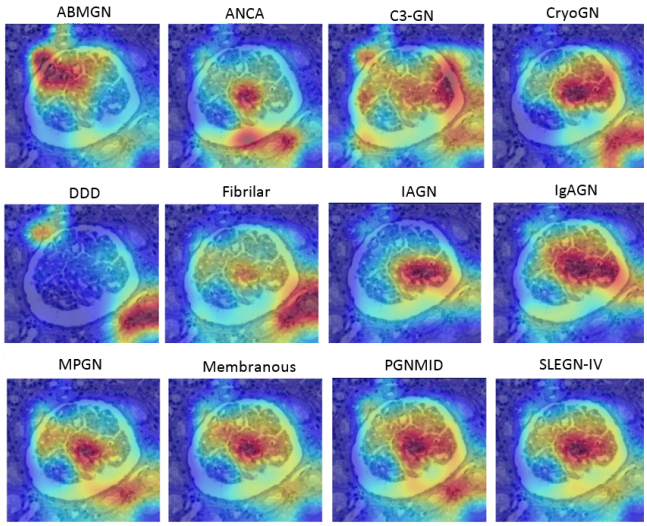
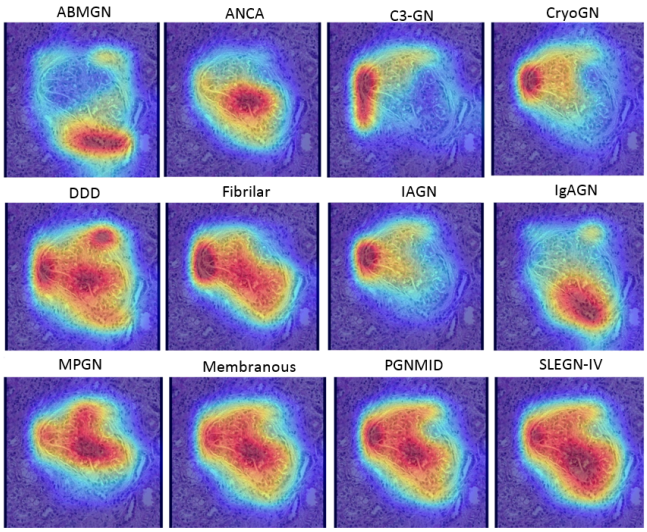


Fig. 3: Activation maps of glomeruli using Grad-CAM++, ScoreCAM, FullGrad, LayerCAM, and XGradCAM. 1st row: DDD-type glomerulus; 2nd row: IAGN-type glomerulus



(a) Cryoglobulinemic-type glomerulus classified as ANCA



(b) DDD-type glomerulus classified correctly as DDD

Fig. 4: Activation maps using Grad-CAM++. Panel a) a Cryoglobulinemic-type glomerulus misclassified as ANCA and b) shows a correctly classified DDD-type glomerulus

mented: first categorizing glomeruli as sclerosed or non-sclerosed, then further dividing them into 12 distinct classes. The findings demonstrated a significant improvement, attaining a balanced accuracy of 0.97 in the initial classification of sclerosed and non-sclerosed glomeruli.

A. Datasets preparation

The preparation of datasets was as follows:

- Set of 12 Glomerulonephritis Pathologies

Sclerotic glomeruli were excluded, leaving a dataset of only non-sclerotic glomeruli crops from dataset A. This results in 12 classes for each GN pathology, without including glomerulosclerosis. The dataset partitions were created using a 5-fold cross-validation technique. However, the number of glomeruli in each set (training, validation, and test) varies across folds. This variation is due to a constraint applied when forming each fold. Specifically, glomeruli from the same WSI must be placed in the same set. This prevents models from learning to extract features shared by glomeruli from the same WSI. It ensures that models classify glomeruli based on pathology-related features, rather than features of the entire image. In addition, biopsies from the same patient are placed in the same set to prevent them from appearing in both the training and test sets.

The final distributions for the folds used in classifying the 12 GN pathologies were based on glomeruli crops from dataset A (see Figure 1), with allocations ranging from [71%, 14%, 15%] to [69%, 15%, 16%] for training, validation, and testing, respectively, across classes and folds.

- Sclerosed and Non-Sclerosed Set

The sclerosed and non-sclerosed glomeruli dataset was created based on the 12-glomerulonephritis dataset. The "Non-Sclerosed" class is a subset of glomeruli from the 12 pathologies, ensuring diverse representation. The partitions from each fold of the original dataset were maintained, adhering to WSI restrictions for consistency and allowing for independent model training. Initially, the test set showed an imbalance by including all glomeruli from the 12 glomerulonephritis pathologies. However, adding sclerosed glomeruli from dataset B balanced the classes. Since sclerosed and non-sclerosed glomeruli can share the same WSIs, WSI restrictions apply to both. This approach resulted in balanced datasets with a representative 'Non-Sclerosed' class. The distribution used for these two classed is shown in Table II for the best fold.

TABLE II: Distribution in training, validation, and test sets for the first fold of the sclerosed and non-sclerosed dataset

Label	Train	Validation	Test	Total
Non-Sclerosed	378	156	1670	2204
Sclerosed	376	156	1289	1821

B. Base Models for Glomeruli Classification

Two base models for glomeruli classification were developed: one for classifying them into sclerosed and non-sclerosed, and another for classifying the non-sclerosed ones into 12 glomerulonephritis pathologies. The best-performing

networks were EfficientNet [14], Swin Transformer [15], and ConvNeXt [16]. Below are the general characteristics and their training parameters. These parameters are the same for training the models for classification into sclerosed and non-sclerosed, as well as into the 12 pathologies. The only difference in the training is the dataset used.

- ConvNeXt

ConvNeXt is a convolutional network that incorporates improvements from transformers by adapting them into a purely convolutional solution. For example, to reduce dimensionality, it uses a convolution with a larger kernel size and a stride of 0, mimicking patch division; it also employs layer normalization instead of batch normalization. The architecture utilized in this project is ConvNeXt Base or ConvNeXt-B. The model is pretrained on ImageNet-1K and subsequently retrained using the glomeruli datasets with specific parameters: it employs the AdamW optimizer with an initial learning rate set to 0.001 and incorporates a weight decay of 0.05. Additional settings include an epsilon value of 10^{-8}, β_1 of 0.9, and β_2 of 0.999. The learning rate policy follows a cosine annealing approach, starting with a warmup phase of 20 epochs using a linear warmup type and a rate of 0.001. The minimum learning rate is established at 0.01, with a maximum training duration of 300 epochs and a batch size of 64. Input images for the model are resized to 224 × 224 pixels and normalized using ImageNet's mean and standard deviation. Training incorporates two data augmentation techniques: RandAugment [17] and RandomErasing [18].

- Swin-Transformers

Swin-Transformers use the sliding window mechanism which allows for better capture of local and global features, as well as improving the efficiency and scalability of transformers. Unlike the basic Vision Transformer (ViT), where the outputs of each layer maintain the same dimension, in Swin the patches become progressively smaller throughout the network (hierarchical structure). The version used in this work is Swin-T (Tiny), with the following parameters: Optimizer AdamW, Initial learning rate 0.001, Weight decay 0.05, ϵ 10^{-8}, β_1 0.9, β_2 0.999, Learning rate policy: Cyclic with cosine annealing, Warm-up epochs 20, Warm-up type: Linear, Warm-up rate 0.001, Minimum learning rate 0.01, Maximum number of epochs 300, and Batch size 64. Additionally, the input images for the model are resized to 224 × 224 pixels and normalized using the mean and standard deviation of ImageNet. For training, two data augmentation policies are also used: RandAugment [17] and RandomErasing [18].

- EfficientNet

EfficientNet focuses on optimizing the size, depth, and number of channels of convolutional networks, providing an optimal balance between performance and computational efficiency. It uses a compound coefficient to scale the size of the network according to the size of the input. The model used in this work is EfficientNet-B, pretrained on ImageNet-1K and retrained using the glomeruli datasets with the following parameters: Optimizer SGD, Initial learning rate 0.1, Momentum 0.9, Weight decay 0.0001, Learning rate policy: Step decay; Steps (epochs) at 30, 60, and 90; Maximum number of epochs 300, and Batch size 4. Additionally, the input images for the model are resized to 224 × 224 pixels and normalized using the mean and standard deviation of ImageNet.

IV. Results

The classification of glomeruli into 12 classes showed improved results in the second stage after classifying and removing the sclerotic glomeruli. Swin-Transformers showed the best performance in the first stage while classifying sclerotic glomeruli versus non-sclerotic, and ConvNeXt showed the best performance in the second one.

A. Base model for classification of glomeruli into sclerosed and non-sclerosed

The performance metrics for the base models classifying glomeruli into sclerosed and non-sclerosed are shown in Table III. These metrics correspond to the averages of the *test* sets from each fold, in order to compare the overall performance of the models, regardless of the partition configuration.

Using *balanced accuracy* for comparison, Swin Transformer performs best with 0.9718 ± 0.0144, compared to 0.8604 ± 0.0360 for ConvNeXt and 0.7780 ± 0.1636 for EfficientNet. Similar results are observed when using *precision* as the reference metric. Regarding the ROC (Receiver Operating Characteristic) and their AUC (Area Under the Curve), Swin Transformer is again the best model, with an AUC of 1.00, followed by ConvNeXt with 0.99 and EfficientNet with 0.79. Figure 5 shows the two best models.

B. Base model for classifying non-sclerosed glomeruli into 12 renal glomerulonephritis pathologies

Figure 6 shows a visual comparison of the base models for classifying glomeruli into the 12 proposed pathologies. These metrics are the average of the 5 *folds*. Although only the model generated in one *fold* is finally used for testing with the hold-out dataset, this average is used to compare different methods, aiming for a comparison independent of the partitions taken. Taking *balanced accuracy* as the reference metric, the best-performing model is ConvNeXt, with an average for the 12 classes of 0.6418 ± 0.0837, compared to 0.6334 ± 0.1080 for Swin Transformer and 0.6009 ± 0.0772 for EfficientNet.

It was also observed that the model correctly identified an average balanced accuracy of 0.795 for all positive classes within the top 3 most likely. Thus, the results were analyzed considering that the model was correct if the true class was among the top-2 and top-3 predicted classes. The complete results for this model are illustrated in Table IV.

If we consider ROC curves and AUC, the difference between the models is higher. Once again, the top model is ConvNeXt, with an average AUC for the 12 classes of 0.7908 ± 0.0831, followed by Swin Transformer with 0.7625 ± 0.0710, and EfficientNet with 0.7317 ± 0.0945. Figure 7 displays the ROC curves and AUC of the top two methods.

TABLE III: Results on the *test* set for the three baseline models for classifying glomeruli as sclerosed and non-sclerosed. The best result for the *sclerosed* class for each metric is highlighted in bold.

Model	Class	Precision	Sensitivity	Specificity	F1-Score	Balanced Accuracy
ConvNeXt	Non-sclerosed	0.8300 ± 0.0372	0.9542 ± 0.0439	0.7666 ± 0.0706	0.8870 ± 0.0281	0.8604 ± 0.0360
	Sclerosed	0.9386 ± 0.0531	0.7666 ± 0.0706	0.9542 ± 0.0439	0.8415 ± 0.0443	0.8604 ± 0.0360
Swin Transf.	Non-sclerosed	0.9702 ± 0.0243	0.9790 ± 0.0055	0.9647 ± 0.0291	0.9745 ± 0.0122	0.9718 ± 0.0144
	Sclerosed	$\mathbf{0.9752 \pm 0.0055}$	$\mathbf{0.9647 \pm 0.0291}$	$\mathbf{0.9790 \pm 0.0055}$	$\mathbf{0.9697 \pm 0.0148}$	$\mathbf{0.9718 \pm 0.0144}$
EfficientNet	Non-sclerosed	0.7725 ± 0.1594	0.8881 ± 0.1293	0.6679 ± 0.2426	0.8219 ± 0.1308	0.7780 ± 0.1636
	Sclerosed	0.8376 ± 0.2122	0.6679 ± 0.2426	0.8881 ± 0.1293	0.7327 ± 0.2089	0.7780 ± 0.1636

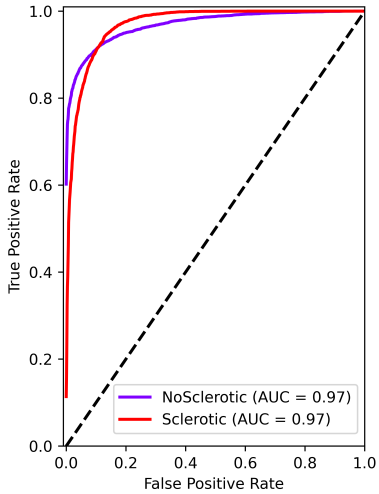
TABLE IV: Complete results on the test set of ConvNeXt for the classification of glomeruli into the 12 pathologies.The worst result is highlighted in red and the best in green.

Class	Top-*k*	Precision	Sensibility	Specificity	F1-Score	Balanced Accuracy
ABMGN	Top-1	0.3293 ± 0.1296	0.1756 ± 0.0444	0.965 ± 0.0204	0.2236 ± 0.0611	0.5703 ± 0.0274
	Top-2	0.5713 ± 0.1634	0.3754 ± 0.1224	0.9737 ± 0.0186	0.4409 ± 0.1262	0.6746 ± 0.0613
	Top-3	0.6971 ± 0.1404	0.5163 ± 0.1797	0.9792 ± 0.0167	0.5731 ± 0.1547	0.7477 ± 0.0878
ANCA	Top-1	0.2117 ± 0.1312	0.2242 ± 0.1364	0.9372 ± 0.0136	0.2171 ± 0.133	0.5807 ± 0.0733
	Top-2	0.3941 ± 0.1428	0.3873 ± 0.1574	0.9562 ± 0.0099	0.3899 ± 0.1495	0.6718 ± 0.0827
	Top-3	0.5735 ± 0.1353	0.5499 ± 0.1907	0.9706 ± 0.0075	0.5589 ± 0.1631	0.7603 ± 0.098
C3-GN	Top-1	0.1497 ± 0.1343	0.1246 ± 0.1106	0.9541 ± 0.023	0.1354 ± 0.1203	0.5393 ± 0.0507
	Top-2	0.3135 ± 0.2444	0.2603 ± 0.2214	0.9674 ± 0.0175	0.2822 ± 0.2287	0.6139 ± 0.1074
	Top-3	0.506 ± 0.1702	0.5105 ± 0.1819	0.9751 ± 0.0138	0.4072 ± 0.2673	0.6778 ± 0.1371
CryoGN	Top-1	0.3745 ± 0.2074	0.3215 ± 0.2428	0.9429 ± 0.0349	0.3142 ± 0.1629	0.6322 ± 0.1112
	Top-2	0.4854 ± 0.1937	0.4286 ± 0.2643	0.9526 ± 0.0305	0.424 ± 0.172	0.6906 ± 0.1223
	Top-3	0.6258 ± 0.1579	0.5595 ± 0.2504	0.9632 ± 0.0264	0.5609 ± 0.1471	0.7613 ± 0.1163
DDD	Top-1	0.5717 ± 0.2014	0.3561 ± 0.2612	0.9778 ± 0.0139	0.4095 ± 0.2523	0.6669 ± 0.1288
	Top-2	0.7347 ± 0.1682	0.4951 ± 0.2882	0.9826 ± 0.0139	0.5566 ± 0.2406	0.7388 ± 0.1436
	Top-3	0.8103 ± 0.1369	0.5808 ± 0.2513	0.986 ± 0.0118	0.6515 ± 0.1896	0.7834 ± 0.1255
Fibrilar	Top-1	0.283 ± 0.0837	0.309 ± 0.1714	0.9383 ± 0.0208	0.2895 ± 0.1231	0.6236 ± 0.0797
	Top-2	0.4588 ± 0.0629	0.4383 ± 0.1506	0.958 ± 0.0101	0.4433 ± 0.1061	0.6982 ± 0.0719
	Top-3	0.6007 ± 0.0354	0.5771 ± 0.1034	0.968 ± 0.007	0.5853 ± 0.0602	0.7726 ± 0.0487
IAGN	Top-1	0.3622 ± 0.2474	0.3069 ± 0.3608	0.9589 ± 0.0169	0.3137 ± 0.307	0.6329 ± 0.1762
	Top-2	0.5588 ± 0.2495	0.4365 ± 0.3709	0.9678 ± 0.0188	0.457 ± 0.3112	0.7022 ± 0.1825
	Top-3	0.6586 ± 0.2653	0.5285 ± 0.3627	0.9735 ± 0.0169	0.557 ± 0.3078	0.751 ± 0.1808
IgAGN	Top-1	0.4383 ± 0.0927	0.6349 ± 0.0676	0.9201 ± 0.0265	0.5158 ± 0.08	0.7775 ± 0.0397
	Top-2	0.5776 ± 0.0896	0.7979 ± 0.09	0.9432 ± 0.0203	0.6675 ± 0.0817	0.8706 ± 0.0481
	Top-3	0.6798 ± 0.0779	0.8795 ± 0.0673	0.96 ± 0.0145	0.7653 ± 0.0674	0.9197 ± 0.0367
MPGN	Top-1	0.1742 ± 0.0848	0.2153 ± 0.073	0.8928 ± 0.0384	0.1876 ± 0.0729	0.5541 ± 0.0446
	Top-2	0.3474 ± 0.1502	0.4255 ± 0.1294	0.9187 ± 0.0297	0.3788 ± 0.1378	0.6721 ± 0.0751
	Top-3	0.4801 ± 0.1589	0.5912 ± 0.1472	0.9369 ± 0.0252	0.5281 ± 0.1542	0.7641 ± 0.0834
Membranous	Top-1	0.6341 ± 0.1504	0.512 ± 0.1902	0.9683 ± 0.0201	0.5401 ± 0.1166	0.7402 ± 0.0899
	Top-2	0.7718 ± 0.1078	0.7412 ± 0.1613	0.9772 ± 0.0146	0.74 ± 0.0758	0.8592 ± 0.076
	Top-3	0.8326 ± 0.0804	0.8413 ± 0.0982	0.9834 ± 0.0095	0.8316 ± 0.0479	0.9123 ± 0.0469
PGNMID	Top-1	0.3614 ± 0.0482	0.5227 ± 0.1037	0.9044 ± 0.0239	0.4219 ± 0.0449	0.7136 ± 0.0452
	Top-2	0.5016 ± 0.0463	0.7389 ± 0.129	0.9244 ± 0.0176	0.5921 ± 0.0459	0.8317 ± 0.059
	Top-3	0.5998 ± 0.0458	0.8418 ± 0.104	0.942 ± 0.0138	0.6959 ± 0.027	0.8919 ± 0.0466
SLEGN-IV	Top-1	0.3324 ± 0.1913	0.4073 ± 0.2555	0.9325 ± 0.024	0.364 ± 0.2163	0.6699 ± 0.1374
	Top-2	0.437 ± 0.2287	0.5048 ± 0.2641	0.9452 ± 0.0268	0.4655 ± 0.2408	0.725 ± 0.1435
	Top-3	0.5334 ± 0.2358	0.6035 ± 0.2465	0.9568 ± 0.0233	0.5642 ± 0.2389	0.7801 ± 0.1337

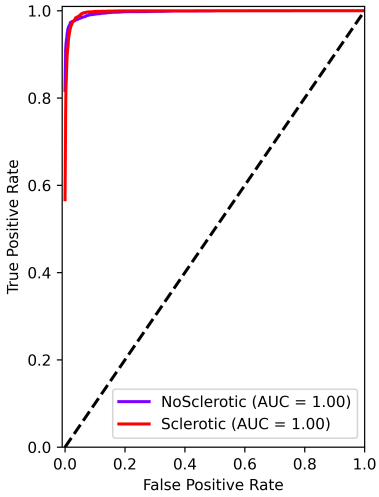
Figure 7 shows superior results compared to Table IV because they assess different aspects with different methods. The ROC curve measures the model's ability to distinguish between positive and negative samples of a class at various classification thresholds. This indicates good ranking ability but not necessarily optimal classification performance at a fixed threshold.

For all three models, performance varies widely across classes, especially when considering other metrics like precision. Taking ConvNeXt's precision as an example across each class, we observe results ranging from 0.63 for membranous nephropathy to 0.15 for C3-GN in top1. The low result for the latter can be justified because diagnosing this pathology typically requires other techniques such as immunofluorescence [19], [20]. However, results for top-3 improve for classification of all GN pathologies. The values increase for all metrics by between 0.21 and 0.35, with the standard deviation remaining consistent. For example, balanced accuracy in top-3 ranges from 0.68 to 0.92. The best-performing classes are Membranous, IgAGN, and PGNMID, where a balance between sensitivity and specificity is maintained, with sensitivity values around 0.84 and 0.87 and specificity ranging from 0.94 to 0.98.

After selecting ConvNeXt as the best model out of the three, the model generated from the best fold is evaluated with the external test set or hold-out. This set is entirely independent from the others and consists of all images from database C. The results are shown in Figure 8. The overall performance is slightly lower, with a balanced accuracy of



(a) ConvNeXt



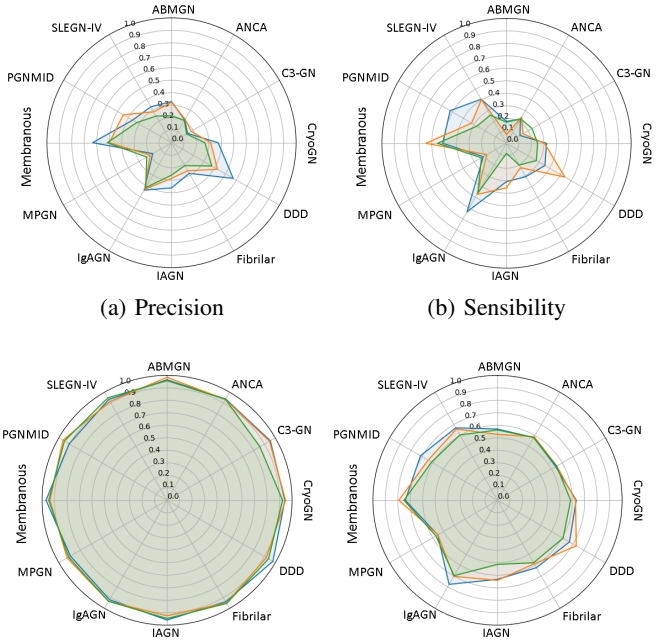
(b) Swin Transformer

Fig. 5: ROC curves and average AUCs on the *test* set across 5 *folds* for the initial stage of glomerulus classification process, highlighting ConvNeXt and Swin Transformer as top performers in classifying Sclerosed and Non-sclerosed glomeruli.

0.5774 ± 0.0739, compared to 0.6418 ± 0.0837 previously. The ROC curves and their respective AUCs are shown in Figure 9. The average AUC values on the hold-out dataset are slightly lower, 0.69 compared to the previously obtained 0.79. However, classification improves slightly in 5 classes, while the class that worsens the most is PGNMID. This variation in performance with this set is expected as it is completely independent from what was used before.

C. Discussion

The results highlight the challenge of identifying key features among the 12 GN pathologies and the need for additional information such as immunofluorescence tests, for accurate diagnosis.. Manual classification with the hold-out dataset by four pathologists was performed to assess inter-pathologist variation and compare with ConvNeXt. The difficulty in classi-



(a) Precision (b) Sensibility

(c) Specificity (d) Balanced Acc.

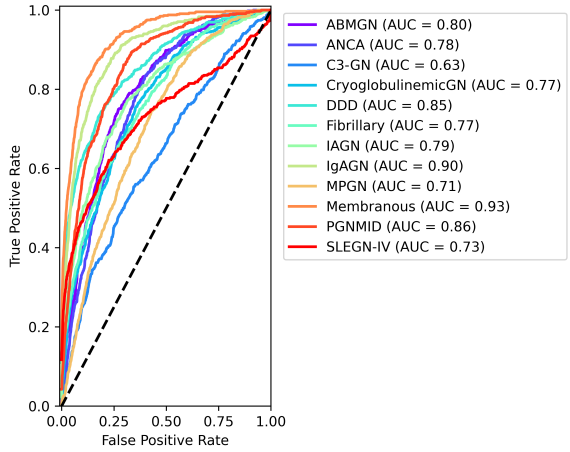
Fig. 6: Average results on the *test* set of the 5 *folds* for the three base models classifying glomeruli into the 12 GN pathologies. In blue, ConvNeXt; in orange, Swin Transformer; and in green, EfficientNet.

fying some pathologies was consistent, with higher agreement in top-3 results. Pathologists noted that some glomeruli crops could be classified into multiple classes visually. Therefore, certain pathologies require the use of additional sample processing techniques, as both methods and experts are unable to distinguish between them using only PAS-stained images.

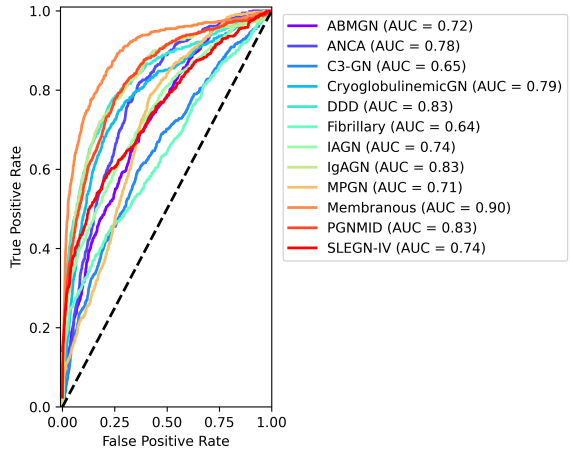
In all methods, performance varies greatly depending on the pathology. Some, such as C3-GN, consistently yield the worst results across all models, unlike Membranous, which always achieved the best results. Pathologies with poorer performance across all models often require additional tests for diagnosis or belong to subtypes within the same group of pathologies, such as MPGN, DDD, and C3-GN [19], citebib20. This reaffirms not only the need to include other types of medical tests in training but also the necessity for a more differentiable taxonomy among pathologies at a medical level.

V. Conclusions

Complete microscopic images were obtained and properly processed for use in deep learning models. Expert annotations were made on complete images, and glomerular crops were extracted and appropriately labeled. Different partitions were prepared using a custom variation of the 5-fold cross-validation technique, which were subsequently used to compare models by evaluating their performances independently of the partition configuration. Specific constraints were applied within the realm of histological images during dataset creation at the patient and WSI levels to prevent data leakage



(a) ConvNeXt



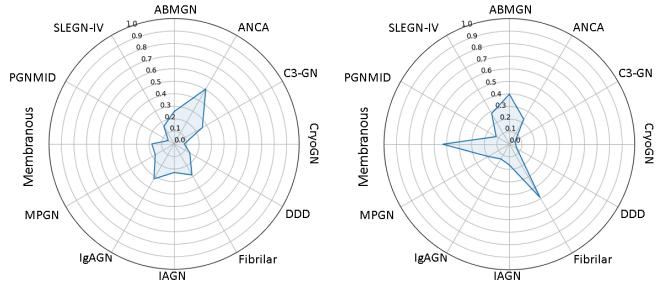
(b) Swin Transformer

Fig. 7: ROC curves and average AUCs on the *test* set across 5 *folds* for the second stage of glomerulus classification process, highlighting ConvNeXt and Swin Transformer as top performers in the 12 glomerulonephritis classification.

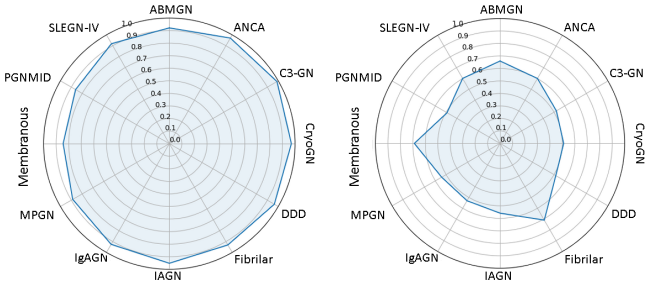
between sets. Additionally, the test sets were enriched with a completely different database, which also helped balance the dataset and test the sequential model implemented. Thus, a robust dataset was successfully created for the training and evaluation of deep learning models.

A sequential model was implemented to first classify glomeruli as sclerosed or non-sclerosed, then further classify non-sclerosed glomeruli into 12 GN pathologies. The models were evaluated using average metrics across 5 folds, including precision, sensitivity, specificity, F1-score, balanced accuracy, and AUC from ROC analysis.

Swin Transformer was chosen as the best model for the first stage due to its superior overall performance, with a *balanced accuracy* of 0.9718 ± 0.0144 and an AUC of 1.00. ConvNeXt was selected as the best model for the second stage to classify the 12 GN types, with an average balanced accuracy of 0.6418 for top-1 and 0.795 for top-3 across the test sets in the 5 folds, along with an AUC of 0.75. In the external test set, this best model achieved an average balanced accuracy of 0.5774



(a) Precision (b) Sensibility



(c) Specificity (d) Balanced Acc.

Fig. 8: Results on the external test set of the best base model for glomerulus classification into 12 GN pathologies. This model is the best fold of ConvNeXt.

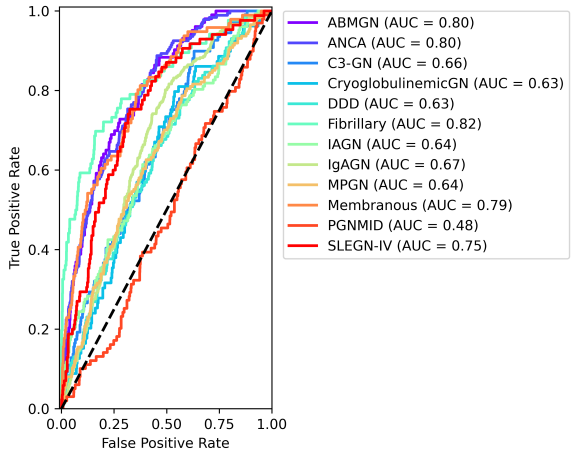


Fig. 9: ROC curves with their respective AUCs on the external test set for the best base method of glomerulus classification across the 12 pathologies. The model is the top fold of ConvNeXt.

and an AUC of 0.6925. The most accurately detected GN types were Membranous and IgAGN. The Membranous class achieved the best values in all metrics from top-1 to top-3 classification, with an average sensitivity of 0.71, specificity of 0.97, balanced accuracy of 0.84 and an average AUC of 0.86 across all datasets.

It can be asserted that the model reliably diagnoses glomerulosclerosis and further distinguishes between 12 types of glomerulonephritis with consistent performance, particularly

identifying the Membranous class.

A QuPath extension has been developed with these models, aiming to ease its integration into clinical practice. The tool classifies glomeruli as sclerosed or non-sclerosed, and predict the most likely GN class from top-1 to top-3 for non-sclerosed cases. This proof-of-concept study establishes a baseline for this difficult classification task, which usually requires immunostains, electron microscopy and even clinical data.

ACKNOWLEDGMENT

We would like to acknowledge the collaboration of Sándor Turkevi-Nagy from the Department of Pathology, Albert Szent-Györgyi Health Center, University of Szeged, Hungary; and Surya V. Seshan from Weill Cornell Medicine, New York, USA, for their assistance with validation and data provision.

REFERENCES

[1] Cazzaniga, G., Rossi, M., Eccher, A., Girolami, I., ... & Pagni, F. (2024). Time for a full digital approach in nephropathology: a systematic review of current artificial intelligence applications and future directions. Journal of Nephrology, 37(1), 65-76.

[2] Mariani, L. H., Pendergraft III, W. F., & Kretzler, M. (2016). Defining glomerular disease in mechanistic terms: implementing an integrative biology approach in nephrology. Clinical Journal of the American Society of Nephrology, 11(11), 2054-2060.

[3] Ronconi, E., Sagrinati, C., Angelotti, M. L., Lazzeri, ... & Romagnani, P. (2009). Regeneration of glomerular podocytes by human renal progenitors. Journal of the American Society of Nephrology, 20(2), 322-332.

[4] Bueno, G., Fernandez-Carrobles, M. M., Gonzalez-Lopez, L., & Deniz, O. (2020). Glomerulosclerosis identification in whole slide images using semantic segmentation. Computer methods and programs in biomedicine, 184, 105273.

[5] Altini, N., Cascarano, G. D., Brunetti, A., De Feudis, I., ... & Bevilacqua, V. (2020). A deep learning instance segmentation approach for global glomerulosclerosis assessment in donor kidney biopsies. Electronics, 9(11), 1768.

[6] Salvi, M., Mogetta, A., Gambella, A., Molinaro, ... & Molinari, F. (2021). Automated assessment of glomerulosclerosis and tubular atrophy using deep learning. Computerized Medical Imaging and Graphics, 90, 101930.

[7] Altini, N., Rossini, M., Turkevi-Nagy, S., Pesce, F., ... & Becker, J. U. (2023). Performance and limitations of a supervised deep learning approach for the histopathological Oxford Classification of glomeruli with IgA nephropathy. Computer Methods and Programs in Biomedicine, 242, 107814.

[8] Cattran, D.C., Coppo, R., Cook, H.T., Feehally, J., ... & Zhang, H. (2009). The Oxford classification of IgA nephropathy: rationale, clinicopathological correlations, and classification. Kidney international, 76(5), 534-545.

[9] Roberts, I.S., Cook, H.T., Troyanov, S., Alpers, C.E., ... & Zhang, H. (2009). The Oxford classification of IgA nephropathy: pathology definitions, correlations, and reproducibility. Kidney International, 76(5):546–556.

[10] Trimarchi, H., Barratt, J., Cattran, D.C., Cook, H.T., ... & Zeng, C. (2017). Oxford classification of IgA nephropathy 2016: an update from the IgA nephropathy classification working group. Kidney international, 91(5), 1014-1021.

[11] Zeng, C., Nan, Y., Xu, F., Lei, Q., ... & Liu, Z. (2020). Identification of glomerular lesions and intrinsic glomerular cell types in kidney diseases via deep learning. The Journal of pathology, 252(1), 53-64.

[12] Liu, X., Wu, Y., Chen, Y., Hui, D., ...& Zheng, W. (2023). Diagnosis of diabetic kidney disease in whole slide images via AI-driven quantification of pathological indicators. Computers in Biology and Medicine, 166, 107470.

[13] Zhuang, K., Wang, W., Xu, C., Guo, X., ... & Cai, G. (2024). Machine learning-based diagnosis and prognosis of IgAN: A systematic review and meta-analysis. Heliyon, 10(12).

[14] Tan, M., & Le, Q. (2019, May). Efficientnet: Rethinking model scaling for convolutional neural networks. In International conference on machine learning (pp. 6105-6114).

[15] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., ... & Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 10012-10022).

[16] Liu, Z., Mao, H., Wu, C. Y., Feichtenhofer, C., Darrell, T., & Xie, S. (2022). A convnet for the 2020s. In Proc. of the IEEE/CVF conference on CVPR (pp. 11976-11986).

[17] Cubuk, E.D., Zoph, B., Shlens, J., & Le, Q.V. (2020). Randaugment: Practical automated data augmentation with a reduced search space. In Proc. of the IEEE/CVF Conference on CVPR (pp. 702-703).

[18] Zhong, Z., Zheng, L., Kang, G., Li, S., & Yang, Y. (2020, April). Random erasing data augmentation. In Proc. of the AAAI Conference on AI (Vol. 34, No. 07, pp. 13001-13008).

[19] Sethi, S., Nester, C. M., & Smith, R. J. (2012). Membranoproliferative glomerulonephritis and C3 glomerulopathy: resolving the confusion. Kidney international, 81(5), 434-441.

[20] Buch, A. C., Sood, S. K., Bamanikar, S. A., Chandanwale, S. S., Kumar, H., & Swapnil, K. (2015). Role of direct immunofluorescence in the diagnosis of glomerulonephritis. Medical Journal of Dr. DY Patil University, 8(4), 452-457.