How to Cite this article: Matthew Basso, Moumita Barua, Rohan John, and April Khademi, Explainable Biomarkers for Automated Glomerular and Patient-Level Disease Classification, Kidney360, Publish Ahead of Print, 10.34067/KID.0005102021

Article Type:

Explainable Biomarkers for Automated Glomerular and Patient-Level Disease Classification

DOI: 10.34067/KID.0005102021

Matthew Basso, Moumita Barua, Rohan John, and April Khademi

Key Points:
*A software tool was developed to perform glomerular and patient-level classification based on clinically relevant biomarkers.
*Ten biomarkers were used for glomerular and patient-level classification that obtained 76.86% and 86.67% accuracies, respectively.
*In the future, these tools can be applied to clinical datasets for glomerular biomarker discovery and for insights into disease mechanisms.

Abstract:

Background Pathologists use multiple microscopy modalities to assess renal biopsies. Besides usual diagnostic features, some changes are too subtle to be properly defined. Computational approaches have potential to systematically quantitate subvisual clues, provide pathogenetic insight, and link to clinical outcomes. To this end, a proof of principle study is presented demonstrating that explainable biomarkers through machine learning can distinguish between glomerular disorders at the light microscopy level.

Methods The proposed system employed image analysis techniques and extracted 233 explainable biomarkers related to color, morphology, and microstructural texture. Traditional machine learning was then used to classify minimal change disease (MCD), membranous nephropathy (MN), and thin-basement membrane nephropathy (TBMN) diseases on a glomerular and patient-level basis.

Results The final model combined the Gini feature importance set and Linear Discriminant Analysis classifier. Six morphological (nuclei-to-glomerular tuft area, nuclei-to-glomerular area, glomerular tuft thickness > 10, glomerular tuft thickness > 3, total glomerular tuft thickness, and glomerular circularity) and four microstructural texture features (luminal contrast using wavelets, nuclei energy using wavelets, nuclei variance using color vector LBP, and glomerular correlation using GLCM) were together the best performing biomarkers. Accuracies of 76.86% and 86.67% were obtained for glomerular and patient-level classification, respectively.

Conclusion Computational methods using explainable glomerular biomarkers have diagnostic value and are compatible with our existing knowledge of disease pathogenesis. Furthermore, this algorithm can be applied to clinical datasets for novel prognostic and mechanistic biomarker discovery.

Disclosures: M. Barua reports the following: Ownership Interest: AstraZeneca Research Funding: Otsuka, Regulus, Sanofi; Honoraria: Natera; Scientific Advisor or Membership: Glomerular Diseases (publication) - Editorial Board. The remaining authors have nothing to disclose.

Funding: Faculty of Arts, Ryerson University: April Khademi; Alport Syndrome Foundation (ASF); Gouvernement du Canada | Canadian Institutes of Health Research (CIHR)

Author Contributions: Matthew Basso: Formal analysis; Investigation; Methodology; Validation; Visualization; Writing - original draft; Writing - review and editing Moumita Barua: Conceptualization; Data curation; Funding acquisition; Validation; Writing - review and editing Rohan John: Conceptualization; Data curation; Funding acquisition; Validation; Writing - review and editing April Khademi: Conceptualization; Funding acquisition; Project administration; Supervision; Validation; Writing - review and editing

Data Sharing Statement: Data cannot be shared: Data cannot be shared due to privacy and confidentiality of the patient-data.

Clinical Trials Registration:

Registration Number:
The information on this cover page is based on the most recent submission data from the authors. It may vary from the final published article. Any fields remaining blank are not applicable for this manuscript.
Explainable Biomarkers for Automated Glomerular and Patient-Level Disease Classification

Matthew Nicholas Basso ¹, Moumita Barua ²,³,⁴,⁵, Rohan John ⁶, April Khademi ¹,⁷,⁸

¹Image Analysis in Medicine Lab (IAMLAB), Department of Electrical, Computer, and Biomedical Engineering, Ryerson University, Toronto, Canada, ²Division of Nephrology, University Health Network, Toronto, Canada, ³Toronto General Hospital Research Institute, Toronto General Hospital, Toronto, Canada, ⁴Department of Medicine, University of Toronto, Toronto, Canada, ⁵Institute of Medical Sciences, University of Toronto, Toronto, Canada, ⁶Department of Pathology, University Health Network, Toronto, Canada, ⁷Keenan research Center for Biomedical Science, St. Michael’s Hospital, Unity Health Network, Toronto, Canada, ⁸Institute for Biomedical Engineering, Science, and Technology (iBEST), a partnership between St. Michael’s Hospital and Ryerson University, Toronto, Canada

Corresponding Author:

Name: Matthew Basso
Address: 24 Maryport Ave, Toronto, Ontario, Canada, M3M 2V9
E-mail address: mnbasso@ryerson.ca
Key Points

- A software tool was developed to perform glomerular and patient-level classification based on clinically relevant biomarkers.

- Ten biomarkers were used for glomerular and patient-level classification that obtained 76.86% and 86.67% accuracies, respectively.

- In the future, these tools can be applied to clinical datasets for glomerular biomarker discovery and for insights into disease mechanisms.
Abstract

Background Pathologists use multiple microscopy modalities to assess renal biopsies. Besides usual diagnostic features, some changes are too subtle to be properly defined. Computational approaches have the potential to systematically quantitate subvisual clues, provide pathogenetic insight, and link to clinical outcomes. To this end, a proof of principle study is presented demonstrating that explainable biomarkers through machine learning can distinguish between glomerular disorders at the light microscopy level.

Methods The proposed system employed image analysis techniques and extracted 233 explainable biomarkers related to color, morphology, and microstructural texture. Traditional machine learning was then used to classify minimal change disease (MCD), membranous nephropathy (MN), and thin-basement membrane nephropathy (TBMN) diseases on a glomerular and patient-level basis.

Results The final model combined the Gini feature importance set and Linear Discriminant Analysis classifier. Six morphological (nuclei-to-glomerular tuft area, nuclei-to-glomerular area, glomerular tuft thickness > 10, glomerular tuft thickness > 3, total glomerular tuft thickness, and glomerular circularity) and four microstructural texture features (luminal contrast using wavelets, nuclei energy using wavelets, nuclei variance using color vector LBP, and glomerular correlation using GLCM) were together the best performing biomarkers. Accuracies of $76.86\%$ and $86.67\%$ were obtained for glomerular and patient-level classification, respectively.

Conclusion Computational methods using explainable glomerular biomarkers have diagnostic value and are compatible with our existing knowledge of disease pathogenesis. Furthermore, this algorithm can be applied to clinical datasets for novel prognostic and mechanistic biomarker discovery.
Introduction

Glomerulonephritides (or glomerulopathies) are a group of rare kidney diseases characterized by injury to the glomerular filtration barrier, and with limited treatment options, often progress to end-stage kidney disease.\(^1\) Renal pathologists examine kidney biopsy tissue by routine light microscopy (LM), immunofluorescence (IF), and electron microscopy (EM), to diagnose, often descriptively, the many glomerular diseases. The disease names are well recognized by treating clinicians, although usually not offering much etiologic or pathogenic insight. Manual approaches are limited in not being able to glean or quantitate any potentially useful sub-visual features, leaving clinically informative pathological findings untapped. Digitization of whole-slide images (WSIs) and novel software tools, including traditional image analysis as well as machine learning can be used to uncover objective biomarkers, besides being tailored to augment pathology workflows. Biomarkers that are designed to be intuitive and explainable relating to the underlying pathology (tissue micro- and macro- structure) give the user confidence in the system and further enhance mechanistic understanding of disease.

Several reports have been published describing machine learning algorithms in mouse and human renal tissue.\(^2\text{-}^4\) Taken together, these approaches demonstrate that automated tools are feasible for renal pathology classification. For this proof of principle study, three glomerular disorders are used: minimal change disease (MCD), membranous nephropathy (MN), and thin-basement membrane nephropathy (TBMN). Though complex in overall pathology they are separated by somewhat simple histologic principles. MCD appears normal by LM, but shows podocyte foot process effacement, which can only be appreciated with EM. MN may show thick glomerular walls on LM, but immune complexes can only be recognized by IF and EM. TBMN shows diffuse thinning of the glomerular wall and can only be reliably identified on EM. MCD and MN are common causes of nephrotic syndrome, while TBMN is the major cause of isolated hematuria. Pathologists often suspect but cannot diagnose these diseases from LM alone. We reasoned that such suspicious features can be better defined and thus quantitated by computational approaches.
An image analysis and machine learning algorithm was designed to identify biomarkers that can be used to distinguish between these disorders using PAS-stained WSIs, which to the best of our knowledge, is novel in the field. An explainable biomarker panel of 233 features was designed with future opportunities to correlate with clinical outcomes, as well as expand to other glomerular diseases. The proposed method has four steps: (1) image preprocessing, (2) automatic glomerular structure segmentation, (3) biomarker feature extraction, and (4) glomerular and patient-level classification. Three machine learning classifiers were compared to pathologist derived assessment that combined three modalities (LM, IF, and EM) as the gold standard.

Methods

Data Preparation

The dataset is from Toronto General Hospital (TGH) in Toronto, Ontario, Canada and has institutional research board approval. This dataset consists of WSIs of renal biopsies from \( n = 45 \) different patients with \( n = 15 \) WSIs per disease (MCD, TBMN, and MN), derived from pathologist (R.J.) assessment using LM, IF, and EM. Biopsies were chosen from cases with classic disease features and without changes associated with other glomerular disease. One slide from each case stained by the standard Periodic acid-Schiff (PAS) method was used for analysis. Slides were scanned at 40x magnification with a resolution of 0.2526 x 0.2526 \( \mu m \). A total of 375 manually cropped regions of interest (ROIs) and glomerular boundary segmentations sized 1500 x 1500 pixels containing glomeruli were segmented using Pathcore’s Sedeen Image Viewer.\(^5\) Glomerular detection and segmentation were not considered in this work as the main focus was analyzing differences in diseases based on explainable biomarkers.

This dataset also has manual glomerular structure annotations for 150 ROIs (50 from each disease) which were used in two validations methods to evaluate the performance of the glomerular structure segmentation algorithm. Every nucleus was manually annotated from the selected glomeruli images. When annotating luminal space (Bowman’s space and capillary lumen) and glomerular tuft
(including GBM and mesangial matrix) structures, a similar annotation approach performed by B. Ginley et al. was used.\textsuperscript{4} K-means ground truths were generated and initial cluster center positions were chosen using the average luminal and glomerular tuft annotation intensities. Annotations were performed by a trained biomedical student (M.B.) and validated by a pathologist (R.J.) for quality control.

**Preprocessing**

Preprocessing was employed to prepare the data for biomarker measurement. Glomeruli can vary in size depending on the level at which the glomerulus sections are sampled.\textsuperscript{5} Glomerular size lower than 1.5 interquartile range (IQR) of Q1 was used to determine outlier glomeruli on a per slide basis and were removed from the analysis. Since specimen preparation can contribute to large variations in the quality of WSIs, color standardization was performed to facilitate texture and color feature comparisons. Each image was color normalized using a modified version of Reinhard's method for color normalization, which decreases color variability in WSIs.\textsuperscript{7-9}

**Glomerular Structure Segmentation**

Sub-glomerular structures were automatically segmented for feature extraction and analysis. Color normalization was not used for segmentation as there was a decrease in performance. Three structures were segmented: (1) luminal (space inside the Bowman’s capsule and the capillary lumen), (2) glomerular tuft (the glomerular basement membrane (GBM) and mesangial matrix), and (3) nuclei. For each structure, the ROIs were transformed into a color representation that was optimal for the given structure, followed by Otsu’s binary thresholding, and a 3x3 median filter to remove noise in the estimated segmented structures.

The estimated luminal, glomerular tuft, and nuclei masks obtained from the previous approaches were compared and five categories were found: (1) unlabeled pixels, (2) luminal pixels, (3) glomerular tuft pixels, (4) nuclei pixels, and (5) double-labeled pixels. A Naïve Bayesian classifier was implemented to determine the class of the unlabeled or double labeled pixels.\textsuperscript{4,10} This classifier was trained using the known pixel classes from the estimated structural masks. A prediction for the unlabeled or double labeled
pixels were found using the trained model.\textsuperscript{4,10} Final glomerular structure segmentation masks were collected for biomarker feature extraction.

**Biomarker Feature Extraction**

Image analysis tools were used to gather color, morphological, and microstructural texture features from glomeruli images forming a total of 233 biomarkers. The RGB color images were converted to the HSV color space, which is similar to the human perception of color.\textsuperscript{11} Color structures were analyzed using histogram mean, variance, skewness, kurtosis, energy, and entropy.\textsuperscript{12} These features describe the amount of structures present in the glomerulus by quantifying the relative hue (color), purity (saturation) and intensity (value) of the image.\textsuperscript{11}

Morphological features were extracted from the sub-glomerular structures to quantify shape and object-based characteristics. These features were organized into four groups: containment features, shape features, interstructural distance features, and intrastructural distance features.\textsuperscript{4} Containment features measure the fraction of one structure's area in comparison within another (e.g. nuclei area divided by glomerular area).\textsuperscript{4} Shape features such as equivalent diameter were computed for each structure and circularity was used to quantify the roundness of the glomerulus. Interstructural distance features were used to assess distance between glomerular structures and describe how structures interact with each other.\textsuperscript{4} These features were measured by finding the centroid of each glomerular component and finding the pairwise distance between two structures. Lastly, intrastructural distance features were used to measure the thickness of a structure (e.g. glomerular tuft max thickness).\textsuperscript{4} To measure thickness, the Euclidean distance transform operator was computed on each glomerular structure, yielding feature images that quantify spatial thickness. From these images, maximum, median, and total thickness features were extracted.

Microstructural texture features were designed to measure spatial relationships between color or gray level pixels and describe glomerular microstructure tissue texture. Local and global texture-based
biomarkers were evaluated using: gray-level co-occurrence matrices (GLCM), color vector local binary patterns (LBP), and wavelet features.\textsuperscript{13–16}

**Classification**

A traditional machine learning approach for glomerular classification was designed to classify glomeruli as either MCD, MN, or TBMN. The dataset was split into training, validation, and testing WSIs based on patients, and glomeruli were individually labeled using the WSI disease label. Five-fold cross-validation was used to examine which features were most important, followed by classifier hyperparameter fine-tuning. Using the optimal configuration for the glomerular classification tool, the held-out testing patients were classified on a patient-level basis.

**Feature Importance**

For each glomerular image in the training and validation set, feature selection was employed to examine which features were most discriminatory. The following four feature selection techniques were examined: using all features, statistical ANOVA F-value feature importance,\textsuperscript{17} Gini feature importance,\textsuperscript{18} and maximum relevance minimum redundancy (mRMR).\textsuperscript{19}

**Glomerular Classification**

Traditional machine learning classifiers were chosen for classification as they are more interpretable and require less training data. After selecting the most relevant features, the following three classifiers were analyzed: Linear Discriminant Analysis (LDA),\textsuperscript{20} Random Forest,\textsuperscript{12} and Logistic Regression.\textsuperscript{21} The output of each classifier is a form of a class conditional probability for each disease which was transformed into hard decisions by taking the maximum probability across disease groups. Glomerular classification performance metrics were then analyzed for all classifier methods.

**Patient-Level Classification Model**

To automatically predict the disease of a patient, each glomerulus from the WSI renal biopsy was first classified using the optimal feature set and model found previously. Three methods were investigated
when performing patient-level classification. In the first method, an average WSI disease diagnosis was represented from all glomeruli predictions. The probabilities found for each glomerulus on a WSI were averaged forming a confidence for each patient. The maximum disease confidence rating then corresponded to the predicted patient diagnosis. Similarly, the second method took the top four glomeruli with the highest probability and averaged them to get an estimated WSI disease diagnosis. In the last method examined, the glomerulus with the highest probability was used to determine the patient-level diagnosis. These methods were compared to how a pathologist visually inspects a patient’s WSI and on quantitative measures.

Performance Evaluation

The dice similarity coefficient (DSC) was used to measure the overlap between a segmented object and ground truth, while extra fraction (EF) was used to measure the false positive rate. The precision and recall were also used to measure the proportion of correctly segmented pixels, and the proportion of the ground truth pixels that were correctly identified by the predicted segmentation, respectively. To quantify glomerular and patient-level classification, accuracy, precision, recall, and F1-scores were investigated. Accuracy measures the fraction of correct predictions over the total number of predictions. F1-score is a combination of both precision and recall and gives an overall accuracy score. High F1-scores indicate the classifier is predicting with high precision and recall.

Results

Table 1 describes the dataset for glomerular and patient-level classification. The dataset was split by patient into 67% (n = 250 glomeruli, n = 30 WSIs) training/validation and 33% (n = 121 glomeruli, n = 15 WSIs) testing to ensure the same patients’ glomeruli were either in the training/validation dataset, or in the testing set with no overlap. Figure S1 illustrates sample WSI needle biopsies and Table S1 details additional clinical information such as age, sex, and disease specific information. The experimental design of the proposed system is shown in Figure 1.
Preprocessing

In total, four glomeruli were found to be outliers (lower than 1.5 IQR of Q1) reducing the TGH dataset from 375 to 371 glomeruli images. See Figure S2 for the distributions in glomerular size and Figure S3 for the images of removed glomeruli. Further analysis was conducted on the reduced set.

Glomerular Structure Segmentation Performance

Sub-glomerular structures were automatically segmented according to luminal, glomerular tuft, and nuclei structures. Sparse rectangular regions were annotated for luminal and glomerular tuft structures, as seen in Figure S4. First, manual annotations and automated segmentations were compared using DSC, EF, precision, and recall found in Table S2.

To further verify the segmentation performance, a semi-supervised k-means approach was used to develop ground truths comparable to gold standard annotations validated in Table S3. Visual results of the automated segmentation and respective k-means ground truths can be found in Figure S5. Figure 2 shows validation metric distributions (DSC, EF, precision, and recall) for each structure and the mean metrics are summarized in Table 2. All three structures had high mean DSC values (>0.80), with the luminal structure having the highest and the nuclei structure having the lowest mean DSC. When analyzing EF results, each structure had relatively low false positive rates. Lastly, for both precision and recall metrics, the segmentation model performed well over all structures. The average segmentation DSC over luminal, glomerular tuft, and nuclei components across all diseases was 0.893±0.057, indicating overall high agreement between all structures and ground truths.

Biomarker Feature Extraction

Exploratory analysis on the 233 biomarkers is visualized in Figure 3 and sample glomeruli are shown in Figure 3A. From the luminal, glomerular tuft, and nuclei structures, average proportions of each structure over all data were found for each respective disease and shown in Figure 3B. Important disease phenotypes were observed as MCD had the highest proportion of nuclei, TBMN had the highest proportion of lumen, and MN had the highest proportion of glomerular tuft, in relation to glomerular area.
These observations reflect what pathologists observe but cannot necessarily quantify as MCD has larger podocyte nuclei reflecting hypertrophy, TBMN has thin glomerular walls (GBM thinning), and MN has diffuse thickening of the GBM causing increased area of the glomerular tuft. From the hue color histogram shown in Figure 3C the TBMN glomerulus had higher hue mean indicating increased luminal structure. Figure 3D illustrates the intrastructural distance feature for the glomerular tuft structure. Visual results in zoomed-in regions show GBM in yellow (high values) for MN indicating slight thickening, while the GBM is dark blue (low values) for TBMN which suggests thinning. Lastly, Figure 3E illustrates the color vector LBP texture map images which quantify ultrastructural spatial relationships between pixels. As can be seen, MCD and TBMN are finer in texture than MN, which is likely from increase in GBM and overlapping mesangial matrix in MN.

Classification Performance

Performance evaluation and hyperparameter tuning was completed on the glomerular training/validation set through five-fold cross-validation. Patient-level classification was then performed on the held-out test set using the optimal model.

Feature Importance

The following four feature selection algorithms were used to reduce the feature set to small subsets of biomarkers: all features (233 features), statistical ANOVA F-value feature importance (10 features), Gini feature importance (10 features), and mRMR (10 features). The top ten features were selected for improved interpretability. See Table S6 for the top selected features from all approaches, along with all features.

Glomerular Classification Performance

Using the four feature sets, three different machine learning classifiers were applied to the glomeruli validation and testing set. Five-fold cross-validation was performed on the validation set for all feature sets and classifiers seen in Table 3. These results indicate that the Gini feature importance set and LDA classifier had the highest performance in cross-validation accuracy of 67.6±8.9%.
Using the GFI-LDA model, glomerular classification performance was measured on the held-out testing set of \( n = 121 \) glomeruli images. Accuracy, precision, recall, and F1-score performance metrics are described in Table 4 resulting in a testing accuracy of 76.86\% and F1-score of 76.47. Table S4 highlights the glomerular confusion matrix and how automated predictions are comparable to the gold standard disease labels. Although classification accuracy is high for MN, lower performance for MCD and TBMN stems from these diseases having more similar appearance under LM.

**Patient-Level Classification**

Using the GFI-LDA model, patient-level classification was performed on the held-out testing set of WSIs \((n = 15\) patients). The performance of the three patient-level classification methods examined is shown in Table 5. All three methods resulted in similar performance with all glomeruli and the top four glomeruli resulting in accuracies and F1-scores of 86.76\% and 85.94 respectively. Using all glomeruli for patient-level classification had the lowest confidence across all patients seen in Figure S6. The method using the top glomerulus was more susceptible to misclassification than the others since only one sample is taken in the WSI. When scanning visually through a slide, a pathologist may look for a few glomeruli to make diagnostic inferences. The top four glomeruli for patient-level classification can therefore mimic how the pathologist would analyze a biopsy. Therefore, this was the method chosen for patient-level classification. As shown in Table S5, this classification model correctly predicted 100\% of the MN WSIs, while predicting 80\% and 75\% of the MCD and TBMN WSIs, respectively. This indicates the model had difficulty differentiating between TBMN and MCD even on the patient-level.

Figure 4 illustrates the confidence rating for each of the held-out patients with respect to disease. All patients except 8 and 43, were correctly classified, where patient 8 was predicted as MN (TBMN ground truth) and patient 43 predicted TBMN (MCD ground truth). Correctly classified and misclassified WSI are shown in Figure 5. Figure 5A shows all glomeruli on the WSI were correctly predicted as MN, with a confidence of 99.65\%. Figure 5A1-A4 shows four glomeruli with high probabilities for MN (> 0.98). In Figure 5B, the WSI label was predicted incorrectly as TBMN but was truly MCD with a
confidence of 62.01%. Figure 5B1-B4 shows the glomeruli with the highest probabilities (>72%). Three out of the four glomeruli were incorrectly predicted as TBMN, and one correctly predicted as MCD.

**Glomerular Biomarker Analysis**

The ten biomarkers selected by the GFI algorithm, are visualized with distributions across diseases in Figure 6. Four microstructural texture and six morphological features were selected as the best to discriminate between the pathologies. Color features were not represented (although two color features were selected when performing mRMR feature selection (Table S5)). This indicates color biomarkers were not as discriminative compared to the morphological and texture counterparts.

The interpretation of the four microstructural texture features will be examined here. Namely, there were nuclei energy using wavelets, nuclei variance using color vector LBP, luminal contrast using wavelets, and glomerular correlation using GLCM. The two nuclei features were based on the color vector LBP and the wavelet energy. The mean color vector LBP was higher for TBMN, and lowest for MN, with MCD in the middle. LBP looks for repeating patterns of lines and edges, and a higher variance in these features indicates there are similar (and repeating) patterns in the objects that are being investigated. Therefore, this feature suggests there are more consistent texture patterns between nuclei in TBMN. The wavelet energy examines the magnitude and prevalence of multiscale edges in the image. If there are many, high contrast edges in the nuclei, this feature will be low. In the analysis, we found the mean wavelet energy of the nuclei to be highest in MN, followed by MCD and TBMN. While these nuclei texture features are not clinically reported, these features may provide mechanistic insights into differences between pathologies which can be a future avenue of investigation regarding disease etiologies and differences. The GLCM correlation biomarker was higher for TBMN compared to MN and MCD. GLCM correlation measures the number of rapid changes of intensity in objects, where homogenous regions would have larger correlation values. Since there are large, continuous luminal regions in TBMN this may explain the larger correlation value for this disease type. The variance seen in
MN could relate to changes caused by irregular subepithelial deposits, influencing glomerular tuft structures. The last feature was the luminal contrast using wavelets. This feature looks at the variation in edge magnitudes in images. Luminal contrast was higher for TBMN, and lowest for MN, with MCD in the middle. Since TBMN has thinner membranes that are on the boundaries and interwoven within the luminal region, there is likely higher and more edge content that is reflected by the luminal contrast metric. MN was found to have less edge content and was more homogeneous in texture. Since MCD looks normal on the light microscopy level, it is interesting that the average texture distribution lies between the two basement membrane diseases.

The glomerular structure proportions (Figure 3B) indicated key pathological findings that were expressed through the morphological features. MCD had the highest proportion of nuclei compared to the glomerular area. This is likely due to the fact that MCD has podocyte nuclear enlargement. TBMN had the highest proportion of luminal area to glomeruli area which can be attributed to thin glomerular walls. Lastly, the mean glomerular tuft composition was higher for MN, and lowest for TBMN, with MCD in the middle. This is seen clinically as MN has thicker glomerular walls, TBMN has thin glomerular walls, and MCD appears normal on LM. The top morphological features were ratio of nuclei-to-glomerular tuft area, ratio of nuclei-to-glomerular area, glomerular tuft thickness > 10, glomerular tuft thickness > 3, total glomerular tuft thickness, and glomerular circularity.

One-way ANOVA and pairwise Tukey post-hoc test was performed to check for any significant differences between the top 10 biomarker means across each disease with a confidence interval of 95%. Table 6 shows the statistical results of the selected biomarkers (in order of GFI importance). All biomarkers were statistically significant across group means. Tukey’s post hoc testing revealed that most features were significant between MCD-MN and MN-TBMN pairs; however, only nuclei-glomerular ratio, luminal contrast using wavelets, and glomerular circularity were significant between MCD-TBMN.
Discussion

A computer-aided diagnosis system applied to PAS images for MCD, MN, and TBMN classification is presented that integrates preprocessing, glomerular structure segmentation, biomarker extraction, and glomerular and patient-level classification. Results showed that from 233 explainable biomarkers, six morphological and four microstructural texture features are enough to obtain high glomerular and patient-level classification accuracy. The top system combined the Gini feature importance and Linear Discriminant Analysis classifier with an accuracy of 67.6±8.9% for the glomeruli cross-validation, and an accuracy and F1-score of, 76.86% and 76.47. For patient-level classification, the model had an accuracy and F1-score of 86.67% and 85.94, respectively.

The interaction between physicians and computer aided systems is important as little or too much trust in a system can impact patient diagnosis. Therefore, the proposed confidence rating can be used to encourage more trust in the computer generated results. Biopsies from subjects in Figure 4 and Figure 5 show that patient-level decisions based on misclassified glomeruli within the same biopsy often had a lower prediction confidence. These cases can be flagged for secondary review through traditional IF/EM approaches. The use of biologically relevant biomarker features is a key advantage in comparison to deep learning methods. Six morphological (ratio of nuclei-to-glomerular tuft area, ratio of nuclei-to-glomerular area, glomerular tuft thickness > 10, glomerular tuft thickness > 3, total glomerular tuft thickness, and glomerular circularity) and four microstructural texture features (luminal contrast using wavelets, nuclei energy using wavelets, nuclei variance using color vector LBP, and glomerular correlation using GLCM) were selected as the best performers to discriminate between the glomerular disorders. The selected morphological features demonstrate that MCD had larger nuclei reflecting podocyte hypertrophy, MN had glomerular tuft thickening, and TBMN had glomerular tuft thinning. Microstructural texture features described TBMN and MCD to be heterogeneous (nuclei and luminal structures), while MN was homogeneous (glomerular tuft thickening). For expanded applications, potentially using more of the 233 biomarker features would be relevant.
Our study has some limitations. The dataset had only three pathologic lesions and in the future, we will apply these tools to other glomerular diseases such as focal and segmental glomerulosclerosis. In addition, the cohort was obtained from a single center and was a small sample size which we hope to bolster in future studies that include more diseases. The intent of this work was to determine whether explainable features from PAS-only biopsies, which are routinely used in clinical practice, are sufficient for classification, and we believe we have satisfied those goals. Further analysis into other stains and pathologic modalities, i.e. IF and EM, could improve biomarker feature selection and analysis.

In conclusion, our work reveals that image analysis algorithms applied to glomerular diseases can quantify biomarkers that are compatible with our existing knowledge of pathogenesis. In the future, these tools can be applied on larger datasets with other glomerular diseases to quantitate subvisual features to seek linkages with clinical outcomes for biomarker discovery and for insights into disease mechanisms.
Disclosures
M. Barua reports the following: Ownership Interest: AstraZeneca Research Funding: Otsuka, Regulus, Sanofi; Honoraria: Natera; Scientific Advisor or Membership: Glomerular Diseases (publication) - Editorial Board. The remaining authors have nothing to disclose.

Funding
Faculty of Arts, Ryerson University: April Khademi; Alport Syndrome Foundation (ASF); Gouvernement du Canada | Canadian Institutes of Health Research (CIHR).

Acknowledgements
We would like to acknowledge Ryerson University’s Dean’s Research Fund (DRF) programs for funding this research. Additionally, we thank KRESCENT, Alport Syndrome Foundation, McLaughlin Centre-University of Toronto, NephCure Kidney International-Neptune, Can-SOLVE CKD Network, the Toronto General Hospital Foundation and CIHR for their support.

Author Contributions
M. Basso performed the formal analysis, investigation, methodology, validation, and writing of the original draft; M. Barua and R. John curated the data and acquired funding; A. Khademi was responsible for the supervision of the study and acquired funding; and all authors conceptualized the study, validated the results, and reviewed and edited the manuscript.

Data Sharing Statement
Data cannot be shared: Data cannot be shared due to privacy and confidentiality of the patient-data.
Supplemental Material

Clinical Information
Figure S1. Sample kidney needle biopsies from TGH dataset. (A) MCD, (B) MN, and (C) TBMN biopsy images.
Table S1. Additional clinical information from the TGH dataset.

Preprocessing: Glomeruli Size Variability
Figure S2. Glomerular size distribution on a patient basis. Outlier glomeruli were found on patients 11, 22, 25, and 27.
Figure S3. Glomeruli that were removed from analysis after performing glomeruli size outlier analysis.

Glomerular Structure Segmentation Performance
Figure S4. Sample manual segmentation ground truths for each disease class and structure.
Table S2. Mean DSC, EF, Precision, and Recall for predicted glomerular structure segmentation with respect to glomerular disease when compared to manual ground truths.
Table S3. Mean DSC, EF, Precision, and Recall for k-means glomerular structure segmentation with respect to glomerular disease when compared to manual ground truths.
Figure S5. Sample segmentation results for each disease class with respect to k-means ground truths.

Classification Performance: Glomerular & Patient-Level
Table S4. Confusion matrix containing the number of correctly classified glomerular images as either minimal change disease (MCD), membranous nephropathy (MN), or thin-basement membrane nephropathy (TBMN).
Table S5. Confusion matrix containing the number of correctly classified WSI’s as minimal change disease (MCD), membranous nephropathy (MN), or thin-basement membrane nephropathy (TBMN).
Figure S6. A comparison between the three patient-level classification methods; all glomeruli, top 4 glomeruli, and top glomerulus for the held-out test set.

Biomarker Feature Extraction: Feature Sets
Table S6. Biomarker features according to feature group (color, morphological, and microstructural texture).
References


13. Ae AK, Krishnan S. Shift-invariant discrete wavelet transform analysis for retinal image classification. doi:10.1007/s11517-007-0273-z


Tables

Table 1. Data configuration for glomeruli and patient-level classification.

<table>
<thead>
<tr>
<th></th>
<th>TGH Dataset</th>
<th>Training/Validation Set (67%)</th>
<th>Testing Set (33%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Patients</td>
<td>Glomeruli</td>
<td>Patients</td>
</tr>
<tr>
<td>MCD</td>
<td>15</td>
<td>103</td>
<td>10</td>
</tr>
<tr>
<td>MN</td>
<td>15</td>
<td>148</td>
<td>9</td>
</tr>
<tr>
<td>TBMN</td>
<td>15</td>
<td>124</td>
<td>11</td>
</tr>
<tr>
<td>Total</td>
<td>45</td>
<td>375</td>
<td>30</td>
</tr>
</tbody>
</table>

TGH dataset was composed of minimal change disease (MCD), membranous nephropathy (MN), and thin-basement membrane nephropathy (TBMN) WSIs. This dataset was split into training/validation and testing on a per patient basis. Glomeruli and patient-level classification was trained/validated and tested using the following configuration.
<table>
<thead>
<tr>
<th>Segmentation Performance Metrics</th>
<th>MCD</th>
<th>MN</th>
<th>TBMN</th>
</tr>
</thead>
<tbody>
<tr>
<td>DSC</td>
<td>0.890 ± 0.056</td>
<td><strong>0.897 ± 0.058</strong></td>
<td>0.893 ± 0.057</td>
</tr>
<tr>
<td>EF</td>
<td>0.123 ± 0.106</td>
<td><strong>0.120 ± 0.108</strong></td>
<td>0.126 ± 0.116</td>
</tr>
<tr>
<td>Precision</td>
<td>0.890 ± 0.086</td>
<td><strong>0.893 ± 0.089</strong></td>
<td>0.888 ± 0.092</td>
</tr>
<tr>
<td>Recall</td>
<td>0.900 ± 0.075</td>
<td><strong>0.909 ± 0.073</strong></td>
<td>0.906 ± 0.070</td>
</tr>
</tbody>
</table>

The dice similarity coefficient (DSC) measures the overlap between the segmented object and ground truth while extra fraction measures the false positive rate. Precision and recall measure the proportion of correctly segmented pixels relative to the ground truth and the proportion of the ground truth pixels that were correctly identified by the predicted segmentation. Bolded values indicate highest segmentation performance.
Table 3. Average classification accuracy from five-fold cross-validation for feature selection and classifier combinations.

<table>
<thead>
<tr>
<th>Feature Selection Method</th>
<th>LDA</th>
<th>RF</th>
<th>LR</th>
</tr>
</thead>
<tbody>
<tr>
<td>All Features</td>
<td>36.0±9.1</td>
<td>63.6±13.2</td>
<td>57.6±37.4</td>
</tr>
<tr>
<td>ANOVA</td>
<td>64.8±19.2</td>
<td>64.0±5.7</td>
<td>63.2±15.3</td>
</tr>
<tr>
<td>GFI</td>
<td>67.6±8.9</td>
<td>66.8±12.0</td>
<td>64.8±30.7</td>
</tr>
<tr>
<td>mRMR</td>
<td>66.8±10.9</td>
<td>66.8±13.3</td>
<td>61.2±23.5</td>
</tr>
</tbody>
</table>

The four feature selection methods analyzed were: all features, ANOVA F-value (ANOVA), Gini feature importance (GFI), and maximum relevance minimum redundancy (mRMR). The three classifiers analyzed were Linear Discriminant Analysis (LDA), Random Forest (RF), and Logistic Regression (LR). Bolded values indicate the model that achieved the best cross-validation performance.
Table 4. Glomerular classification performance for top feature sets and classifiers on the held-out test set.

<table>
<thead>
<tr>
<th>Classification Metrics</th>
<th>Feature Selection Method</th>
<th>Glomerular Classification</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>LDA</td>
</tr>
<tr>
<td>Accuracy</td>
<td>All Features</td>
<td>60.33</td>
</tr>
<tr>
<td></td>
<td>ANOVA</td>
<td>64.46</td>
</tr>
<tr>
<td></td>
<td>GFI</td>
<td><strong>76.86</strong></td>
</tr>
<tr>
<td></td>
<td>mRMR</td>
<td>60.33</td>
</tr>
<tr>
<td>Precision</td>
<td>All Features</td>
<td>58.34</td>
</tr>
<tr>
<td></td>
<td>ANOVA</td>
<td>64.03</td>
</tr>
<tr>
<td></td>
<td>GFI</td>
<td>76.55</td>
</tr>
<tr>
<td></td>
<td>mRMR</td>
<td>58.52</td>
</tr>
<tr>
<td>Recall</td>
<td>All Features</td>
<td>59.64</td>
</tr>
<tr>
<td></td>
<td>ANOVA</td>
<td>62.94</td>
</tr>
<tr>
<td></td>
<td>GFI</td>
<td><strong>76.40</strong></td>
</tr>
<tr>
<td></td>
<td>mRMR</td>
<td>58.91</td>
</tr>
<tr>
<td>F1-Score</td>
<td>All Features</td>
<td>58.98</td>
</tr>
<tr>
<td></td>
<td>ANOVA</td>
<td>63.48</td>
</tr>
<tr>
<td></td>
<td>GFI</td>
<td><strong>76.47</strong></td>
</tr>
<tr>
<td></td>
<td>mRMR</td>
<td>58.71</td>
</tr>
</tbody>
</table>

The four feature selection methods analyzed were: all features, ANOVA F-value (ANOVA), Gini feature importance (GFI), and maximum relevance minimum redundancy (mRMR). The three classifiers analyzed were: Linear Discriminant Analysis (LDA), Random Forest (RF), and Logistic Regression (LR). Accuracy, precision, recall, and F1-score classification performance metrics were evaluated. Bolded values indicate the model that achieved the best classification performance.
Table 5. Performance of three patient-level classification methods on the held-out test set.

<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy</th>
<th>Precision</th>
<th>Recall</th>
<th>F1-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>All glomeruli</td>
<td>86.67</td>
<td>86.9</td>
<td>85</td>
<td>85.94</td>
</tr>
<tr>
<td>Top 4 glomeruli</td>
<td>86.67</td>
<td>86.9</td>
<td>85</td>
<td>85.94</td>
</tr>
<tr>
<td>Top glomerulus</td>
<td>80</td>
<td>81.9</td>
<td>78.33</td>
<td>80.08</td>
</tr>
</tbody>
</table>

The three patient-level classification methods analyzed were using all glomeruli, top four glomeruli, and top glomerulus. Accuracy, precision, recall, and F1-score classification performance metrics were evaluated. Bolded values indicate the model that achieved the best classification performance.
Table 6. Biomarkers sorted by Gini feature importance (GFI) with one-way ANOVA and Tukey’s post-hoc statistical analysis across biomarker feature groups.

<table>
<thead>
<tr>
<th>Gini Importance</th>
<th>Feature</th>
<th>$F$</th>
<th>$Pr &gt; F_{crit}$</th>
<th>MCD vs MN</th>
<th>MCD vs TBMN</th>
<th>MN vs TBMN</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0350</td>
<td>Nuclei-Glomerular Tuft Ratio</td>
<td>125.78</td>
<td>2.36e-42</td>
<td><strong>0.001</strong></td>
<td>0.3716</td>
<td><strong>0.001</strong></td>
</tr>
<tr>
<td>0.0198</td>
<td>Nuclei-Glomerular Ratio</td>
<td>71.15</td>
<td>7.54e-27</td>
<td><strong>0.001</strong></td>
<td><strong>0.001</strong></td>
<td><strong>0.001</strong></td>
</tr>
<tr>
<td>0.0162</td>
<td>Glomerular Tuft Thickness &gt;10</td>
<td>37.16</td>
<td>2.00e-15</td>
<td><strong>0.001</strong></td>
<td>0.4521</td>
<td><strong>0.001</strong></td>
</tr>
<tr>
<td>0.0157</td>
<td>Wavelet: Luminal Contrast</td>
<td>27.50</td>
<td>7.39e-12</td>
<td><strong>0.0088</strong></td>
<td><strong>0.001</strong></td>
<td><strong>0.001</strong></td>
</tr>
<tr>
<td>0.0138</td>
<td>Wavelet: Nuclei Energy</td>
<td>39.56</td>
<td>2.74e-16</td>
<td><strong>0.001</strong></td>
<td>0.6407</td>
<td><strong>0.001</strong></td>
</tr>
<tr>
<td>0.0122</td>
<td>Glomerular Tuft Thickness &gt;3</td>
<td>63.12</td>
<td>2.71e-24</td>
<td><strong>0.001</strong></td>
<td>0.3075</td>
<td><strong>0.001</strong></td>
</tr>
<tr>
<td>0.0115</td>
<td>Color Vector LBP: Nuclei Variance</td>
<td>31.53</td>
<td>2.30e-13</td>
<td><strong>0.001</strong></td>
<td>0.5403</td>
<td><strong>0.001</strong></td>
</tr>
<tr>
<td>0.0103</td>
<td>Glomerular Tuft Thickness</td>
<td>59.81</td>
<td>3.22e-23</td>
<td><strong>0.001</strong></td>
<td>0.3446</td>
<td><strong>0.001</strong></td>
</tr>
<tr>
<td>0.0098</td>
<td>GLCM: Glomerular Correlation</td>
<td>11.02</td>
<td>2.26e-05</td>
<td>0.0849</td>
<td>0.0642</td>
<td><strong>0.001</strong></td>
</tr>
<tr>
<td>0.0097</td>
<td>Glomerular Circularity</td>
<td>21.26</td>
<td>1.84e-09</td>
<td><strong>0.0048</strong></td>
<td><strong>0.001</strong></td>
<td><strong>0.001</strong></td>
</tr>
</tbody>
</table>

Bolded values indicate disease group distributions means are significant and $p$-value is less than 0.05.
Figures and Figure Legends

Figure 1. Overview of our experimental design.

Figure 2. Automated glomerular segmentation performance compared to k-means ground truths. (A) Dice similarity coefficient (DSC) measures overlap between segmented object and ground truth, (B) extra fraction (EF) measures false positive rate, (C) precision measures the proportion of correctly segmented pixels relative to the ground truth, and (D) recall measures the proportion of the ground truth pixels that were correctly identified by the predicted segmentation.

Figure 3. Visual representation of biomarker features extracted. (A) Columns represent sample minimal change disease (MCD), membranous nephropathy (MN), and thin-basement membrane nephropathy (TBMN) diseases. (B) Bar and scatter plots visualizing glomerular structure proportions according to structure and disease. (C) Color features: displays the hue histogram for the following three sample images, with their corresponding mean values. (D) Morphological features: glomerular tuft intrastructural distance feature maps for each corresponding sample image. Thicker structures are represented as red or orange in color, while thinner structures are green and blue in color. (E) Microstructural texture features: texture maps for sample glomeruli using color vector LBP. Scale bars, 100 µm.

Figure 4. Patient-level confidence results per testing patient are shown. Each patient was predicted with a certain confidence corresponding to thin-basement membrane nephropathy (TBMN), membranous nephropathy (MN), and minimal change disease (MCD). The four highest glomerular probabilities were averaged and then used to get a patient-level confidence. The disease with the largest confidence determined the patients predicted diagnosis. Above the confidence bars indicates whether the patient was predicted correctly with a checkmark and incorrectly predicted with an X.

Figure 5. Correctly classified and misclassified patient WSI, and glomeruli with the highest probabilities. (A) Correctly predicted patient 16 as membranous nephropathy (MN) with 99.65% confidence. (A1-A4) Top four glomeruli ROIs with highest probabilities from (A). (B) Patient 43 misclassified as thin-basement membrane nephropathy (TBMN) while truly diagnosed as minimal change disease (MCD) with 62.01% confidence. (B1-B4) Top four glomeruli ROIs with highest probabilities from (B).

Figure 6. Top four microstructural texture features and top six morphological feature and their respective disease group distributions corresponding to thin-basement membrane nephropathy (TBMN), membranous nephropathy (MN), and minimal change disease (MCD).
Figure 6

- Nuclei-GT Ratio
- Nuclei Glomerular Ratio
- GT Thickness>10
  - Wavelet: Luminal Contrast
  - Wavelet: Nuclei Energy
- GT Thickness>3
  - CV-LBP: Nuclei Variance
- GT Total Thickness
  - GLCM: Glomerular Correlation
  - Glomerular Circularity