Computers in Biology and Medicine 196 (2025) 110719

Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

Unsupervised learning for labeling global glomerulosclerosis

Hrafn Weishaupt [a,*], Justinas Besusparis [a], Cleo-Aron Weis [b], Stefan Porubsky [c], Arvydas Laurinavičius [d,e], Sabine Leh [a,f]

[a] Department of Pathology, Haukeland University Hospital, Bergen, 5021, Norway
[b] Institute of Pathology, University Medical Centre Mannheim, University of Heidelberg, Mannheim, 68167, Germany
[c] Institute of Pathology, University Medical Center of the Johannes Gutenberg University Mainz, Mainz, 55131, Germany
[d] Translational Health Research Institute, Faculty of Medicine, Vilnius University, Vilnius, LT-03101, Lithuania
[e] National Center of Pathology, Affiliate of Vilnius University Hospital Santaros Clinics, Vilnius, LT-08406, Lithuania
[f] Department of Clinical Medicine, University of Bergen, Bergen, 5021, Norway

ARTICLE INFO

Keywords:
Glomeruli
Glomerulosclerosis
Deep learning
Clustering
Unsupervised learning
Nephropathology

ABSTRACT

Background: Labeling images for supervised learning in nephropathology is highly time-consuming and dependent on domain-expertise. Unsupervised strategies have been suggested for mitigating this bottleneck. For instance, previous work suggested that clustering/grouping of glomeruli based on image features might enable a more semi-automated labeling of morphological classes or even a completely unsupervised training. However, even for the most basic separation between globally sclerosed and non-globally sclerosed glomeruli, the performance of clustering approaches has not yet been fully elucidated. The current study sought to fill this gap by extensively evaluating the accuracy and limitations of capturing these two classes via clustering.
Methods: Clustering was investigated across 10 labeled datasets with diverse compositions and histological stains and across the feature embeddings produced by 34 different pre-trained CNN models.
Results: As demonstrated by the study, clustering of globally and non-globally sclerosed glomeruli is generally highly feasible, yielding accuracies of over 95% in most datasets.
Conclusions: While further work will be required to expand these experiments towards the clustering of additional glomerular lesion categories, the study clearly demonstrates that clustering might serve as a highly accurate means of pre-labeling glomeruli. Importantly, these findings strongly support clustering as a solid basis for downstream interactive labeling approaches or unsupervised learning approaches. Together, these results might greatly improve the possibilities and lookout for the establishment of clinically applicable glomerular classification models in the community. Further improvements in this area might be achieved by exploring more domain-specific feature extractors through contrastive learning or established foundation models.

1. Introduction

With the advent of digital nephropathology, many diagnostic tasks in the evaluation of a kidney biopsy can now be addressed using artificial intelligence (AI), ushering in a new era of computer-assisted diagnosis [1]. For example, one of the most central tasks in the diagnosis of chronic kidney disease is the characterization of glomeruli, clusters of capillaries that act as the basic filtration units of the kidney and that can be affected by a myriad of disease-related morphological changes [2]. The classification of these structures, when done manually, can be quite time consuming and difficult, and it has therefore been the focus of recent developments of deep learning (DL)-based nephropathology applications [3–8].

However, while clearly demonstrating the potential of DL for automatic classification of glomerular lesions, these works remain mostly proof-of-concept, often lacking sufficiently large and diverse training datasets to achieve clinical applicability. In fact, while whole slide scanners have enabled the rapid and feasible collection of vast amounts of histological image data, few of these data can be readily used to build DL models. Specifically, the training of such models is still mostly conducted in a supervised fashion, requiring additional image labels. However, the labeling of histological images is typically an extremely time-consuming task and highly dependent on domain expertise and thus remains a critical barrier to the development of AI tools.

Possible solutions for overcoming these bottlenecks include the use of more interactive labeling strategies [9–12] or alternative "not-so-supervised" learning regimes [13], i.e. training strategies not explicitly relying on extensively labeled data. Out of such approaches,

* Corresponding author.
E-mail address: hrafn.holger.weishaupt@helse-bergen.no (H. Weishaupt).

https://doi.org/10.1016/j.compbiomed.2025.110719
Received 26 November 2024; Received in revised form 5 June 2025; Accepted 2 July 2025
Available online 1 August 2025

unsupervised and self-supervised learning strategies appear particularly appealing, because they can operate on completely unlabeled data and in the absence of supervision of a pathologist. Instead, utilizing a data-driven approach to identify a potentially meaningful separation between histological images, such methods can help to (i) group images and patches [14–17] or to (ii) pretrain models [18–21], which in turn can be used as feature extractors or be repurposed for downstream tasks via transfer learning.

However, while a plethora of unsupervised and self-supervised learning paradigms has been developed in the computer vision field [22–29], very little research has so far focused on the application of such approaches in the field of glomerular lesions. Liu et al. [30] employed self-supervised learning on glomerular images, but only for the downstream classification between patches with and without glomeruli. Yao et al. [31] utilized self-supervised learning to pretrain a CNN on web-mined images of glomeruli, but subsequently still followed with a supervised fine-tuning using thousands of labeled images. Sato et al. [15] asked the question of whether a purely unsupervised definition of glomerular classes might provide a clinically meaningful basis for classifying glomerular changes in the absence of labels. While the study produced some promising concepts and initial results pertaining to the clustering of glomerular images, the separation between morphological classes and the purity of clusters appeared only limited and was not fully evaluated. Furthermore, the study utilized only a very restricted dataset, including only a single histological stain (Hematoxylin & Eosin; HE), and investigated only a single convolutional neural network (CNN) for feature extraction. Consequently, there is still a substantial lack of understanding regarding the full potential of clustering for the unsupervised distinction between glomerular lesions.

In theory, the clustering of morphological lesions might yield better results when performed on more extensive image datasets and/or when using a stain such as Periodic Acid Schiff (PAS) that better highlights relevant glomerular structures [32]. However, the majority of glomerular lesions might also be inherently difficult to distinguish via clustering, as the morphological changes in these cases are often segmental [33,34], i.e. affecting only part of the glomerulus, while the remaining structure would be normal or displaying other lesions. Thus, an unsupervised separation might be most successful for glomeruli with global lesions, i.e. with a morphological change affecting the entire glomerulus.

Consequently, towards investigating the feasibility of clustering glomerular lesions, it would be prudent to demonstrate it first in the context of such a well delineated task, i.e. a separation of the most distinct lesion categories, enabling a systematic proof-of-concept evaluation of methodology and performance. Specifically, the current study hypothesized that the best separation would probably be achieved for globally sclerosed glomeruli (GS), as they represent a global pattern without any normal structures. This assumption is aligned with the observation that it is likely the class most easily classified [3–5]. The notion is further supported by the study by Altini et al. [7], who illustrated several CNN-based feature embeddings of glomerular images, where the most substantial separation appeared to manifest between the GS and non-globally sclerosed glomeruli (nonGS). Similarly, another recent publication showed that the unsupervised clustering of kidney biopsy image patches could theoretically distinguish between patches containing (globally) sclerosed and patches containing other glomeruli [14].

Beyond serving as a proof-of-concept for ongoing clustering efforts, the unsupervised separation of GS and nonGS would also provide an avenue for semi-automatic labeling opportunities enabling a more straightforward collection of DL datasets for the training of GS classifiers. However, while the previous studies have demonstrated the general feasibility of such an approach, to the best of our knowledge, there is virtually no research systematically evaluating the clustering of GS and nonGS, including (i) an investigation of the most suitable strategy for clustering GS images, and (ii) explicitly documenting the accuracy with which such an unsupervised strategy can distinguish globally sclerosed glomeruli.

Accordingly, the current study thoroughly investigated the hypothesized possibility of detecting global glomerulosclerosis through clustering, by (i) utilizing larger datasets from varied sources and across different histological stains, (ii) evaluating a large number of feature extraction methods, (iii) carefully measuring the separation of classes in the associated feature embeddings, and (iv) explicitly evaluating the performance of clustering in capturing the classes.

2. Material and methods

The project was conducted in four major stages as outlined in Fig. 1. Utilizing various repositories, a large collection of glomerular image patches was compiled (Fig. 1A) and preprocessed (Fig. 1B) for downstream analyses. Subsequently, numerous CNN models were employed to extract image features from the glomerular image patches, and the class separation between GS and nonGS in the resulting feature embedding was evaluated (Fig. 1C). Finally, it was investigated how accurately the GS and nonGS classes could be captured by clustering of the images in the feature space (Fig. 1D). A more detailed account of the individual steps is given below and in the Supplemental Material.

2.1. Data

2.1.1. Glomerular image datasets

The current study utilized glomerular image patches from seven sources: Besusparis et al. [5] (Besusparis2023, n = 3993), Bueno et al. [36,37] (Bueno2020, n = 946), Gallego et al. [38] (two datasets: Gallego2021-HE/Gallego-PAS, n = 611/527) [39], the Kidney Precision Medicine Project (KPMP, n = 5978) [35], the Human BioMolecular Atlas Program (HuBMAP, n = 4130) [40], the Norwegian Renal Registry (three datasets: NRR-PAS/NRR-HE/NRR-SIL, n = 250/555/568), and Weis et al. [8] (Weis2022, n = 5210). The sources displayed different properties, e.g. with respect to image formats, histological stains, availability of glomerular segmentations and/or class labels, which diagnoses and morphological lesions (beyond GS) are present in the dataset, and the proportion of GS and nonGS images, which are further described in the supplemental material and in supplementary tables 1 and 2.

2.1.2. Image patch generation and preprocessing

Depending on the source, the kidney biopsy images containing glomeruli came in different formats and were thus subjected to a sequence of preprocessing steps (Fig. 1A-B) to generate the final glomerular image patches.

Specifically, where not already available, glomeruli were segmented from the surrounding tissue and classified as either GS or nonGS (Fig. 1A). The segmentation annotations were utilized to compute the bounding box for each glomerulus, the image region within which was then extracted (Fig. 1A).

The resulting raw glomerular image patches were first rescaled to 224 × 224 pixels. Then, similar to the procedure documented by Sato et al. [15], image patches within each dataset were subjected to a stain-normalization step utilizing the Macenko method [41] (Fig. 1B). To evaluate the robustness of the feature embedding and clustering results with respect to the choice of reference image used during stain-normalization, 20 reference images with substantial color differences were selected per stain (Supp. Fig. 1) and utilized to generate 20 different stain-normalized versions of each dataset.

Finally, after image normalization, the surrounding tissue (pixels outside of the annotated glomerulus) in each image patch was masked out by replacing it with black color (Fig. 1B) [5,7].

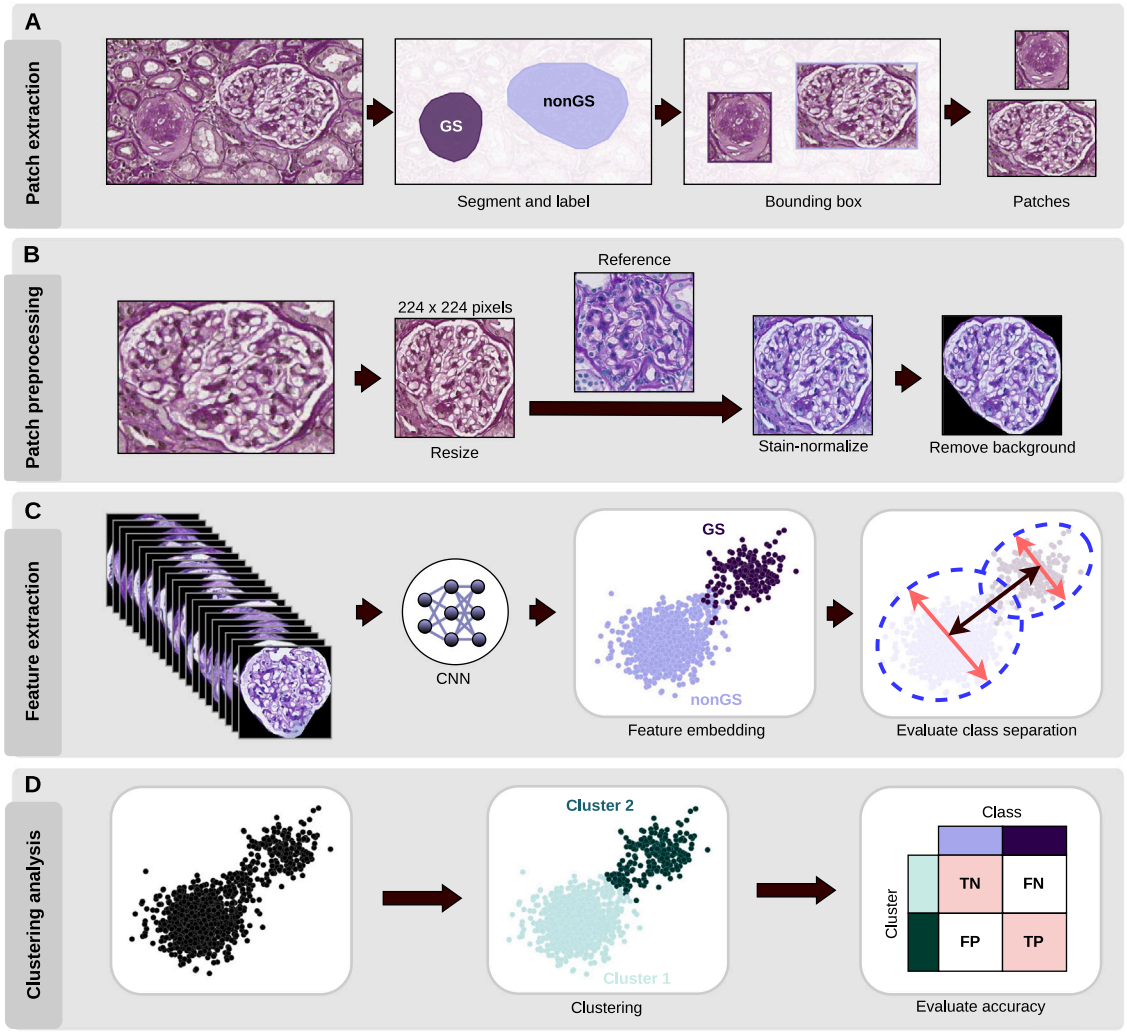

Fig. 1. Workflow illustrating the overall procedure employed for data generation and analysis. (**A**) Gathering glomerular patches from WSIs or other images: where necessary, glomeruli were segmented and labeled (here shown on part of a WSI from the Kidney Precision Medicine Project (KPMP) [35]); subsequently, glomerular images where cropped using the bounding box. (**B**) Glomerular image patch preprocessing: All image patches were first resized to 224 × 224 pixels, then stain-normalized, and finally the surrounding tissue background was replaced by black pixels. (**C**) Feature embedding: Each glomerular image patch was fed into a CNN for feature extraction, and in the resulting feature embedding for a single dataset, the separation between GS and nonGS images was evaluated quantitatively. (**D**) Clustering of glomerular image patches: For each dataset, the images were clustered in the feature embedding, and the cluster assignments were compared to the corresponding class labels.

2.2. Feature extraction and evaluation

2.2.1. CNN models used for feature extraction

The current study compared the separation of GS and nonGS groups of glomeruli in the embedding of features obtained by different CNN models (Fig. 1C). A total of 34 CNNs were utilized (Supp. table 2), largely overlapping the models investigated for glomerular classification by Weis et al. [8] and Altini et al. [7], and also including the NASNetLarge model used in the previous clustering study by Sato et al. [15]. The CNN models were run either using the implementations available from Keras [42] or from the *classification_models* (https://github.com/qubvel/classification_models) library (Supp. table 3). All models were utilized without the top (classification) layers, the input tensor size was set to 224 × 224 × 3, weights pretrained on ImageNet were used, and a global average pooling was applied to the output of the final convolutional layer. All models were utilized with the existing pre-trained weights without any further fine-tuning.

2.2.2. Evaluating class separation in feature embedding

The separation between nonGS and GS glomeruli in a feature embedding was evaluated visually and through standard cluster evaluation metrics. Specifically, visual assessment involved the inspection of scatter plots after a Uniform Manifold Approximation and Projection (UMAP) [43] of the high-dimensional feature space down to two dimensions. For a quantitative assessment, the nonGS and GS labels were instead interpreted as cluster assignments, and then three internal cluster validity indices (CVIs) were applied to measure the quality of this "clustering" in the high-dimensional feature embedding: the Silhouette score [44], the C-index [45], and the Dunn index [46]. For the C-index, a better clustering would result in a lower score, while for the other two methods, a better clustering would result in a higher score.

2.3. Clustering analyses and evaluation

To cluster the glomerular images in the feature embedding (Fig. 1D), the project utilized either a Gaussian mixture model (GMM) strategy, similar to the study by Sato et al. [15], or a Leiden clustering [47] approach, in both cases aiming for exactly two clusters. The performance of clustering in capturing the nonGS and GS classes (Fig. 1D) was then evaluated (i) using confusion matrices, displaying the association between glomerular classes and the identified clusters, (ii) by one metric measuring the similarity between two clusterings, i.e. the

adjusted Rand index (ARI) [48], and (iii) by one metric measuring classification performance, i.e. accuracy (ACC).

2.4. Hardware

Experiments were conducted on a HP Z4 G4 workstation, with an Intel® Xeon® W-2245 processor, 128 GB DDR4 RAM, and two NVIDIA® RTX™ A4500 graphic cards.

2.5. Code availability

Representative code, demonstrating the preprocessing of patches, feature extraction, and clustering, is available on GitHub: https://github.com/patologiivest/GlobalGlomerulosclerosisClustering.

3. Results

3.1. CNN-based feature embedding separates GS and nongs

To evaluate the ability of automatically distinguishing between GS and nonGS glomerular images via unsupervised learning, the study first aimed to assess the separation of these classes in the feature space provided by different CNN models. Towards this end, ten datasets of glomerular images were selected, referred to respectively as Besusparis2023, Bueno2020, Gallego2021-HE, Gallego2021-PAS, HuBMAP, KPMP-PAS, NRR-HE, NRR-PAS, NRR-SIL, and Weis2022, covering various histological stains and glomerular lesion categories (Supp. tables 1-2). Each dataset was labeled for GS and nonGS and was further stain-normalized with 20 different reference images (Supp. fig. 1), producing a total of 200 datasets for testing. Subsequently, 34 different CNNs (Supp. table 3) were selected by screening for models evaluated in recent glomerular classification or clustering studies [7,8,15] and with available pretrained weights accessible in Keras. The models were then utilized to extract image features from each of the datasets.

Evaluating the relationship between class affiliations and feature embeddings via internal CVIs, i.e. the Silhouette score [44], the C-index [45], and the Dunn index [46], the models displayed substantial differences in the separation achieved between the two classes (Fig. 2A-C, Supp. fig. 2). Specifically, the MobileNet, DenseNet169, DenseNet201, SE-ResNet101, and Xception were among the best performing models, while SE-ResNeXt50, EfficientNetV2M, EfficientNetV2L, VGG16, and VGG19 ranked generally the lowest.

When visually evaluated in respective two-dimensional UMAP embeddings, models with different ranks according to the CVIs also displayed differences in the separation between GS and nonGS (Fig. 2D-F), suggesting that the ranking provides some insight into how well the model-related features capture differences between the two classes. In addition, the normalization of datasets with different reference images also led to slight variations of the embeddings (Supp. fig. 3).

Finally, visualizing the resulting UMAP embeddings for two of the most promising feature extractors, i.e. the MobileNet and the DenseNet169, a marked separation between the GS and nonGS classes was observed across all datasets (Fig. 3, Supp. fig. 4), in comparison to the UMAP embeddings produced by some of the worst performing models such as the SE-ResNeXt50 or the EfficientNetV2L model (Supp. fig. 5-6).

3.2. Unsupervised learning captures GS and nongs classes

Having demonstrated that CNN-derived features can enable a visually highly pronounced separation between the GS and nonGS classes, the next question was then whether a clustering strategy could also automatically detect the two glomerular classes from the respective feature embedding. Towards this goal, the project first adopted a strategy equivalent to what was proposed by Sato et al. [15], i.e. a UMAP projection followed by GMM clustering. Applied to the ten datasets, this approach generally produced good clustering results (data not shown). However, when investigating the robustness of this finding across the different stain-normalized versions of each dataset, there were cases in which GMM clustering failed to pick up the two visually apparent groups (Supp. fig. 7).

To overcome this issue, the project instead adopted a Leiden clustering strategy [47], which has recently also been shown highly feasible for the clustering of histological images in other studies [16]. Leiden clustering produced a similar separation of glomeruli (data not shown), but also managed to pick up the visually apparent groups of glomeruli in the cases in which GMM clustering failed (Supp. fig. 7, Supp. fig. 8). Consequently, downstream evaluations of the clustering performance between GS and nonGS were based solely on Leiden clustering.

Specifically, for each dataset, clustering was then performed on the features extracted by each of the 34 different CNNs, followed by the computation of the adjusted Rand index (ARI), comparing the ground truth labels (GS and nonGS) to the resulting cluster assignments (C1 and C2). The models displayed a wide range of ARI values (Supp. fig. 9 A, and data not shown), which could not be explained by differences in the number of trainable parameters (Supp. fig. 9B) or the number of returned features (Supp. fig. 9C). Similarly, for each model, ARI variations where also observed between the different stain-normalized variants of the respective dataset (Supp. fig. 9A, and data not shown). An investigation of ARI values with respect to the stain-normalization reference image among all PAS datasets suggested that the choice of reference image might play some role in these differences (Supp. fig. 10). Specifically, ARI values were often worse in datasets stain-normalized with a reference image with a fainter or more bluish stain appearance.

Upon ranking models based on the observed ARI values (Fig. 4A), it was found that the MobileNet [49] might generally produce the best clustering performance across datasets. Particularly, assuming the minor cluster (C2) to capture the GS glomerular patches, and inspecting the corresponding confusion matrices for one stain-normalized version of each dataset, it was found that the use of the MobileNet in combination with Leiden clustering led generally to few false-negative (FN, ≤11%) and very few false-positive (FP, ≤3.69%) detections (Fig. 4B-K). These findings were confirmed by inspecting the ARI (Fig. 4L) and balanced accuracy (Fig. 4M) values across all 20 stain-normalized versions for each dataset, with the mean balanced accuracies exceeding 93% in the Bueno2020 and NRR-SIL datasets, and exceeding 97% in all other datasets.

3.3. Characterization and detection of misclustered cases

To understand the limitations of the clustering performances, a detailed investigation of the misclustered cases was conducted. Specifically, the analysis utilized a single stain-normalized version of each dataset, the MobileNet for feature extraction, and the Leiden algorithm for clustering. Subsequently, images labeled as GS and assigned to the major cluster (C1) where considered FNs, while images labeled as nonGS and assigned to the minor cluster (C2) were considered FPs. A preliminary inspection of these images across all ten datasets (Supp. fig. 11-13) revealed six different categories of glomerular images: (i) truly misclustered cases, (ii) mislabeled cases, (iii) borderline cases with advanced sclerosis, (iv) globally sclerosed glomeruli with holes or split tissue, (v) images with potentially insufficient information for accurate labeling, e.g. tangential sections of glomeruli, and (vi) artifacts.

In addition, in the UMAP embedding, many of the misclustered images were also located on the border between the two clusters (Supp. fig. 11-13). We anticipated that these might predominantly be borderline cases either on the gradient between segmental and global sclerosis or otherwise difficult to distinguish from global sclerosis. Accordingly, we hypothesized that the cluster affiliation of these cases might not be robust across different clustering runs, which would enable their detection as uncertain cases. To investigate this question, we

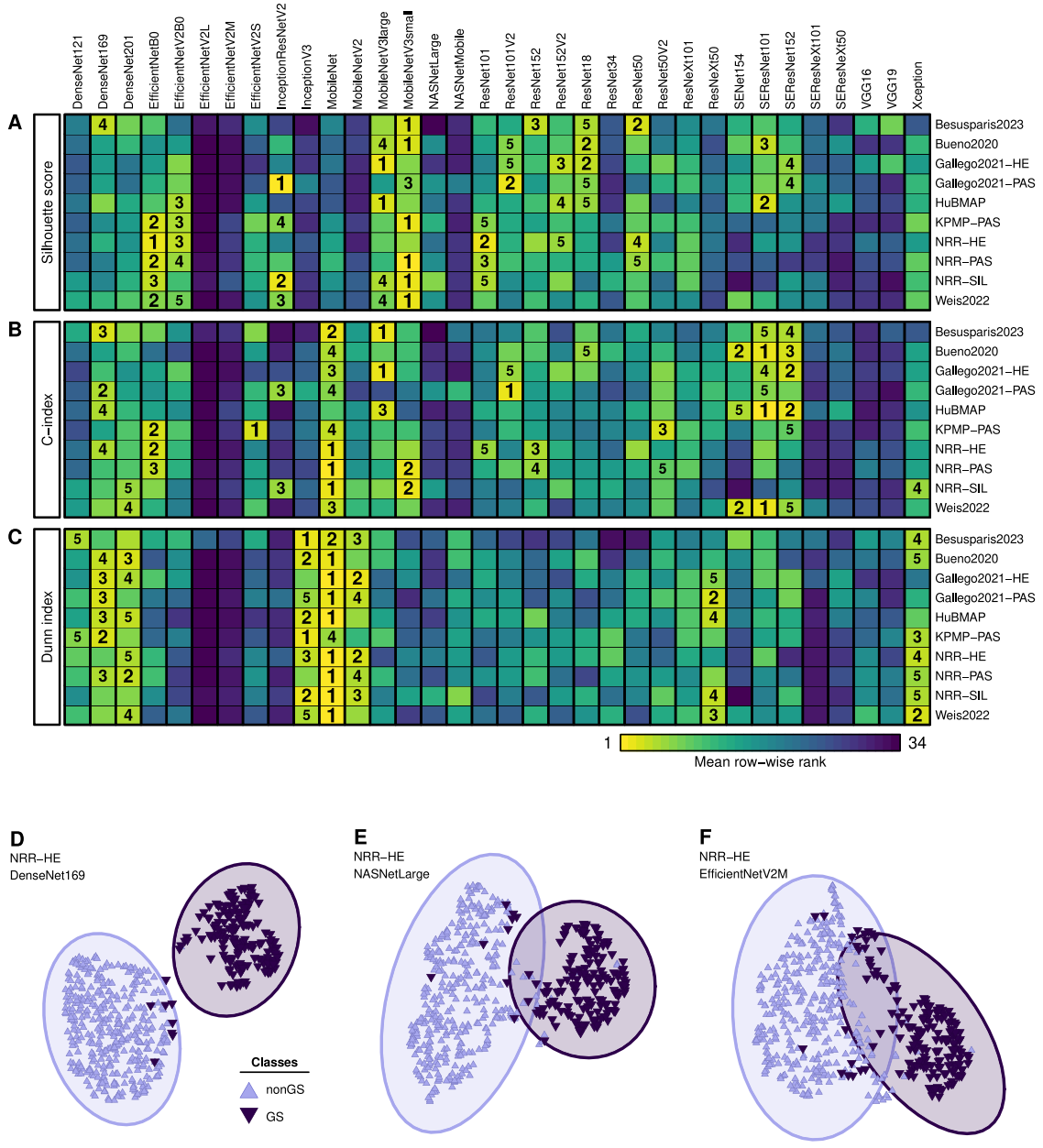

Fig. 2. Evaluating the separation of GS and nonGS image patches between different CNN feature extractors. (**A-C**) Heatmaps displaying the ranking of models based on internal cluster validity indices (CVIs), i.e. Silhouette score (A), C-index (B), and Dunn index (C). For each dataset, the models were ranked 20 times, once per stain-normalized version of the dataset, and the heatmap was generated based on the mean across these 20 rankings. (**D-F**) UMAP embedding of the NRR-HE dataset using features from three different models, i.e. DenseNet169 (D), NASNetLarge (E), and EfficientNetV2M (F), chosen based on decreasing scores in all three CVIs. The shaded ellipses represent the 95% confidence ellipse for each class in the respective embedding.

conducted two additional experiments, in which we compared either (i) the clustering of the 20 stain-normalized variants of the KPMP-PAS dataset after feature extraction with the MobileNet (Fig. 5A), or (ii) the clustering of one stain-normalized variant of the KPMP-PAS dataset processed with three different CNNs, i.e. DenseNet169, MobileNet, and MobileNetV3Small (Supp. fig. 14A). Uncertain cases were then identified as those not always assigned to the same cluster across the different runs, and they covered many of the previously misclustered cases. Following a detailed re-labeling of these uncertain cases and the remaining FP and FN cases (those not absorbed into the uncertain group) by the consensus of two experienced nephropathologists, it was found that, in addition to distinct GS and nonGS images, they constituted many images with borderline GS, GS glomeruli with extensive white areas due to holes/split tissue, images with insufficient detail on the glomerulus for adequate labeling, or different types of artifacts (Fig. 5B). In addition, the re-labeling also suggested that many of the FN and FP clusterings were caused by a mislabeling in the initial labeling round (Fig. 5C-D, Supp. fig. 14B-C). The uncertain glomeruli were only very seldom mislabeled (Fig. 5E-F, Supp. fig. 14D-E). Glomeruli with more difficult appearances (borderline sclerosis, GS with holes, glomeruli with too little detail, or artifacts) were clearly present at varying percentages among all four groups of misclustered cases (Fig. 5C-F, Supp. fig. 14B-E), potentially explaining inconsistent clustering results for these cases.

4. Discussions

The detection of global glomerulosclerosis plays a clear role in CKD diagnostics since it is one of the most prominent lesions examined during biopsy reporting. Thus, the automatic classification of GS glomeruli

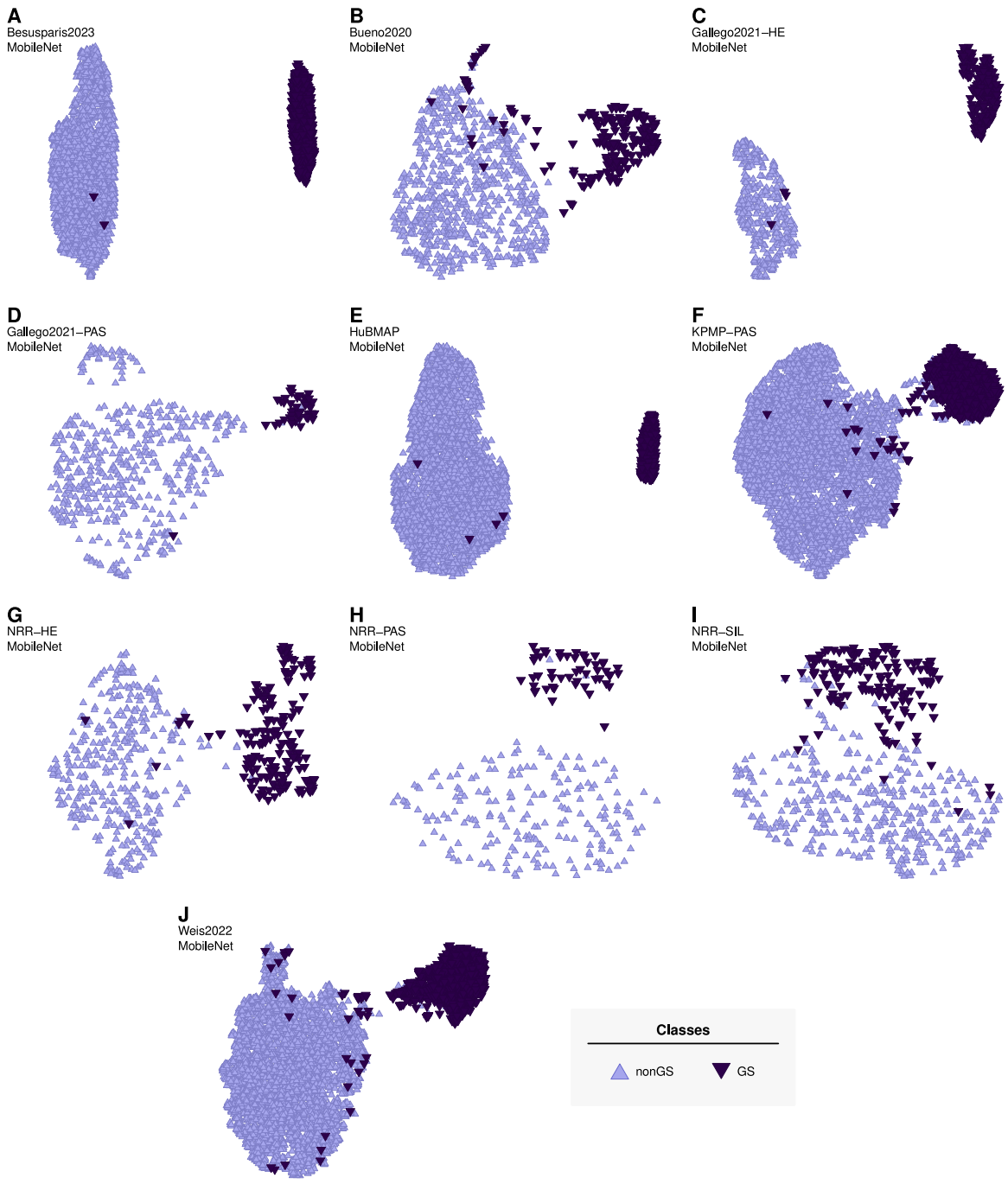

Fig. 3. Scatterplots illustrating the separation of GS and nonGS glomerular image patches following feature extraction with the MobileNet and UMAP embedding in two dimensions. Each scatterplot depicts the feature embedding and classes of one stain-normalized variant of each of the ten datasets: Besusparis2023 (A), Bueno2020 (B), Gallego2021-HE (C), Gallego2021-PAS (D), HuBMAP (E), KPMP-PAS (F), NRR-HE (G), NRR-PAS (H), NRR-SIL (I), and Weis2022 (J). Light and dark purple colors indicate nonGS and GS glomerular images, respectively.

represents a key objective in the development of computer-assisted diagnostic tools [4,50]. However, current DL models for the classification of glomerular lesions are almost exclusively trained via supervised learning [3–8], which relies on extensively labeled image datasets, the establishment of which is often very costly. The current project demonstrated that, across several large collections of glomerular images and histological stains, the GS and nonGS classes could be consistently separated using clustering, enabling the labeling of glomerular images through a purely data-driven, unsupervised approach.

The current findings are closely related to the publication by Sato et al. [15], which appears to be one of the first documented studies on clustering glomeruli. However, while the authors outlined a framework for the clustering of glomeruli based on CNN-derived image features, they did not document any quantitative measure of how well this clustering separated between different glomerular lesions. Instead, the authors utilized the cluster affiliations as labels for downstream supervised fine-tuning of the network (NASNetLarge), and the softmax output of the final model was then interpreted as a score that could be evaluated in terms of association with clinical parameters. Furthermore, the authors only utilized a single glomerular dataset, and did not evaluate clustering with respect to different CNN models or different reference images for stain-normalization. In contrast, Altini et al. [7], compared a large panel of DL models for the classification of glomerular lesions, and showed the feature embeddings

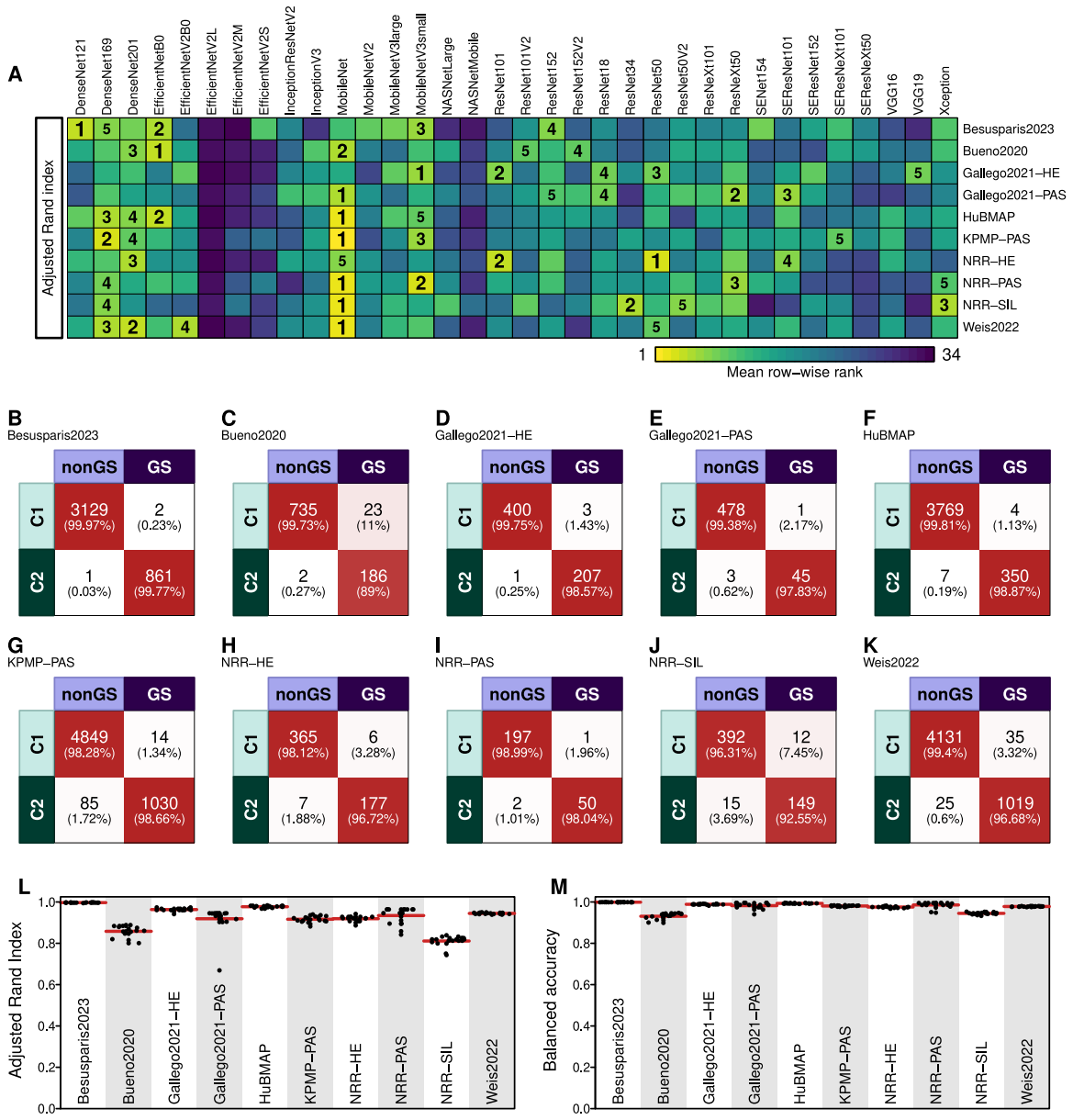

Fig. 4. Evaluating the agreement between cluster and class labels. (**A**) Heatmap displaying the ranking of CNN models within each dataset based on the adjusted Rand index (ARI). The ARI was computed between the class labels and the cluster assignments produced by Leiden clustering on the features produced by the respective CNN model. For each dataset, the models were ranked 20 times, once per stain-normalized version of the dataset, and the heatmap was generated based on the mean across these 20 rankings. (**B-K**) Confusion matrices illustrating association between cluster (C1, C2) and class (nonGS,GS) memberships for the ten datasets: Besusparis2023 (B), Bueno2020 (C), Gallego2021-HE (D), Gallego2021-PAS (E), HubMap (F), KPMP-PAS (G), NRR-HE (H), NRR-PAS (I), NRR-SIL(J), and Weis2022 (K). (**L-M**) Quantitative evaluation of the agreement between cluster affiliations and class labels as measured by the adjusted Rand index (ARI, L) and the balanced accuracy (ACC, M). Each data point represents the measurement performed on one of the 20 different stain-normalized variants of the respective dataset. The red lines indicate the mean across the 20 values for each dataset.

of the best performing models. Importantly, highlighting the class labels of glomeruli in the embedding plots, the results suggested a notable separation between GS and nonGS glomeruli even when only using the original pretrained model without further fine-tuning. However, beyond an illustration of feature embeddings in relation to glomerular class labels, the authors did not conduct any clustering experiments to evaluate the unsupervised separation between glomerular lesions.

Of note, the presented approach is likely not suited for use in a diagnostic setting because a single biopsy section typically contains too few glomeruli to warrant clustering. Instead, the outlined approach might help in easing the manual labeling requirements and/or enable downstream unsupervised learning strategies [15], pushing towards more robust and clinically relevant GS detection models. Specifically, given the high accuracy with which GS images could be identified in the current study, the method appears highly suited for approaching a (semi-)automatic labeling of GS images, which would greatly simplify the collection of DL training datasets.

Furthermore, while the clustering of GS serves only as a first proof-of-concept, the accuracies achieved in the current study also represent a promising starting point to investigate the further subclustering of other glomerular lesions. For instance, a similar clustering strategy applied and evaluated on the GS cluster alone might provide insight into the separation between obsolescent, solidified, and disappearing GS [31,51]. In addition, it could also be investigated whether the clustering of sclerosis and non-sclerosis can be extended to smaller image patches from within glomeruli, as suggested by Sato et al. [15], which would potentially enable the separation between nonGS, GS, and segmentally sclerosed glomeruli (SS). Finally, the separation of other glomerular lesions might also be possible by evaluating feature embedding and clustering strategies on the nonGS cluster alone. However, achieving

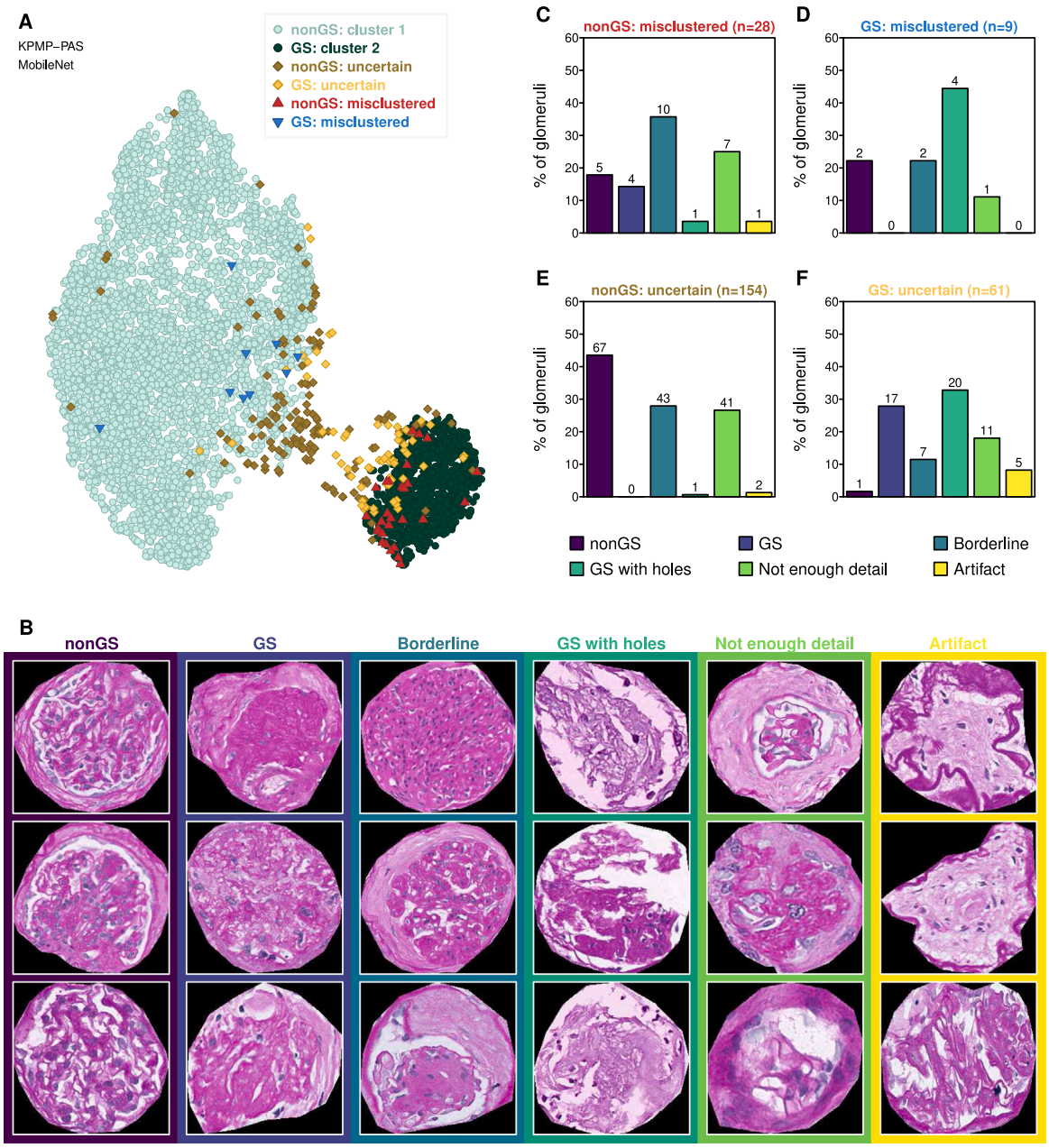

Fig. 5. Evaluation of incorrectly clustered image patches: voting over multiple stain-normalized variations of the same dataset. (**A**) Consensus of the clustering of the KPMP-PAS dataset following 20 different stain-normalizations and feature extraction using the MobileNet. Light and dark green circles indicate, respectively, glomeruli that are nonGS and always in cluster 1 and GS glomeruli that are always in cluster 2. Red and blue triangles indicate glomeruli that are always false-positives (nonGS in cluster 2) and false-negatives (GS in cluster 1), respectively. Brown and yellow diamonds indicate nonGS and GS glomeruli, respectively, that switch clusters between the different stain-normalizations and can thus be identified as uncertain cases. (**B**) Three examples for each of the six different categories, i.e. nonGS, GS, borderline glomeruli with advanced sclerosis, GS with holes, glomeruli with insufficient detail for labeling, and artifacts, employed during relabeling of misclustered glomeruli. (**C-F**) Barplots representing the percentages (absolute number above each bar) of these six groups among the glomeruli from each of the four misclustered categories, i.e. false-positives (C), false-negatives (D), and uncertain nonGS (E) and GS (F) glomeruli.

such a subclustering of meaningful nonGS lesion categories would likely require further efforts in identifying or establishing suitable feature extraction models able to capture features relevant to the separation of other morphological lesions. Possible avenues towards such improved feature extractors are for instance the use of existing general-histology foundation models [18–21] or the use of self-supervised learning on kidney biopsy images [30,31].

Importantly, unsupervised approaches should be approached with care. For instance, the current project always assumed that all datasets contained a subset of globally sclerosed glomeruli and that this subset constitutes the minor cluster. However, such a cluster-to-class assignment might fail if no globally sclerosed glomeruli are present or if they are more abundant than the non-globally sclerosed glomeruli. Possible solutions might be to either inspect a few randomly sampled, representative cases from each cluster, or to include a set of external reference GS and nonGS images with known labels. Specifically, when processed together with the unlabeled dataset, including stain-normalization, feature extraction, and clustering, the cluster affiliation of the reference images might be used to automatically identify the GS and nonGS clusters.

Furthermore, the clustering of GS and nonGS might only work if a sufficient total number of glomeruli and a sufficient number of GS image patches is included. The clustering might also be substantially affected by the composition of the dataset, including e.g. the

types of diseases from which glomeruli were extracted, the types of glomerular lesions included, and the amount of pruning and curation performed on the image patches. Specifically, the best separations of the GS and nonGS classes were observed in the Besusparis2023 and HuBMAP datasets. While the former comprised a variety of different lesion categories, image patches were highly curated, being subjected to multiple rounds of validation to only include the image patches with the cleanest and most discernible lesions, thus potentially leading to the exceptional separation between GS and nonGS images. The HuBMAP data on the other hand appeared to lack, apart from globally sclerosed glomeruli, a large diversity of other lesions, thus leading to a clustering of almost exclusively GS and normal glomeruli, again resulting in a superb separability. The NRR, KPMP-PAS, Weis2022, and Bueno2020 datasets on the other hand were known or assumed to comprise both a broad range of diseases and/or lesion categories and were collected without much curation or pruning, thus representing a more natural distribution of glomerular appearances in both the GS and nonGS classes and consequently also a potentially more difficult task for clustering. In addition, the images in the Weis2022 dataset were also less clean with respect to glomerular extraction, containing sometimes only incompletely cropped glomeruli. The worst clustering performance was observed with respect to the NRR-SIL dataset. However, this finding was somewhat anticipated, considering that the pathologist also experienced difficulties labeling some of the images solely based on an inspection of the Periodic Acid Schiff Methenamine Silver (SIL)-stained glomerulus itself, instead requiring a cross-check with corresponding PAS-stained images. With respect to the evaluation of clustering performance, it should also be noted that the distinction between GS and SS might often be subjective. Specifically, the development of GS is a gradual process [52], and the definition of a consistent cut-off between SS and GS might be difficult. In light of this consideration, a more complete evaluation of clustering performance might benefit from the labeling of glomeruli with any level of sclerosis, to be able to distinguish misclustered cases into true mistakes (e.g. non-sclerotic glomeruli in the GS cluster) and borderline cases. Importantly, the observation of images that occur on the boundary between the two clearly clustered groups of glomeruli is also expected and highly consistent with the results published by Walker et al. [12]. Ultimately, when using the clustering as a starting point for labeling glomeruli, the glomeruli on the boundary between the two clusters would then likely require more detailed inspections and labeling by pathologists, while the glomeruli clearly associated to either one of the clusters would probably have more easily verifiable labels and would often only require a superficial screening for any obvious mistakes [12]. Considering the accuracy of the clustering and the low number of boundary images documented in the current study, such a strategy could then result in a drastic speedup in labeling.

Importantly, while all employed CNNs were pretrained on ImageNet data, they performed often quite different in terms of feature embedding and clustering accuracy. The study found no evident association between performance and the number of trainable parameters or the number of extracted features. Other factors might be differences in preprocessing, the dimension of input images, or the particular training procedure, all of which might affect the feature outputs of the network. Thus, fully elucidating the causes for such discrepancies might be difficult and would likely require a much more extensive investigation of differences between model features and their impact on glomerular class separation. Ultimately, there is likely no one-fits-all CNN when it comes to feature embedding for clustering of GS and nonGS. Specifically, while the MobileNet appeared to be a generally good choice for feature extraction in combination with Leiden clustering, it was not the best choice across all datasets, with differences likely determined by variations in dataset composition, image properties, and histological stain. Thus, when applied to a new, unlabeled dataset, a strategy to find the feature extractor resulting in the best clustering might be to test different CNNs and then to evaluate the respective clustering qualities via internal CVIs.

However, as indicated in the present study, selecting the best feature embedding and/or clustering based on internal CVI is not always straightforward. Specifically, there exists a vast number of internal CVI methods in the literature [53], each of which evaluates a specific objective function that might or might not align with the objective function optimized in the chosen clustering approach. In fact, in the current study, the ARI results for clustering with the Leiden algorithm did not correlate very well with the internal CVIs applied to the class separation in the feature embedding and discrepancies were also observed between the different CVIs. While the high-dimensional data makes it difficult to visualize and fully investigate the cause for such discrepancies, a likely explanation might still be base differences in what these CVIs evaluate. For instance, while the Silhouette score measures cluster quality by determining how well each individual point fits into its own cluster as compared to other clusters, the Dunn index is computed on more global estimates of inter-cluster distance and intra-cluster similarity [54]. Furthermore, CVIs might work differently depending on the skewness, noise, or subclusters in the data [53,55].

The current study was also subject to some technical and methodological limitations. Specifically, given the numerous combinations of models, datasets, and stain-normalizations utilized, the UMAP and GMM functions were run with default parameters to make them generically applicable rather than fine-tuning them separately for each individual case. In addition, only a single stain-normalization method was utilized, but it would be relevant to investigate how alternative approaches [56,57] might affect clustering. The study also demonstrated that the choice of stain-normalization reference can affect the outcome of the clustering and discussed a possible reason for some of this effect. Specifically, in the PAS datasets, a reference image with a fainter or more bluish stain appearance often resulted in a lower ARI value, possibly due to a lower contrast between the PAS-positive, PAS-negative, and/or nuclear compartments. These preliminary findings warrant further investigations into the effect of reference image choice on clustering and how to select reference images accordingly.

Furthermore, it is not known how changes to the other preprocessing steps, i.e. the cropping and scaling or the removal of non-glomerular background, would influence clustering results. For instance, globally sclerosed glomeruli are often smaller than normal glomeruli [52], but by cropping and rescaling all glomeruli based on a bounding box rather than just using a fixed-sized crop, information about glomerular sizes is lost. Lastly, the study evaluated only the use of pre-trained CNN-based feature extractors, but did not explore other techniques for feature embedding such as autoencoders [15,58] or feature extractors trained on histological data [18–21]. Given the results of the current study, the clustering of GS versus nonGS might already be approaching a limit in achievable performance even when only using models pretrained on ImageNet, since the remaining misclustered glomeruli often just represent artifacts, outliers, mislabelings, or glomeruli that are also difficult to label for a pathologist. However, in order to achieve an acceptable clustering of other glomerular lesions, fine-tuning or domain-specific pretraining, e.g. through contrastive learning, might likely be required to achieve image representations and feature extractors more adapted to the underlying histology.

5. Conclusion

In summary, the current study clearly demonstrated that an unsupervised learning approach, utilizing CNNs pre-trained solely on ImageNet, can separate globally and non-globally sclerosed glomerular images with very high proficiency. Specifically, among the evaluated CNN models, the MobileNet [49] was found to generally perform best for clustering nonGS and GS glomeruli, achieving accuracies of over 95% in most datasets.

An inspection of the presumably misclustered glomeruli suggested that many such cases were either mislabeled during dataset generation or represented borderline glomeruli that are inherently difficult to classify even for a pathologist. Furthermore, many such misclustered/uncertain glomeruli could be automatically detected by clustering the data multiple times, either using different stain-normalization references or different CNN models for feature extraction. Following this approach, glomeruli that require manual attention by a pathologist can easily be highlighted, while clearly clustered glomeruli could potentially be screened more rapidly for any obvious mistakes, resulting in a speedup as compared to manual labeling from scratch.

We are convinced that the outlined strategy will aid researchers (i) in achieving a more labeling-efficient generation of datasets for the training of deep-learning GS classifiers, and (ii) by providing a basis for ongoing research into the clustering of glomerular lesions. Specifically, considering the findings of the present work as a proof-of-concept, it appears promising to further explore the potential of unsupervised learning towards the detection of a broader range of glomerular lesions patterns.

CRediT authorship contribution statement

Hrafn Weishaupt: Writing – review & editing, Visualization, Validation, Software, Resources, Project administration, Methodology, Writing – original draft, Investigation, Formal analysis, Data curation, Conceptualization. **Justinas Besusparis:** Writing – review & editing, Data curation. **Cleo-Aron Weis:** Writing – review & editing, Resources. **Stefan Porubsky:** Writing – review & editing, Resources. **Arvydas Laurinavičius:** Writing – review & editing, Resources. **Sabine Leh:** Writing – review & editing, Supervision, Resources, Funding acquisition, Data curation.

Ethical declaration

Processing of NRR datasets: The collection and processing of data was approved by the Regional Committee for Medical and Health Research Ethics (REK number 517496).

Processing of Weis2022 dataset: The collection and processing of data was conducted in accordance with a vote from the ethics commission II of the Heidelberg university (vote 2020-847R).

Processing of Besusparis2023 dataset: The collection and processing of data was conducted with the permission from the Vilnius Regional Biomedical Research Ethics Committee (No. 2019/6-1148-637).

Funding

The project was funded by The Western Norway Health Authority (strategic research fund F-12563).

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Cleo-Aron Weis reports administrative support was provided by Heidelberg University Hospital Institute of Pathology. Sabine Leh reports a relationship with Sectra AB that includes: travel reimbursement. Sabine Leh reports a relationship with European Society of Digital and Integrative Pathology that includes: board membership. Sabine Leh reports a relationship with The Royal Society of Medicine that includes: travel reimbursement. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The results here are in part based upon data generated by (i) the Kidney Precision Medicine Project (https://www.kpmp.org), (ii) the HuBMAP program (https://hubmapconsortium.org), and (iii) the Norwegian Renal Registry (https://www.nephro.no/nnr.html). We would further like to thank Mindaugas Morkūnas for help in compiling the Besusparis2023 dataset.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.compbiomed.2025.110719.

Data availability

Data is available from the corresponding author on reasonable request.

References

[1] R.D. Büllow, J.N. Marsh, S.J. Swamidass, J.P. Gaut, P. Boor, The potential of artificial intelligence-based applications in kidney pathology, Curr. Opin. Nephrol. Hypertens. 31 (2022) 251–257.

[2] M. Haas, S.V. Seshan, L. Barisoni, K. Amann, I.M. Bajema, J.U. Becker, K. Joh, D. Ljubanovic, I.S. Roberts, J.J. Roelofs, 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 (2020) 1120–1134.

[3] E. Uchino, K. Suzuki, N. Sato, R. Kojima, Y. Tamada, S. Hiragi, H. Yokoi, N. Yugami, S. Minamiguchi, H. Haga, et al., Classification of glomerular pathological findings using deep learning and nephrologist–ai collective intelligence approach, Int. J. Med. Inform. 141 (2020) 104231.

[4] C.-K. Yang, C.-Y. Lee, H.-S. Wang, S.-C. Huang, P.-I. Liang, J.-S. Chen, C.-F. Kuo, K.-H. Tu, C.-Y. Yeh, T.-D. Chen, Glomerular disease classification and lesion identification by machine learning, Biomed. J. 45 (2022) 675–685.

[5] J. Besusparis, M. Morkunas, A. Laurinavicius, A spatially guided machine-learning method to classify and quantify glomerular patterns of injury in histology images, J. Imaging 9 (2023) 220.

[6] R. Yamaguchi, Y. Kawazoe, K. Shimamoto, E. Shinohara, T. Tsukamoto, Y. Shintani-Domoto, H. Nagasu, H. Uozaki, T. Ushiku, M. Nangaku, et al., Glomerular classification using convolutional neural networks based on defined annotation criteria and concordance evaluation among clinicians, Kidney Int. Rep. 6 (2021) 716–726.

[7] N. Altini, M. Rossini, S. Turkevi-Nagy, F. Pesce, P. Pontrelli, B. Prencipe, F. Berloco, S. Seshan, J.-B. Gibier, A.P. Dorado, et al., Performance and limitations of a supervised deep learning approach for the histopathological oxford classification of glomeruli with iga nephropathy, Comput. Methods Programs Biomed. 242 (2023) 107814.

[8] C.-A. Weis, J.N. Bindzus, J. Voigt, M. Runz, S. Hertjens, M.M. Gaida, Z.V. Popovic, S. Porubsky, Assessment of glomerular morphological patterns by deep learning algorithms, J. Nephrol. 35 (2022) 417–427.

[9] N. Bouteldja, B.M. Klinkhammer, R.D. Bülow, P. Droste, S.W. Otten, S.F. von Stillfried, J. Moellmann, S.M. Sheehan, R. Korstanje, S. Menzel, et al., Deep learning–based segmentation and quantification in experimental kidney histopathology, J. Am. Soc. Nephrol.: JASN 32 (2021) 52.

[10] B. Lutnick, B. Ginley, D. Govind, S.D. McGarry, P.S. LaViolette, R. Yacoub, S. Jain, J.E. Tomaszewski, K.-Y. Jen, P. Sarder, An integrated iterative annotation technique for easing neural network training in medical image analysis, Nat. Mach. Intell. 1 (2019) 112–119.

[11] R. Miao, R. Toth, Y. Zhou, A. Madabhushi, A. Janowczyk, Quick annotator: an open-source digital pathology based rapid image annotation tool, J. Pathol.: Clin. Res. 7 (2021) 542–547.

[12] C. Walker, T. Talawalla, R. Toth, A. Ambekar, K. Rea, O. Chamian, F. Fan, S. Berezowska, S. Rottenberg, A. Madabhushi, et al., Patchsorter: a high throughput deep learning digital pathology tool for object labeling, Npj Digit. Med. 7 (2024) 164.

[13] V. Cheplygina, M. de Bruijne, J.P. Pluim, Not-so-supervised: a survey of semi-supervised, multi-instance, and transfer learning in medical image analysis, Med. Image Anal. 54 (2019) 280–296.

[14] J. Lee, E. Warner, S. Shaikhouni, M. Bitzer, M. Kretzler, D. Gipson, S. Pennathur, K. Bellovich, Z. Bhat, C. Gadegbeku, et al., Unsupervised machine learning for identifying important visual features through bag-of-words using histopathology data from chronic kidney disease, Sci. Rep. 12 (2022) 4832.

[15] N. Sato, E. Uchino, R. Kojima, M. Sakuragi, S. Hiragi, S. Minamiguchi, H. Haga, H. Yokoi, M. Yanagita, Y. Okuno, Evaluation of kidney histological images using unsupervised deep learning, Kidney Int. Rep. 6 (2021) 2445–2454.

[16] B. Liu, M. Polack, N. Coudray, A.C. Quiros, T. Sakellaropoulos, A.S. Crobach, J.H.J. van Krieken, K. Yuan, R.A. Tollenaar, W.E. Mesker, et al., Self-supervised learning reveals clinically relevant histomorphological patterns for therapeutic strategies in colon cancer, BioRxiv (2024) 2024–2002.

[17] H. Xu, L. Liu, X. Lei, M. Mandal, C. Lu, An unsupervised method for histological image segmentation based on tissue cluster level graph cut, Comput. Med. Imaging Graph. 93 (2021) 101974.

[18] X. Wang, S. Yang, J. Zhang, M. Wang, J. Zhang, W. Yang, J. Huang, X. Han, Transformer-based unsupervised contrastive learning for histopathological image classification, Med. Image Anal. 81 (2022) 102559.

[19] X. Wang, Y. Du, S. Yang, J. Zhang, M. Wang, J. Zhang, W. Yang, J. Huang, X. Han, Retccl: clustering-guided contrastive learning for whole-slide image retrieval, Med. Image Anal. 83 (2023) 102645.

[20] O. Ciga, T. Xu, A.L. Martel, Self supervised contrastive learning for digital histopathology, Mach. Learn. Appl. 7 (2022) 100198.

[21] E. Zimmermann, E. Vorontsov, J. Viret, A. Casson, M. Zelechowski, G. Shaikovski, N. Tenenholtz, J. Hall, D. Klimstra, R. Yousfi, et al., Virchow2: Scaling self-supervised mixed magnification models in pathology, 2024, arXiv preprint arXiv:2408.00738.

[22] M. Caron, P. Bojanowski, A. Joulin, M. Douze, Deep clustering for unsupervised learning of visual features, in: Proceedings of the European Conference on Computer Vision, ECCV, 2018, pp. 132–149.

[23] M. Caron, P. Bojanowski, J. Mairal, A. Joulin, Unsupervised pre-training of image features on non-curated data, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 2959–2968.

[24] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, A. Joulin, Emerging properties in self-supervised vision transformers, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 9650–9660.

[25] T. Chen, S. Kornblith, M. Norouzi, G. Hinton, A simple framework for contrastive learning of visual representations, in: International Conference on Machine Learning, PMLR, 2020, pp. 1597–1607.

[26] W. Chen, S. Pu, D. Xie, S. Yang, Y. Guo, L. Lin, Unsupervised image classification for deep representation learning, in: Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16, Springer, 2020, pp. 430–446.

[27] X. Chen, K. He, Exploring simple siamese representation learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 15750–15758.

[28] J. Zbontar, L. Jing, I. Misra, Y. LeCun, S. Deny, Barlow twins: Self-supervised learning via redundancy reduction, in: International Conference on Machine Learning, PMLR, 2021, pp. 12310–12320.

[29] L. Schmarje, M. Santarossa, S.-M. Schröder, R. Koch, A survey on semi-, self-and unsupervised learning for image classification, IEEE Access 9 (2021) 82146–82168.

[30] X. Liu, X. Zhu, X. Tian, T. Iwasaki, A. Sato, J.J. Kazama, Renal pathological image classification based on contrastive and transfer learning, Electronics 13 (2024) 1403.

[31] T. Yao, Y. Lu, J. Long, A. Jha, Z. Zhu, Z. Asad, H. Yang, A.B. Fogo, Y. Huo, Glo-in-one: holistic glomerular detection, segmentation, and lesion characterization with large-scale web image mining, J. Med. Imaging 9 (2022) 052408.

[32] K. Amann, C.S. Haas, What you should know about the work-up of a renal biopsy, Nephrol. Dial. Transplant. 21 (2006) 1157–1161.

[33] S. Sethi, M. Haas, G.S. Markowitz, V.D. D'Agati, H.G. Rennke, J.C. Jennette, I.M. Bajema, C.E. Alpers, A. Chang, L.D. Cornell, et al., Mayo clinic/renal pathology society consensus report on pathologic classification, diagnosis, and reporting of gn, J. Am. Soc. Nephrol. 27 (2016) 1278–1287.

[34] L. Barisoni, C.C. Nast, J.C. Jennette, J.B. Hodgin, A.M. Herzenberg, K.V. Lemley, C.M. Conway, J.B. Kopp, M. Kretzler, C. Lienczewski, et al., Digital pathology evaluation in the multicenter nephrotic syndrome study network (neptune), Clin. J. Am. Soc. Nephrol. 8 (2013) 1449–1459.

[35] I.H. de Boer, C.E. Alpers, E.U. Azeloglu, U.G. Balis, J.M. Barasch, L. Barisoni, K.N. Blank, A.S. Bomback, K. Brown, P.C. Dagher, et al., Rationale and design of the kidney precision medicine project, Kidney Int. 99 (2021) 498–510.

[36] G. Bueno, M.M. Fernandez-Carrobles, L. Gonzalez-Lopez, O. Deniz, Glomerulosclerosis identification in whole slide images using semantic segmentation, Comput. Methods Programs Biomed. 184 (2020) 105273.

[37] G. Bueno, L. Gonzalez-Lopez, M. Garcia-Rojo, A. Laurinavicius, O. Deniz, Data for glomeruli characterization in histopathological images, Data Brief 29 (2020) 105314.

[38] J. Gallego, Z. Swiderska-Chadaj, T. Markiewicz, M. Yamashita, M.A. Gabaldon, A. Gertych, A u-net based framework to quantify glomerulosclerosis in digitized pas and h & e stained human tissues, Comput. Med. Imaging Graph. 89 (2021) 101865.

[39] Z. Swiderska-Chadaj, J. Gallego, A. Gertych, Kidney glomeruli-rois extracted from histological slides stained with he or pas. (v1.0) [data set]. zenodo., 2020, http://dx.doi.org/10.5281/zenodo.4299694, (Online; Accessed 27 May 2024).

[40] M.P. Snyder, S. Lin, A. Posgai, M. Atkinson, A. Regev, J. Rood, O. Rozenblatt-Rosen, L. Gaffney, A. Hupalowska, R. Satija, et al., The human body at cellular resolution: the nih human biomolecular atlas program, Nature 574 (2019) 187–192.

[41] M. Macenko, M. Niethammer, J.S. Marron, D. Borland, J.T. Woosley, X. Guan, C. Schmitt, N.E. Thomas, A method for normalizing histology slides for quantitative analysis, in: 2009 IEEE international symposium on biomedical imaging: from nano to macro, IEEE, 2009, pp. 1107–1110.

[42] F. Chollet, et al., Keras, 2015, https://keras.io (Online; Accessed 30 August 2024).

[43] L. McInnes, J. Healy, J. Melville, Umap: Uniform manifold approximation and projection for dimension reduction. arxiv 2018, 2018, arXiv preprint arXiv:1802.03426.

[44] P.J. Rousseeuw, Silhouettes: a graphical aid to the interpretation and validation of cluster analysis, J. Comput. Appl. Math. 20 (1987) 53–65.

[45] L.J. Hubert, J.R. Levin, A general statistical framework for assessing categorical clustering in free recall., Psychol. Bull. 83 (1976) 1072.

[46] J.C. Dunn, Well-separated clusters and optimal fuzzy partitions, J. Cybern. 4 (1974) 95–104.

[47] V.A. Traag, L. Waltman, N.J. Van Eck, From louvain to leiden: guaranteeing well-connected communities, Sci. Rep. 9 (2019) 5233.

[48] L. Hubert, P. Arabie, Comparing partitions, J. Classification 2 (1985) 193–218.

[49] A.G. Howards, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, H. Adam, Mobilenets: efficient convolutional neural networks for mobile vision applications (2017), 2017, arXiv preprint arXiv:1704.04861 126.

[50] N. Altini, G.D. Cascarano, A. Brunetti, F. Marino, M.T. Rocchetti, S. Matino, U. Venere, M. Rossini, F. Pesce, L. Gesualdo, et al., Semantic segmentation framework for glomeruli detection and classification in kidney histological sections, Electronics 9 (2020) 503.

[51] Y. Lu, H. Yang, Z. Asad, Z. Zhu, T. Yao, J. Xu, A.B. Fogo, Y. Huo, Holistic fine-grained global glomerulosclerosis characterization: from detection to unbalanced classification, J. Med. Imaging 9 (2022) 014005.

[52] M.D. Hughson, K. Johnson, R.J. Young, W.E. Hoy, J.F. Bertram, Glomerular size and glomerulosclerosis: relationships to disease categories, glomerular solidification, and ischemic obsolescence, Am. J. Kidney Dis. 39 (2002) 679–688.

[53] O. Arbelaitz, I. Gurrutxaga, J. Muguerza, J.M. Pérez, I. Perona, An extensive comparative study of cluster validity indices, Pattern Recognit. 46 (2013) 243–256.

[54] L.E. Ekemeyong Awong, T. Zielinska, Comparative analysis of the clustering quality in self-organizing maps for human posture classification, Sensors 23 (2023) 7925.

[55] Y. Liu, Z. Li, H. Xiong, X. Gao, J. Wu, Understanding of internal clustering validation measures, in: 2010 IEEE International Conference on Data Mining, IEEE, 2010, pp. 911–916.

[56] S. Roy, A. kumar Jain, S. Lal, J. Kini, A study about color normalization methods for histopathology images, Micron 114 (2018) 42–61.

[57] M.Z. Hoque, A. Keskinarkaus, P. Nyberg, T. Seppänen, Stain normalization methods for histopathology image analysis: A comprehensive review and experimental comparison, Inf. Fusion (2023) 101997.

[58] B. Lutnick, R. Yacoub, K.-Y. Jen, J.E. Tomaszewski, S. Jain, P. Sarder, Deep variational auto-encoders for unsupervised glomerular classification, in: Medical Imaging 2018: Digital Pathology, vol. 10581, SPIE, 2018, pp. 88–94.