Diabetic Complications Consortium

Application Title: Computational Imaging of Renal Structures for Diagnosing Diabetic Nephropathy

Principal Investigator: Pinaki Sarder, PhD

Project Accomplishments:

We have developed extensive computational methods to segment glomerular and tubular features from renal tissue histology images, specifically brightfield microscopy image data of DN patients. We also have conducted the first study investigating whether the quantified image features in DN and other renal biopsies have any correlation with disease outcome. The preliminary result obtained from this award was essential to secure a 5-year R01 grant from NIDDK.

Specific Aims:

Specific Aim 1: Quantification of glomerular microstructure from histologically prepared diabetic nephropathy (DN) renal tissue images

Results: We have developed extensive computational strategies for segmenting glomerular micro-compartments from histologically stained renal tissue images diagnosed with DN or control tissue images with no apparent histological abnormalities. The generalized nature of these computational methods allows us to develop extended functionality to segment interstitial fibrosis and tubular atrophy (IFTA) features in DN renal tissue images. Below we briefly describe the developed computational methods and key results.

Human Artificial Intelligence Loop based Segmentation of Renal Micro-Structures: Neural networks promise to bring robust quantitative analysis to medical fields, but adoption is limited by the technicalities of training these networks. To address the translation gap between medical researchers and neural networks in the field of pathology, we have created an intuitive interface, utilizing the commonly used whole slide image (WSI) viewer, Aperio ImageScope (Leica Biosystems Imaging, Inc.), for the annotation and display of neural network predictions on WSIs. Our platform uses a human-in-the-loop strategy to reduce the burden of WSI annotation. To explore the functionality of this pipeline, we track network performance improvements as a function of iteration and quantify segmentation performance of histologic findings on WSIs. The human artificial intelligence loop (H-AI-L) method is outlined in Fig. 1. As a first study we have analyzed the performance of H-AI-L for segmentation of renal micro-compartments, as well as demonstrated multi-class segmentation in human renal tissue (Fig. 2).

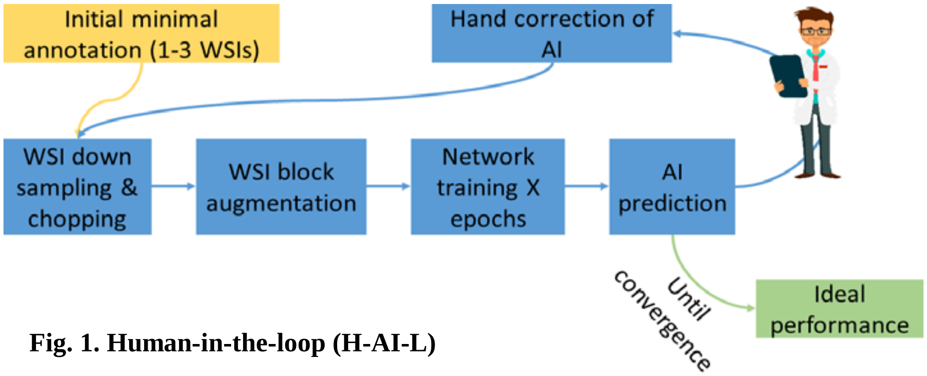
Initial minimal annotation (1-3 WSIs)
Hand correction of AI
WSI down sampling & chopping
WSI block augmentation
Network training X epochs
AI prediction
Until convergence
Ideal performance

Fig. 1. Human-in-the-loop (H-AI-L)

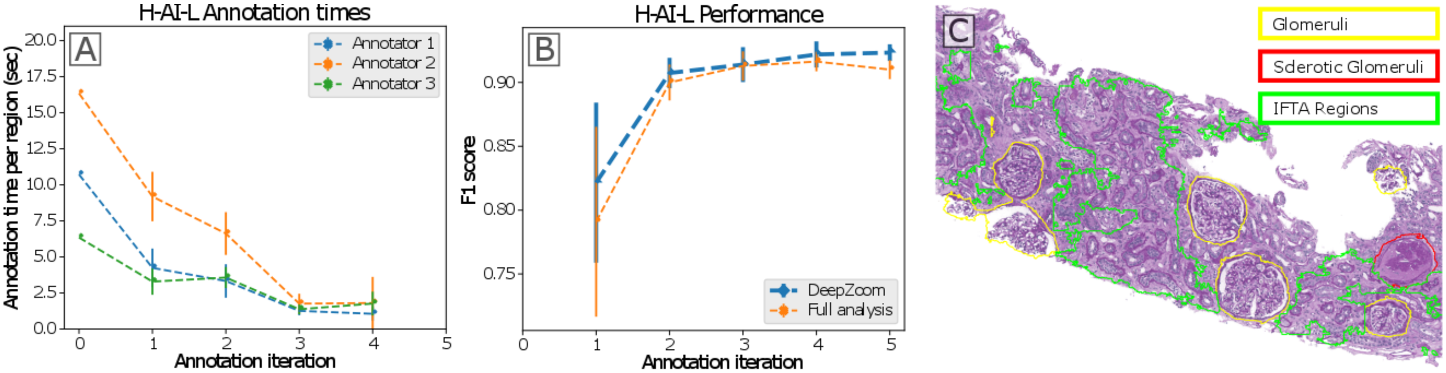

Fig. 2. H-AI-L pipeline performance. (A) Annotation times per glomerulus in a renal WSI as a function of annotation iteration for three annotators. The 0th iteration was performed without preexisting predicted annotations. Subsequent iterations use network predictions as an initial annotation prediction that can be corrected by the annotator. (B) F1 score of glomerular segmentation as a function of training iteration using two different versions of H-AI-L, DeepZoom and Full analysis, details of which are omitted for brevity. (C) Multiclass segmentation in human DN renal WSI.

The PI and the team heavily use the developed H-AI-L pipeline for segmenting images of renal tissues diagnosed with DN. A journal article on this topic is in review for publication in Nature Machine Intelligence while writing this report. This unpublished article is submitted as 'SupplementaryDocumentA.pdf' with this report.

<u>Glomerular Localization in DN Renal Biopsies:</u> While the H-AI-L method proposed above can reproducibly locate glomeruli and segment the respective boundaries in DN renal biopsies, we have developed another alternative method for localizing glomeruli in WSIs of renal tissue sections marked with histological stains (Fig. 3A-B). Multi-radial color local binary pattern features extracted from glomerular and non-glomerular regions are used to train a support vector machine (SVM), deployed in tandem with a deep convolutional neural network trained for glomerular recognition. Precision (positive predictive value), recall, and F1 score for five different histological stains (2 WSIs per stain) were computed to be 0.98, 0.64, and 0.76, respectively (Fig. 3C). The F1 score measures accuracy; it is 1 at best precision and recall and worst at 0. Fig. 3D shows performance for 5 patients with DN (five WSIs) and 3 patients with normal glomeruli (nine WSIs). Precision, recall, and F1 score for both cases were ~0.9, ~0.76, and ~0.83, respectively. Using 5 Intel(R) Core(TM) i7-4790 CPUs with 40 GB RAM, the method takes ~2 min to extract glomeruli from a biopsy WSI (~10^6 pixels). Publication # 1 below was generated based on this work.

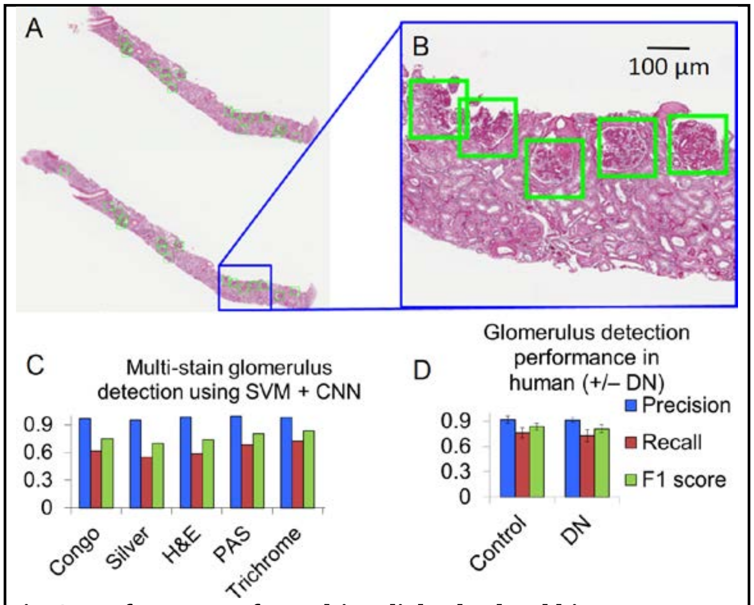

Fig. 3. Performance of a multi-radial color local binary pattern (mrcLBP) feature extraction method to localize glomerular locations in renal tissue images. (A-B) Human DN renal biopsy images with automatically detected glomeruli locations (green squares), using mrcLBP feature vector to train a support vector machine (SVM). (C) Precision, recall, and F1 score of glomeruli detection using renal tissue sections stained with five different histological stains. An integrated approach comprising SVM and convolutional neural network (CNN) models in series was used. (D) Performance of the method on human samples with and without DN.

<u>Glomerular Boundary Segmentation:</u> <u>*Method 1:*</u> Figs. 4A-D describe our method to segment glomerular boundaries from murine renal histology images using a combination of Gabor filter bank, statistical F-testing, and spatial weighting. This method can segment rat glomerular boundaries, compared to a resident pathologist

in our department, with sensitivity/specificity ~0.88/0.96 on n = 1000 glomeruli images, for 5 different histological stains, at 5X faster speed. This result was generated prior to the DiaComp award, and we include this result here for completeness; see Ginley *et al., JMI-SPIE*, 2017 and see Ginley *et al., Proc. of SPIE–Medical Imaging 2017: Digital Pathology*. Method 2: We discovered that tissue auto-fluorescence (AF) facilitates glomerular boundary segmentation (Figs. 4E-H). The glomerulus does not undergo AF as compared to the tubules, thus providing a distinction in glomerular intensity levels compared to background when the normalized AF image is subtracted from 1 (Fig. 4F). The tubular AF signal is found to be uniform and emits light with lower spatial frequency than the signal coming from intra-glomerular regions in the AF image. We exploited this variation in AF signal using a band-pass filter to extract the glomerular boundary (Fig. 4H). We detect glomeruli with a sensitivity/specificity 0.92/0.86 on n = 40 images, each containing one or more murine glomeruli. Publication # 2 below was generated based on this work.

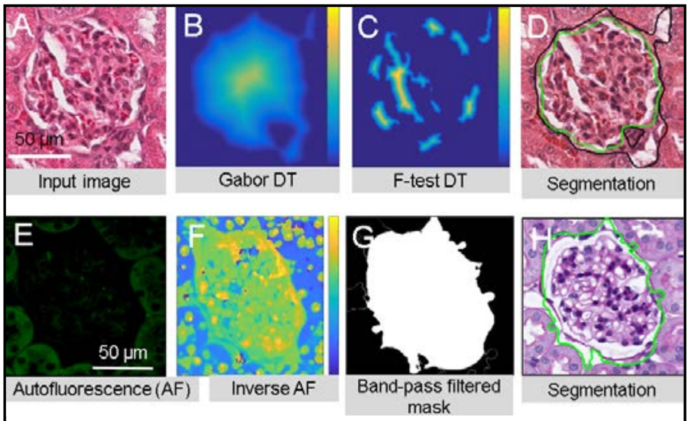

Fig. 4. Glomerular boundary segmentation. (A-D) Integrated method using Gabor filter and statistical F-testing. (A) Histological image of a murine glomerulus, (B) Gabor filtered output after distance transform (DT), (C) Foreground pixels in the grayscaled glomerulus image at the high intensity regions of the image in Fig. 4B were compared with the background using an F-test, and the DT of the output was computed. (D) Thresholding Fig. 4B offers the glomerular boundary (black line) and thresholding averaged intensity of 4B and 4C defines refined glomerular boundary (green line). (E-H) Tissue autofluorescence based method. (E) Autofluorescence image of glomerulus and surrounding tubule. (F) Inverse intensity image of Fig. 4E. (G) Mask generated upon band-pass filtering the image in Fig. 4F. (H) Glomerular boundary generated from Fig. 4G overlaid on the histological image of the glomerular tissue shown in Fig. 4E.

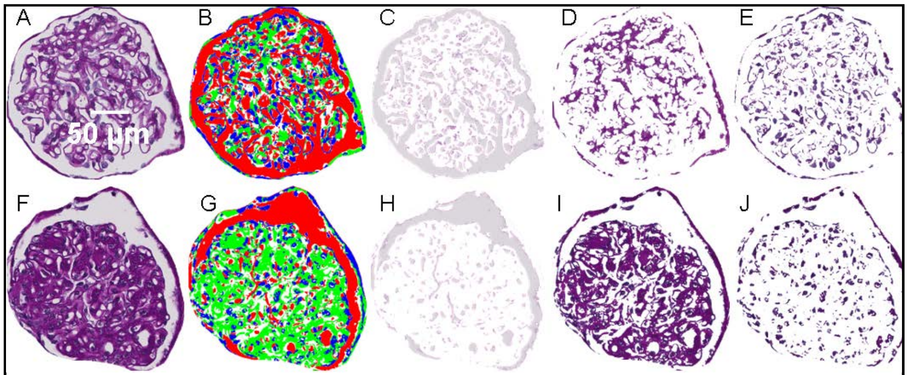

Fig. 5. Multi-scale segmentation of glomeruli from human renal biopsies. (A, F) Human glomeruli images: (A) Normal/healthy, (F) With moderate DN. (B, G) Multi-scale segmentation of glomeruli shown in (A) and (F), respectively. Colors represent different compartments. From (B) & (G), respective segments corresponding to Bowman's & luminal spaces (C & H), mesangial space (D & I), and nuclei (E & J).

Intra-Glomerular Compartment Segmentation: We developed an efficient method to simultaneously separate glomerular structures from histological images in multi-scale without using any parameters or training the computer. Segmentation is achieved by solving an energy optimization problem. Representing the image as a graph, nodes (pixels) are grouped by minimizing a Potts model Hamiltonian function, adopted from theoretical physics, modeling interacting electron spins. Pixel relationships (modeled as edges) are used to update the energy of the partitioned graph. By iteratively improving the clustering, the optimal number of segments is revealed. By tuning sensitivity to image background, a resolution metric reveals pertinent structures in multi-scale. With this method, segmentation of images with 10^6 pixels requires 5 sec. Fig. 5 shows the segmentation result of all glomerular

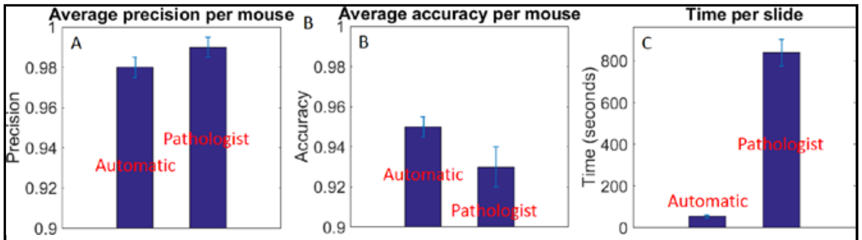

Fig. 6. Performance comparison between automatic and manual methods in segmenting glomerular compartments in murine renal tissues. (A) Precision, (B) accuracy, and (C) the analysis time. Error-bars for precision and accuracy metrics indicate standard deviation (SD) across mice. Error-bars for evaluation time indicates SD across slides.

structural scales, including Bowman's and luminal spaces, mesangial area, and nuclei, from healthy and moderate DN human renal glomeruli. We analyzed the glomerular compartmentalization performance discussed here using renal tissue histology images from male wild-type C57BL/6J mice. Renal tissue sections from 3 mice were histologically stained and imaged. Images were analyzed using our glomerular compartmentalization method (Fig. 5). 10 randomly selected glomeruli images from each mouse were used. Ground-truth compartments were generated by co-I Dr. Tomaszewski. We compared the performance of the automatic computation against combined masks generated from manual segmentations of one resident pathologist in the Pathology & Anatomical Sciences department and our collaborator Dr. Rabi Yacoub (Medicine – Nephrology division, University at Buffalo). Fig. 6 compares the precision, accuracy, and evaluation time of the automatic and manual methods. Average precision and accuracy metrics were computed across mice and time was noted per slide. Via two-sample t-test, we found that automation provides significantly faster speed and reproducibility in glomerular compartmentalization, while offering precision and accuracy comparable to the manual method. Partial result of the glomerular compartmentalization presented here was discussed in the DiaComp grant proposal submitted for funding. The *Journal article in revision* # 1 is a result of this work, which is submitted as 'SupplementaryDocumentB.pdf' with this report. Lutnick *et al.*, *Proc. of SPIE–Medical Imaging 2017: Digital Pathology*, provides partial results of this work, which was completed prior to the DiaComp award.

<u>Glomerular Compartmentalization (Alternative Method) and Feature Extraction:</u> This method extracts glomerular features from whole slide DN renal biopsies. Glomerular localization and segmentation are done using the CNN method described above. Fig. 7 describes our pipeline. A journal article is in preparation for publication in *Journal of American Society in Nephrology* on this work, and the current version of the manuscript is submitted as 'SupplementaryDocumentC.pdf' with this report. Nuclei segmentation from segmented glomeruli is done via a deep convolutional neural network (CNN) approach; see 'SupplementaryDocumentC.pdf'. In general, our network was able to achieve high performance for nuclear segmentation in human control data (Table 1). For glomerular nuclei of human DN cases, our network demonstrated moderate sensitivity at 0.79. The reason for this finding is that the human annotator tended to over-segment during training set generation while the network tended to under-segment the nuclear boundaries, creating a persistent bias in performance analysis. However, the precision of the network to identify nuclei was 0.99 on average, which was sufficient to analyze nuclear structure within the scope of this study. Comparable high network performance of nuclear segmentation was seen for test images prepared and stained at different institutions (Fig. 8), which illustrates the robustness of our network in terms of overcoming variations between laboratories. Nuclear predictions on a human glomerulus are also shown in Fig. 7C. Note that quantification of subsequent features restricted nuclei using the glomerular boundary to exclude tubular nuclei.

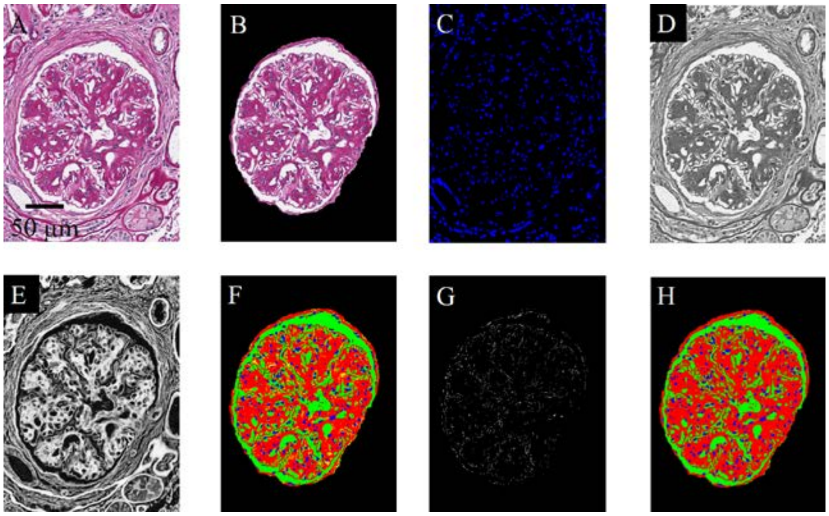

Fig. 7. Pipeline for glomerular compartment segmentation. A. Example image of PAS stained human glomerulus. B. Segmented glomerular boundary. C. CNN segmentation of nuclei. D. Grayscale image depicting the lightness component of $L^*a^*b^*$ color space, delineating luminal spaces. E. Grayscale image depicting stain deconvolution for PAS components, delineating mesangium and

basement membranes. F. Preliminary compartment segmentation generated by CNN segmentation of nuclei and global thresholding of D and E. A naïve Bayesian classifier is trained using these pixels. G. Pixels from F which do not yet have a label. The naïve Bayesian classifier predicts these labels. H. Final, complete segmentation of all three compartments after naïve Bayesian segmentation correction.

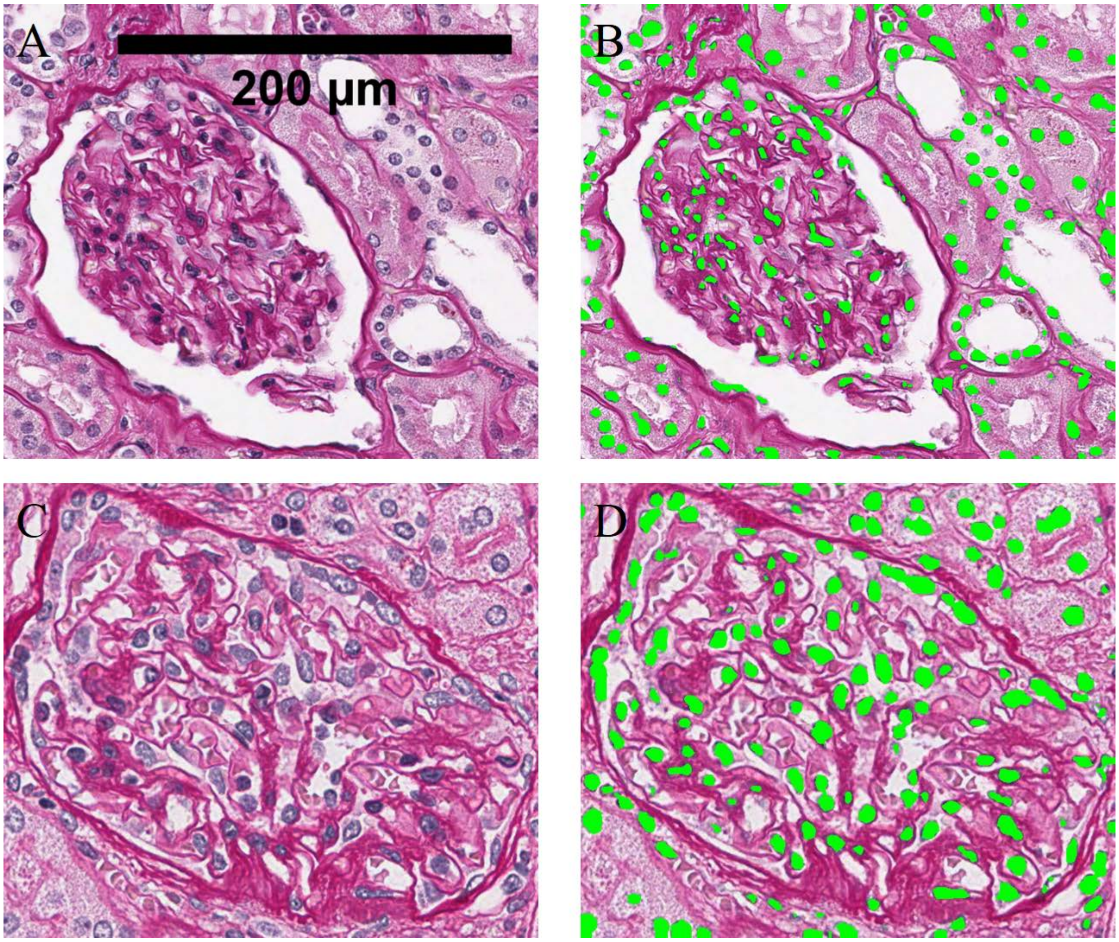

Fig. 8. Nuclear segmentation by deep CNN. A. Example of a human glomerulus. B. Segmentation of nuclei from image in A by deep CNN. C. Glomerulus image which was prepared in a separate institute than glomerulus from image A. D. Segmentation of nuclei from C.

Luminal and Periodic acid-Schiff positive (PAS+) regions are identified in two steps, first in a rough preliminary fashion using unsupervised thresholding, afterwards in a final fashion with a naïve Bayesian classification approach. These two methods combined complement each other: thresholding techniques are fast and unsupervised, while relatively imprecise; naïve Bayesian classification is capable of achieving high precision when there is a clear measurable difference in the class distributions, but needs training data. This approach makes it very convenient to identify various components of glomeruli which are similar in color with high performance, and is adaptive to slight shifts in stain variation (the naïve Bayesian model is re-trained on each image). Fig. 7D shows the lightness (*L**) component of *L*a*b** color space. *L*a*b** transformation is designed to be a color space which is perceptually uniform to human color vision. The lightness component, as it suggests, transforms pixel values such that the brightest whites have highest value and darkest blacks have lowest value. Otsu's thresholding is a technique to automatically determine image foreground from background by maximizing inter-class variance. Thresholding of the lightness component generates preliminary segmentation masks of luminal spaces. Fig. 7E shows stain deconvolution for PAS+ components (mesangium, basement membranes, capsule). Stain deconvolution is a technique to rotate an image's color space axes so that they are along the directions of the stain color. The PAS deconvolved image is also thresholded with Otsu's method to create a preliminary segmentation mask of PAS+ objects. The combined preliminary segmentation masks are shown in Fig. 7F. CNN nuclear predictions are shown in blue, preliminary PAS+ components in red, and luminal components in green. Fig. 7G shows unlabeled pixels from the preliminary segmentation. The

PAS+ component and luminal component labels are used to train a naïve Bayesian classifier to predict the unlabeled pixels, into either the PAS+ class or the luminal class, and results in a final segmentation; see Fig. 7H. The performance analysis of the glomerular compartmentalization is shown in Table 1.

Table 1. Performance of glomerular sub-compartmentalization.

Compartment	Sensitivity	Specificity	PPV	NPV	MCC
Control human PAS+	0.98 ± 0.02	1	1	0.96	0.97
Control human lumen	0.99 ± 0.01	1	1	0.94	0.96
Control human nuclei	0.76 ± 0.08	1	1	0.98 ± 0.01	0.87 ± 0.05
Disease human PAS+	0.99	0.99	0.987 ± 0.08	0.988 ± 0.04	0.98
Disease human lumen	0.99	1	1	0.95 ± 0.1	0.96
Disease human nuclei	0.79 ± 0.1	1	0.99	1	0.88 ± 0.06

Based on our glomerular compartment analysis of DN biopsies, a set of computational features were extracted to describe the pathological structural progression of glomeruli in DN. These image features are based on texture, morphology, intra-compartmental distance, and glomerular structural conformation. The 51 extracted features are listed in Table 2.

Table 2. List of quantified glomerular features.

Feature No.	***Distance features***
1	Average distance of lumina center from glomerular center
2	Averaged average distance between lumina and glomerular boundaries
3	Average maximum distance between lumina and glomerular boundaries
4	Average minimum distance between lumina and glomerular boundaries
5	Averaged average distance between luminal regions
6	Average maximum distance between luminal regions
7	Average minimum distance between luminal regions
8	Average distance of PAS+ from glomerular center
9	Averaged average distance of PAS+ from glomerular boundaries
10	Average maximum distance of PAS+ from glomerular boundaries
11	Average minimum distance of PAS+ from glomerular boundaries
12	Averaged average distance between PAS+ regions
13	Average maximum distance between PAS+ regions
14	Average minimum distance between PAS+ regions
15	Average distance of nuclei from glomerular center
16	Averaged average distance of nuclei from glomerular boundaries
17	Average maximum distance of nuclei from glomerular boundaries
18	Average minimum distance of nuclei from glomerular boundaries
19	Averaged average distance between nuclei
20	Average maximum distance between nuclei
21	Average minimum distance between nuclei
	Containment features
22	Average PAS+ area contained in convex luminal boundary
23	Average nuclear area contained in convex luminal boundary
24	Average luminal area contained in convex PAS+ boundaries
25	Average nuclear area contained in convex PAS+ boundaries
26	Average nuclear overlap with lumina
27	Average nuclear overlap with PAS+
Feature No.	***Texture features***
28	Nuclear gray-level spatially dependent contrast
29	Nuclear gray-level spatially dependent correlation
30	Nuclear gray-level spatially dependent energy

31	Nuclear gray-level spatially dependent homogeneity
32	Luminal gray-level spatially dependent contrast
33	Luminal gray-level spatially dependent correlation
34	Luminal gray-level spatially dependent energy
35	Luminal gray-level spatially dependent homogeneity
36	PAS+ gray-level spatially dependent contrast
37	PAS+ gray-level spatially dependent correlation
38	PAS+ gray-level spatially dependent energy
39	PAS+ gray-level spatially dependent homogeneity
	Morphological features
40	Average convexity of lumina
41	Sum total area of luminal space
42	Mean area of luminal spaces
43	Median area of luminal spaces
44	Average convexity of PAS+ components
45	Sum total area of PAS+ components
46	Mean area of PAS+ components
47	Median area of PAS+ components
48	Sum total nuclear area
49	Mean nuclear areas
50	Mode nuclear areas
51	Total glomerular area

Textural features were computed in aggregate for each glomerulus based on the glomerular compartment. For example, the total Bowman and capillary luminal space within a single glomerulus was analyzed as one unit comprising the luminal compartment. Glomerular compartment-specific textural descriptions identified as gray level entropy, energy, correlation, and homogeneity were extracted from the respective matrices. Morphological features were calculated per individual compartment object. Summary statistics were taken along the glomerulus dimension, e.g., mean nuclear area refers to the mean area of all nuclei in a particular glomerulus. These features included the mean, median, and mode of areas, and the average convexity for identified compartmental objects. Compartmental containment features define the amount of one compartment contained within the boundaries of another. Specifically, it is a ratio, where one part is the convex area of the containing compartment object, and the other part is the area of the contained compartment. Distance features are comprised of averaged distances between compartments and other identical glomerular compartments, or glomerular landmarks. Glomerular landmarks include the estimated glomerular centroid and the estimated glomerular boundary points. The following distance features are extracted for each object of each glomerulus: 1) the distance between that object's centroid and identically labeled objects' centroids, 2) the average distance to the glomerular boundary, and 3) the distance to the glomerular centroid. A total of 51 features were extracted from a total of $n = 613$ human glomeruli obtained from DN biopsies.

We next used a neighborhood component analysis (NCA) to compare the relative usefulness of the hand-crafted structural features in describing structural progression of glomeruli. NCA is a method for selecting features which maximize the prediction accuracy of classification models. It was discovered that only 16 of the derived features were useful in classifying the DN stage of human glomeruli.

Singular value decomposition (SVD) was performed on the original 51 features to reduce the feature dimensionality and correlation. This can help improve classification by removing extraneous information. It was found that only 25 singular vectors were needed to account for 99% of variance in the original feature space. These compressed features were used to train a naïve Bayesian classifier to classify DN structural states. All classifiers used 50% of data as holdout and 50% as training data. The performance of each classifier to make a binary decision between disease states is shown in Table 3. Further, it can be concluded that stage IIb is generally the most difficult to identify, as the performance scores are notably lower for this classes comparisons. This is likely because the class distinction is based on whether or not the mesangial area appears to exceed the mean area of the capillary lumen, which is somewhat subjective dependent on the observer. It can also be noted that the further away two stages are from each other, the more easily they can be classified (e.g., stage IIa is much easier to classify when compared against stage IV than stage IIb).

Table 3. Performance of mouse and human structural stage classification.

Classes compared	Human, optimized	
	Specificity	Sensitivity
I-IIa	1	1
I-IIb	0.6528	0.996
I-III	0.8281	0.996
I-IV	0.8634	0.9919
IIa-IIb	1	0.8085
IIa-III	0.9219	0.7872
IIa-IV	0.9508	0.9574
IIb-III	0.75	0.5833
IIb-IV	0.9508	0.9028
III-IV	0.918	0.8906

Graph Based Glomerular Features Describing DN Pathology: Fig. 9 shows glomerular nuclear change in murine DN from a graph-theoretical standpoint, using minimal spanning trees (MSTs) and distance matrices. A MST is a graph network where points of the network are joined using the shortest total connection distance. MSTs have already proven useful for characterization of subtle structural changes in diverse fields. Using MSTs, we studied minimum inter-nuclear edge distances and full nuclear connectivity within glomeruli for DN and healthy cases. Fig. 9 suggests that DN is marked by a subtle but significant *decrease* in the *average minimum distance* between nuclear bodies and a slight *increase* in the *average inter-nuclear distance*. These effects can be explained by the progressive hypertrophy and proliferation of mesangial cells coupled with sclerosis and reduction of the capillary endothelium. Publication # 3 was generated based on this result.

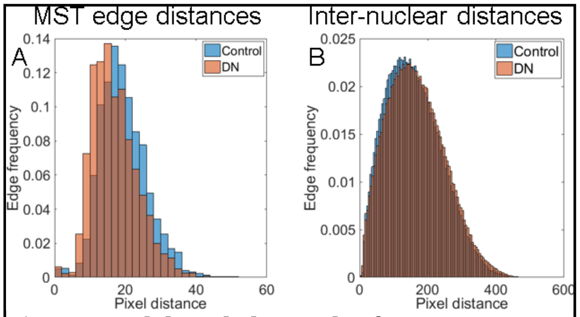

Fig. 9. Graph based glomerular feature quantification, using nuclei as nodes and inter-nuclear distance as edges, for human control glomeruli (*n* = 20) and DN glomeruli (*n* = 20). Edge distributions obtained using (A) minimum spanning trees of all the glomerular nuclei, and (B) full graphs formed by the glomerular nuclei.

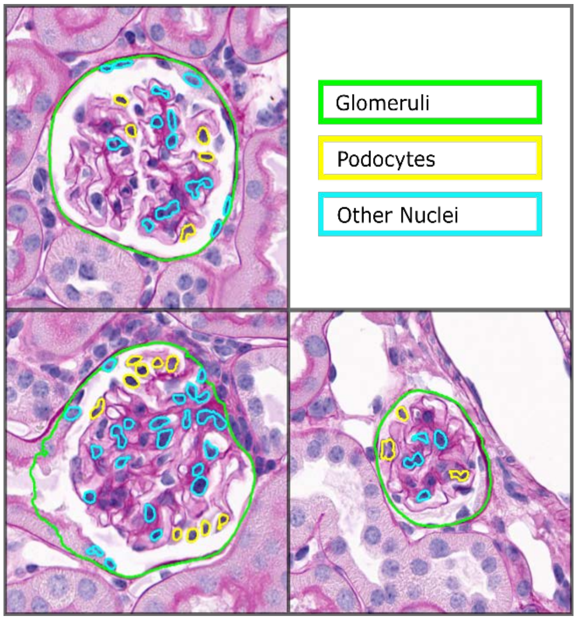

Fig. 10. Multiclass nuclei prediction on mouse WSI. Several examples of multiclass nuclei predictions are visualized on a mouse WSI. Computation was done using the H-AI-L method described above.

Podocyte Counting in Renal Tissue: Loss of podocytes indicates DN progression, and thus is an important pathobiological marker of DN. Extending the H-AI-L pipeline described above (see 'SupplementaryDocumentA.pdf'), we have developed a computational method to count all the podocytes from whole slide images of renal tissues. A training set was generated via immunofluorescence (IF) labeling of podocytes in renal tissues. In this study, we post-stain the same tissue sample using Periodic acid-Schiff and counterstain the sample using hematoxylin (PAS-H). Brightfield microscopy imaging of the resultant histological image is used for development of the podocyte detection algorithm. Fig. 10 shows our preliminary podocyte counting result for murine renal tissues. Here training was done using 5 registered WSI images (IF and brightfield images) for 5 epochs. The dataset was augmented 5X via random color shifting, flipping, and piecewise affine transformations. Podocytes, non-podocyte nuclei, and glomeruli boundaries were predicted using the above H-AI-L platform. We are now generating more training datasets for this project, and are continuing to optimize the method using murine renal tissue samples. Upon accomplishing this task, we will apply the method in human DN renal biopsies for performance evaluation.

Segmentation and Quantification of Interstitial Fibrosis and Tubular Atrophy: Upon accomplishing segmentation tasks of important renal glomerular micro-compartments, we extended our computational pipeline to renal tubules. There are several histological changes in the tubular region of DN renal biopsies, one of the most important of which is the abnormal accumulation of extracellular matrix material within the interstitium. This change is coupled with the progressive reduction of tubular cross-sectional area termed tubular atrophy. Together, these changes are commonly known as interstitial fibrosis and tubular atrophy (IFTA). IFTA is typically assessed by scanning a histologically stained biopsy at low resolution, and determining the percentage of the cortical regions within the biopsy that display IFTA. This process can be tedious, imprecise, and subject to intra/inter-rater reliability. Modern digital image analysis algorithms have the ability to mitigate these issues by increasing precision and speed of analysis, and reducing manual labor. We have implemented a convolutional neural network approach for the segmentation of IFTA.

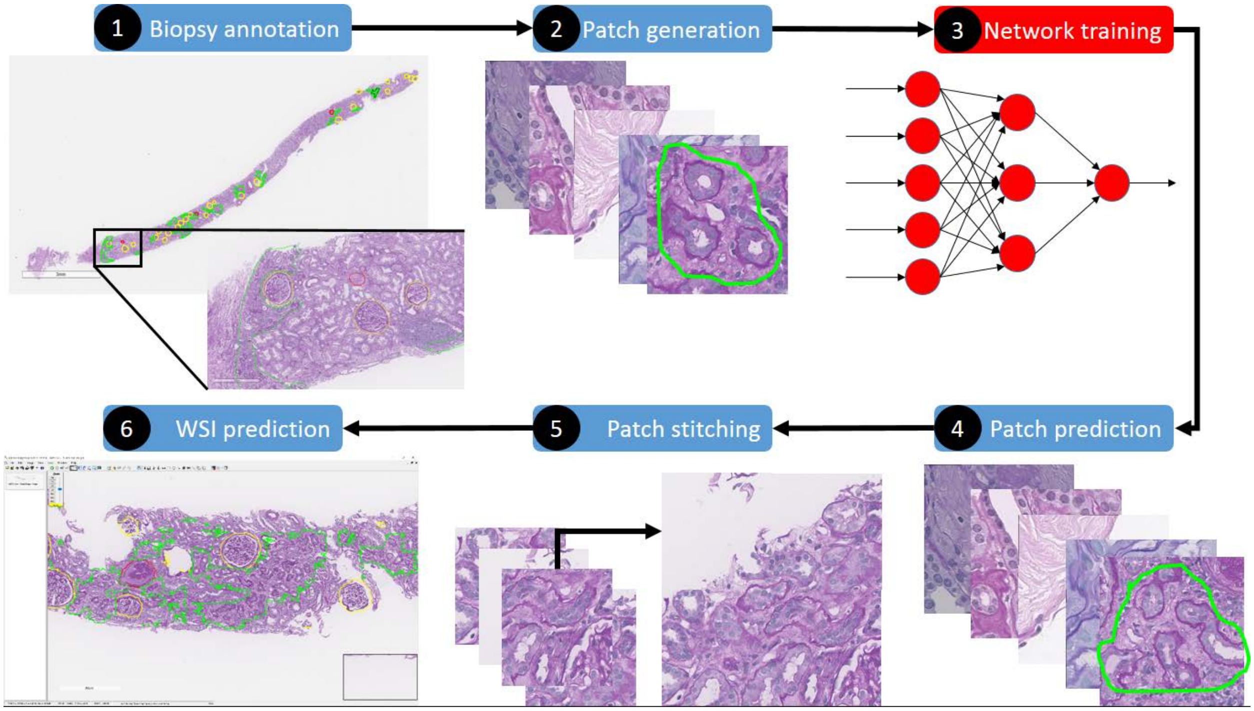

Fig. 11. Schematic of the proposed IFTA segmentation pipeline

Schematic of the proposed IFTA segmentation pipeline is provided in Fig. 11. WSIs of histologically stained renal biopsies are annotated for IFTA by Collaborator Dr. Kuang-Yu Jen (Pathology, UC Davis). Random patches are generated from the annotated images. The annotated regions are used to create IFTA masks corresponding to the patch images. Together the renal tissue histology image patches and the masks are used to train a fully connected convolutional neural network. The trained network segments the IFTA location in the patches, which are stitched back to obtain the renal biopsy WSI with segmented IFTA overlaid on top.

In the IFTA prediction work, we are also aiming to “open the black box” in order to visualize, understand, and analyze the image regions which neural networks utilize to make decisions. As a preliminary analysis, we have extracted the first layer of convolutional parameters and filtered image regions using these learned kernels. Below in Figure 12, on the left, is an image region of IFTA. In the middle, the convolved output of the original image and a kernel of the CNN is displayed, which potentially targets tubular basement membranes. On the right is the output after convolution of the original image with another learned filter. It appears that this particular filter may target regions of interstitial fibrosis.

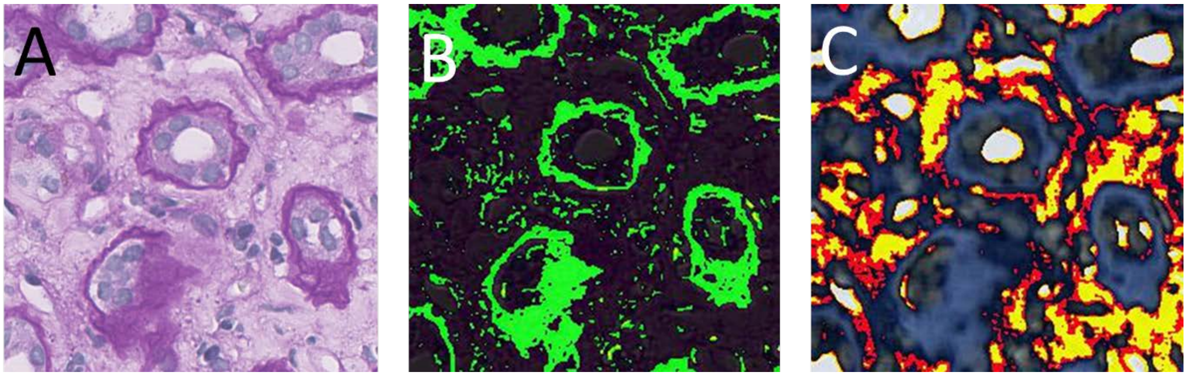

Fig. 12. CNN filters' outputs in IFTA segmentation. (A) An image patch of renal biopsy with IFTA. (B) One learned CNN filter output, shown in green, depicting tubular basement membranes. (C) Another learned CNN filter output, shown using RGB colormap, depicting interstitial fibrosis.

Our current work uses 25 renal biopsy WSIs for training and 10 WSIs for testing. IFTA segmentation sensitivity/specificity of our pipeline is currently 0.87/0.98. We are working on performance evaluation of IFTA segmentation in larger DN renal biopsy datasets, and a journal article is in preparation for publication in *Journal of American Society in Nephrology* on this work.

<u>Histological and Infrared Microscopy Data Fusion in Diabetic Nephropathy:</u> We investigated infrared microscopy to measure biochemical alternations in DN murine renal tissues. Since biochemical alterations precede structural changes (Varma *et al.*, Kidney International, 2016), IR microscopy tools can also be potentially used for early DN diagnosis. Our goal is to fuse the biochemical information with histological image information, and investigate if the fused dataset has better predictive power of disease progression than histological image data alone. We have optimized the protocol of IR imaging in our facilities while writing this report. Namely, IR imaging of tissues requires optimizing tissue thickness, the deparaffinization protocol, and the microscopy imaging set-up. Below we discuss our preliminary study in this direction.

Optimization of tissue thickness: Two adjacent formalin fixed paraffin embedded (FFPE) tissue sections were mounted on a barium fluoride (BaF_2) substrate. Two such slides were prepared, each with a different tissue thickness – 5 µm and 8 µm. The ideal thickness of tissue for Fourier transform infrared (FTIR) microscopy is documented to be between 5 and 10 µm (Varma *et al.*, Kidney International, 2016). The 5 µm thickness resulted in a poor signal-to-noise ratio (SNR) image. We hypothesize that this low signal is due to lack of sufficient tissue thickness for the IR waves to penetrate. However, the 8 µm tissues had high SNR with respect to the background and thus, for this study, we proceed with tissues of 8 micron thickness.

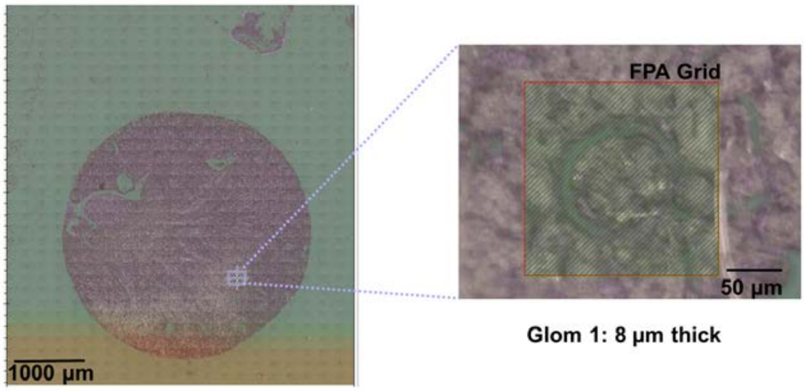

Fig. 13. Selection of glomerulus for analysis. Five glomeruli were manually selected for the FTIR analysis.

Deparaffinization: The tissues were deparaffinized by heating at 60 ^{0}C in an oven for 10 mins, followed by xylene immersion for 5 minutes at room temperature. This process was repeated twice. The tissues were then immersed in acetone for 5 minutes at room temperature and were subsequently allowed to air-dry.

FTIR microscopy imaging: Fourier transform infrared spectra of these tissues were collected in reflectance mode with gold background using a Hyperion 3000 microscope interfaced to a Vertex 70 FT-IR bench (Bruker) and equipped with a focal plane array (FPA) detector with pixel size of 2.7 microns. The obtained spectra represent an average of 400 scans in the mid-IR wavenumber range 900–4000 cm^{-1} with a spectral resolution of 8 cm^{-1}. Background signals were eliminated by obtaining scans from a region outside the sample field and was subtracted from the sample spectrum. The imaging was conducted in the laboratory of our collaborator Dr. Frank Bright (expert in IR imaging, University at Buffalo). Five different glomeruli were manually selected from each tissue section for analysis (Fig. 13). The hyperspectral cube was baseline corrected and normalized using the system's inbuilt min-max normalization function.

Spectral analysis: The major biomolecules – lipids, proteins, carbohydrates, and DNA, were analyzed. Carbohydrates dominate the mid-IR spectral region 1,300–700 cm^{-1}. The median of absorbance values between this range of wavenumbers was obtained and displayed as a heat map (Fig. 14A-B). In DNA, absorption bands appears near 1710 cm^{-1} from the carbonyl vibration of bases whereas, phosphate vibration is characterized by absorption bands near 1225 and 1080 cm^{-1} which arise from asymmetric and symmetric PO_2 vibrations, respectively. The median of absorbance values between these wavenumbers was obtained and displayed as a heat map (Fig. 14C-D). Lipids dominate the mid-IR spectral region 3,050–2,800 cm^{-1}. The median of absorbance values between this range of wavenumbers was obtained and displayed as a heat map (Fig. 14E-F). Proteins dominate the mid-IR spectral region 1,700–1,600 cm^{-1}. The median of absorbance values between this range of wavenumbers was obtained and displayed as a heat map (Fig. 14G-H). As evident below, the control murine tissue sample and DN murine tissue sample display different heatmap signatures.

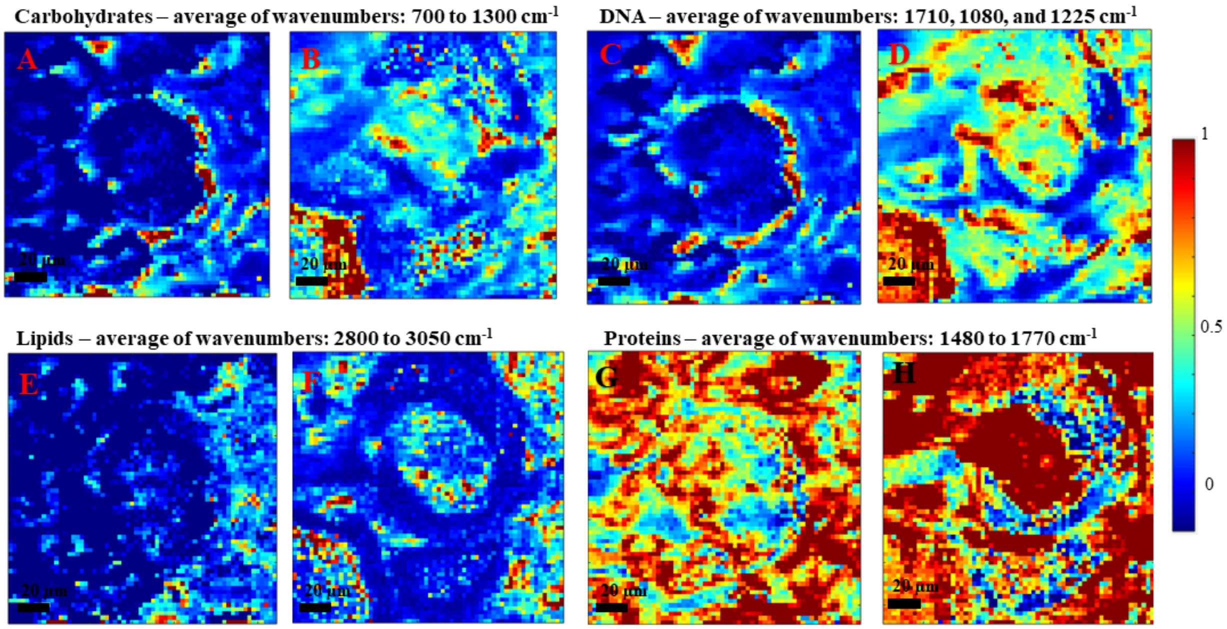
Carbohydrates – average of wavenumbers: 700 to 1300 cm-1
DNA – average of wavenumbers: 1710, 1080, and 1225 cm-1
Lipids – average of wavenumbers: 2800 to 3050 cm-1
Proteins – average of wavenumbers: 1480 to 1770 cm-1
A
B
C
D
E
F
G
H
20 µm
1
0.5
0

Fig. 14. FTIR spectral analysis. Carbohydrate heat map (A) control, (B) disease. DNA heat map (C) control, (D) disease. Lipid heat map (E) control, (F) disease. Proteins heat map (G) control, and (H) disease.

We are in the process of optimizing a post-staining method for tissue samples, with PAS, after conducting IR imaging. Due to the section thickness, tissues often get detached from slides during PAS staining, and this issue is currently being investigated in our lab. Upon accomplishing this task, we will continue working with murine DN models to develop a data fusion algorithm to integrate RGB color image of PAS-H stained tissue and multi--spectral IR images of the same tissue. The whole process will be repeated with human DN biopsies, and we will investigate the value of the fused data features in predicting DN progression.

Specific Aim 2. Predict DN disease severity from quantified glomerular features.

Results: We derived simple features from segmented glomeruli in DN histological datasets in a limited number of cases while writing this report. For our study, we employed *n* = 7 DN patient cases, *n* = 6 IgAN cases as controls, and *n* = 3 cases from renal cell carcinoma patients with tissues with no apparent histological damage as another set of controls. We correlated the derived glomerular features with clinical features such as eGFR drop rate to investigate if the image features contain any information on disease progression. Our preliminary result below motivates our current process of extending the study with more patient cases. By July 2019, we expect to total *n* = 35 DN patient cases, and as control *n* = 6 IgAN cases, n = 15 lupus nephritis cases, and *n* = 12 cases from renal cell carcinoma with tissues with no apparent histological damage. This feat will be an integrated effort summing resources from DiaComp project and other PI grant funding. Note that the DiaComp project has been extended until April 30, 2019, when we expect to provide another update on this aim.

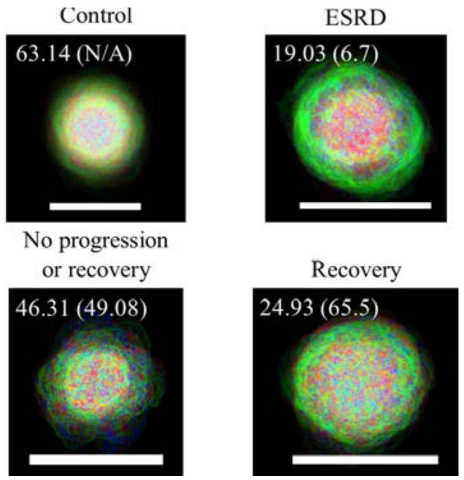

Fig. 15. Average glomerulus segmentation probability maps for various disease outcome classes. White numbers in upper left indicate eGFR at time of biopsy, and later eGFR in parentheses. All scale bars are 200 µm.

The prediction of a patient's outcome from a set of structural information derived from microscopy images is at the heart of many clinical assessments. While other cofounding factors are known to influence outcome, such as diet, exercise, and alcohol and tobacco use, a great amount of information on the patient's status can be determined by quantifying the extent to which their cellular structure is deformed from the normal or expected structure. As an initial study to determine if a patient's outcome can be predicted solely from image structure in a digital format, we derived a set of image features on glomeruli derived from biopsies of patients with diagnosed diabetic nephropathy. In the first step of this technique, all glomeruli from a patient's biopsy are computationally identified and segmented into three corresponding compartments: luminal compartments, Periodic acid-Schiff positive components, and nuclei. Segmentations of glomerular images are transposed onto a central landmark point, and averaged, to yield an average probability map of glomerular compartment locations (Fig. 15). Next, a set of features along the line drawn from the centroid of the average glomerulus map and extending just past the glomerular boundary are calculated, where gradients are subsequently taken along the radial, angular, and RGB axes (Fig. 16). This results in an enormous number of features, which are compressed by binning into histograms and averaging. Plotting the first three principal axes of this dataspace by patient progression status demonstrates that structural image features derived digitally correlate with clinical outcome. Namely, Fig. 17 describes each dot as a single patient corresponding to the principal component features, and the patients are colored/ marked depending on whether they are coming from the control population, those who progressed to end-stage renal disease, those who recovered, those who did not experience any progression, and those with no follow-up data. This preliminary finding suggests data separation of classes. When conjoining the three principal components together using another principal component feature analysis, which we term as structural risk score of a patient, we obtain Fig. 18, where we see a correlation between the respective patient and eGFR drop rate per day. Due to the limited dataset, we see the data general trend as a whole using all the patients data, and we must note that all the diabetic patients progressed to renal failure here. However, this study motivates us to believe that the image may contain useful features on disease progression, and in our current study while the DiaComp grant is still active and using our resource from other funding we are conjoining image features with the clinical outcome data using statistical longitudinal analysis. We will also investigate if we can use a deep convolution neural network based method by training the system using image patches classed based on outcome.

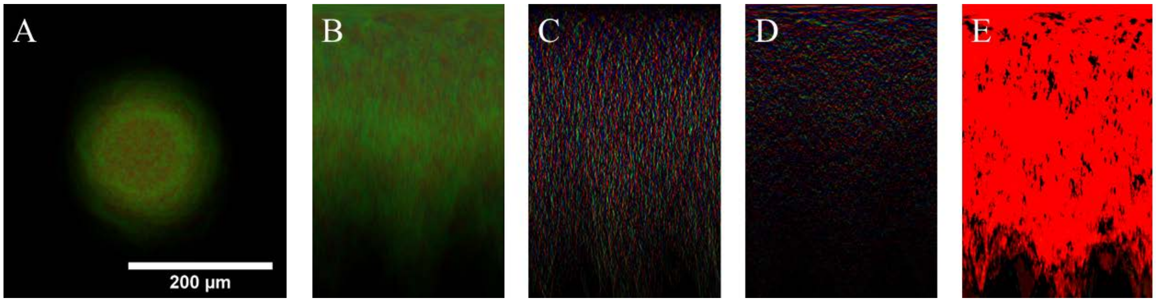

Fig. 16. Average glomerulus map, polar transform average map, and gradients along radial, angular, and RGB axes.

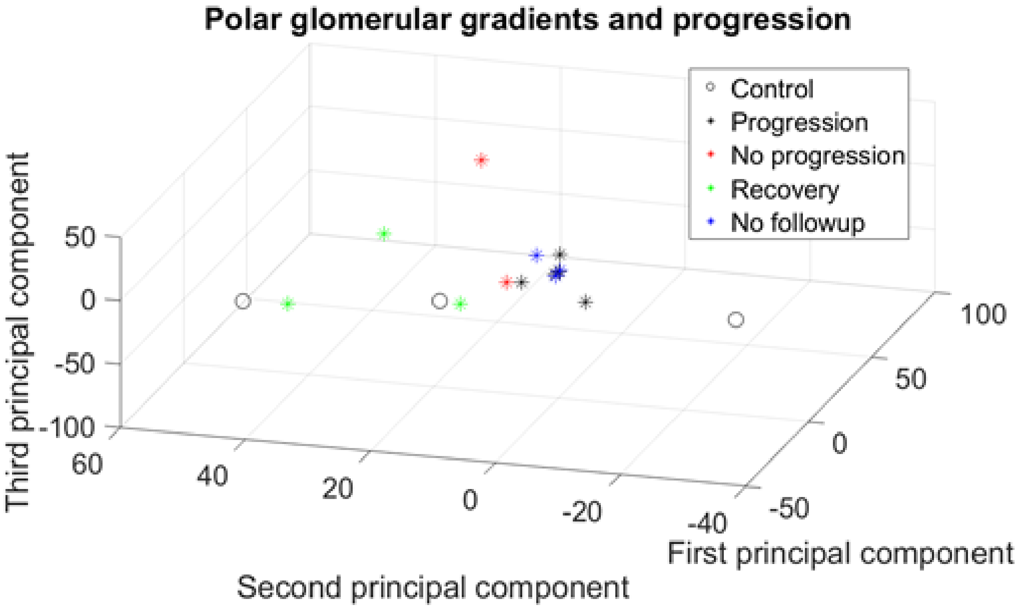

Fig. 17. Principal axes of features derived for progression prediction. Each dot represents a patient.

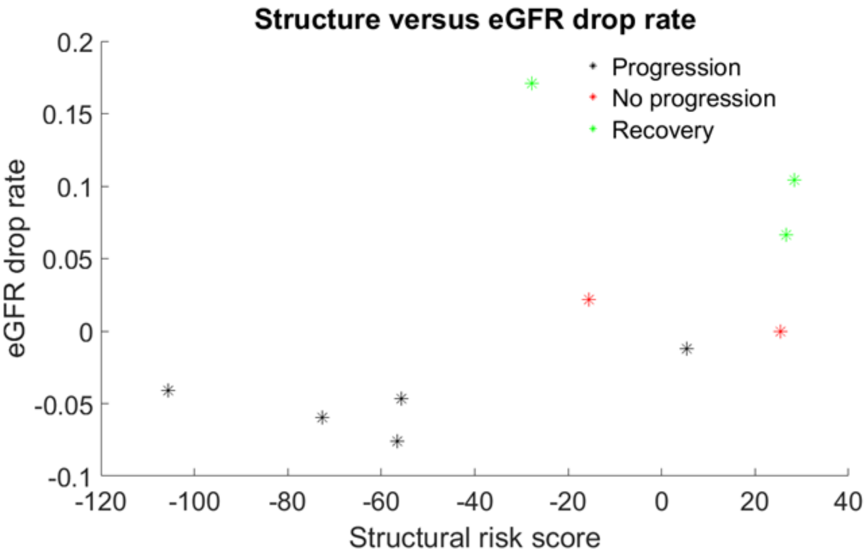

Fig. 18. Comparison between eGFR drop rate per day and the computationally derived structural risk score. Each dot represents a patient.

Publications:

Published

1. O. Simon, R. Yacoub, S. Jain, J. E. Tomaszewski, and P. Sarder, "Multi-radial LBP features as a tool for rapid glomerular detection and assessment in whole slide histopathology images," *Scientific Reports - Nature,* vol. 8, pp. 2032:1–11, Feb. 2018.

2. D. Govind, B. Ginley, B. Lutnick, J. E. Tomaszewski, and P. Sarder, "Glomerular detection and segmentation from multimodal microscopy images using a Butterworth band-pass filter," *Proc. of SPIE–Medical Imaging 2018: Digital Pathology,* vol. 10581, pp. 1058114:1–7, Houston, Texas, USA, Feb. 2018.

3. O. Simon, R. Yacoub, S. Jain, J. E. Tomaszewski, and P. Sarder, "Examining structural changes in diabetic nephropathy using inter-nuclear distances in glomeruli," *Proc. of SPIE–Medical Imaging 2018: Digital Pathology,* vol. 10581, pp. 105810B:1–10, Houston, Texas, USA, Feb. 2018.

In revision

1. B. Lutnick and P. Sarder, "Unsupervised community detection using Potts model Hamiltonian, an efficient algorithmic solution, and application in digital pathology," in revision for *Journal of Medical Imaging - SPIE.*

Iterative annotation to ease neural network training: Specialized machine learning in medical image analysis

Brendon Lutnick[1], Brandon Ginley[1], Darshana Govind[1], Sean D. McGarry[2], Peter S. LaViolette[2], Rabi Yacoub[3], Sanjay Jain[4], John E. Tomaszewski[1], Kuang-Yu Jen[5], and Pinaki Sarder[1*]

[1]Department of Pathology & Anatomical Sciences, SUNY Buffalo, USA. [2]Department of Radiology, Medical College of Wisconsin. [3]Department of Medicine, Nephrology, SUNY Buffalo. [4]Department of Medicine, Nephrology, Washington University School of Medicine. [5]Department of Pathology, University of California, Davis Medical Center.

[*]Address all correspondence to: Pinaki Sarder
Tel: (716)-829-2265, e-mail: pinakisa@buffalo.edu

Abstract

Neural networks promise to bring robust, quantitative analysis to medical fields, but adoption is limited by the technicalities of training these networks. To address this translation gap between medical researchers and neural networks in the field of pathology, we have created an intuitive interface which utilizes the commonly used whole slide image (WSI) viewer, Aperio ImageScope (Leica Biosystems Imaging, Inc.), for the annotation and display of neural network predictions on WSIs. Leveraging this, we propose the use of a human-in-the-loop strategy to reduce the burden of WSI annotation. We track network performance improvements as a function of iteration and quantify the use of this pipeline for the segmentation of renal histologic findings on WSIs. More specifically, we present network performance when applied to segmentation of renal micro compartments, and demonstrate multi-class segmentation in human and mouse renal tissue slides. Finally, to show the adaptability of this technique to other medical imaging fields, we demonstrate its ability to iteratively segment human prostate glands from radiology imaging data.

Introduction

In the current era of artificial intelligence, robust automated image analysis is attained using supervised machine learning algorithms. This approach is gaining considerable ground in virtually every domain of data analysis, mainly under the advent of neural networks [2-5]. Neural networks are a broad range of algorithms which can take many different forms, but all are considered graphical models, whose nodes can be variably activated by a non-linear operation on the sum of their inputs [4, 6]. The connections between nodes are modulated by weights, which can be adjusted to dampen or amplify the power of contribution of that node to the output of the network. These weights can be iteratively tuned via back propagation so that the input of a particular type of data leads to a desired output (usually a classification of the data) [7]. Particularly useful for image analysis are convolutional neural networks (CNNs) [3, 4], a specialized subset of neural networks which leverage convolutional filters to learn spatial representations of image regions specific to the desired image classification. This allows high dimensional filtering operations to be learned automatically, a task which has traditionally been done through hand-engineering. Neural networks are problematic in certain applications, as they require significant amounts of annotated data (on the order of tens of thousands) in order to provide generalized high performance, yet their potential exceeds other machine learning techniques [8].

Work to ease the burden of data annotation is arguably as important as the generation of state-of-the-art network architectures, which without sufficient data are unusable [9, 10]. Many large-scale modern machine learning applications are indeed based on cleverly designed crowd sourced active learning pipelines, which in the era of constant firmware updates, comes in the form of human-in-the-loop training [11-13]. Initiated by low classification probabilities, machine learning applications such as automated teller machine character recognition, self-driving cars, and Facebook's automatic tagging, all rely on user refined training sets for fine tuning neural network applications post deployment [4]. These 'active learning' techniques require users to 'correct' the predictions of a network, therefore identifying gaps in network performance [14].

The adoption of neural networks to biological datasets has largely lagged behind adoption in computer science [15, 16]. While computational strategies for image analysis are seeing ever increasing translation to biological research, the late adoption of CNN-based methods for classification in biology is largely due to the lack of centrally curated and annotated training sets [17]. Due to the specialized nature of medical datasets, annotation by experts necessary for generation of training sets is less feasible than traditional datasets [18]. This issue creates challenges when trying to apply CNNs to medical imaging databases where domain-expert knowledge is required to perform image annotation, but domain-expert annotation is difficult to acquire because it is expensive, time consuming, labor intensive, and there are no technical mediums which enable easy transference of this information from clinical practice to training sets [19].

Despite of the above mentioned challenges in digital pathology, segmentation and classification of tissue slides by neural networks will not only aid clinical diagnosis based on current guidelines and practice, but will likely facilitate the creation of refined and improved future diagnostic guidelines using quantitative computational metrics. Additionally, neural networks can generate searchable data repositories [20], providing practicing clinicians and students access to collections of domain knowledge, such as labeled images and associated clinical outcomes that were not previously available [21-23]. While this end goal necessitates a combination of curated pathological datasets, machine learning classifiers [4], automatic anomaly detection [24, 25], and efficient searchable data hierarchies [22]; pipelines for creating easily viewable annotations on pathology images are a necessary first step. Towards this aim, we have developed an iterative interface between the successful semantic segmentation network DeepLab v2 [26] and the widely used WSI viewing software Aperio ImageScope [27], which we have termed Human A.I. Loop (H-AI-L) (Figure 1). Put simply, the algorithm converts annotated regions stored in XML format (provided in ImageScope) into image region masks. These masks are used to train the semantic segmentation network, whose predictions are converted back to XML format for display in ImageScope. This graphical display of network output is an ideal visualization tool for segmentation predictions on WSI, with the ability to view the entire tissue slide, pan and zoom functionalities, as well as the efficient JPG2000 decompression [28] of WSI files provided by ImageScope. Using this open sourced pipeline, a supervising domain expert can correct the network predictions and initiate further training using the newly annotated data. This enables networks to be trained "on demand", or as the data is available. Using H-AI-L, we are able to significantly reduce the annotation effort required to learn robust segmentations of large microscopy images [28]. Adaptation of this technique to other modes of medical imaging is highly feasible, which we demonstrate using MRI imaging data.

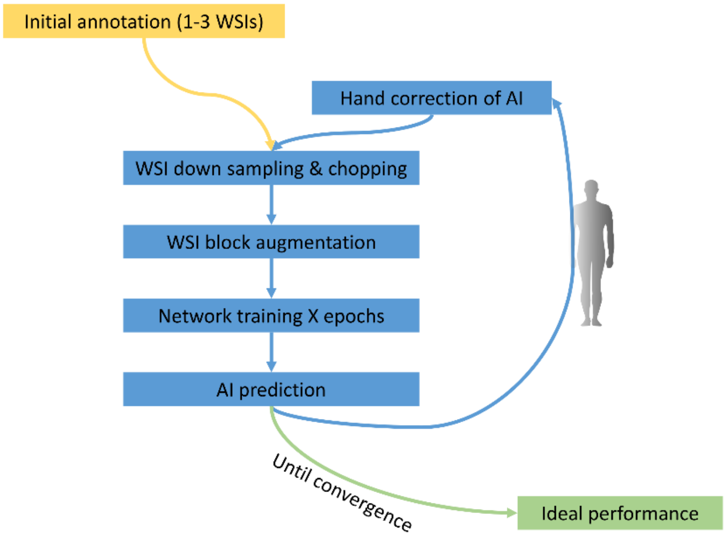

Figure 1 Iterative Human A.I. Loop **(*H-AI-L*)** ***pipeline overview.***

Schematic representation of H-AI-L pipeline for training semantic segmentation of WSI. Several rounds of training are performed using human expert feedback in order to optimize ideal performance, resulting in improved efficiency in network training with limited numbers of initial annotated WSIs.

Results

To evaluate the utility of H-AI-L, we first quantified its performance and efficiency with histologic sections of kidney tissue, the first being glomerular localization in mouse kidney WSIs [5, 29-32]. This glomeruli segmentation network was trained for 5 iterations, using a combination of periodic acid-Schiff (PAS) and hematoxylin and eosin (H&E)-stained murine renal sections. For more data variation, streptozotocin (STZ) induced diabetic nephropathy [1, 33-35] murine data was included in iteration 4 (Table 1). To validate the performance of our network, we use 4 holdout WSIs, including one STZ induced WSI.

During the training process, we observed approximately 4 to 10-fold increases in average glomerular annotation speed between the initial and end iterations (Figure 2a). This represents time savings of 81.4%, 82%, and 72.7% for three annotators, annotator-1, -2, and -3, respectively, when compared to each annotator's baseline speed. This results in the prediction performance increase shown in Figure 2b, where the network reaches nearly perfect performance on a holdout dataset by annotation iteration 4. One side effect of using iterative annotation is the intuitive qualification of network performance it provides after each interaction; that is, an expert interacts with the network predictions after each training round, visualizing network biases and shortcomings on holdout data. Two examples of evolving network predictions are highlighted in Supplemental Figure 1.

In order to improve network prediction efficiency we designed a multi-resolution approach, which uses two segmentation networks: identifying hot spot regions at 1/16th scale before segmenting them at the highest resolution. This approach, which we call DeepZoom, obtains better F-measure (F1 score) [36, 37] (Figure 2b) versus a full resolution pass, as well as approximately 4.5-times faster predictions (Figure 2c). An overview of this method can be found in Supplemental Figure 2.

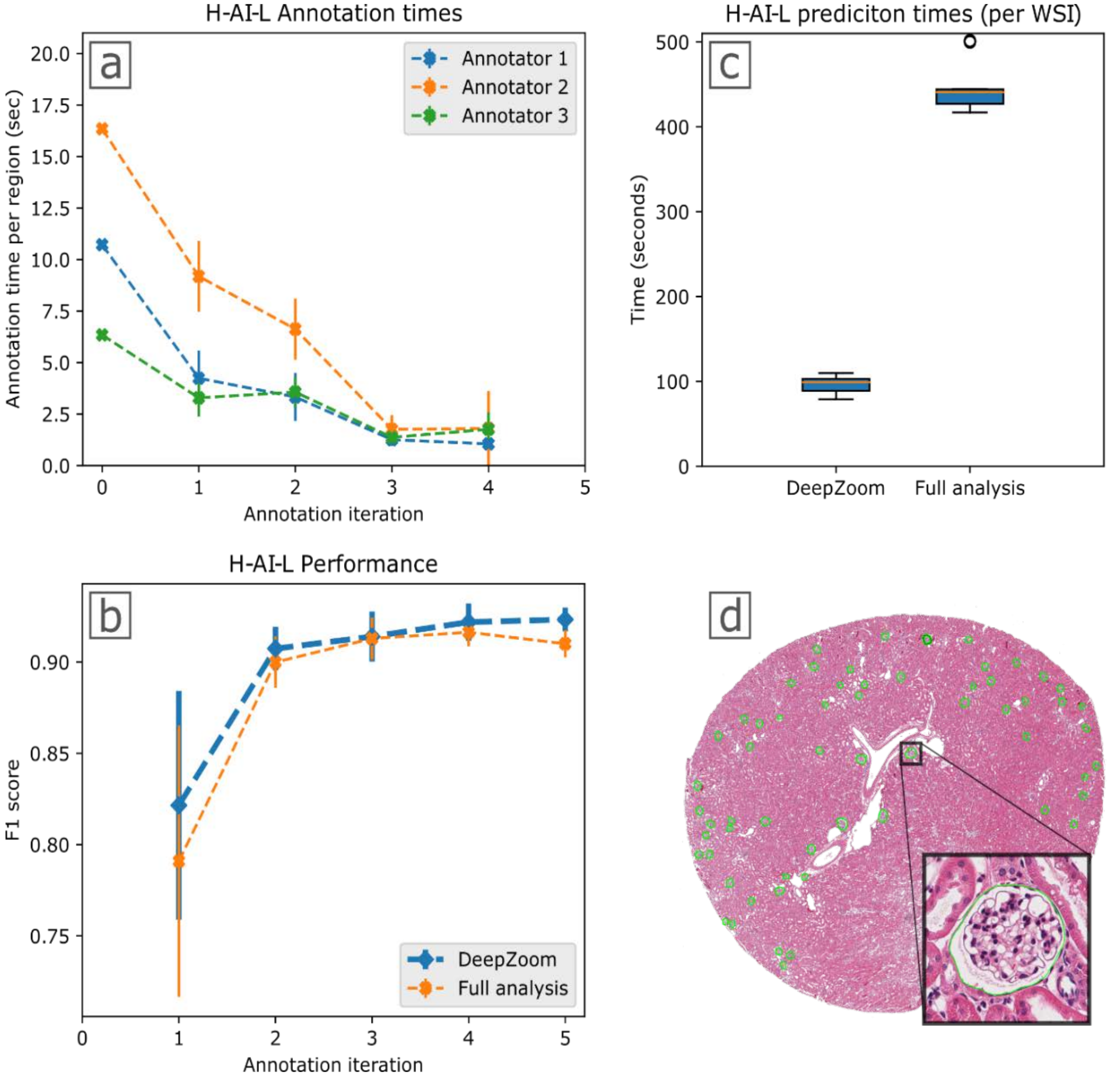

Figure 2 H-AI-L pipeline performance: glomerular segmentation on holdout mouse WSI.

(a) Annotation times per glomerulus as a function of annotation iteration. The 0th iteration was performed without preexisting predicted annotations, whereas subsequent iterations use network predictions as an initial annotation prediction that can be corrected by the annotator. **(b)** F1 score of glomerular segmentation of 4 holdout mouse renal WSIs as a function of training iteration. **(c)** Runtimes for glomerular segmentation prediction on holdout mouse renal WSIs using H-AI-L with DeepZoom (multi-resolution segmentation) versus full resolution segmentation. **(d)** Example of a mouse WSI with segmented glomeruli. Network predictions are outlined in green. Error bars indicate ±1 standard deviation.

H-AI-L Data Set							
Annotation iteration		0	1	2	3	4	*Test*
WSI added		1	2	4	6	4	4
Total glomeruli	*Normal*	32	84	86	418	0	138
	STZ	0	0	0	0	293	96

Table 1 H-AI-L segmentation mouse WSI training and testing datasets.

Mouse WSI training set used to train the glomerular segmentation network. Data presenting structural damage from streptozotocin (STZ) induced diabetes [1] was introduced in iteration 4. The test dataset included 3 normal and 1 STZ WSI.

Quantification of the performance achieved by our method in WSI is a challenge due to the imbalance between class distributions [38]. Therefore, we choose to report F-measure which considers both precision and recall (sensitivity) simultaneously [36], as specificity and accuracy are always high due to the large percentage of negative region with respect to the positive class. This is particularly important considering the performance characteristics of DeepZoom. During testing we found that DeepZoom trades segmentation sensitivity for increased precision, while outperforming full analysis overall with improved F1 score (Figure 2). This performance gap is due to a lower false positive rate achieved by DeepZoom as a result of the low resolution network pre-pass, which limits the amount of background region seen by the high resolution network. Overall, on four holdout WSIs, our network achieved its best performance after the 5th iteration of training using DeepZoom with sensitivity 0.92 ± 0.02, specificity 0.99 ± 0.001, precision 0.93 ± 0.14, and accuracy 0.99 ± 0.001.

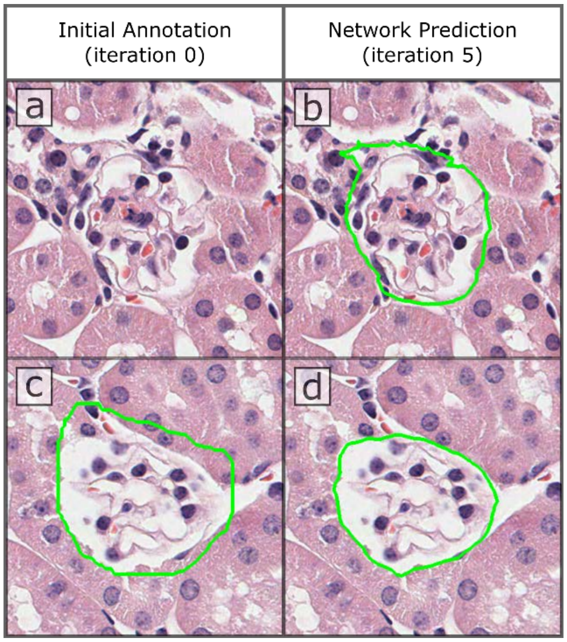

Figure 3 H-AI-L human annotation errors (mouse data).

Comparison of initial manual annotations from iteration 0 (a and c) with their respective final network predictions from iteration 5 (b and d). These examples were selected due to poor manual annotation, where the glomerulus **(a)** was not annotated or **(c)** showed poorly drawn boundaries.

Network performance analysis is further complicated by human annotation errors. We note several instances where network predictions outperformed human annotators, despite being trained using flawed annotations. This phenomenon is highlighted in Figure 3, where glomerular regions annotated manually in iteration 0 are compared to the prediction by the iteration 5 network. Such errors are more prevalent in WSIs annotated in early iterations, where network predictions need the most correction.

To qualitatively prove the effectiveness and extendibility of our method, we show its extension to multi class detection by segmenting glomerular nuclei types [39, 40], interstitial fibrosis and tubular atrophy (IFTA) [41, 42], as well as differentiating sclerotic and non-sclerotic glomeruli [43] in mouse kidney and human renal biopsies. Figure 4 shows the glomeruli detection network from Figure 2 adapted for nuclei detection. This was done by re-training the high resolution network using a set of 143 glomeruli with labeled podocyte and non-podocyte nuclei, marked via immunofluorescence labeling. For this analysis, the low resolution network from Figure 2 was kept unchanged to identify the glomerular regions in the mouse WSI.

Due to the non-sparse nature of IFTA regions in some human WSI we forgo our DeepZoom approach to generate the results shown in Figure 5. The development of this IFTA network has been limited due to the biological expertise required to produce these multiclass annotations. However, preliminary segmentation results on holdout WSI show promising results despite using only 15 annotated biopsies for training (Figure 5). We note that this is a small training set, as human biopsy WSIs contain much less tissue area than the mouse kidney sections used to train the glomerular segmentation network above.

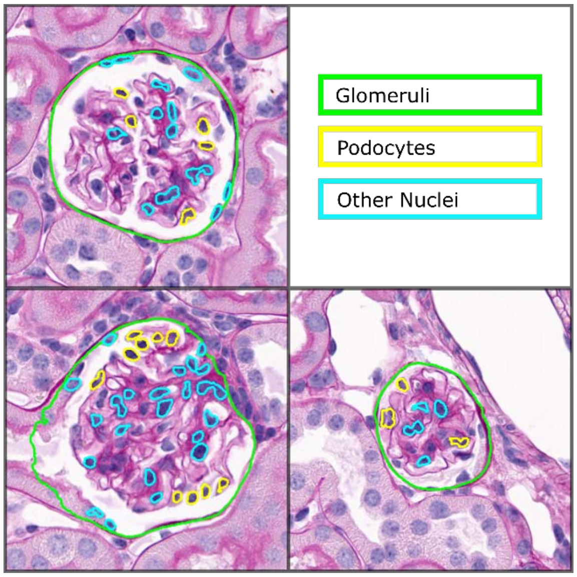

Figure 4 Multiclass nuclei prediction on mouse WSI.

Several examples of multiclass nuclei predictions are visualized on a mouse WSI. Here transfer learning was used to adapt the high resolution network from above (Figure 2) to segment nuclei classes. This network was trained using 143 labeled mouse glomeruli. The low resolution network was kept unchanged for the initial detection of glomeruli. We expect the results to significantly improve using more labeled training data.

Finally, to show the adaptability of the H-AI-L pipeline to other medical imaging modalities, we quantify the use of our approach for the segmentation of human prostate glands from T2 MRI data. This data was orientated and normalized as described in [44] and saved as a series of TIFF image files, which can be opened in ImageScope and are compatible with our H-AI-L pipeline. This analysis was completed using a training set of data from 39 patients with an average of 32 slices per patient (512 x 512 pixels) (Figure 6d). Iterative training was completed by adding data from 4 patients to the training set prior to each iteration. Data from the remaining 7 patients was used as a holdout testing set. The newly annotated/corrected training data was augmented 10-times and a full resolution network was trained for 2 epochs during each iteration: the results of this training are presented in Figure 6. While the network performs well after just 1 round of training, the performance on holdout patient data continues to improve with the addition of training data (Figure 6a), achieving sensitivity of 0.88 ± 0.04, specificity of 0.99 ± 0.001, precision of 0.9 ± 0.03, and accuracy of 0.99 ± 0.001. This trend is also loosely reflected in the networks prediction on

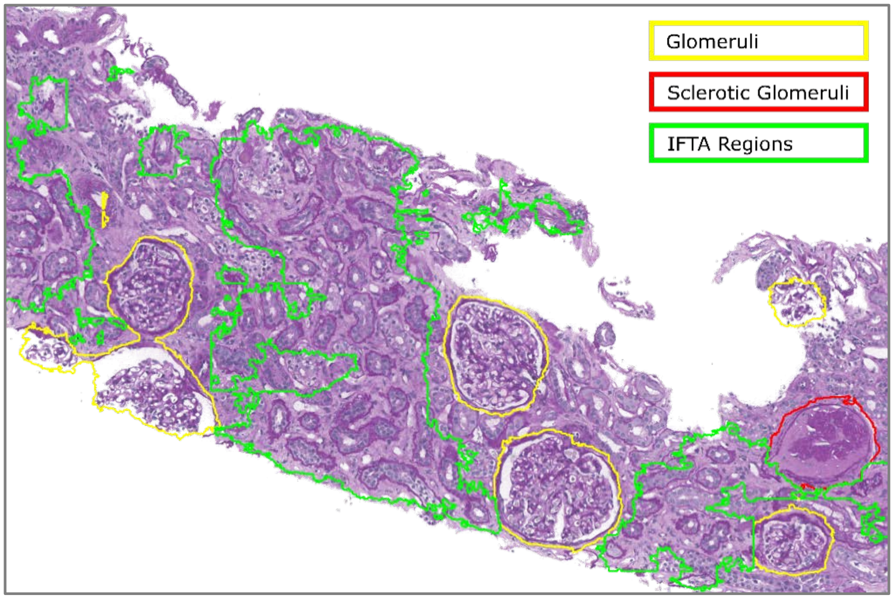

Figure 5 Multiclass IFTA prediction on a holdout human renal WSI.

Segmentation of healthy and sclerotic glomeruli, as well as IFTA regions from human renal biopsy WSI. Due to the non-sparse nature of IFTA regions, these predictions were made using only a high resolution pass. This is a screenshot of Aperio ImageScope which we use to interactively visualize the network predictions.

newly added training data, where an upward trend in prediction performance is observed in Figure 6b. Notably, when our iterative training pipeline is applied to this dataset, annotation is reduced by approximately 90 percent after the second iteration, where only 10 percent of MRI slices containing prostate fall below our segmentation performance threshold (Figure 6c).

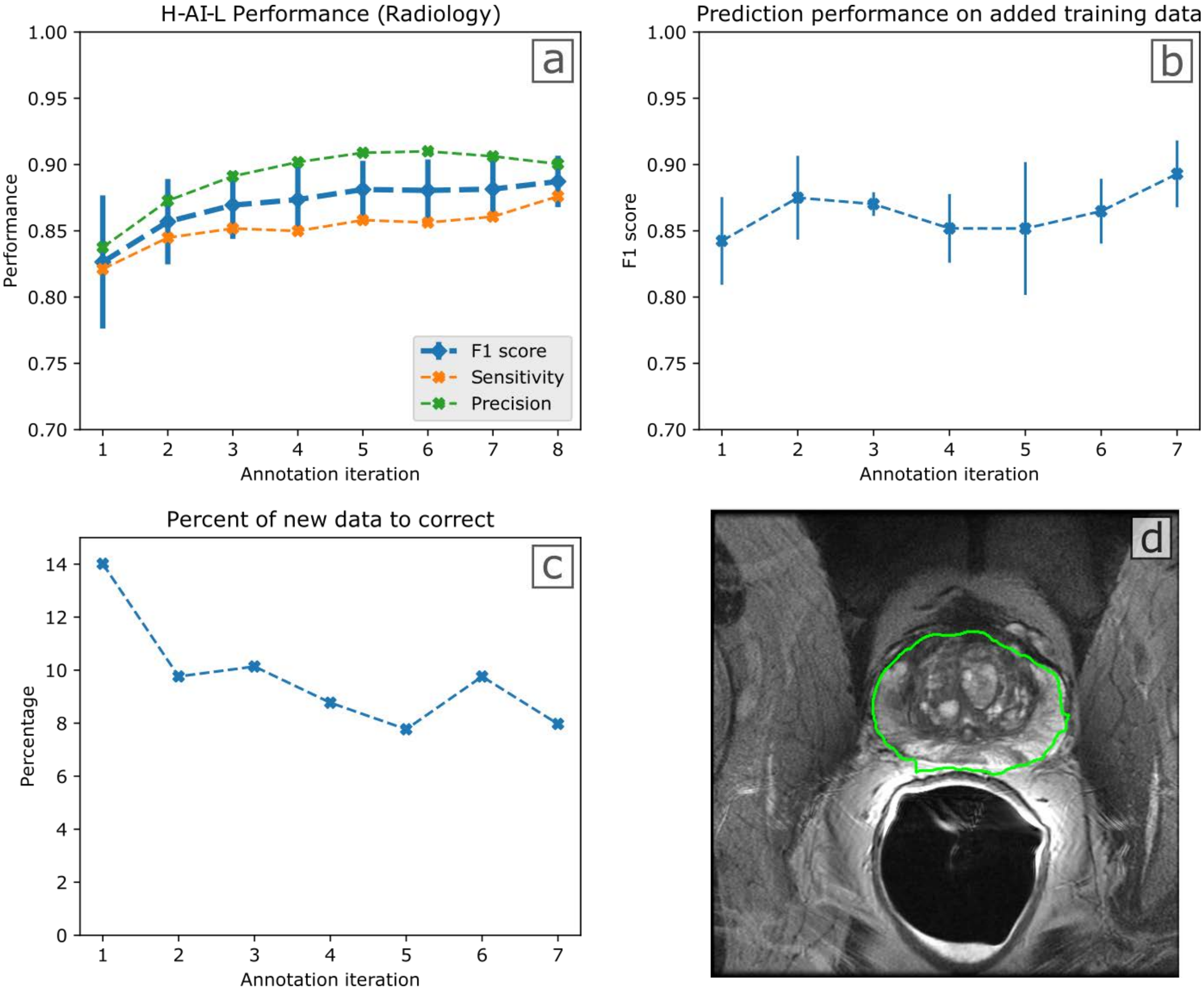

Figure 6   H-AI-L method performance: human prostate segmentation from T2 MRI slices.

(a) Segmentation performance as a function of training iteration, evaluated on 7 patient holdout MRI images (224 slices). Performance was evaluated on a patient basis. We note that despite the decline in network precision after iteration 6, the F1 score improves as a result of increasing sensitivity. **(b)** The prediction performance on added training data. This figure shows the prediction performance on newly added data w.r.t. the expert corrected annotation, and is evaluated on a patient basis (data from 4 new patients was added at the beginning of each training iteration). **(c)** The percentage of prostate regions where network prediction performance (F1 score) fell below an acceptable threshold (percentage of slices which needed expert correction) as a function of training iteration. We define acceptable performance as F1 score > 0.88. Using this criteria, expert annotation of new data is reduced by 92% by the fifth iteration. **(d)** A randomly selected example of a T2 MRI slice with segmented prostate: the network predictions are outlined in green. Error bars indicate ±1 standard deviation.

Conclusions

We have developed an intuitive pipeline for segmentation of structures from WSI commonly used in pathology, a field where there is often a large disconnect between domain experts and engineers. We aim to bridge this gap by making the robust data analytics provided by state-of-the-art neural networks accessible to pathologists. Along this direction we have developed an intuitive library for the adaptation of DeepLab v2 [26], a semantic segmentation network, to whole slide imaging data, commonly used in the field. This library uses annotation tools from the common WSI viewing software Aperio ImageScope [27] for annotation and display of the network predictions. Training, prediction and validation of the network is done via a single python script with a command line interface, where data management is as simple as dropping data into a pre-determined folder structure.

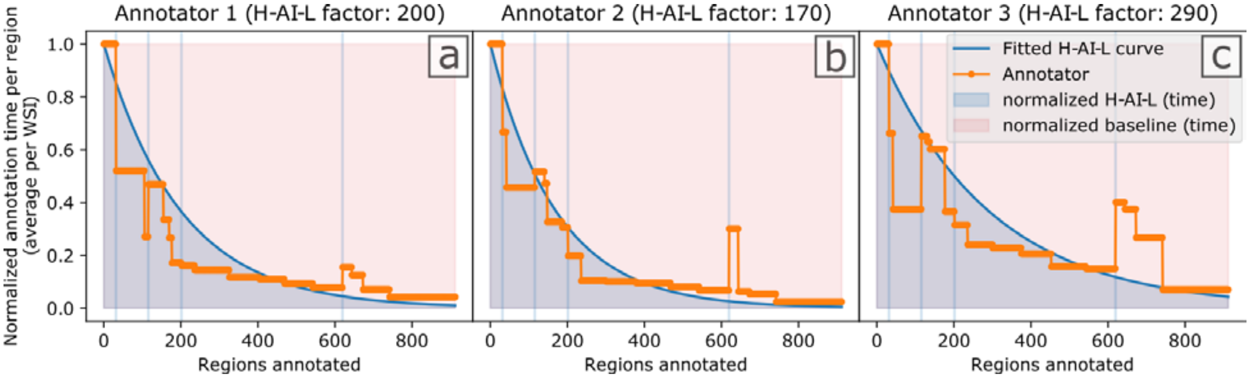

Figure 7 Annotation time savings using the H-AI-L method: compared to baseline segmentation speed (Figure 2a).

H-AI-L plots showing the annotation time per region normalized with respect to the baseline annotation speed of each annotator. An exponential decay distribution (H-AI-L curve) is fitted to each annotator, where the H-AI-L factor is the exponential time constant: a derivation can be found in the methods section. The vertical lines are gaps between iterations (where the network was trained). The area under the H-AI-L curve represents the normalized annotation time per annotator. This can be compared to the area of the normalized baseline region, which represents the normalized annotation time without the H-AI-L method. **(a)** The time savings by annotator 1 (calculated to be 81.3 percent) when creating the training set used to train the glomerular segmentation network in Figure 2. **(b)** Annotator 2 was 82.0 percent faster. **(c)** Annotator 3 was 72.7 percent faster. While the y-axis in these plots is not a direct measure of network performance, it is highly correlated. The spike in annotation time seen at 600 regions is data from a WSI with severe glomerular damage from DN. We believe that plots like these will offer insight into optimal iterative training strategies in the future, with a goal of reducing annotation burdens for expert annotators.

Using our iterative, human in the loop training allows considerably faster annotation of new WSIs (or similar imaging data), as network predictions can easily be corrected in ImageScope before incorporation into the training set. This approach allows the qualitative assessment of network performance after each iteration, as newly added data acts as a holdout validation set, where predictions are easily viewed during correction. The theoretical performance achievable by this method is bounded by the training set used, and is therefore the same as the current state-of-the-art (manual annotation of all training data). However, due to the increased speed of annotation, and intuitive visualization of network performance (allowing selection of poorly predicted new data after each iteration), we argue that H-AI-L training has the potential to converge to the upper bound of performance more efficiently than the traditional method; achieving state-of-the-art segmentation performance much faster than traditional methods, which are limited by data annotation speed (Figure 7). To our knowledge, our approach: displaying network predictions in ImageScope, is the first of its kind. It offers an ideal viewing environment for network predictions on WSIs, using the fast pan and zoom functionality provided by ImageScope [28], improving the accuracy and ease of expert annotation.

The ability to transfer parameters from a trained network (repurposing it for a different task), ensures that segmentation of tissue structure can be tailored to any clinical or research definition, including other biomedical imaging modalities. Our multiresolution (DeepZoom) analysis allows rapid prediction of sparse regions from large WSIs, without sacrificing accuracy due to low resolution analysis alone. Inspired by the way pathologists scan tissue slides, multiresolution approaches have been successfully used in digital pathology literature for the detection of cell nuclei [45]. We believe that this technique offers the perfect compromise between speed and specificity, producing high resolution sparse segmentations ideal for display in ImageScope. The use of our method for non-sparse segmentation of WSI is achievable by foregoing DeepZoom analysis. However, in the future we plan to change the way that the class hierarchy is defined in our algorithm, offering easy functionality to search for low resolution regions with high resolution sub-compartments.

In the future we will undergo extensive testing of our method in a clinical research setting. This testing would involve evaluation of the segmentation performance as well as ergonomic aspects which pertain to a clinician's ease of use. We will extend our method to provide anomaly detection, defining a confidence metric and threshold where WSIs are flagged for further evaluation. To compliment this, we will create an algorithm to predict the optimal amount of annotation in each iteration (to optimize expert time) using a curve fitting similar to figure 6. We will also adapt our method for native use with a DICOM viewer, allowing easier workflows for segmentation of Radiology datasets. Given these tools, we foresee a segmentation approach similar to our H-AI-L method acting as a cornerstone of efforts to build searchable databases of digital pathology slides [22], and other medical imaging datasets.

Methods

All animal tissue sections were collected in accordance with protocols approved by the Institutional Animal Care and Use Committee at University at Buffalo, and are consistent with federal guidelines and regulations and in accordance with recommendations of the American Veterinary Medical Association guidelines on euthanasia. Renal biopsy samples were collected from the Kidney Translational Research Center at Washington University School of Medicine, directed by co-author Dr. Jain, following a protocol approved by the Institutional Review Board at University at Buffalo prior to commencement. Digital MRI images of human prostate glands were provided by co-author Dr. LaViolette, following a protocol approved by the Institutional Review Board at Medical College of Wisconsin. All methods were performed in accordance with the relevant federal guidelines and regulations. All patients provided written informed consent, and basic demographic information was collected.

In the H-AI-L pipeline, an annotator labels one whole slide image using annotation tools in ImageScope [27], which provides the input for network training. The resulting trained network is then used to predict the annotations on a new WSIs. These predictions are used as rough annotations, which are corrected by the annotator and sent back for incorporation into the training set; improving network performance and optimizing the amount of expert annotation time required. Because this technique makes the adaptation of network parameters to new data easy, adapting a trained network to new data generated in different institutions is extremely feasible. We have made our code openly available online: goo.gl/Wr6qYE.

At the heart of H-AI-L is the conversion between mask and XML [46] formats which are used by DeepLab v2 [26] and ImageScope [27], respectively. Training any semantic segmentation architecture relies on pixel-wise image annotations which are input to the network for training and output after network predictions as mask images. In the case of DeepLab, the mask images take the form of indexed greyscale 8 bit PNG files, where each unique value pertains to an image class. On the other hand, annotations done in ImageScope are saved in text format, as XML files [46], where each region is saved as a series of boundary points or vertices. Determining the vertices of a mask image is a common image processing task, known as image contour detection [47, 48]. As opposed to edge detection, contour detection can have hierarchal classifications [48], lending itself ideally to conversion into the hierarchal XML format used by ImageScope.

To facilitate the transfer between ImageScope XML and greyscale mask images, we use the *OpenCV-Python* library (*cv2*) [47], using the function *cv2.findContours* to convert from masks to contours. Using this function, we are able to automatically convert DeepLab predictions to XML format which can be viewed in ImageScope, easily evaluating and correcting network performance. Additionally, we have written a library for converting an XML file into mask regions, using *cv2.fillPoly.* This library follows the *OpenSlide-Python* [49] conventions for reading WSI regions, returning a specified mask region from the WSI.

OpenSlide [49] and our XML to mask libraries allow for efficient chopping of WSI into overlapping blocks for network training and prediction; similar sliding window approaches are common practice for predicting semantic segmentations on large medical images [50, 51]. To simplify the iterative training process, and compliment the easy annotation pipeline proposed, we have created a callable function which handles operations automatically, prompting the user to initiate the next step. This function needs two flags [--*option*] and [--*project*] which are the parameters identifying the iterative step and project one would like to train respectively. Initially created using [--*option*] *'new'*, a new project is trained iteratively by alternating the [--*option*] flag between '*train'* and '*test'*. Our algorithm uses our DeepZoom approach by default, but full-resolution analysis is achievable by setting the [--*one_network*] flag to '*True'* during training and prediction.

Training:

To streamline the training process, we created a pipeline where a user places new WSIs and XML annotations in a project folder structure, then calls a function to train the project. This automatically initiates data chopping and augmentation, then loads parameters from the most recently trained network (if available) before starting to train. For faster convergence, we utilize transfer learning, automatically pulling a pre-trained network file whenever a new project is created, which is used to initialize the network parameters prior to training. We have also included functionality to specify a pre-trained file from an existing project using the [--*transfer*] flag. For ease of use, the network hyper-parameters can be changed using command line flags, but are set automatically by default.

When [--*option*] *'train'* is specified, WSIs and XML annotations are chopped into a training set containing 500 x 500 blocks with 50% overlap. This training set is then augmented via: random flipping, hue and lightness shifts, as well as piecewise affine transformations; accomplished using the *imgaug* python library [52]. To keep the network unbiased, the total number of blocks containing each class is tabulated and used to augment less frequent classes with a higher probability [53]. Once augmented, the network is trained for the specified number of epochs, and the user is prompted to upload new WSIs and run the [--*option*] *'predict'* flag. This produces XML predictions which can be corrected using ImageScope before incorporation into the training set.

Prediction:

Due to the sparse nature of the structures we attempt to segment from renal WSI, we limit the search space, using a low resolution pass to determine hotspot regions before segmentation at full resolution (DeepZoom). This is accomplished in two ways: Firstly, thresholding and morphological processing are used to determine which WSI blocks contain tissue, eliminating background regions. Secondly, down-sampled blocks (1/16th resolution) are tested using a semantic segmentation network (DeepLab) to roughly segment structures. The output predictions of the preprocessing steps are then stitched back into a hotspot map, which identifies important regions at this resolution. Using this map, full size hotspot indices are calculated, and the regions are extracted using OpenSlide for pixel-wise segmentation by a second network.

Validation:

While the performance of network is easily visualized after prediction on new WSI, we have included functionality for explicit evaluation of performance metrics and prediction time on a holdout dataset. This is accomplished using the [--*option*] *'validate'* flag. When called, it evaluates the network performance on holdout images for every annotation iteration by pulling the latest models automatically. To perform this performance comparison, ground truth XML annotations of the holdout set are required for the calculation of sensitivity, specificity, accuracy, and precision performance metrics [37].

Estimating H-AI-L performance (Figure 7):

To quantify the time savings of our H-AI-L method, we plot the normalized annotation time per region vs the number of regions annotated. Here we define the normalized annotation time per region A as:

$$A = \frac{t}{t_0},$$

where t is the annotation time per region (averaged per WSI), and t_0 is the average annotation time per region in iteration 0. A is bounded from $[0,1]$ where 1 is the normalized time it takes to annotate one region fully. While the annotation time is reduced as a piecewise function of training iteration, in Figure 7 we use a continuous exponential decay distribution to approximate $A(r)$:

$$A(r) = e^{-\frac{r}{\tau}},$$

where r is the number of regions annotated, and τ is the exponential time constant which we call the H-AI-L factor.

The normalized annotation time of our H-AI-L method (H) can therefore be estimated as:

$$H = \int_0^R A(r)\mathrm{d}r = \tau\left[1 - e^{\frac{-R}{\tau}}\right],$$

where R is the total number of regions annotated. Like-wise, the normalized baseline annotation time (B) can be calculated as:

$$B = \int_0^R 1\mathrm{d}r = R$$

Therefore the time savings performance (P) of our H-AI-L method can be estimated as a percentage using:

$$P = \left(1 - \frac{H}{B}\right) * 100 = \left(1 + \frac{\tau}{R}\left[e^{\frac{-R}{\tau}} - 1\right]\right) * 100.$$

The H-AI-L factor τ reflects the effectiveness of iterative network training, where lower values of τ represent training curves that decay faster. In the future, algorithms to select the optimal amount of annotation and identify data outliers to be annotated at each iteration will improve the performance of the H-AI-L method by reducing τ.

References

1. *STZ-induced diabetes*. Available from: https://www.jax.org/jax-mice-and-services/find-and-order-jax-mice/surgical-and-preconditioning-services/stz-induced-diabetes
2. Krizhevsky, A., I. Sutskever, and G.E. Hinton, *ImageNet classification with deep convolutional neural networks.* Commun. ACM, 2017. **60**(6): p. 84-90.
3. LeCun, Y. and Y. Bengio, *Convolutional networks for images, speech, and time series*, in *The handbook of brain theory and neural networks*, A.A. Michael, Editor. 1998, MIT Press. p. 255-258.
4. LeCun, Y., Y. Bengio, and G. Hinton, *Deep learning.* Nature, 2015. **521**(7553): p. 436-44.
5. Pedraza, A., et al., *Glomerulus Classification with Convolutional Neural Networks*, in *Medical Image Understanding and Analysis: 21st Annual Conference, MIUA 2017, Edinburgh, UK, July 11–13, 2017, Proceedings*, M. Valdés Hernández and V. González-Castro, Editors. 2017, Springer International Publishing: Cham. p. 839-849.
6. Schmidhuber, J., *Deep learning in neural networks: an overview.* Neural Netw, 2015. **61**: p. 85-117.
7. Bottou, L., *Large-scale machine learning with stochastic gradient descent*, in *Proceedings of COMPSTAT'2010*. 2010, Springer. p. 177-186.
8. Szegedy, C., et al. *Going Deeper with Convolutions*. ArXiv e-prints, 2014. **1409**.
9. Swingler, K., *Applying neural networks: a practical guide*. 1996: Morgan Kaufmann.
10. Ronneberger, O., P. Fischer, and T. Brox. *U-net: Convolutional networks for biomedical image segmentation*. in *International Conference on Medical image computing and computer-assisted intervention*. 2015. Springer.
11. Zhang, T. and M. Nakamura, *Neural network-based hybrid human-in-the-loop control for meal assistance orthosis.* IEEE transactions on neural systems and rehabilitation engineering, 2006. **14**(1): p. 64-75.
12. Krogh, A. and J. Vedelsby. *Neural network ensembles, cross validation, and active learning*. in *Advances in neural information processing systems*. 1995.
13. Cohn, D., L. Atlas, and R. Ladner, *Improving generalization with active learning.* Machine learning, 1994. **15**(2): p. 201-221.
14. Gosselin, P.H. and M. Cord, *Active learning methods for interactive image retrieval.* IEEE Transactions on Image Processing, 2008. **17**(7): p. 1200-1211.
15. Shi, L. and X.-c. Wang. *Artificial neural networks: Current applications in modern medicine*. in *Computer and Communication Technologies in Agriculture Engineering (CCTAE), 2010 International Conference On*. 2010. IEEE.
16. Madabhushi, A. and G. Lee, *Image analysis and machine learning in digital pathology: Challenges and opportunities.* Med Image Anal, 2016. **33**: p. 170-5.
17. Baxevanis, A.D. and A. Bateman, *The importance of biological databases in biological discovery.* Current protocols in bioinformatics, 2015. **50**(1): p. 1.1. 1-1.1. 8.
18. Cheplygina, V., et al., *Early experiences with crowdsourcing airway annotations in chest CT*, in *Deep Learning and Data Labeling for Medical Applications*. 2016, Springer. p. 209-218.
19. Szolovits, P., R.S. Patil, and W.B. Schwartz, *Artificial intelligence in medical diagnosis.* Annals of internal medicine, 1988. **108**(1): p. 80-87.

20. Orthuber, W., et al., *Design of a global medical database which is searchable by human diagnostic patterns.* The open medical informatics journal, 2008. **2**: p. 21.
21. Smeulders, A.W., et al., *Content-based image retrieval at the end of the early years.* IEEE Transactions on pattern analysis and machine intelligence, 2000. **22**(12): p. 1349-1380.
22. Müller, H., et al., *A review of content-based image retrieval systems in medical applications—clinical benefits and future directions.* International journal of medical informatics, 2004. **73**(1): p. 1-23.
23. Gong, T., et al. *Automatic pathology annotation on medical images: A statistical machine translation framework*. in *Pattern Recognition (ICPR), 2010 20th International Conference on*. 2010. IEEE.
24. Abe, N., B. Zadrozny, and J. Langford. *Outlier detection by active learning*. in *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*. 2006. ACM.
25. Doyle, S. and A. Madabhushi. *Consensus of Ambiguity: Theory and Application of Active Learning for Biomedical Image Analysis*. 2010. Berlin, Heidelberg: Springer Berlin Heidelberg.
26. Chen, L.-C., et al., *Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs.* IEEE transactions on pattern analysis and machine intelligence, 2018. **40**(4): p. 834-848.
27. Biosystems, L. *Aperio Imagescope*. Available from: https://www.leicabiosystems.com/digital-pathology/manage/aperio-imagescope/.
28. Skodras, A., C. Christopoulos, and T. Ebrahimi, *The JPEG 2000 still image compression standard.* IEEE Signal processing magazine, 2001. **18**(5): p. 36-58.
29. Ginley, B., J.E. Tomaszewski, and P. Sarder. *Automatic computational labeling of glomerular textural boundaries*. 2017.
30. Kato, T., et al., *Segmental HOG: new descriptor for glomerulus detection in kidney microscopy image.* BMC Bioinformatics, 2015. **16**: p. 316.
31. Sarder, P., B. Ginley, and J.E. Tomaszewski. *Automated renal histopathology: digital extraction and quantification of renal pathology*. 2016.
32. Simon, O., et al. *Multi-radial LBP Features as a Tool for Rapid Glomerular Detection and Assessment in Whole Slide Histopathology Images*. ArXiv e-prints, 2017. **1709**.
33. Goyal, S.N., et al., *Challenges and issues with streptozotocin-induced diabetes - A clinically relevant animal model to understand the diabetes pathogenesis and evaluate therapeutics.* Chem Biol Interact, 2016. **244**: p. 49-63.
34. Kitada, M., Y. Ogura, and D. Koya, *Rodent models of diabetic nephropathy: their utility and limitations.* Int J Nephrol Renovasc Dis, 2016. **9**: p. 279-290.
35. Wu, K.K. and Y. Huan, *Streptozotocin-induced diabetic models in mice and rats.* Curr Protoc Pharmacol, 2008. **Chapter 5**: p. Unit 5 47.
36. Hripcsak, G. and A.S. Rothschild, *Agreement, the f-measure, and reliability in information retrieval.* J Am Med Inform Assoc, 2005. **12**(3): p. 296-8.
37. Sokolova, M., N. Japkowicz, and S. Szpakowicz. *Beyond accuracy, F-score and ROC: a family of discriminant measures for performance evaluation*. in *Australasian joint conference on artificial intelligence*. 2006. Springer.
38. Japkowicz, N. and S. Stephen, *The class imbalance problem: A systematic study.* Intelligent data analysis, 2002. **6**(5): p. 429-449.
39. Bariety, J., et al., *Parietal podocytes in normal human glomeruli.* J Am Soc Nephrol, 2006. **17**(10): p. 2770-80.
40. Pavenstadt, H., W. Kriz, and M. Kretzler, *Cell biology of the glomerular podocyte.* Physiological Reviews, 2003. **83**(1): p. 253-307.

41. Solez, K., et al., *Banff 07 classification of renal allograft pathology: updates and future directions.* American journal of transplantation, 2008. **8**(4): p. 753-760.
42. Mengel, M., *Deconstructing interstitial fibrosis and tubular atrophy: a step toward precision medicine in renal transplantation.* Kidney international, 2017. **92**(3): p. 553-555.
43. Wang, X., et al., *Glomerular pathology in dent disease and its association with kidney function.* Clinical Journal of the American Society of Nephrology, 2016. **11**(12): p. 2168-2176.
44. McGarry, S.D., et al., *Radio-pathomic Maps of Epithelium and Lumen Density Predict the Location of High-Grade Prostate Cancer.* International Journal of Radiation Oncology*Biology*Physics, 2018. **101**(5): p. 1179-1187.
45. Janowczyk, A., et al., *A resolution adaptive deep hierarchical (RADHicaL) learning scheme applied to nuclear segmentation of digital pathology images.* Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization, 2016: p. 1-7.
46. Bray, T., et al., *Extensible markup language (XML).* World Wide Web Journal, 1997. **2**(4): p. 27-66.
47. Bradski, G., *The opencv library (2000).* Dr. Dobb's Journal of Software Tools, 2000.
48. Klette, R., et al., *Computer vision*. 1998: Springer-Verlag New York.
49. Goode, A., et al., *OpenSlide: A vendor-neutral software foundation for digital pathology.* Journal of pathology informatics, 2013. **4**.
50. Lu, C. and M. Mandal. *Automated segmentation and analysis of the epidermis area in skin histopathological images*. in *2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. 2012.
51. Govind, D., et al., *Automated erythrocyte detection and classification from whole slide images.* Journal of Medical Imaging, 2018. **5**(2): p. 027501.
52. Jung, A. *imgaug*. 2017; Available from: http://imgaug.readthedocs.io/en/latest/.
53. Zhou, Z.-H. and X.-Y. Liu, *Training cost-sensitive neural networks with methods addressing the class imbalance problem.* IEEE Transactions on Knowledge and Data Engineering, 2006. **18**(1): p. 63-77.

Acknowledgements

The project was supported by the faculty startup fund from the Pathology and Anatomical Sciences Department, University at Buffalo, the University at Buffalo IMPACT grant, the DiaComp Pilot and Feasibility Program grant #32307-5, and NIDDK grant R01 DK114485. We thank NVIDIA Corporation for the donation of the Titan X Pascal GPU used for this research (NVIDIA, Santa Clara, CA).

Author contributions

B.L. conceived the H-AI-L method, analyzed the data, and wrote the paper. The code was written by B.L. and B.G. D.G. contributed in generating results for Figure 4. S.D.M. and P.S.L. provided the radiology data and annotations for Figure 6. R.Y. implemented the mouse model. S.J. provided human renal biopsy data. J.E.T. evaluated renal pathology segmentation as a domain expert. K.Y.J. provided the IFTA annotation for Figure 5. P.S. is responsible for the overall coordination of the project, mentoring and formalizing the image analysis concept, and oversaw manuscript preparation.

Competing interests

The authors declare they have no competing interests.

Unsupervised Community Detection Using Potts Model Hamiltonian, an Efficient Algorithmic Solution, and Application in Digital Pathology

Brendon Lutnick[a], **Pinaki Sarder**[a,b,c,*]

[a]Department of Pathology and Anatomical Sciences

[b]Department of Biomedical Engineering

[c]Department of Biostatistics, University at Buffalo – The State University of New York, Buffalo, New York 14203,United States

***Pinaki Sarder, Tel : 716-829-2265, E-mail : pinakisa@buffalo.edu**

Abstract. Unsupervised segmentation of large datasets using a Potts model Hamiltonian[1] is unique in that segmentation is governed by a resolution parameter which scales the sensitivity to small clusters. Input data, represented as a graph is clustered by minimizing a Hamiltonian cost function. However, there exists no closed form solution, and using traditional iterative algorithmic solution techniques,[1] the problem scales with $(InputLength)^2$. Therefore, while Potts model clustering gives accurate segmentation, it is grossly underutilized as an unsupervised learning technique. Considering only distinct nodes while utilizing a fast statistical down-sampling of input data, we propose a fast and reproducible algorithmic solution, and demonstrate the application of the method in computational renal pathology in segmenting glomerular micro-environment. Our method is input size independent, scaling only with the number of features used to describe the data. This aspect makes our method uniquely suited for use in image segmentation tasks, giving it the ability to determine pixel specific segmentations from large 3-channel images ($\approx 10^8$ pixels) in seconds, $\approx 150000\times$ faster than previous implementations. However, our method is not limited to image segmentation, and using information theoretic measures,[2,3] we show that our algorithm outperforms K-means[4–8] and spectral clustering[9,10] on a synthetic dataset segmentation task.

Keywords: Potts model, Machine learning, Unsupervised segmentation, Glomerulus, Renal pathology..

1 Introduction

Segmentation is crucial in any large-scale data analysis problem. In unsupervised learning,[11] data clusters are learned from data relationships (determined from features) with no prior knowledge, meaning that segmentation is determined from the structure of the input data without any bias from predetermined class labels. This is particularly useful in data exploration,[12] as large training sets are not required for segmentation, which can reveal communities that are not immediately apparent.

Representing a dataset us-ing graph the-ory ap-proaches[13,14] (data-points are nodes, and data-relationships edges), unsupervised segmentation can be achieved by minimizing a Potts model Hamiltonian cost function.[1] This energy function is adopted from theoretical physics where it is used to describe the electrons in ferromagnetic materials by modeling the interactions of their spins.[15] Relationships (modeled as edges) are used to update the energy of the partitioned graph, which is iteratively improved until convergence. Unique to this segmentation method is the ability to tune a resolution parameter and modulate the sensitivity to small structures.[16] The Potts model is known in the literature to give precise and accurate segmentations[17] due to its ability to perform automatic model selection.[1,18]

For segmentation of large datasets, Hamiltonian optimization quickly becomes computationally unmanageable and its iterative solution limits parallelization, leading to long run times to reach convergence.[17,19] The computational challenges of Hamiltonian based segmentation are 2-fold: both graph generation and Hamiltonian optimization for large input datasets have large computational overheads. Pixel-scale image segmentation requires an input node for each pixel, and the calculation of a fully connected set of edges (pixel relations). Large edge matrices quickly overwhelm the memory limits of modern hardware, and are intensive to calculate. Hamiltonian optimization has been described as NP-Hard,[20,21] with the number of possible solutions scaling as $2^{nodes} - nodes$. There is no closed form solution for the Potts model Hamiltonian, requiring algorithmic solution which previously used relatively inefficient algorithms to achieve convergence.[1,19,22]

In this paper, we propose a new algorithmic approach that quickly converges, with tunable down sampling methods, able to segment large images exponentially faster than previous implementations,[1,19,22] extendable to any data-set with a discrete feature set. While we foresee applications of our method in diverse fields such as genomics, security, and social media, we test the performance quantitatively for image segmentation tasks, particularly, in segmenting glomerular microcompartments in renal pathology, and qualitatively for clustering of a synthetic dataset. Utilizing information theoretic measures,[2,3] we find that our method outperforms classical K-means[4] and spectral clustering[9,10] in segmentation of synthetic data. We then compare our method with two recent state-of-the-art implementations of K-means outlined in[5,6] and[7,8] respectively. Here we reiterate the automatic model selection provided by using the Hamiltonian cost, which is unique to Potts model segmentation.

This paper is organized as follows. In Section 2, we describe our clustering algorithm and present the mathematics which define its segmentation. In Section III we present results of segmentation using our Potts model algorithm, as well as K-means, and spectral clustering. In Section 4, we discuss the results presented in Section 3, and conclude in Section 5.

2 Method

We discuss the mathematical basis of Potts model segmentation, as well as our algorithmic approach to its solution. Here we discuss segmentation in terms of a general dataset with data-points and the related features.

2.1 Overview

Algorithmic advancements to iterative Hamiltonian cost function optimization enable scalable, multi-resolution segmentation on relevant timescales,[23] making the use of unsupervised Potts model based segmentation a feasible technique.

Traditionally graph based image segmentation methods include all image data-points as nodes, and segmentation of large datasets quickly becomes unmanageable as the number of edges in the fully connected graph increases with $\approx nodes^2/2$ where each data-point is a node.[24] To increase computational efficiency, methods may only use a random subset of edges, or create graphs connected by region adjacency.[24] Depending on application, we find that this can lead to segmentation errors, and inaccurate region dependent segmentation for applications in image segmentation.

Our method maintains a fully connected graph (full edge set) instead, choosing only distinct nodes, and for large data-sets, down-sampling before determining the distinct nodes. This greatly increases the clustering efficiency for data with constrained features, leading to a higher order reduction in possible edges. Traditional algorithmic solutions to the Potts model Hamiltonian involve optimizing data clustering by improving node-node relationships within each segment, our algorithmic approach allows a level of parallelization while reducing the computational load by considering node-cluster relationships to determine optimal clustering.

Any dataset with associated features can be clustered with a user specified resolution, which governs the sensitivity to small clusters. To optimize speed and memory allocation, the user can choose to limit the maximum number of nodes considered by our algorithm. Segmentation is achieved by our method in three steps:

1. *Graph generation* - Data down-sampling followed by determination of the distinct nodes to be used in the graph.

2. *Graph segmentation* - An iterative improvement of the Potts model cost function by altering the class labels of the graph.

3. *Full segmentation application* - Apply the segmentation determined above to the full graph for visualization.

The following sections describe these steps as follows. Graph generation is detailed in Sections 2.2 to 2.6. Graph segmentation is detailed in Section 2.7, and the full segmentation application in Section 2.8.

2.2 Dataset definition

Any vectorized dataset of length $M > 1$ with $N \geq 1$ associated features can be clustered by our algorithm, defined by:

$$\mathbf{D} = \left(\begin{array}{cccc} d_{11} & d_{12} & \cdots & d_{1N} \\ d_{21} & d_{22} & \cdots & d_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ d_{M1} & d_{M2} & \cdots & d_{MN} \end{array} \right) \tag{1}$$

where $\mathbf{D}$ is the data, containing M data-points, and $d_{\{.\}}$ are the data features, associated with each point. A 3-channel RGB image of size k-by-l can be input for clustering after vectorization, where an image is represented as a list of image pixels length $M = k \times l$, with each of R, G, and B value is an integer in $[0, 255]$. Each of the M data-points has $N = 3$ associated features ($R, G, \& B$ values) represented by rows in the vectorized image. A class labeled list of the same length M is output after clustering.

2.3 Graph definition

To perform clustering, the vectorized input data must be represented as a graph, where $\mathbf{V}$ are nodes, each of which corresponds to a data-point in $\mathbf{D}$, given as,

$$\mathbf{V} = \{\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_m\}, \tag{2}$$

where $1 < m \leq M$, and $\mathbf{E}$ are edges,

$$\mathbf{E} = \{e_{12}, e_{13}, \ldots, e_{1m}, e_{23}, e_{24}, \ldots, e_{2m}, \ldots, e_n\}, \tag{3}$$

where

$$n = \binom{m}{2}. \tag{4}$$

We do not employ all the data-points as nodes in the above graph, and use a highly efficient node selection process as described next to optimize the clustering speed of our proposed algorithm.

2.4 Node selection

The node selection process has the ability to incorporate 2 levels of reduction, determined by the size and complexity of the dataset. The first reduction excludes redundant data, and the second performs a down-sampling operation to simplify the resulting graph. This lowers the effective m, ensuring the graph is computationally manageable for modern hardware, essentially performing an over-segmentation, which has been shown in the literature to provide good results.[25]

2.4.1 Data selection

For datasets with $N > 1$ features, a Cantor pairing operation is used to determine distinct data-points.[26,27] This operation produces a unique number at every point in the N-dimensional space. The Cantor pairing output between the first two features in the i^{th} row is given as:

$$\pi_i(d_{i1}, d_{i2}) = \frac{1}{2}(d_{i1} + d_{i2})(d_{i1} + d_{i2} + 1) + d_{i2}, \tag{5}$$

where $i \in \{1, 2, ..., m\}$. Next Cantor pairing output between $\pi_i(d_{i1}, d_{i2})$ and d_{i3} is computed, and this process is subsequently repeated for $n - 1$ times for all the features in the i^{th} row, reducing the $M \times N$ dataset $\mathbf{D}$ to an M-by-1 representation. Unique $(\pi_i(\cdot)|i \in \{1, 2, ..., m\})$ in this reduced dataset are identified, and a reduced $\tilde{\mathbf{D}}$ is defined using the corresponding features from original dataset $\mathbf{D}$ as,

$$\tilde{\mathbf{D}} = \begin{pmatrix} \tilde{d}_{11} & \tilde{d}_{12} & \cdots & \tilde{d}_{1N} \\ \tilde{d}_{21} & \tilde{d}_{22} & \cdots & \tilde{d}_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ \tilde{d}_{m'1} & \tilde{d}_{m'2} & \cdots & \tilde{d}_{m'N} \end{pmatrix}, \tag{6}$$

where $\tilde{d}_{ij}$ are the features of the reduced dataset, containing m' data-points. Reducing a dataset to its distinct entries makes our algorithm scale with the feature-space occupancy rather than the dataset size m. For a dataset with discrete bounded features (pixel R, G, B values) this method often represents a significant reduction in the considered data, ensuring that the size of the input graph for Potts model based segmentation is not dependent on M, the input data size, rather the occupancy of its N-dimensional feature-space. In RGB image segmentation tasks, this removes the dependency on image size, however, there are still 255^3 possible nodes.

2.4.2 Data Down-sampling

To further speed clustering, statistical down sampling can be used to reduce the number of included nodes. In the case that $\tilde{\mathbf{D}}$ is large, we use a modified K-means algorithm, which quickly partitions the data into K discrete bins, to down-sample $\tilde{\mathbf{D}}$, defining a new $\tilde{\mathbf{D}}'$ before assignment as nodes. Here K is a user defined parameter defining the maximum number of nodes $\mathbf{V}$ can be used in Eq. (2). Practically, K should be much larger than the expected number of segments present in the image, allowing faster segmentation while inducing negligible error. Effectively this process bins the data, where error is minimal due to a large ratio of data bins to data segments.

Because the exact number of user specified groups K is not critical, we use a modified K-means algorithm optimized for speed. Rather than traditional K-means,[4] which groups data-points, our algo-rithm per-forms K-means clus-ter-ing in-di-vid-u-ally for each fea-ture di-men-sion in the N-dimensional feature-space, determining bins in each feature dimension independently. This technique surveys the feature-space rather than the data-structure to determine revised $\approx K$ data-points.

Using this technique, the number of groups in each feature is defined as:

$$k_j \in \{k_1, k_2, ..., k_N\}\,, \tag{7}$$

where k_j satisfies:

$$K \geq \prod_j k_j, \tag{8}$$

effectively breaking the data feature-space into $\leq K$ bins. Here k_j is determined by the ratio of the data variance in each feature, which is given by:

$$\sigma_j^2 = \frac{\sum_{i=1}^{m'} \left[\tilde{d}_{ij} - \mu_j\right]^2}{m'}, \tag{9}$$

where μ_j is the average of the data along j^{th} feature. Therefore the initial number of bins in each feature k_j can be determined by:

$$k_j = \left\lfloor \frac{\sigma_j^2}{\sum_j \sigma_j^2} \left(\frac{K}{\left(\prod_j \sigma_j^2\right)^{1/N}} \right) + \frac{1}{2} \right\rceil, \tag{10}$$

where $\lfloor \bullet \rceil$ is the rounding operation.

K-means clustering[4] is performed independently on each feature dimension, and for the j^{th} feature dimension, is given by:

$$\hat{\mathbf{b}}_{lj} | l \in \{1, 2, ..., k_j\} = \underset{\mathbf{b}_{lj} | l \in \{1,2,...,k_j\}}{\text{arg min}} \sum_{l=1}^{k_j} \sum_{i | \tilde{d}_{ij} \in \mathbf{b}_{lj}} (\tilde{d}_{ij} - \mu_{lj})^2, \tag{11}$$

where $\mathbf{b}_{lj}$ represents the $l = \{1, 2, ..., k_j\}^{\text{th}}$ cluster containing data from the j^{th} feature dimension in $\tilde{\mathbf{D}}$, μ_{lj} the mean of the corresponding data, and $\hat{\mathbf{b}}_{lj}$ the K-means clustering output.

To construct $\tilde{\mathbf{D}}'$, first the mean of the data in $\hat{\mathbf{b}}_{lj}$ is derived, and denoted as $\hat{\mu}_{lj}$. In each row of $\tilde{\mathbf{D}}'$, $\hat{\mu}_{lj}$ is used as the j^{th} feature, and rows are formed using all possible combinations of $\hat{\mu}_{lj}$, $\forall l \in \{1, 2, ..., k_j\}$. Further $\tilde{\mathbf{D}}'$ is reduced by selecting unique rows using Cantoor pairing operation as described above. This technique however does not guarantee always K rows to be present in the resulting $\tilde{\mathbf{D}}'$.

If user prefers to attain total number of data-points in $\tilde{\mathbf{D}}'$ to be K, k_j is iteratively updated, until $\approx K$ data-points are obtained in $\tilde{\mathbf{D}}'$. We constrain this approximate criteria using a parameter α to define the lower bound of the resulting data-points in $\tilde{\mathbf{D}}'$, defined as:

$$\alpha K \leq \prod_j k_j \leq K. \tag{12}$$

Further discussion about α can be found in Section 4. We use a proportional-integral-derivative (PID) control system[28,29] to iteratively change k_j (the number of groups in each feature dimension), quickly fulfilling the criteria defined by Eq. (12). This way the algorithm quickly attains the desired $\approx K$ in a generalizable fashion. We define the adjusted numbers of groups as k'_j, given by:

$$k'_j = \left\lfloor k_j \left(K_p e(t) + K_i \int_0^t e(t)dt + K_d \frac{de(t)}{dt} + 1 \right) \right\rceil . \tag{13}$$

Here t defines the iteration, and $e(t)$ represents the error at this iteration, given as:

$$e(t) = 1 - \frac{\prod_j k'_j}{K}. \tag{14}$$

A discussion of the tuning parameters K_p, K_i, and K_d can be found in Section 4.

Our modified K-means algorithm, as described above, is iteratively repeated, substituting k'_j for k_j, and updating $\tilde{\mathbf{D}}'$, until the criteria defined in Eq. (12) is met. The i^{th} ($\forall i \in \{1, 2, \ldots, m\}$) data-point in $\mathbf{V}$, is given in vector form as:

$$\mathbf{v}_i = [v_{i1}, v_{i2}, \cdots, v_{iN}], \tag{15}$$

where v_{ij} is corresponds to a data value in j^{th} ($\forall j \in \{1, 2, \ldots, N\}$) feature dimension.

This technique provides a non-linear down-sampling of the input data, which we have found to converge significantly faster than traditional K-means clustering.[4] However, in the case that the algorithm fails to converge, we limit the number of K-means iterations to a maximum of K.

2.5 Edge Calculation

Once the nodes $\mathbf{V}$ have been determined, a full set of edges $\mathbf{E}$ are calculated. We define an edge e_{ij} as the Euclidean distance between two nodes $\mathbf{v}_i$ and $\mathbf{v}_j$, $\forall j > i$ in Eq. (3). Edges are calculated for all combinations of nodes. The down-sampling done before node selection ensures that the number of edges is computationally manageable for a typical desktop computer with an Intel Core i7-4790 and 8 Gb RAM.

2.6 Average Edge Calculation

To determine attraction/repulsion in data relationships, the average edge value is used as a reference value in the Potts model optimization discussed below. Due to the non-uniform down-sampling used prior to calculation of $\mathbf{E}$ in Section 2.4.2, the mean of $\mathbf{E}$ is not representative of true edge average value. To avoid calculating edges for the full data set, we employ a uniform down-sampling of the data, reducing it to a maximum length of m'', where $m \leq m'' \leq M$ is satisfied. A full set of edges is computed using this reduced data-points using method described in Section 2.5, and is denoted as $\bar{e}$. Given m'' is sufficiently large, $\bar{e}$ will asymptotically approximate the true average edge closely, while greatly reducing the algorithmic overhead. Further discussion about m'' can be found in Section 4.

2.7 Graph Segmentation

The graph is clustered by minimizing a modified Potts model Hamiltonian:[1]

$$\mathcal{H} = \sum_{j=i+1}^{m} \sum_{i=1}^{m} (e_{ij} - \bar{e}) \left[\Theta(\bar{e} - e_{ij}) + \gamma \Theta(e_{ij} - \bar{e}) \right] \delta(S_i, S_j), \tag{16}$$

where the Heaviside function[30] determines which edges are considered, given by:

$$\varTheta(e_{ij} - \bar{e}) = \begin{cases} 1, & e_{ij} > \bar{e}, \\ 0, & \text{otherwise.} \end{cases} \tag{17}$$

The resolution parameter γ is used to tune the clustering. Decreasing γ results in clusters with lower intra-community density, revealing larger communities. The Kronecker delta[31] is given by:

$$\delta(S_i, S_j) = \begin{cases} 1, & S_i = S_j, \\ 0, & \text{otherwise,} \end{cases} \tag{18}$$

where $\delta(S_i, S_j)$ ensures that each node (spin) $\mathbf{v}_i$ interacts only with nodes in its own segment. Here S_i defines the segment identity of the i^{th} node. The segment identities are optimized by minimizing the Hamiltonian energy $\mathcal{H}$, thus, giving the segmented graph $S = \{S_1, S_2, \ldots, S_m\}$, where S_i is an integer and $\in \{1, 2, \ldots, s\}$, with total number of segments as s.

2.7.1 Algorithmic Energy Optimization

There exists no closed form solution to the Hamiltonian, given in Eq. (16). Moreover, the number of possible solutions to the clustering problem scales with $2^n - n$. Our major contribution to the field, is our algorithmic approach, which reduces both computational overhead, as well as user specified parameters, in comparison to the previous algorithmic approaches.[1,19,22]

To accomplish this above, we randomly initialize the starting node identities into s segments. While clustering can be completed with all nodes initialized in the same segment, we have found that random initialization helps to perturb the system leading to more accurate clustering.

Moving node-wise, our algorithm improves clustering through node-segment relationship optimization. The identity of the i^{th} node $\mathbf{v}_i$ is optimized by changing its identity and comparing the associated cost, given by:

$$S_{i\_\text{new}} = \underset{S_i \in \{1,2,\ldots,s+1\}}{\text{arg min}} \mathcal{H}, \tag{19}$$

where $s + 1$ is a newly formed segment containing only $\mathbf{v}_i$. The ability to form new segments enables our algorithm to automatically determine the optimal number of segments in the dataset, updated for every node. This node-wise energy update is iteratively repeated to optimize $\mathcal{H}$.

2.7.2 *Iterative Approach*

To sufficiently minimize $\mathcal{H}$, the energy update process in Eq. (19) is repeated for η iterations. We define an iteration as one cycle, with random order, through all nodes by Eq. (19).

Here η is determined when the segmentation fails to change after one full iteration, given as:

$$S_\eta \equiv S_{\eta-1}. \tag{20}$$

This ensures that no energetically favorable move exists for the system, and all nodes are in the optimal segment. This optimally segmented graph is then up-sampled, reversing the down-sampling performed in Section 2.4, to determine the segmented dataset.

2.8 *Segmentation Up-sampling*

The iterative segmentation approach, described in Section 2.7, is used to cluster the down-sampled graph. Using the segmented graph, node identities are assigned to the represented data-points, thus reversing the k-means, and Cantor data down-sampling described in Section 2.4. The resulting data segment-identity labels, output as a m-by-1 vector, correspond to the input data-points. An overview of the algorithmic pipeline and iterative solution is detailed in Fig. 1.

3 Results

Our optimized Potts model segmentation method was evaluated for speed and segmentation performance using benchmark images,[32] as well as segmentation of histologically stained murine glomerular microscopy images. For simplicity, we use image color (R, G, & B values) as $N = 3$

image features. Additionally, synthetic dataset was used for a more rigorous quantitative validation of the method. Here we compare the performance against other unsupervised segmentation methods: spectral clustering[9, 10] and K-means clustering.[4–8]

3.1 Image Segmentation

It is difficult to quantitatively evaluate method performance in the context of image segmentation as fully annotated ground truth segmentations are not readily available. However, we present our findings to give insight into the use of Potts model clustering for image segmentation. The most apparent advantage using a Potts model Hamiltonian for large data mining and clustering is the automatic model selection provided by this approach. Unlike spectral or K-means clustering, Potts model clustering automatically determines the optimal number of segments due to its algorithmic implementation.

3.1.1 Benchmark Image Segmentation

To validate our method on an independent dataset, clustering was performed on the Berkeley segmentation dataset.[32] The results of high/low resolution segmentation of four benchmark images are presented in Fig. 2. We found that using $m = 300$ nodes gave good clustering performance while optimizing algorithmic speed; taking an average of 3.87 sec to segment each 481×321 pixel image. Quantitative evaluation using this dataset is limited, as our algorithm was given pixel RGB values as image features, however, it can be seen that higher resolution gives more specific segmentation.

3.1.2 Glomerular Segmentation

To validate Potts model segmentation on an independent dataset, we used images of glomerular regions extracted from histologically stained whole slide murine renal tissue slices. The glomerulus is the blood-filtering unit of the kidney; a normal healthy mouse kidney typically contains thousands of glomeruli.[33,34] Basic glomerular compartments are Bowman's and luminal spaces, mesangial matrix, and resident cell nuclei.[35] In this paper, we demonstrate the feasibility of our proposed method in correctly segmenting these three biologically relevant glomerular regions, see Fig. 3. Image pixel resolution was 0.25 μm in this study. Once again we found that using $m \approx 300$ nodes gave good clustering performance, for segmenting $\approx 499 \times 441$ pixel glomerular RGB image, while optimizing algorithmic speed.

We analyzed the performance of our three basic glomerular compartment segmentation as stated above using renal tissue histopathology images from three wild-type mice.[36] Testing was done on five glomerular images per mouse, and evaluated against ground-truth segments generated by renal pathologist Dr. John E. Tomaszewski (University at Buffalo). The performance of our method was compared against compartmental segments jointly generated by two junior pathologists' (Dr. Buer Song and Dr. Rabi Yacoub) manual segmentations. Fig. 4 compares the precision and accuracy,[37] per time, of the Potts model and manual methods. Average precision and accuracy per unit segmentaion time (automatic or manual) were computed across mice, and standard deviations of these metrics over mice were computed. Comparison indicates Potts model segmentation significantly outperforms manual annotation with high efficiency.

Potts model segmentation was also compared against spectral[9,10] and classical K-means clustering;[4] see Fig. 6. Two different realizations of the segmentation using identical parameters are represented for each method. Qualitatively, we found the Potts model to give the best, and most reproducible segmentation. Spectral clustering also performs well, but gives less reproducible segmentation, while K-means does a poor job distinguishing glomerular compartments. With no optimization the Potts model automatically determines the three image classes using the baseline resolution ($\gamma = 1$), resembling the three biological compartments depicted in Fig. 3.

3.2 Synthetic Data Segmentation

To quantitatively evaluate our method, synthetic dataset was generated, and segmentation performance was evaluated using information theoretic measures.[2,3]

3.2.1 Synthetic Data Generation

To ensure that the synthetic dataset was easily separable in its given feature space, clusters were defined as 3-dimensional Gaussian distributions with dimension independent mean and variance. Altering the x, y, and z mean and variance for each cluster (distribution) controlled the separability of the clusters in the feature space. The number of nodes in each cluster was also altered. An example of this synthetic data is given in Fig. 7. For evaluation, the mean and variance values of the synthetic clusters were changed periodically to ensure robustness. However, all datasets were designed to give a small amount of overlap between classes to complicate the segmentation task.

3.2.2 Evaluation Metric

To quantitatively evaluate each methods clustering performance on the synthetic data, we utilized information theoretic measures.[2,3] Specifically method performance was evaluated by calculating the Normalized Mutual Information (NMI = I_{N}) between the clustered data c, and ground truth labels g, given as:

$$I_{\mathrm{N}}(c, g) = \frac{2I(c, g)}{H_c + H_g} \tag{21}$$

where $0 \leq I_{\mathrm{N}} \leq 1$. Here H is the Shannon entropy, and $I(c, g)$ is the mutual information between c and g. These metrics are given as:

$$H_c = -\sum_{k=1}^{s_c} \frac{N_k}{M} \log_2 \frac{N_k}{M} \tag{22}$$

and

$$I(c,g) = \sum_{k_1=1}^{s_c} \sum_{k_2=1}^{s_g} \frac{N_{k_1k_2}}{M} \log_2 \frac{N_{k_1k_2}M}{N_{k_1}N_{k_2}}. \tag{23}$$

Here N_k represents the cardinality of the k^{th} segment. Likewise, $N_{k_1k_2}$ denotes the common pixels in the k_1^{th} segment of c and k_2^{th} segment of g, and M is the total number of data-points.

3.2.3 Potts Model Performance

For synthetic data clustering, the Potts model was allowed to discover the number of data clusters. The resolution, γ, was tuned to optimize clustering, and the maximum number of nodes, m, was altered to study its effect on clustering performance, shown in Fig. 8. We find that for this dataset, Potts model clustering performs best at $\gamma \approx 0.02$ (Fig. 9), and performance increased with increasing nodes m. However, at $m \approx 300$, performance gains begin to have diminishing returns, as highlighted in Fig. 10. To optimize clustering performance and time, m should be given as ≈ 350. This is an acceptable compromise between method performance (Fig. 10) and speed (Fig. 11). In practice, the average clustering times presented in Fig. 11 would be significantly faster with proper resolution selection, as the clustering time increases with number of classes (s) determined by the algorithm (Fig. 12). Finally, to show our method's robustness, we present our algorithm's performance as a function of the number of random initial classes. While the method occasionally suffers as a result of convergence to a sub-optimal local minima, Fig. 13 shows that performance is consistent regardless of initialization.

3.2.4 Method Comparison

To compare clustering performance, segmentation was performed on the generated synthetic data (Fig. 7) first using classical K-means,[4] and spectral clustering.[9,10] For both methods, the correct number of classes was specified. For Potts model segmentation, the resolution was set to the optimal value, $\gamma = 0.02$. The results of 100 clustering realizations are presented in Fig. 14. The Potts model outperforms the classical K-means and spectral clustering, having the best mean and maximal NMI. Here the Potts mean NMI ≈ 0.96 matches the optimal one as shown in Fig. 9 and Fig. 10. Additionally, Fig. 15 presents the computational time taken by each method when clustering synthetic data. Here clustering times for the Potts model contain all combinations of γ and m, leading to a higher variation than spectral, and K-means clustering. However, while the Potts model is slower than the other methods, average clustering time is comparable across all three methods.

3.3 Comparison with Modern K-means

Improvements to the K-means clustering algorithm have been proposed in recent literature.[5–8] We evaluate our method against two such recent implementations of K-means[5] and.[7] Namely, these recently developed methods address the primary issue of the classical K-means method which is computationally inefficient and does not scale well with increasing data size. These recent K-means algorithms, referred to by the authors in their original codes as eakmeans[5] and kMeansCode[7] are compared to Potts model segmentation in Fig. 16. We find that the eakmeans algorithm outperforms the Potts model segmentation when the correct number of data classes is specified. However, fair comparison of Potts model and eakmeans performance is challenging, as the Potts model Hamiltonian cost performs automatic model selection.[1,18] Unlike eakmeans (or any K-means method), Potts model clustering does not require knowledge about the number of clusters present. To fairly compare these algorithms, we therefore present eakmeans-Poisson and eakmeans-uniform in Fig. 16. Here the number of clusters specified to the eakmeans algorithm is randomly sampled from a Poisson and uniform distribution respectively, drastically reducing the eakmeans method performance.

3.4 Data Sharing for Reproducibility

All of the source code and images used to derive the results presented within this manuscript are made freely available from `https://goo.gl/V3NatP`.

4 Discussion

The primary intention of this paper is to provide an overview of our proposed method for Potts model based segmentation, where the results above are merely applications to validate our method when applied to specific segmentation tasks. These analyses were performed on the raw data, with no pre-processing enhancements or feature selection; as a result the image segmentations presented in Sections 3.1.2 and 3.2 may be sub-optimal. We expect image segmentation to be limited without the use of high level contextual features, or image pre-processing. However these examples help evaluate the computational performance and scalability of our method. For future applications in image segmentation we would like to apply automated methods for feature selection such as sparse auto-encoders to represent image data in more meaningful dimensions.[38,39] The synthetic data presented in Fig. 7 is more representative of actual separable data containing abstract features, and while our method provides an approximation to the optimal clustering (due to the incomplete graph used), it outperforms both standard K-means and spectral clustering. Unlike these algorithms, the Potts model performs automatic model selection, selecting the number of clusters as a result of Hamiltonian optimization. While a modern implementation of K-means[5,6] outperforms Potts segmentation in Fig. 16, this is contingent on knowing the correct number of data-clusters. Unlike the Hamiltonian, where cost is regularized by the γ parameter which tunes cost penalization, K-means has no regularizing term to its cost function making it impossible to compare the clustering costs for different numbers of clusters.

To enhance algorithmic performance, we utilize several statistical assumptions, which reduce the complexity of computationally intensive problems. Namely the inclusion of modified K-means down-sampling in Section 2.4.2, and the uniform down-sampling in Section 2.6. In Section 2.4.2 Eq. (12), we propose α, a parameter which broadens the criteria for convergence of the modified K-means iteration. Practically we have set $\alpha = 0.95$ to ensure that the number of nodes selected, ϕ, is within 95% of the user specified value. We have found that this encourages fast convergence while maintaining an acceptable level of accuracy. Additionally in Eq. (13) we define the PID tuning parameters K_p, K_i, and K_d which have been assigned 0.5, 0.05, and 0.15, respectively. We find that these values provide fast settling times, while minimizing overshoot, satisfying Eq. (12) quickly. Likewise, in Section 2.6 we propose m'', a parameter which determines the maximum data-points used in the calculation of $\bar{e}$. Practically we define $m'' = 5000$, ensuring the calculation of $\bar{e}$ is fast. We have found that using $m'' = 5000$ gives accurate and reliable estimation of $\bar{e}$. Computing $\bar{e}$ with the down-sampling resulted in $\approx 1\%$ error, while exponentially increasing algorithmic speed.

The algorithmic solution we present quickly converges to stable solutions, but is not immune to poor initialization. While the algorithm automatically determines the correct number of segments, poor initializations often converge to sub-optimal local minima. Practically this occurs when no energetically favorable move exists for any node in the system, there may be a better solution, but to find it would require moves that increase the Hamiltonian cost. The effects of poor initialization are presented in Fig. 13, where the number of initial classes has no discernible trend on clustering performance, but poor initialization likely leads to occasional performance loss. The simplest solution is to repeat the segmentation several times, selecting the one with the lowest cost, $\mathcal{H}$. Alternatively, future study of optimal initialization techniques could help discover a computationally easier work around. The effects of the resolution parameter are not yet fully understood and in future work we plan to develop a theoretical framework for these effect through empirically study. Additionally we plan to study the effects of system perturbations on Hamiltonian optimization. Addition of robust perturbation functions to disturb system equilibrium, will likely benefit clustering performance.

5 Conclusion

The Potts model provides a unique approach to large scale data mining, its tunable resolution parameter and automatic model selection provide useful tools for cluster discovery. Unlike other unsupervised approaches, the number of clusters is determined by leveraging the data structure.

Previously, use of the Potts model was limited due to inefficient algorithmic optimization of the Hamiltonian cost function. Our approach circumvents this problem by offering an innovative iterative solution which is independent of initialization, and utilizing statistical simplifications of input data. This down-sampling employed by our method serves to approximate the optimal solution as segmentation of large datasets would be unfeasible without such assumptions. As a result, our method inherently scales with the N-dimensional feature space of the data, specifically with the number of distinct data-points. The algorithmic overhead can be further reduced using a modified K-means down-sampling method to sample the feature space prior to segmentation. In practice, the resolution (γ) and down-sampling (m) can be altered to optimize segmentation and speed, respectively, allowing the use of our method for data mining and discovery on any dataset with a discrete feature set.

Acknowledgment

This project was supported by the UB Impact Award, DiaComp Pilot and Feasibility Program grant #32307-5, and faculty start-up funds from the Pathology & Anatomical Sciences Department, Jacobs School of Medicine and Biomedical Sciences, University at Buffalo. We thank Dr. John E. Tomaszewski, Dr. Rabi Yacoub, and Dr. Buer Song for the pathological annotations (see Fig. 3 and Fig. 4) and biological guidance they provided which was essential to the glomerular segmentation results presented in Section 3.1.2, as well as the histology core facility (Pathology & Anatomical Sciences, University at Buffalo) for performing histopathological staining of tissues.

References

1 P. Ronhovde and Z. Nussinov, “Local resolution-limit-free potts model for community detection,” *Phys. Rev. E*, vol. 81, pp. 046 114: 1–15, 2010.

2 G. Bouma, “Normalized (pointwise) mutual information in collocation extraction,” in *From Form to Meaning: Processing Texts Automatically, Proceedings of the Biennial GSCL Conference 2009*, vol. 156, Tübingen, 2009, pp. 31–40.

3 G. N. Nair, “A nonstochastic information theory for communication and state estimation,” *IEEE Trans. Automatic Control*, vol. 58, pp. 1497–1510, 2013.

4 T. Kanungo, D. M. Mount, N. S. Netanyahu, C. D. Piatko, R. Silverman, and A. Y. Wu, “An efficient k-means clustering algorithm: analysis and implementation,” *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 24, pp. 881–892, 2002.

5 J. Newling and F. Fleuret, “Fast k-means with accurate bounds,” in *Proceedings of The 33rd International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, M. F. Balcan and K. Q. Weinberger, Eds., vol. 48, New York, New York, USA, 2016, pp. 936–944.

6 ——, “Nested mini-batch k-means,” in *Proceedings of NIPS*, 2016.

7 M. Shindler, A. Wong, and A. Meyerson, “Fast and accurate k-means for large datasets,” in *Proceedings of the 24th International Conference on Neural Information Processing Systems*, ser. NIPS’11. USA: Curran Associates Inc., 2011, pp. 2375–2383.

8 V. Braverman, A. Meyerson, R. Ostrovsky, A. Roytman, M. Shindler, and B. Tagiku, “Streaming k-means on well-clusterable data,” in *Proceedings of the Twenty-second Annual ACM-SIAM Symposium on Discrete Algorithms*, ser. SODA ’11. Philadelphia, PA, USA: Society for Industrial and Applied Mathematics, 2011, pp. 26–40.

9 A. Y. Ng, M. I. Jordan, and Y. Weiss, “On spectral clustering: Analysis and an algorithm,” in *Advances in Neural Information Processing Systems 14*, T. G. Dietterich, S. Becker, and Z. Ghahramani, Eds. MIT Press, 2002, pp. 849–856.

10 L. Ding, F. M. Gonzalez-Longatt, P. Wall, and V. Terzija, “Two-step spectral clustering controlled islanding algorithm,” *IEEE Trans. Power Systems*, vol. 28, pp. 75–84, 2013.

11 Q. V. Le, “Building high-level features using large scale unsupervised learning,” in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013, pp. 8595–8598.

12 A. Wasay, M. Athanassoulis, and S. Idreos, “Queriosity: Automated data exploration,” in *2015 IEEE International Congress on Big Data*, June 2015, pp. 716–719.

13 E. R. Scheinerman and D. H. Ullman, *Fractional graph theory: A rational approach to the theory of graphs*. Mineola, NY: Dover Publications, 2011.

14 M. T. Pham, G. Mercier, and J. Michel, “Change detection between sar images using a pointwise approach and graph theory,” *IEEE Trans. on Geoscience and Remote Sensing*, vol. 54, pp. 2020–2032, 2016.

15 F. Y. Wu, “The potts model,” *Rev. Mod. Phys.*, vol. 54, pp. 235–268, 1982.

16 Y. Yu, Z. Cao, and J. Feng, *Continuous Potts Model Based SAR Image Segmentation by Using Dictionary-Based Mixture Model.* New York, NY: Springer International Publishing, 2014, pp. 577–585.

17 D. Hu, P. Ronhovde, and Z. Nussinov, “Replica inference approach to unsupervised multiscale image segmentation,” *Phys. Rev. E*, vol. 85, pp. 016 101: 1–25, 2012.

18 J. L. Castle, J. A. Doornik, and D. F. Hendry, “Evaluating automatic model selection,” *Journal of Time Series Econometrics*, no. 1, pp. 1–31, 2011.

19 D. Hu, P. Sarder, P. Ronhovde, S. Orthaus, S. Achilefu, and Z. Nussinov, “Automatic segmentation of fluorescence lifetime microscopy images of cells using multiresolution community detection–a first study,” *Journal of Microscopy*, vol. 253, no. 1, pp. 54–64, 2014.

20 I. Kovtun, “Partial optimal labeling search for a np-hard subclass of (max,+) problems,” in *Pattern Recognition: 25th DAGM Symposium, Magdeburg, Germany, September 10-12, 2003, Proceedings*, B. Michaelis and G. Krell, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2003, pp. 402–409.

21 Y. Liu, C. Gao, Z. Zhang, Y. Lu, S. Chen, M. Liang, and L. Tao, “Solving np-hard problems with physarum-based ant colony system,” *IEEE/ACM Trans. Comput. Biol. Bioinformatics*, vol. 14, no. 1, pp. 108–120, 2017.

22 D. Hu, P. Sarder, P. Ronhovde, S. Achilefu, and Z. Nussinov, “Community detection for fluorescent lifetime microscopy image segmentation,” *Proc. SPIE*, vol. 8949, pp. 89 491K: 1–13, 2014.

23 R. Y. Levine and A. T. Sherman, “A note on bennetts time-space tradeoff for reversible computation,” *SIAM Journal on Computing*, vol. 19, pp. 673–677, 1990.

24 B. Peng, L. Zhang, and D. Zhang, “A survey of graph theoretical approaches to image segmentation,” *Pattern Recognition*, vol. 46, pp. 1020–1038, 2013.

25 R. C. Amorim, V. Makarenkov, and B. G. Mirkin, “A-ward$_{p\beta}$: Effective hierarchical clustering using the minkowski metric and a fast k-means initialisation,” *Information Sciences*, vol. 370, pp. 343–354, 2016.

26 J. E. Hopcroft, R. Motwani, and J. D. Ullman, *Introduction to Automata Theory, Languages, and Computation*, 3rd ed. Boston, MA: Addison-Wesley Longman Publishing Co., Inc., 2006.

27 S. Pigeon, “Pairing function,” MathWorld—A Wolfram Web Resource. [Online]. Available: http://mathworld.wolfram.com/PairingFunction.html

28 K. H. Ang, G. Chong, and Y. Li, “Pid control system analysis, design, and technology,” *IEEE Trans. Control Systems Technology*, vol. 13, pp. 559–576, 2005.

29 H. M. Hasanien, “Design optimization of pid controller in automatic voltage regulator system using taguchi combined genetic algorithm method,” *IEEE Systems Journal*, vol. 7, pp. 825–831, 2013.

30 E. W. Weisstein, “Heaviside step function,” MathWorld—A Wolfram Web Resource. [Online]. Available: http://mathworld.wolfram.com/HeavisideStepFunction.html

31 ——, “Kronecker delta,” MathWorld—A Wolfram Web Resource. [Online]. Available: http://mathworld.wolfram.com/KroneckerDelta.html

32 D. Martin, C. Fowlkes, D. Tal, and J. Malik, “A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics,” in *Proc. 8th Int’l Conf. Computer Vision*, vol. 2, 2001, pp. 416–423.

33 B. Young, G. O’Dowd, and P. W. P, *Wheater’s functional histology: A text and colour atlas*, 6th ed. London, United Kingdom: Churchill Livingstone, 2013.

34 M. R. Pollak, S. E. Quaggin, M. P. Hoenig, and L. D. Dworkin, “The glomerulus: The sphere of influence,” *Clin. J. Am. Soc. Nephrol.*, vol. 9, pp. 1461–1469, 2014.

35 W. Kriz, N. Gretz, and K. V. Lemley, “Progression of glomerular diseases: is the podocyte the culprit?” *Kidney International*, vol. 54, pp. 687–697, 1998.

36 G. Tesch, H. Greg, and T. J. Allen, “Rodent models of streptozotocin-induced diabetic nephropathy (methods in renal research),” *Nephrology*, vol. 12, pp. 261–266, 2007.

37 R. H. Fletcher and S. W. Fletcher, *Clinical epidemiology: The essentials*, 4th ed. Baltimore, MD: Lippincott Williams & Wilkins, 2005.

38 A. Ng, "Sparse autoencoder," *CS294A Lecture Notes*, vol. 72, pp. 1–19, 2011.

39 J. Xu, L. Xiang, Q. Liu, H. Gilmore, J. Wu, J. Tang, and A. Madabhushi, "Stacked sparse autoencoder for nuclei detection on breast cancer histopathology images," *IEEE Trans. on Medical Imaging*, vol. 35, pp. 119–130, Jan 2016.

40 C. Li and A. C. Bovik, "Content-partitioned structural similarity index for image quality assessment," *Signal Processing: Image Communication*, vol. 25, pp. 517–526, 2010.

41 Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Processing*, vol. 13, pp. 600–612, 2004.

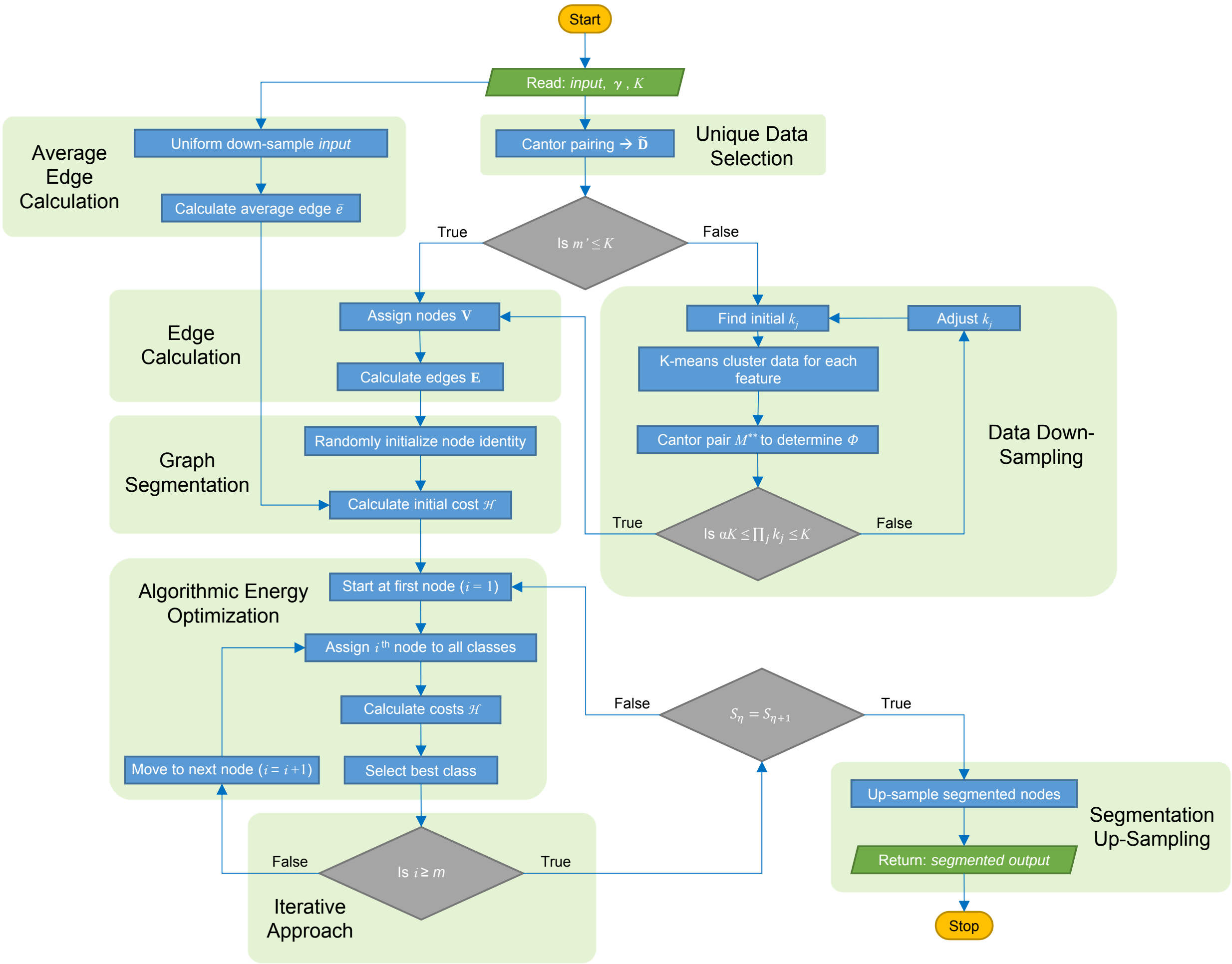
Start
Read: *input*, γ , K
Average Edge Calculation
Uniform down-sample *input*
Calculate average edge $\bar{e}$
Cantor pairing → $\widetilde{\mathbf{D}}$
Unique Data Selection
True
Is $m' \leq K$
False
Edge Calculation
Assign nodes $\mathbf{V}$
Calculate edges $\mathbf{E}$
Find initial k_j
Adjust k_j
K-means cluster data for each feature
Cantor pair M^{**} to determine Φ
Data Down-Sampling
Graph Segmentation
Randomly initialize node identity
Calculate initial cost $\mathcal{H}$
True
Is $\alpha K \leq \prod_j k_j \leq K$
False
Algorithmic Energy Optimization
Start at first node $(i = 1)$
Assign i^{th} node to all classes
Calculate costs $\mathcal{H}$
Move to next node $(i = i + 1)$
Select best class
False
$S_\eta = S_{\eta+1}$
True
Up-sample segmented nodes
Return: *segmented output*
Segmentation Up-Sampling
Stop
False
Is $i \geq m$
True
Iterative Approach

Fig 1 Algorithm flow chart detailing our iterative solution pipeline. The steps are labeled by section.

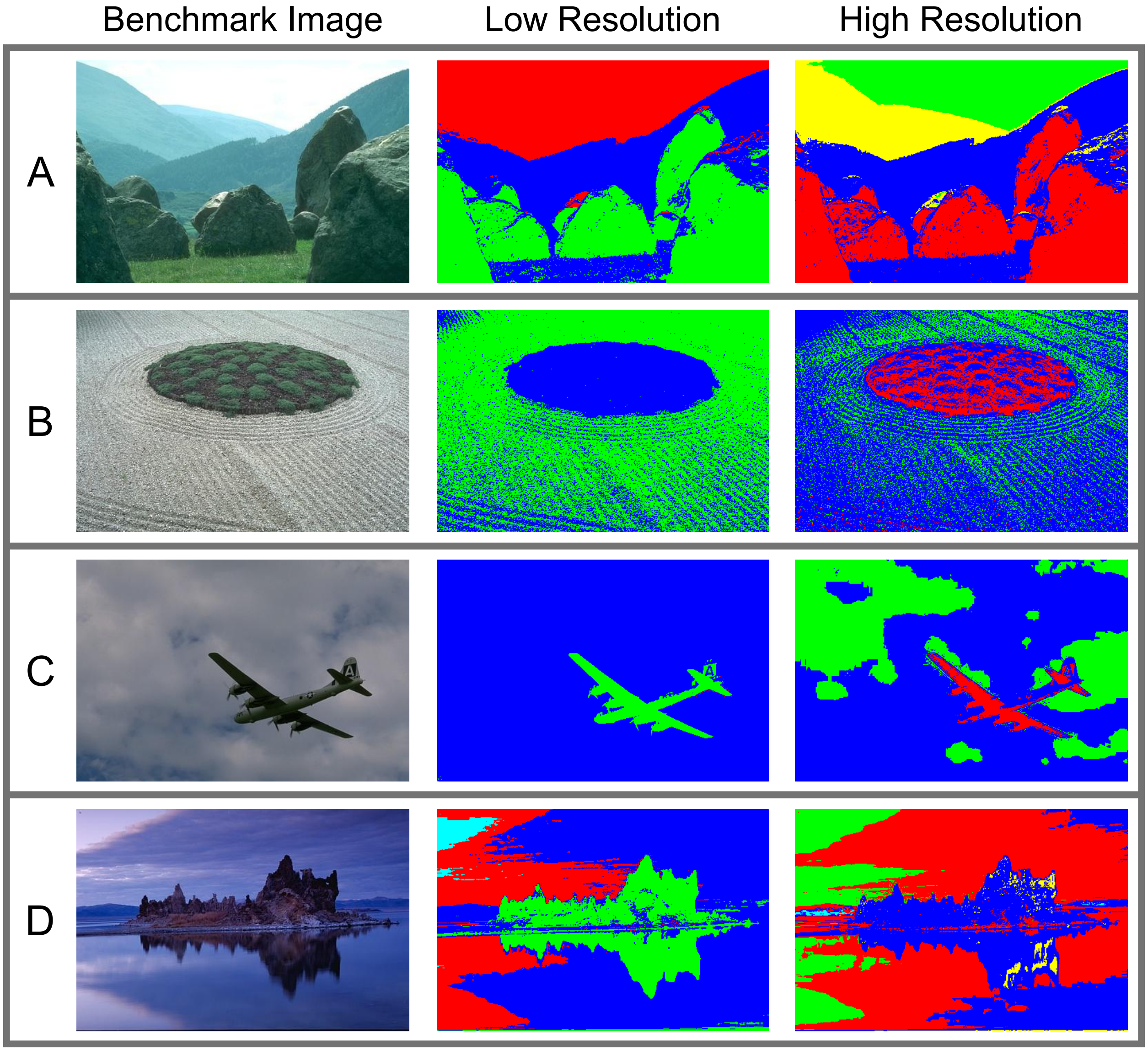

Fig 2 Benchmark image segmentation using a Potts model Hamiltonian. Segmentation at low resolution ($\gamma = 5$) and high resolution ($\gamma = 50$) were performed on four benchmark images. Raw pixel RGB values ($n = 3$) were used as image features, no pre-processing was done to enhance segmentation. Low resolution segmentations were completed in $\eta \approx 4$ iterations, with high resolution longer to converge, $\eta \approx 6$. Here color represents segments as determined by the algorithm.

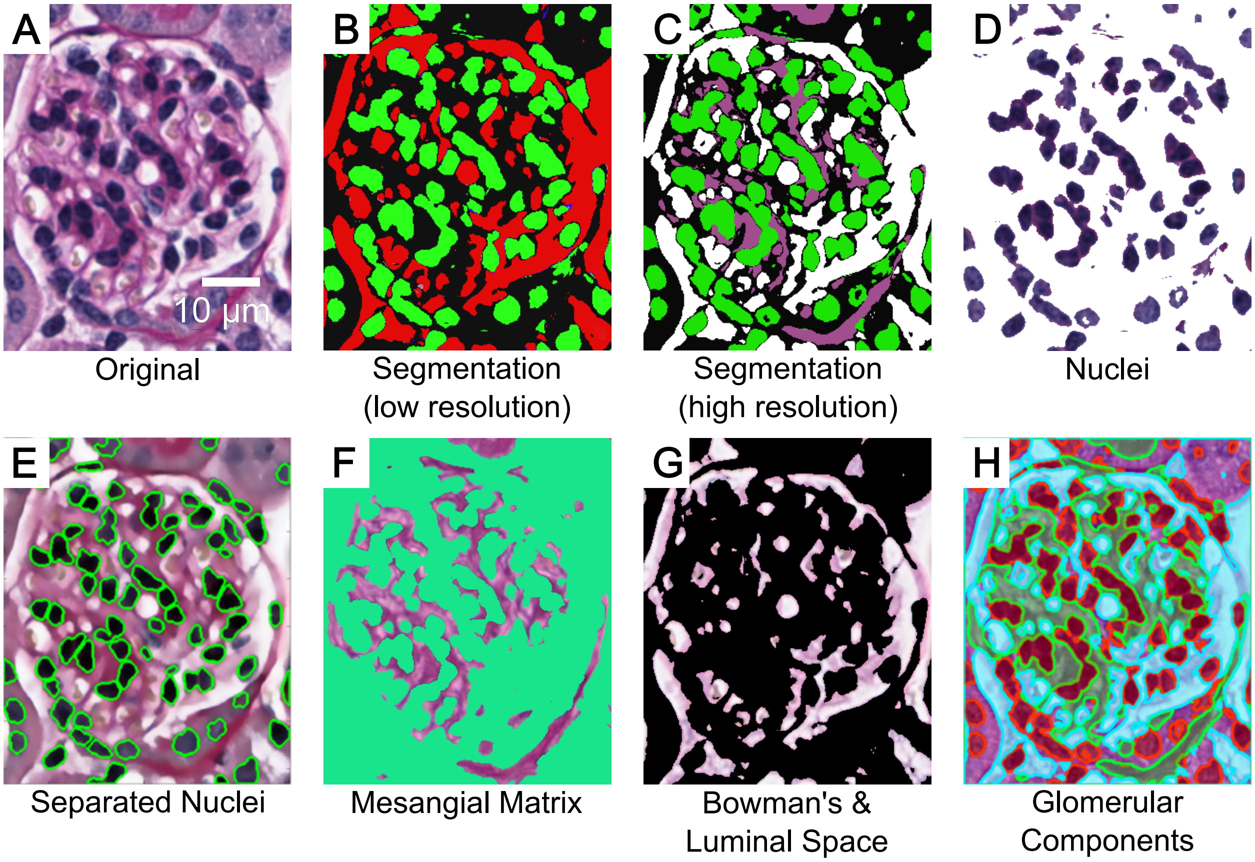

Fig 3 Murine renal glomerular compartment segmentation using a Potts model Hamiltonian. (A) The original glomerulus image, (B) low resolution ($\gamma \approx 0.5$) segmentation, and (C) high resolution ($\gamma \approx 5$) segmentation, (D) segmented nuclei, (E) separated nuclei superimposed on (A) using morphological processing, (F) segmented mesangial matrix, and (G) segmented Bowman's/luminal space. Compartment segments depicted in (D, F, and G) were obtained at optimally chosen γ values where the respective compartment segmentations were verified by a renal pathologist (Dr. John E. Tomaszewski, University at Buffalo). (H) All three segmented components (D, F, and G) overlaying on the original image. All segmentations use $m = 350$ nodes. Here color is used to signify segments, but is not conserved between panels, and the background colors in (D, F, and G) were chosen to enhance contrast in the image for visualization.

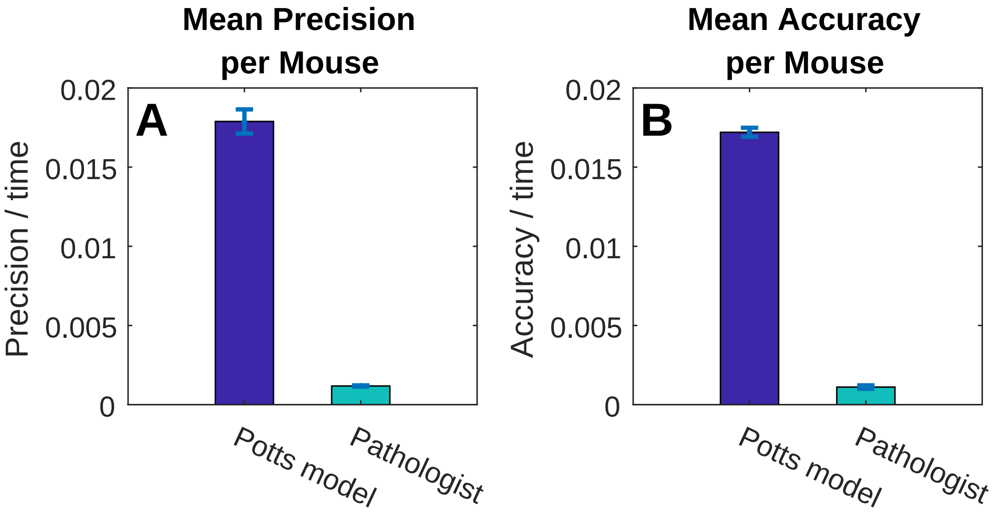

Fig 4 Comparison of performance between Potts model and manual methods in segmenting murine intra-glomerular compartments. (A) Precision per time. (B) Accuracy per time. Five glomeruli images per mouse from three normal healthy mice were used. Error-bars for the precision and accuracy metrics indicate standard deviation. Potts model based segmentation significantly outperforms manual method.

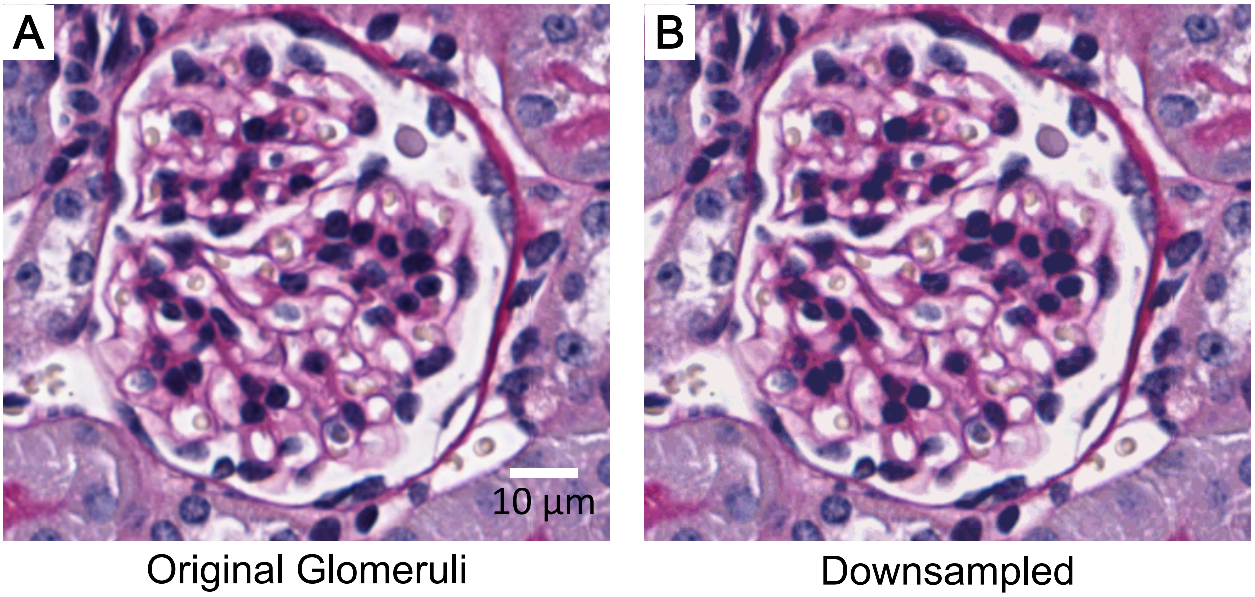

Fig 5 A common PAS stained glomeruli image, containing 499×441 RGB pixels. (A) The original cropped image containing 137034 distinct colors. (B) Depicts the same image down-sampled to $m = 350$ colors. The structural similarity index[40,41] between (A) and (B) is 0.97, despite a 391.5% reduction in colors information.

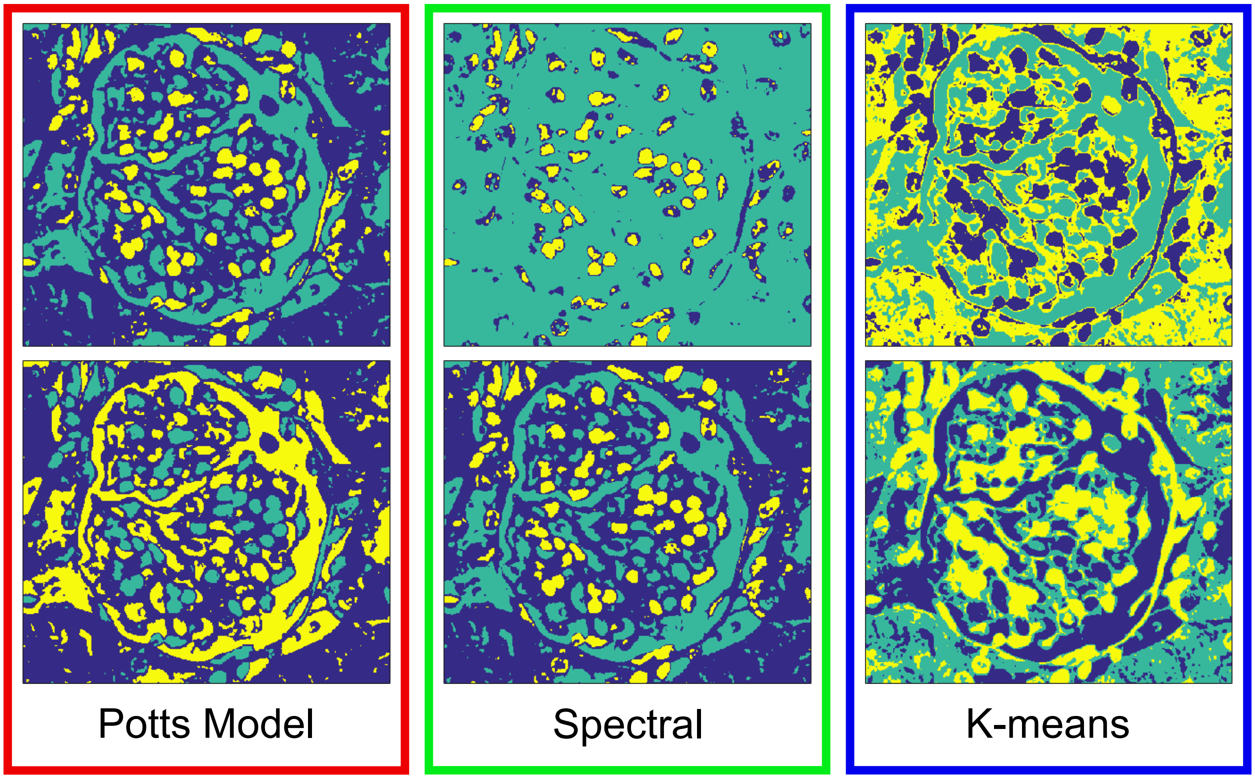

Fig 6 Glomerular compartment segmentation using different clustering methods. The original image is depicted in Fig. 5-A. Identical parameters were used to generate both segmentations for each method: Potts model resolution was $\gamma = 1$, spectral and K-means employed three classes for the investigation. For all segmentation, pixel RGB values were used as the three image features. For potts model clustering, $m = 350$ was used, as depicted in Fig. 5-B.

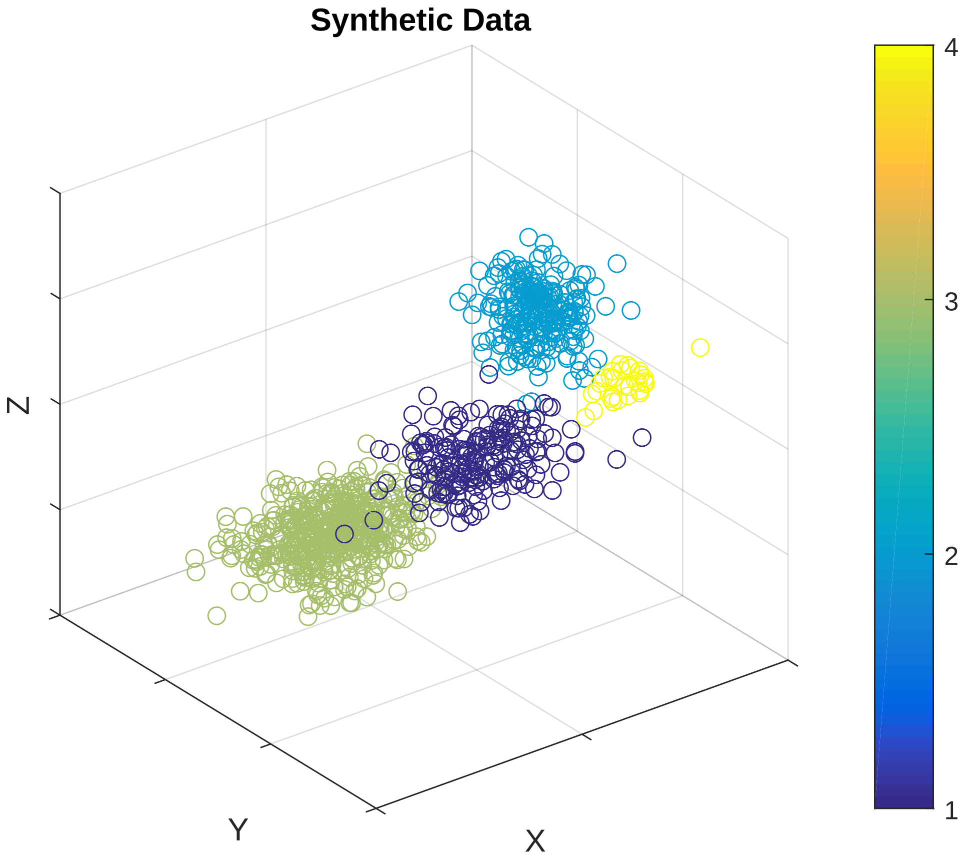

Fig 7 Synthetic Data, containing 1000 points in 4 normally distributed classes.

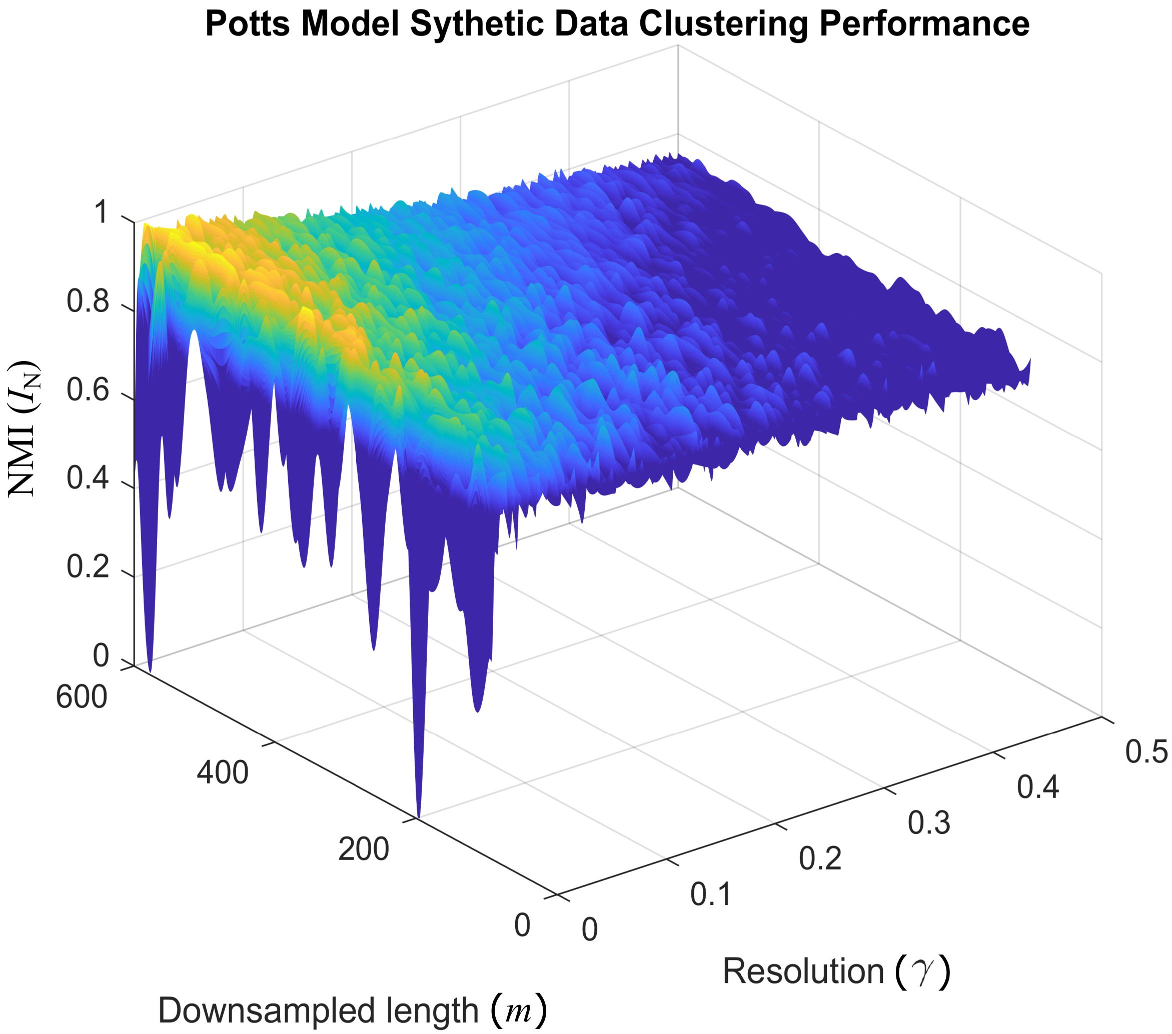

Fig 8 $I_{\mathrm{N}}(\gamma, m)$ - Potts model performance (NMI) as a function of resolution and nodes, when clustering synthetic data (shown in Fig. 7). Due to the simple nature of the dataset clustered, we observe optimal performance at a low resolution $\gamma = 0.02$. While the clustering performance increases with the number of included nodes, m, near optimal performance is observed when $m \geq 350$.

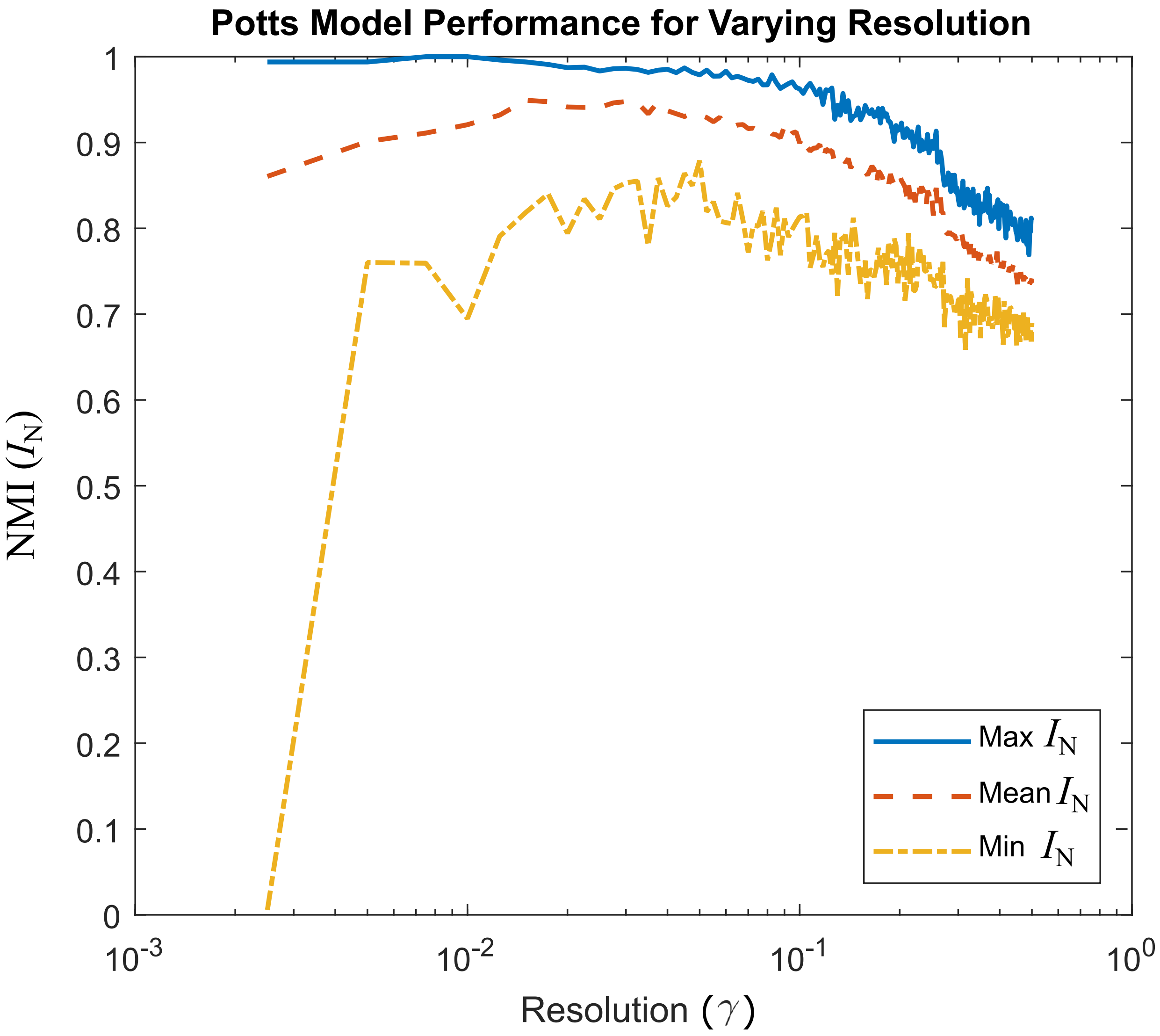

Fig 9 Potts model performance as a function of resolution γ, when clustering synthetic data (shown in Fig. 7). The best performance is achieved at $\gamma = 0.02$. The result represents 36 realizations at each resolution, varying m between $350 \leq m \leq 600$. We observe optimal performance at $\gamma = 0.02$.

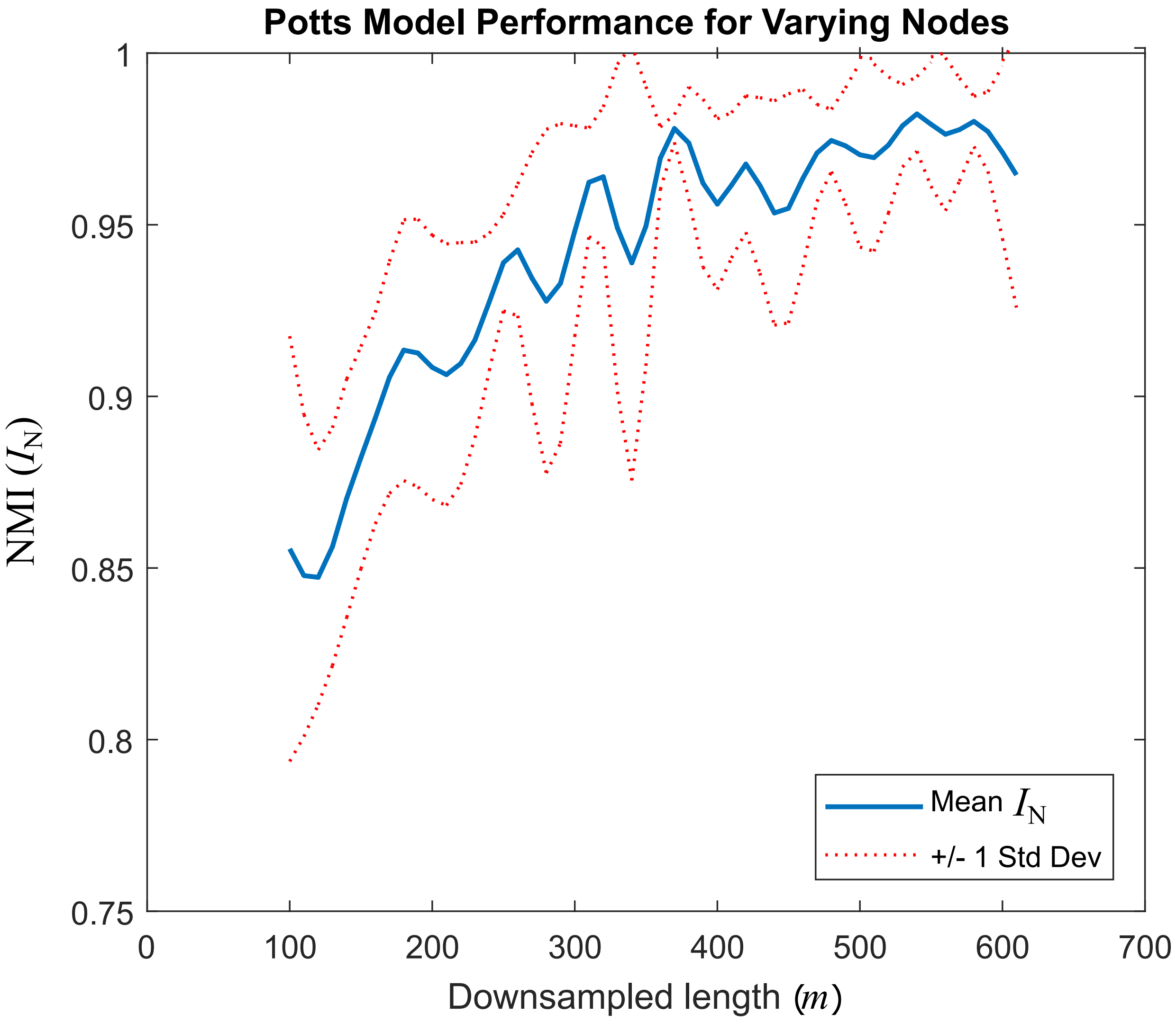

Fig 10 Average Potts model performance as a function of the number of nodes m, when clustering synthetic data (shown in Fig. 7). Clustering at $0.01 \leq \gamma \leq 0.03$ are included in the averaged performance. The result represents 15 realizations at each down-sampled length. We observe that clustering performance is consistent for $350 \leq m \leq 600$.

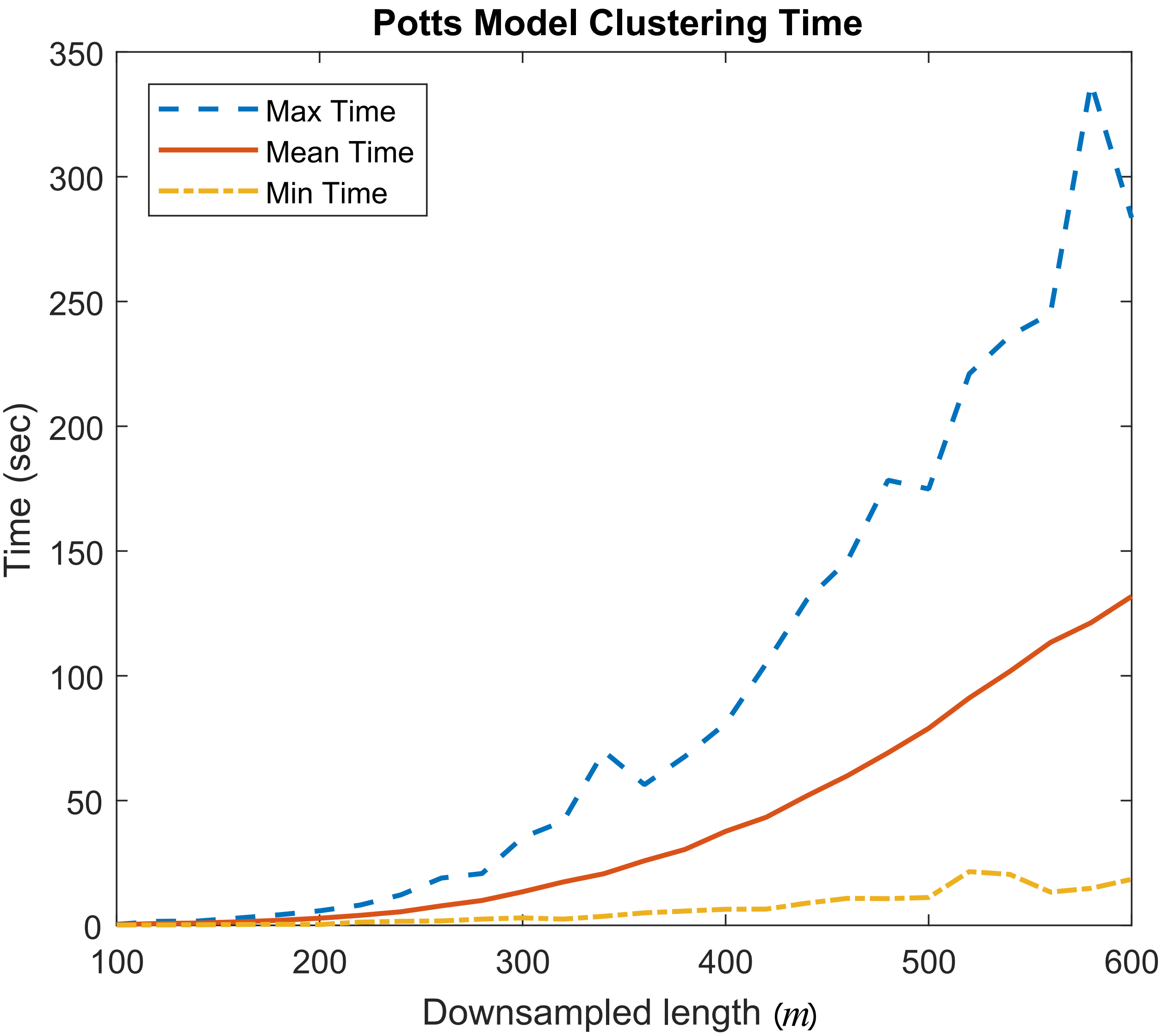

Fig 11 Potts model convergence times as a function of the number of nodes m, when clustering synthetic data (shown in Fig. 7). We observe that our algorithm scales with $\approx m^2 - m$ as expected, highlighting the effect of reducing m.

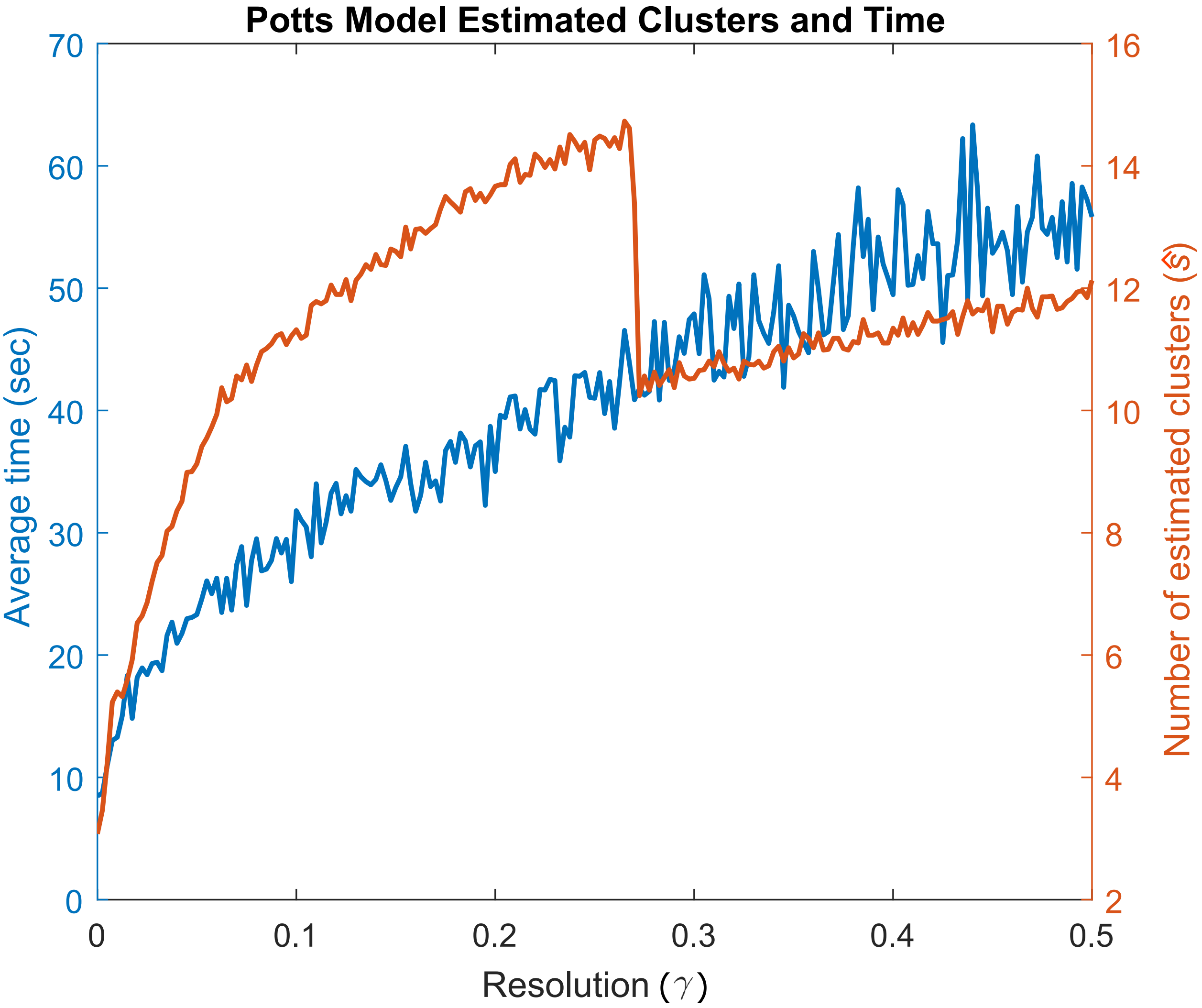

Fig 12 Average convergence time and number of estimated clusters, $\hat{s}$, as a function of Potts model resolution (γ), when clustering synthetic data (shown in Fig. 7). Results were averaged over all down-sampled lengths $100 \leq m \leq 600$. The algorithm determines the correct number of clusters ($\hat{s} = 4$) at $\gamma \approx 0.02$. The jump seen in the number of clusters at $\gamma \approx 0.275$ resembles those seen in previous work[19] and occurs at unstable resolutions.

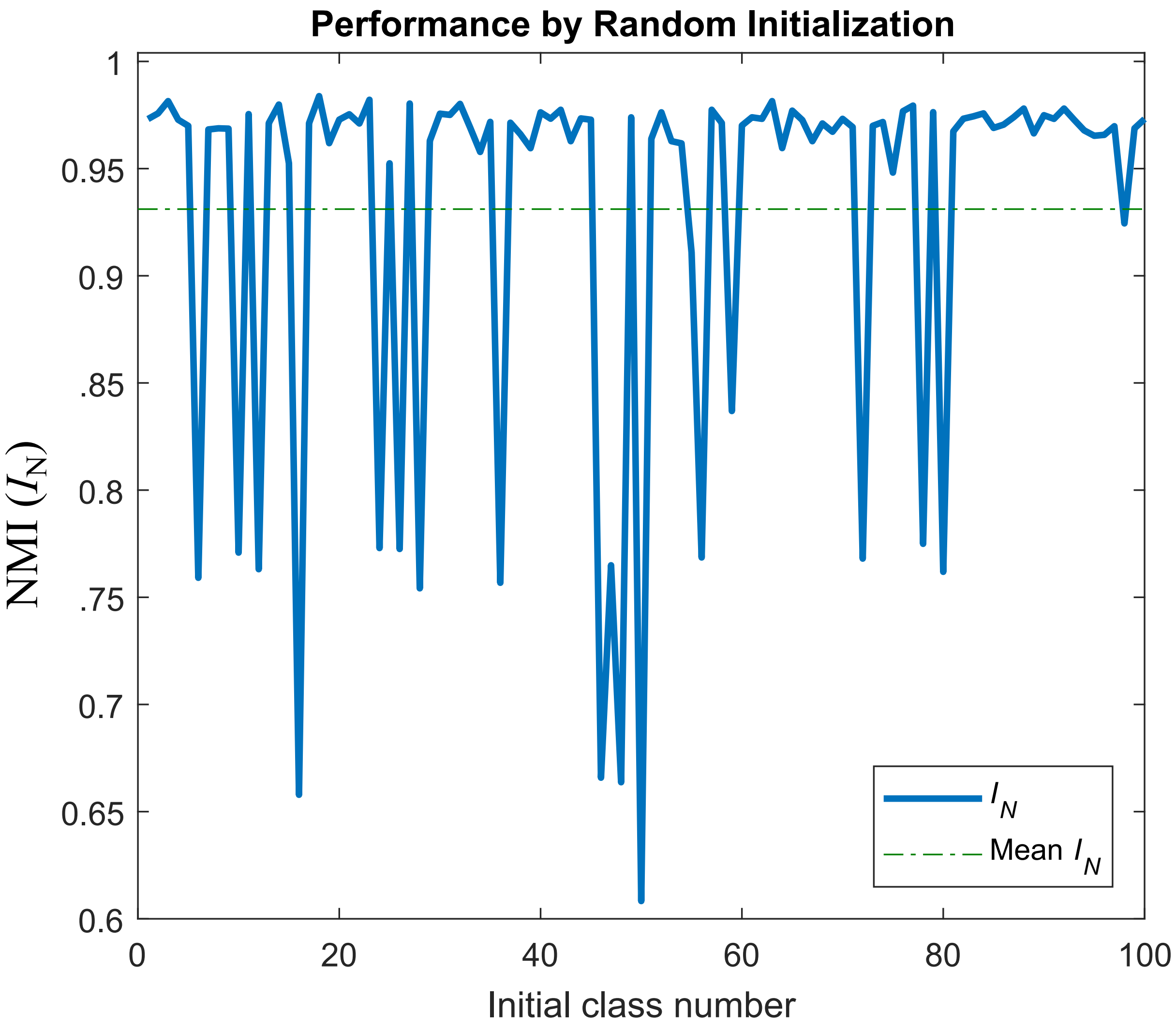

Fig 13 Potts model clustering performance as a function of the number of randomized initial classes (s), when clustering a synthetic dataset. Result was generated using $\gamma = 0.02$ and $m = 250$. Overall we find no correlation between the number of random initialization classes and method performance, indicating that our algorithm is capable of converging to an optimal solution independent of initialization. We do observe random drops in performance, which we attribute to convergence to suboptimal local minima.[19]

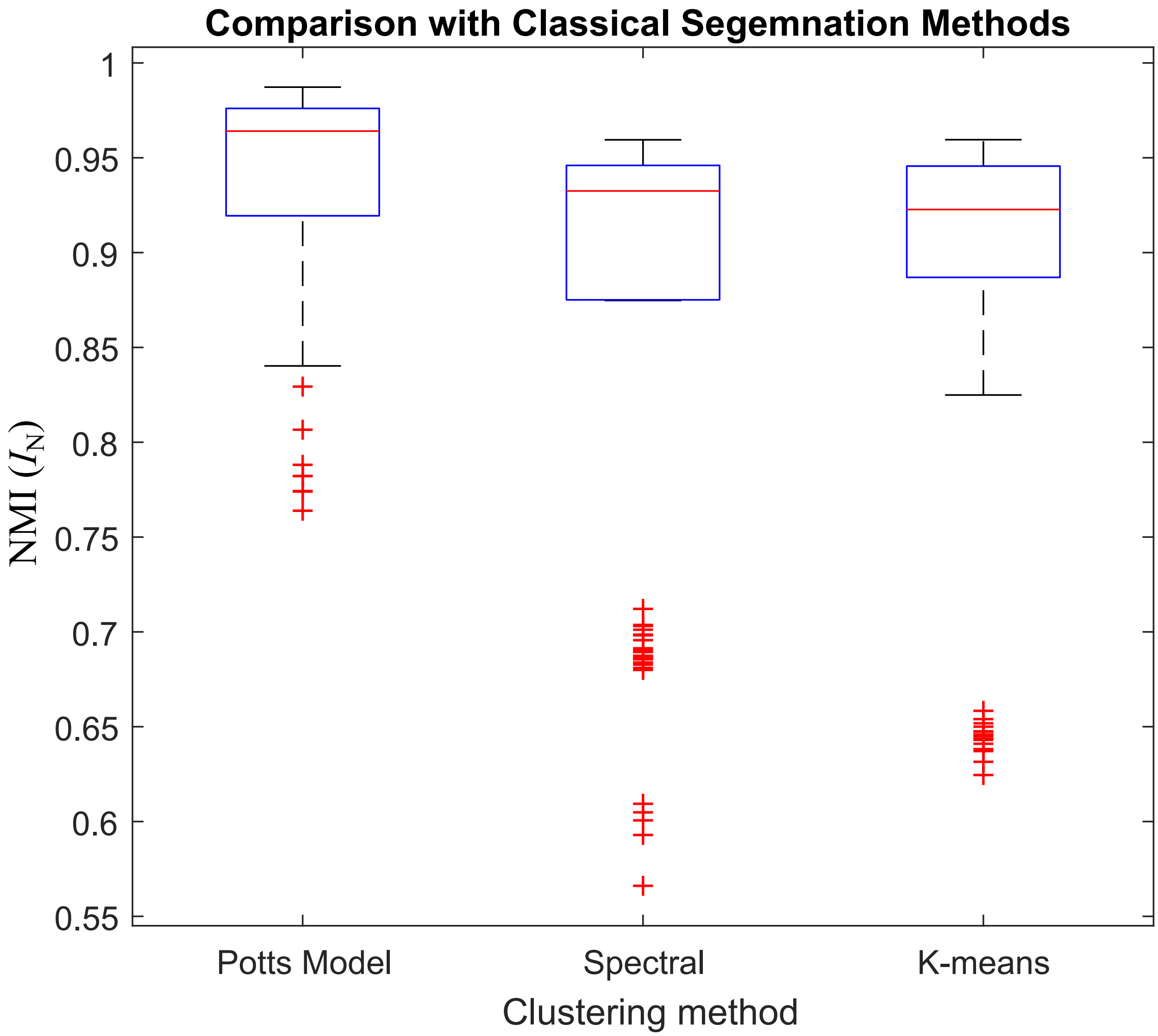

Fig 14 Potts model segmentation performance with respect to classical segmentation methods – Spectral and K-means clustering. For spectral and K-means, the number of classes employed to investigate was four. Potts model clustering was performed at $\gamma = 0.02$ and $m = 350$. The Potts model outperforms the other two methods.

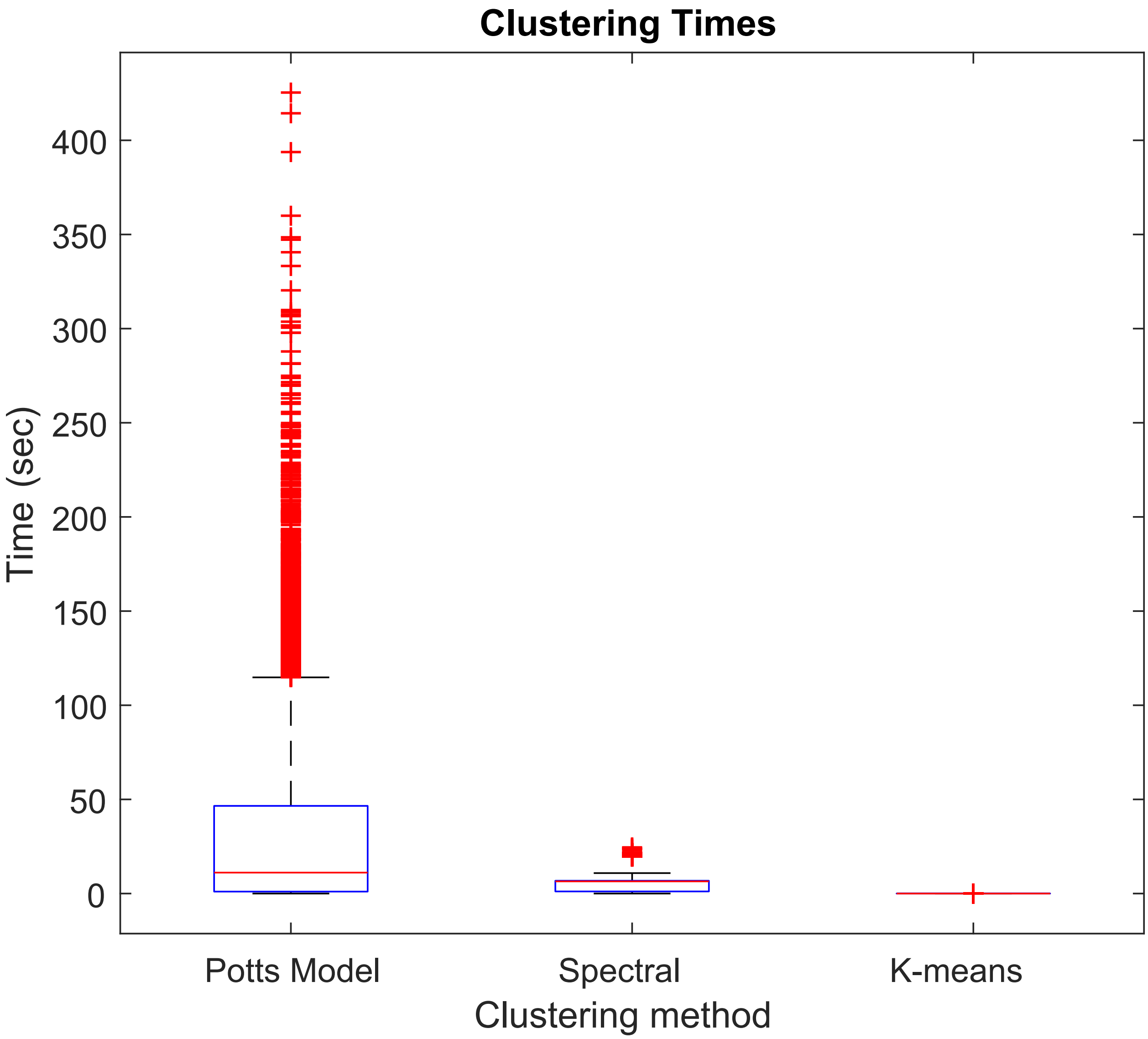

Fig 15 Clustering time by method. For Spectral and K-means, the number of classes employed to investigate was four. Potts model clustering was performed at $0 \leq \gamma \leq 0.5$ and $100 \leq m \leq 600$. This a statistical representation of 10000 realizations per method. Clustering using the Potts model is longer than the other methods, however, the Potts model result is likely to be skewed by the inclusion of clustering where $m > 350$.

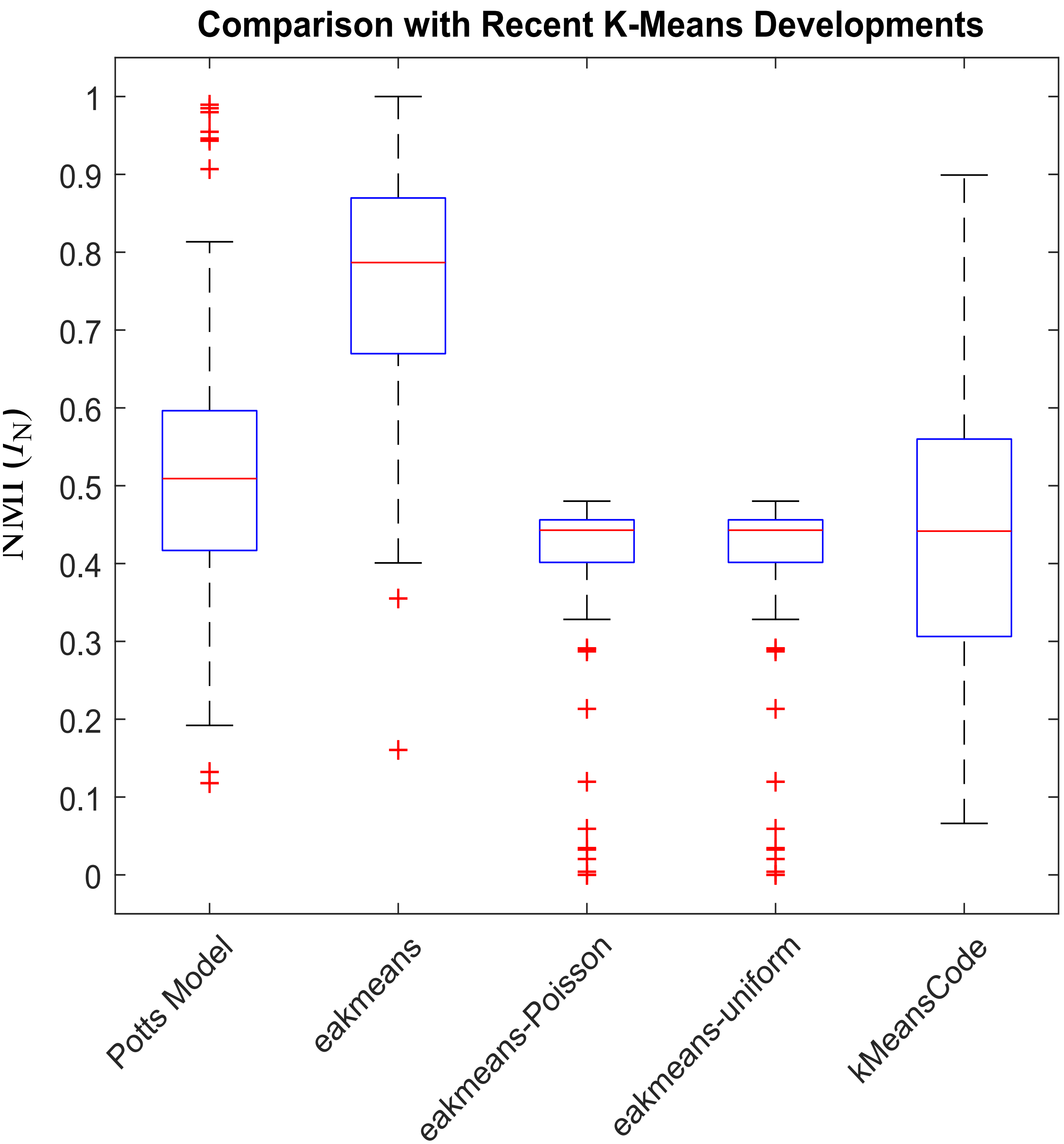

Fig 16 Method-wise clustering performance, evaluated on 100 realizations of 5000 synthetic data-points. Potts model segmentations were generated by selecting the lowest $\mathcal{H}$ for $0 < \gamma \leq 1$. Here eakmeans and kMeansCode represent methods described in[5,6] and[7,8] respectively. The segmentations for eakmeans and kMeansCode were generated using four classes. To fairly compare the Potts model's automatic model selection to K-means, performance of the eakmeans algorithm was evaluated for a non-specific number of classes; eakmeans-Poisson and eakmeans-uniform segmentations represent the average NMI for 100 realizations of eakmeans when the number of classes was randomly sampled from a Poisson and uniform distribution. These distributions were constructed of numbers ranging from 1 to 10, with the Poisson mean set to be the correct number of classes. The synthetic data-sets used in this comparison contain more intra-class variance to exemplify method performance.

Supplementary document C

Computational Segmentation and Classification of Diabetic Glomerulosclerosis

Brandon Ginley[1], John E. Tomaszewski[1,5], Brendon Lutnick[1], Sanjay Jain[3], Diane Salamon[3], Rabi Yacoub[2], Agnes Fogo[4], Kuang-Yu Jen[6], and Pinaki Sarder[1,7,8,*]

[1]Departments of Pathology & Anatomical Sciences, [2]Medicine – Nephrology,
[5]Biomedical Informatics, [7]Biostatistics, and [8]Biomedical Engineering
University at Buffalo – The State University of New York,

[3]Department of Medicine – Nephrology
Washington University School of Medicine

[4]Vanderbilt University, Nashville, Tennessee, USA

[6]Department of Pathology and Laboratory Medicine
University of California, Davis Medical Center

[*]Address all correspondence to: Pinaki Sarder
Tel: 716-829-2265; E-mail: pinakisa@buffalo.edu

Fix references formatting

Manuscript is too long

Significance statement- Pathologic classification of diabetic nephropathy is most commonly based on glomerular pathology as defined by Tervaert's consensus classification system. Although diagnostic guidelines are well established, interobserver reproducibility remains an issue especially for less advanced stages of the disease. Modern image analysis algorithms and artificial intelligence have the potential to automate as well as provide accurate and precise classification. Furthermore, digital algorithms are able to extract novel features which may be relevant to disease progression and prognosis. In this study, we used image analysis and machine learning algorithms to digitally segment, quantify, and classify glomerular images from patients with diabetic nephropathy. To our knowledge, this study is the first to demonstrate an automated method using machine learning to classifying glomerular disease.

Abstract

Background. Pathologic diagnosis of diabetic nephropathy is based on examination and identification of glomerular lesions on renal biopsies, commonly according to Tervaert's consensus classification system. Traditionally, these histopathologic lesions are classified by visual inspection, identification, and interpretation by the observing pathologist, which is susceptible to interobserver variability. Digital algorithms have deterministic outputs which can be leveraged to reduce this variability by providing a unified interpretation of image structure. We have developed a complete digital pipeline to quantify glomerular lesions of diabetic nephropathy.

Methods. To compartmentalize the glomerulus, we simplified glomerular structure to a three class system consisting of nuclei, luminal and Bowman spaces, and Periodic Acid-Schiff positive compartments (mesangium, basement membranes). The glomerular boundary and glomerular nuclei are identified using the DeepLab V2 ResNet convolutional neural network. The other classes are identified using color deconvolution, automated thresholding, and naive Bayesian refinement. We identify a set of task-specific features for characterization of glomerular lesions and use these features to computationally classify their pathological stage.

Results. Our deep neural network identifies glomerular boundaries with average 0.96 sensitivity and 0.96 specificity. Our deep neural network for renal nuclear segmentation performs with average 0.78 sensitivity, 0.99 specificity, and 0.99 precision. Our unsupervised method to segment simplified glomerular compartments performs with average 0.95 sensitivity and 0.99 specificity. In regards to classification, we are able to identify glomerular disease stages with average 0.89 sensitivity and 0.88 specificity.

Conclusions. Computationally derived, class-specific histological image features hold significant diagnostic information that can be mined for clinical application.

INTRODUCTION

In the United States, an estimated 9.4% of the population has diabetes mellitus. Of these individuals, slightly over one-third will develop diabetic nephropathy (DN), making diabetes the leading cause of chronic kidney disease and end stage kidney disease. The impact of diabetes and DN on public health will only intensify as the Centers for Disease Control projects 1 in 3 Americans will have diabetes by 2050 if current trends continue[1].

Although confirmatory biopsies are rarely performed to definitively establish the diagnosis of DN, there is typically good correlation between the clinical stages of DN and the renal morphologic changes seen on biopsy. A consensus pathologic classification system has been developed that divides DN into four hierarchical categories based on glomerular morphologic findings. In this system, class I is the mildest form of DN and is characterized by glomerular basement membrane thickening based on electron microscopic evaluation. Such biopsies should show no significant glomerular changes by light microscopy that would otherwise qualify the biopsy for a more advanced class. Class II disease is defined by mesangial widening in at least 25% of the observed mesangium, with class IIa representing cases with mild mesangial expansion and class IIb representing cases with severe mesangial expansion. Biopsies showing at least one convincing Kimmelstiel-Wilson nodule qualifies for class III DN and represents the classic form of nodular diabetic glomerulosclerosis. Class IV disease indicates advanced diabetic glomerulosclerosis with global glomerulosclerosis in >50% of the sampled glomeruli.

An important strength for this DN classification system is that the morphologic variables are well-defined, which results in improved interobserver reproducibility. Furthermore, the DN classes are based on readily recognizable glomerular lesions that are not time-consuming for proper evaluation and classification. However, in practice, reproducibility remains an issue. Additionally, a broad morphologic spectrum of glomerular lesions can still be seen within each class in terms of extent of glomerular involvement by the class-defining lesions. Thus, given the hierarchical structure of this DN classification system and the concrete definitions of the class-defining glomerular lesions, automated morphometric analysis using computer algorithms to analyze and classify biopsies with DN may be feasible and may significantly increase reproducibility and accuracy. Moreover, automated quantitative analysis may be a tool to discover clinically and prognostically meaningful morphologic findings in biopsies of DN. Automatic, concrete numerical measurements taken digitally on biopsy images may provide a much higher degree of diagnostic information in a smaller amount of time, thereby improving diagnosis and clinical utility.

In this study, our goal was to engineer an automated computational pipeline using convolutional neural networks in order to define glomerular boundaries from digitized biopsy slides, segment and quantify glomerular compartments, and classify glomerular lesions. This process was applied to renal tissue obtained from a streptozotocin (STZ) mouse model of diabetes mellitus as well as biopsies of patients with DN in order to classify the pathological findings under the DN classification system by Tevaert and colleagues[2]. These techniques can be extended for application to other glomerular diseases such as focal segmental glomerulosclerosis, lupus nephritis, and IgA nephropathy, among others. Our study suggests that image-based features derived computationally from histological images hold significant diagnostic information that can be mined for clinical application.

METHODS

Human data collection followed protocols approved by the Institutional Review Board at the University at Buffalo (UB). All methods were performed according to federal guidelines and regulations. All animal studies were performed according to protocols approved by the UB Animal Studies Committee Procedures and the Institutional Animal Care and Use Committee. Source code, a subset of 50 images (10 from each DN class), and network models presented within this manuscript are made freely available to the public. The data can be found at

Image data

Histological whole slide images (WSIs) from human ($n = 34$ patient core needle biopsies), mouse ($n = 25$ whole kidneys), and rat ($n = 5$ whole kidneys) were used for this study. Human tissues were from control (renal cell carcinoma nephrectomies with no apparent histological damage), DN, and IgA nephropathy cases. Mouse kidneys were from normal or STZ-treated animals. Rat kidneys were from normal animals. Human and mouse tissues were cut at 2μm thickness, and stained with periodic acid-Schiff (PAS) with hematoxylin as counterstain. Rat tissue preparation is described in our previous work[3]. WSIs were captured at 0.25 μm/pixel with a brightfield whole slide scanner (Aperio®, Leica). Additional information is available in S1.1.

Glomerular boundary segmentation

We trained the DeepLab V2 ResNet[4,5] network to segment glomerular boundaries from image patches that were extracted manually. The training dataset encompassed 3984 unique glomerulus images, of which $n = 1973/1011/1000$ were human / mouse / rat. Full resolution patches were chopped into $n = 36252$ sub-blocks size

$512x512$ and augmented to generate $n = 269458$ images. Glomerular boundary label was annotated manually. Network training specifications can be found in S1.2.

Glomerular compartmentalization

A textual overview of the compartment segmentation pipeline is available in Fig. S2B. To reduce the complexity and difficulty of glomerular compartment segmentation, glomerular structure was simplified into three classes, based on presentation in a PAS-hematoxylin stain. Namely, they are: 1) luminal compartment consisting of Bowman space and capillary lumina, 2) PAS-positive (PAS+) compartment consisting of mesangium, glomerular basement membranes, and Bowman capsule, and 3) nuclear compartment.

The luminal and PAS+ compartments were identified using a unique two step strategy (Figure 2). First, a preliminary segmentation of the compartments was developed by automated thresholding of grayscale images. Color deconvolution[6] was used to delineate PAS+ structures in high intensity. The *L** component of the *L*a*b* color space was used to delineate luminal structures in high intensity. These grayscale images were thresholded with Otsu's method[7] to yield rough, sparse binary label masks. These were used to train a naïve Bayesian classification model[8] implemented in MATLAB (MathWorks, Natick, MA) to predict remaining pixels based on RGB value.

For segmentation of nuclei in human tissues, the DeepLab V2 ResNet CNN was trained on glomerular nuclei partially annotated by computer and partially by human. Specifically, images were first partially segmented by one of two unsupervised methods (color deconvolution[6] or color gradient threshold, see S1.3.1), and then corrected by human annotator. In total, 410 human glomeruli were used to train the network, chopped into $n = 3186$ unique $500x500$ sub-blocks and augmented to result in $n = 31860$ images. Extended network training information can be found in S1.3.2.

For segmentation of nuclei in mouse tissues, we used color deconvolution[6] and Otsu's[7] method, described in our previous work[9]. Complex algorithms were not required to yield appropriate segmentations because the staining of these images was well controlled.

Segmentation performance analysis

Performance was assessed against manual annotation of holdout images. Tables 1 and 2 show glomerular boundary and compartment segmentation performance respectively. Performance is reported as $\mu \pm \sigma$, rounded to 2 decimal places. Asterisks (*) mark standard deviations which appear to make the performance greater than 1, though this is only a byproduct of the reported format, and no performance was greater than 1. Performance metrics are sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and Matthews correlation coefficient (MCC)[10]. For glomerular boundary segmentation, all classes of human and mouse glomeruli received an MCC of greater than 0.9. For segmentation of glomerular compartments, the human nuclear class was hardest to recognize with MCC of 0.87 in control and 0.88 in disease. All other compartment segmentations scored > 0.95 MCC. More details in S1.4.

Feature extraction

We derived four types of features for classification of glomerular structures: textural, morphological, distance, and containment. Texture (PAS+, lumina, and nuclei) is computed via gray-level co-occurrence[11] (measures sub-visual compartment changes). Morphological features included area and convexity of compartments (measures expansion or collapse of compartments). Distance features assess the distance between same-class objects (e.g. nuclei) and their distance to glomerular landmarks (measures how compartments move relatively, e.g. average nuclear distance from glomerular boundary. Containment features quantify the relative amount of one segmented compartment within another (identifies how compartments are expanding or collapsing in regards to each other, e.g., nuclear area contained within a mesangial segment). A full list of all 51 features is available in Table 3. Extended description of the extracted features is available in S1.5.1.

DN classification

We classified $n = 613$ images of individual human glomeruli into their corresponding DN stages. Traditional classification of DN is performed on a whole biopsy. However, DN biopsy data is scarce, and we did not have enough cases to make a feasible comparison. Therefore we performed classification on individual glomeruli (whole biopsy stage can still be determined by taking an aggregate score of all glomeruli). Glomeruli were extracted from $n = 15$ DN biopsies or $n = 6$ control nephrectomies (from renal cell carcinoma, only from areas with no apparent histological damage). Glomeruli were annotated for DN stage by a pathologist, see section S1.6. Control images had surplus glomeruli so $n = 297$ were selected randomly. All disease glomeruli were utilized, and $n = 247/47/72/64/183$ were annotated as stage I, IIa, IIb, III, or IV, respectively, by coauthor and renal pathologist Kuang-Yu Jen according to a slightly modified pathologic classification of DN as presented by Tervaert et al[2] (see S1.6).

We also classified $n = 1011$ mouse glomeruli from a diabetes mellitus (DM) STZ model by disease severity (measured by fasting glucose) and time of sacrifice (in weeks). Glomeruli were extracted from 25 kidney section whole slides of unique mice. Glucose cutoffs were specified as control, mild, or moderate for 90-100, 160-170, and 250+ mg/dL respectively. Sacrifice times were 15, 20, and 25 weeks. Additional mouse model information in S1.1.2.

The task-specific features derived were highly correlated, and therefore we used singular value decomposition (SVD) to reduce feature dimensionality and correlation[12]. 25 singular vectors represented 99% of original feature variance. These 25 vectors and pathologist class labels were used to train a naïve Bayesian classifier to differentiate consecutive disease states. The classification model was implemented in MATLAB, with hyper parameters (data distribution model, kernel smoothing window width and kernel smoothing function) iteratively selected using an optimization procedure which minimizes the model's cross-validated loss[13]. All training was performed using 50% of the available data as holdout.

RESULTS

Automated identification of glomeruli

The first step in automating analysis of glomerular disease is to automate glomerular identification. To achieve this goal, we trained a semantic CNN (DeepLab V2[4]) to segment glomerular boundaries from image patches. Other groups have shown similar segmentation of glomeruli using deep networks[14,15]; however, we demonstrate the use of an alternative network that is able to perform segmentation on multiple species and disease states. Importantly, our network models are openly available. This architecture was selected because it achieves high segmentation performance, allows many output classes for future expansion, and is easily ported to new segmentation tasks using pretrained models. Furthermore, it is designed for semantic segmentation, where every pixel in the image is assigned a class (useful for future networks which compartmentalize all aspects of renal tissue). Examples of our network's performance to delineate glomerular boundaries from manually selected images from human, mouse, and rat glomeruli are shown in Figure 1. Examples include human DN (Figure 1B-1F), murine DN (Figure 1H), and rat control from different histochemical stains aside from PAS such as Masson trichrome (Figure 1J). Our network showed high performance for glomerular boundary segmentation as compared to human analysis, with sensitivity ranging between 0.91 – 0.98 and specificity ranging from 0.95 – 1 (Table 1) Overall, across all classes, the network performed with an average Matthews correlation coefficient[10] of 0.93. The most difficult boundaries for the network to identify are those with Bowman capsule reduplication. In many cases, however, the network provides a more exact-fit of glomerular boundary to the Bowman capsule on a pixel-by-pixel basis. Performance was not calculated for rat glomerular identification since these WSIs were not used for any subsequent experiments. Though we have trained this network on square crops of glomerular regions, this technique can be extended to localize glomeruli from whole slide images by creating image patches from the WSI, testing all the patches in the network, and stitching the image back together, as we have shown in our previous work.

Glomerular compartmentalization

A visual exhibition of our pipeline for glomerular compartmentalization can be found in Figure 2.

Nuclear image analysis

Information on acquisition of nuclear training data can be found in the supplementary document sections S1.3.1 and S2.2. In general, our network was able to achieve high performance for nuclear segmentation in human data (Table 2). Our unsupervised approach for murine nuclear detection (color deconvolution[6] and thresholding[7]) also achieved high performance (Table 2). The one exception was for glomerular nuclei of human disease cases, which demonstrated moderate sensitivity at 0.79. The reason for this finding is that the human annotator tended to over-segment while the network tended to under-segment the nuclear boundaries, creating a persistent bias in performance analysis. However, the precision of the network to identify nuclei was 0.99 on average, which was sufficient to analyze nuclear structure within the scope of this study. Comparable high network performance of nuclear segmentation was seen for test images prepared and stained at different institutions (Figure 3), which illustrates the robustness of our network in terms of overcoming variations between laboratories. Nuclear predictions on a human glomerulus is also shown in Fig. 2C. Note that nuclear predictions were restricted using the glomerular boundary to exclude tubular nuclei. Overall, even though the CNN was not trained on as many samples as is optimal (likely in the thousands), it still provided significant improvements to the segmentation of nuclei over a color deconvolution approach, and is significantly helpful in analyzing nuclear properties of histology images.

Mouse glomerular nuclei were identified using color deconvolution[16] because the mouse images were less complex, had little stain variation, and color deconvolution provided adequate segmentation of nuclei for structural analysis. Interestingly, the CNN was able to identify mouse nuclei with high accuracy even without undergoing training on images of murine glomeruli (data not shown).

Luminal and PAS+ image analysis

Luminal and PAS+ regions are identified in two steps, first in a rough preliminary fashion using unsupervised thresholding, afterwards in a final fashion with a naive Bayesian approach. These two methods combined complement each other: thresholding techniques are fast and unsupervised, while relatively imprecise; naïve Bayesian classification is capable of achieving high precision when there is a clear measurable difference in the class distributions, but needs training data. This approach makes it very convenient to identify various components of glomeruli which are similar in color with high performance, and is adaptive to slight shifts in stain variation (the naive Bayesian model is re-trained on each image). Fig. 2D shows the lightness (L^*) component of $L^*a^*b^*$ color space. $L^*a^*b^*$ transformation is designed to be a color space which is perceptually uniform to human color vision. The lightness component, as it suggests, transforms pixel values such that the brightest whites have highest value and darkest blacks have lowest value. Otsu's thresholding is a technique to automatically determine image foreground from background by maximizing inter-class variance. Thresholding of the lightness component generates preliminary segmentation masks of luminal spaces. Fig. 2E shows stain deconvolution[17] for PAS+ components (mesangium, basement membranes, capsule). Stain deconvolution is a technique to rotate an image's color space axes so that they are along the directions of the stain color. The PAS deconvolved image is also thresholded with Otsu's method to create a preliminary segmentation mask of PAS+ objects. The combined preliminary segmentation masks are shown in Fig. 2F. CNN nuclear predictions are shown in blue, preliminary PAS+ components in red, and luminal components in green. Fig. 2G shows unlabeled pixels from the preliminary segmentation. The PAS+ component and luminal component labels are used to train a naïve Bayesian classifier to predict the unlabeled pixels, into either the PAS+ class or the luminal class, and results in a final segmentation; see Fig. 2H. The performance analysis of the glomerular compartmentalization discussed in this section is discussed in the methods and is shown in Table 2.

Glomerular feature extraction, separation, and ranking

Based on our glomerular compartment analysis of DN biopsies, a set of computational features were extracted to describe the pathological structural progression of glomeruli in DN. These image features are based on texture, morphology, intra-compartmental distance, and glomerular structural conformation. The 51 extracted features are listed in Table 3.

Feature extraction

Textural features were computed in aggregate for each glomerulus based on the glomerular compartment. For example, the total Bowman and capillary luminal space within a single glomerulus was analyzed as one unit comprising the luminal compartment. Images were transformed from RGB to grayscale, then converted to respective gray-level co-occurrence matrices. Glomerular compartment specific textural descriptions identified as gray level entropy, energy, correlation, and homogeneity were extracted from the respective matrices[11].

Morphological features were calculated per individual compartment object. Summary statistics were taken along the glomerulus dimension, e.g., mean nuclear area refers to the mean area of all nuclei in a particular glomerulus. These features included the mean, median, and mode of areas, and average convexity for identified compartmental objects. Summary statistics were taken along the glomerulus dimension, e.g., if a glomerulus contained 20 nuclei, the area of each nucleus was measured, and the mean, median, and mode of the set of 20 nuclei was taken.

Compartmental containment features define the amount of one compartment contained within the boundaries of another. Specifically, it is a ratio, where one part is the convex area of the containing compartment object, and the other part is the area of the contained compartment.

Distance features are comprised of averaged distances between compartments and other identical glomerular compartments, or glomerular landmarks. Glomerular landmarks include the estimated glomerular centroid and the estimated glomerular boundary points. The following distance features are extracted for each object of each glomerulus: 1) the distance between that object's centroid and identically labeled object's centroids, 2) the average distance to the glomerular boundary, and 3) the distance to the glomerular centroid.

See section S1.5.1 for extended details on extraction of all features. A total of 51 features were extracted from a total of $n = 613$ human glomeruli or $n = 1011$ mouse glomeruli. A specific summary of all features is provided in Table 3.

Feature ranking

We next used a neighborhood component analysis (NCA)[18] to compare the relative usefulness of the hand-crafted structural features in describing structural progression of glomeruli. NCA is a method for selecting features which maximize the prediction accuracy of classification models. It was discovered that only 16 of the derived features were useful in classifying the DN stage of human glomeruli, and only 19 were useful in classifying the DN stage of mouse glomeruli. Interestingly, we found that 10 features co-described both human and mouse DN pathology. A specific listing of which features were important for which classification can be found in the supplementary methods section S2.3, and the weight associated with each feature in supplementary Fig. S4.

Naïve Bayesian classification of glomerular structure

SVD was performed on the original 51 features to reduce the feature dimensionality and correlation. This can help improve classification by removing extraneous information. A visualization of the top three singular dimensions according to disease state can be seen in Fig. S5.

It was found that only 25 singular vectors were needed to account for 99% of variance in the original feature space. These compressed features were used to train a naïve Bayesian classifier to classify DN structural states of glomeruli as annotated according to the procedure outlined in the methods section. All classifiers used 50% of data as holdout and 50% as training data. The performance of each classifier to make a binary decision between disease states is shown in Table 4. Table 4A shows the performance before optimization of hyper parameters (data distribution model, kernel smoothing window width and kernel smoothing function), and Table 4B shows the performance after optimization of hyper parameters. It appears that optimizing the hyper parameters of the naïve Bayesian classifier significantly improved the performance. Further, it can be concluded that stage IIb is generally the most difficult to identify, as the performance scores are notable lower for these classes comparisons. This is likely because this distinction is based on whether or not the mesangial area appears to exceed the mean area of the capillary lumen, which is somewhat subjective dependent on the observer. It can also be noted that the farther away two stages are from each other, the more easily they can be classified (e.g., stage IIa is much easier to classify when compared against stage IV than stage IIb).

Table 4C shows the optimized performance of attempting to classify mouse data by time of sacrifice. Table 4D shows the optimized performance of attempting to classify mice by severity of DM. On average, it is easier to determine stages of human glomeruli than mouse glomeruli. This is likely due to human DN stages having been categorized strictly by image presentation and mouse stages categorized more loosely by experimental conditions.

DISCUSSION

In this manuscript we have shown it is feasible to automatically classify structural classes of glomeruli using a combination of common features and hand crafted features. In this section we will discuss both our motivations for and criticisms of these methods.

The detection and segmentation of glomeruli from renal tissue architecture is a highly complex task. In health, the glomerulus appears as a coalescing patch of capillaries bound together by mesangium, laminated with basement membrane, coated with podocytes, and surrounded by an ample section of Bowman space that is bound by the glomerular capsule. In disease, any or all of these structures can enlarge, shrink, distort, or disappear, utterly altering the complete appearance of the glomerulus. Not only this, but the glomerulus is embedded within a multitude of other renal compartments such as vasculature, interstitium, and tubules, all of which display similar staining to glomeruli with different structural organization. Because of this complexity, it is difficult to construct an algorithm which both correctly labels all types of glomeruli while simultaneously excluding any other similar renal structures. We selected deep learning to solve this task as deep learning is capable of deriving non-linear relationships between pixels to identify classes of objects with a very high dimensional representation. This means the deep learner is easily capable of describing an object with many different appearances as the same class, while also being highly specific at excluding other structures. This is ideal for describing the wide range of morphologic changes observed in glomeruli. Further, the semantic CNN we use is ideal for describing objects that can be identified using contextual information, such as the surroundings, through the use of atrous convolution. Specifically, in renal tissue, an object's identity may be highly dependent on its surroundings (e.g., podocyte identification), which is more efficiently captured using atrous convolutions. Lastly, the network we selected is very large and much more complex than one would expect for performing a simple binary classification task. However, we decided to use this network so that we can add additional renal tissue classes in the future without having to modify our architecture.

The segmentation of glomerular compartments is similarly challenging, perhaps even more. In truth, the identification of glomerular compartments is likely an easy task for a deep learning algorithm, but annotation challenges inhibit the development of such models. Specifically, not only does annotation of glomerular compartments require significant domain expertise, it is exceptionally tedious to annotate thousands of such compartments. Realistically, this is far more hours of work than a specialized domain expert can provide. It is for this reason we decided to pursue unsupervised segmentation of glomerular compartments and further utilize these segmentations to train deep network models. However, unfortunately, there are no developed algorithms for robustly identifying glomerular compartments such as mesangium, basement membranes, capillaries, etc. In the current state of this work the unsupervised segmentation comes at a cost of a simplified model of glomerular compartments. In future works, we will look towards using these simplified segmentations to jumpstart glomerular compartment annotations that can be used to train a deep networks for segmentation of glomerular compartments including mesangium, capillary loops, basement membrane, lumen, Bowman space, and Bowman capsule. We will also aim to develop a classification scheme for resident cell nuclei.

After segmentation of glomerular compartments we also experimented with hand-crafting features to describe DN projection. These features are extensively discussed in S1.5.1 and ranked for usefulness in S2.3. These features were designed to reflect the understood pathological progression of glomeruli in DN. Based on the results presented in Table 4 and Figs. S4 & S5, it is clear these features have predictive merit in distinguishing structural stages of DN glomeruli.

The major goal of this sort of work is to motivate the shift of pathological diagnoses from discrete categories extrapolated from visual characterizations to continuous risk models derived from structural quantification. A more rigorous future computational study will attempt to predict disease structure as a multi-class continuum rather than binary comparisons. However, future studies should also aim to compare task-specific hand-crafted feature sets against task-agnostic feature sets, in order to determine which of the two is more useful, and if hand-crafted features are even necessary.

Author contributions

B.G. coded and performed the computational analysis of the proposed work, trained the deep networks, and wrote the manuscript. R.Y. produced the STZ mouse model. B.L. assisted in set-up and training of deep networks. S.J. curates the renal tissue database from which the human data in this manuscript is taken. D.S. contributed in organization of clinical data associated with each biopsy case. J.E.T. did…. A.F. provided pathologically labeled diabetic nephropathy biopsies used to train segmentation networks. K.J. annotated glomerular images. P.S. coordinated the image analysis protocols, conceptualized the project direction, and assisted in manuscript preparation.

Acknowledgments

The project was supported by the faculty startup funds from the Jacobs School of Medicine and Biomedical Sciences, University at Buffalo, the DiaComp Pilot and feasibility Program Grant #32307-5, as well as the University at Buffalo IMPACT award. We thank NVIDIA Corporation for the donation of the Titan X Pascal GPU used for this research (NVIDIA, Santa Clara, CA). We also acknowledge the assistance of the Multispectral Imaging Suite in the Dept. of Pathology & Anatomical Sciences, School of Medicine and Biomedical Sciences, University at Buffalo for microscopic imaging of tissue sections. We are thankful to Dr. Tracey Ignatowski for the rat tissue data provided by her lab.

References

1 Prevention, C. f. D. C. a. *Diabetes Report Card*, <https://www.cdc.gov/diabetes/library/reports/congress.html> (2014).

2 Tervaert, T. W. *et al.* Pathologic classification of diabetic nephropathy. *J Am Soc Nephrol* **21**, 556-563, doi:10.1681/ASN.2010010010 (2010).

3 Brandon Ginley, J. E. T., Rabi Yacoub, Feng Chen, Pinaki Sarder. Unsupervised Labeling of Glomerular Boundaries using Gabor Filters and Statistical Testing in Renal Histology. *Journal of Medical Imaging* (2017).

4 Wang, Z. & Ji, S. in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery &#38; Data Mining* 2486-2495 (ACM, London, United Kingdom, 2018).

5 Chen, L. C., Papandreou, G., Kokkinos, I., Murphy, K. & Yuille, A. L. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. *Ieee T Pattern Anal* **40**, 834-848, doi:10.1109/Tpami.2017.2699184 (2018).

6 Ruifrok, A. C. & JOHnston, D. A. Quantification of histochemical staining of color deconvolution. *Anal. Quant. Cytol. Histol.* **23**, 291-299 (2001).

7 Otsu, N. A Threshold Selection Method from Gray-Level Histograms. *IEEE Transactions on Systems* **9** (1976).

8 Hastie, T., R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. 2nd edn, (Springer, 2008).

9 Sarder, P., Ginley, B. & Tomaszewski, J. E. 97910F-97910F-97912.

10 Matthews, B. W. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim Biophys Acta* **405**, 442-451 (1975).

11 Haralick, R. M., Shanmugam, K. & Dinstein, I. Textural Features for Image Classification. *Ieee T Syst Man Cyb* **Smc3**, 610-621, doi:Doi 10.1109/Tsmc.1973.4309314 (1973).

12 Golub, G. H. & Kahan, W. Calculating the singular values and pseudo-inverse of a matrix. *Journal of the Society for Industrial and Applied Mathematics: Series B, Numerical Analysis* **2**, 205–224 (1965).

13 MATLAB Statistics and Machine Learning Toolbox.

14 Bukowy, J. D. *et al.* Region-Based Convolutional Neural Nets for Localization of Glomeruli in Trichrome-Stained Whole Kidney Sections. *J Am Soc Nephrol* **29**, 2081-2088, doi:10.1681/ASN.2017111210 (2018).

15 Pedraza, A. *et al.* in *Medical Image Understanding and Analysis: 21st Annual Conference, MIUA 2017, Edinburgh, UK, July 11–13, 2017, Proceedings* (eds María Valdés Hernández & Víctor González-Castro) 839-849 (Springer International Publishing, 2017).

16 Ruifrok, A. C. & Johnston, D. A. Quantification of histochemical staining by color deconvolution. *Anal Quant Cytol* **23**, 291-299 (2001).

17 *Stain Normalisation Toolbox*, <http://www2.warwick.ac.uk/fac/sci/dcs/research/combi/research/bic/software/sntoolbox/> (

18 Yang, W., K. Wang, W. Zuo. Neighborhood Component Feature Selection for High-Dimensional Data. *Journal of Computers* **7** (2012).

Table 1. Performance of glomerular boundary segmentation on select classes of images.

Annotated stage	Sensitivity	Specificity	Positive predictive value	Negative predictive value	Matthews correlation coefficient
Human control (*n* = 173)	0.96 ± 0.03	0.98 ± 0.02	0.97 ± 0.03	0.97 ± 0.03	0.94 ± 0.03
Human DN I (*n* =13)	0.96 ± 0.06	0.96 ± 0.04	0.95 ± 0.05	0.96 ± 0.05	0.91 ± 0.05
Human DN IIa (*n* = 35)	0.98 ± 0.02	0.95 ± 0.04	0.94 ± 0.06	0.98 ± 0.02	0.92 ± 0.04

Human DN IIb (*n* = 97)	0.97 ± 0.03	0.96 ± 0.04	0.95 ± 0.04	0.97 ± 0.03	0.93 ± 0.04
Human DN III (*n* = 100)	0.97 ± 0.03	0.96 ± 0.04	0.95 ± 0.05	0.97 ± 0.03	0.92 ± 0.04
Human DN IV (*n* = 94)	0.94 ± 0.05	0.95 ± 0.05	0.94 ± 0.06	0.95 ± 0.04	0.90 ± 0.07
Control mouse (*n* = 30)	0.95 ± 0.04	0.99 ± 0.01	0.99 ± 0.02	0.97 ± 0.02	0.95 ± 0.04
DM mouse (*n* = 45)	0.91 ± 0.06	1 ± 0.007*	0.99 ± 0.01	0.95 ± 0.03	0.93 ± 0.04

Table 2. Performance of glomerular subcompartmentalization.

Compartment	Sensitivity	Specificity	PPV	NPV	MCC
Control mouse PAS+	0.97 ± 0.02	1 ± 0	1 ± 0	0.99 ± 0.008	0.98 ± 0.02
Control mouse lumen	0.97 ± 0.03	1 ± 0	1 ± 0	0.93 ± 0.07	0.95 ± 0.04
Control mouse nuclei	0.98 ± 0.02	1 ± 0.001*	0.99 ± 0.03*	0.99 ± 0.002	0.99 ± 0.02*
DM mouse PAS+	0.97 ± 0.03	0.99 ± 0.002*	0.99 ± 0.008	0.98 ± 0.02	0.97 ± 0.03
DM mouse lumen	0.98 ± 0.02	1 ± 0	1 ± 0	0.94 ± 0.06	0.96 ± 0.04
DM mouse nuclei	0.97 ± 0.02	1 ± 0	0.99 ± 0.005	0.99 ± 0.004	0.98 ± 0.01
Control human PAS+	0.98 ± 0.02	1 ± 0.001*	1 ± 0	0.96 ± 0.06*	0.97 ± 0.04*
Control human lumen	0.99 ± 0.01	1 ± 0	1 ± 0	0.94 ± 0.1*	0.96 ± 0.06*
Control human nuclei	0.76 ± 0.08	1 ± 0.0002*	1 ± 0.002*	0.98 ± 0.01	0.87 ± 0.05
Disease human PAS+	0.99 ± 0.03*	0.99 ± 0.004*	0.987 ± 0.08	0.988 ± 0.04	0.98 ± 0.06*
Disease human lumen	0.99 ± 0.02*	1 ± 0	1 ± 0	0.95 ± 0.1	0.96 ± 0.07*
Disease human nuclei	0.79 ± 0.1	1 ± 0.001*	0.99 ± 0.03*	1 ± 0.006*	0.88 ± 0.06

Table 3. List of quantified glomerular features.

Feature No.	***Distance features***
1	Average distance of lumina center from glomerular center
2	Averaged average distance between lumina and glomerular boundaries
3	Average maximum distance between lumina and glomerular boundaries
4	Average minimum distance between lumina and glomerular boundaries
5	Averaged average distance between luminal regions
6	Average maximum distance between luminal regions
7	Average minimum distance between luminal regions
8	Average distance of PAS+ from glomerular center
9	Averaged average distance of PAS+ from glomerular boundaries
10	Average maximum distance of PAS+ from glomerular boundaries
11	Average minimum distance of PAS+ from glomerular boundaries
12	Averaged average distance between PAS+ regions
13	Average maximum distance between PAS+ regions
14	Average minimum distance between PAS+ regions
15	Average distance of nuclei from glomerular center
16	Averaged average distance of nuclei from glomerular boundaries
17	Average maximum distance of nuclei from glomerular boundaries
18	Average minimum distance of nuclei from glomerular boundaries
19	Averaged average distance between nuclei
20	Average maximum distance between nuclei
21	Average minimum distance between nuclei

	Containment features
22	Average PAS+ area contained in convex luminal boundary
23	Average nuclear area contained in convex luminal boundary
24	Average luminal area contained in convex PAS+ boundaries
25	Average nuclear area contained in convex PAS+ boundaries
26	Average nuclear overlap with lumina
27	Average nuclear overlap with PAS+
Feature No.	***Texture features***
28	Nuclear gray-level spatially dependent contrast
29	Nuclear gray-level spatially dependent correlation
30	Nuclear gray-level spatially dependent energy
31	Nuclear gray-level spatially dependent homogeneity
32	Luminal gray-level spatially dependent contrast
33	Luminal gray-level spatially dependent correlation
34	Luminal gray-level spatially dependent energy
35	Luminal gray-level spatially dependent homogeneity
36	PAS+ gray-level spatially dependent contrast
37	PAS+ gray-level spatially dependent correlation
38	PAS+ gray-level spatially dependent energy
39	PAS+ gray-level spatially dependent homogeneity
	Morphological features
40	Average convexity of lumina
41	Sum total area of luminal space
42	Mean area of luminal spaces
43	Median area of luminal spaces
44	Average convexity of PAS+ components
45	Sum total area of PAS+ components
46	Mean area of PAS+ components
47	Median area of PAS+ components
48	Sum total nuclear area
49	Mean nuclear areas
50	Mode nuclear areas
51	Total glomerular area

Table 4. Performance of mouse and human structural stage classification.

Classes compared	**A.** Human, not optimized		**B.** Human, optimized	
	Specificity	Sensitivity	Specificity	Sensitivity
I-IIa	0.1702	0.9474	1	1
I-IIb	0.4444	0.9352	0.6528	0.996
I-III	0.6719	0.9514	0.8281	0.996
I-IV	0.765	0.9312	0.8634	0.9919
IIa-IIb	0.7222	0.617	1	0.8085
IIa-III	0.7656	0.7872	0.9219	0.7872

IIa-IV	0.8798	0.7447	0.9508	0.9574
IIb-III	0.4375	0.5139	0.75	0.5833
IIb-IV	0.9126	0.7917	0.9508	0.9028
III-IV	0.9344	0.8438	0.918	0.8906

Sacrifice time (weeks)	**C.** Mouse, optimized	
	Specificity	Sensitivity
15-20	0.6495	0.8497
15-25	0.4932	0.9663
20-25	0.602	0.9245

	D. Mouse, optimized	
DM class	Specificity	Sensitivity
Control	0.8211	0.703
Mild	0.7847	0.7228
Moderate	0.733	0.8035

Figure 1.

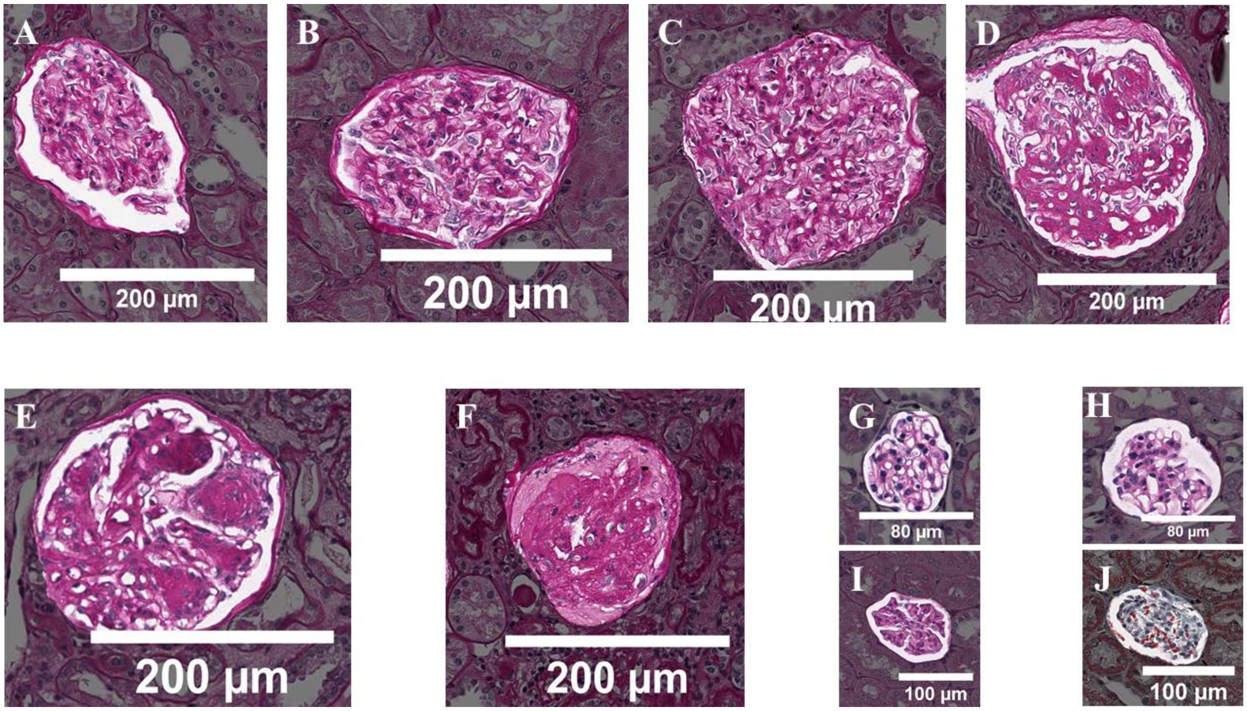

Figure 2.

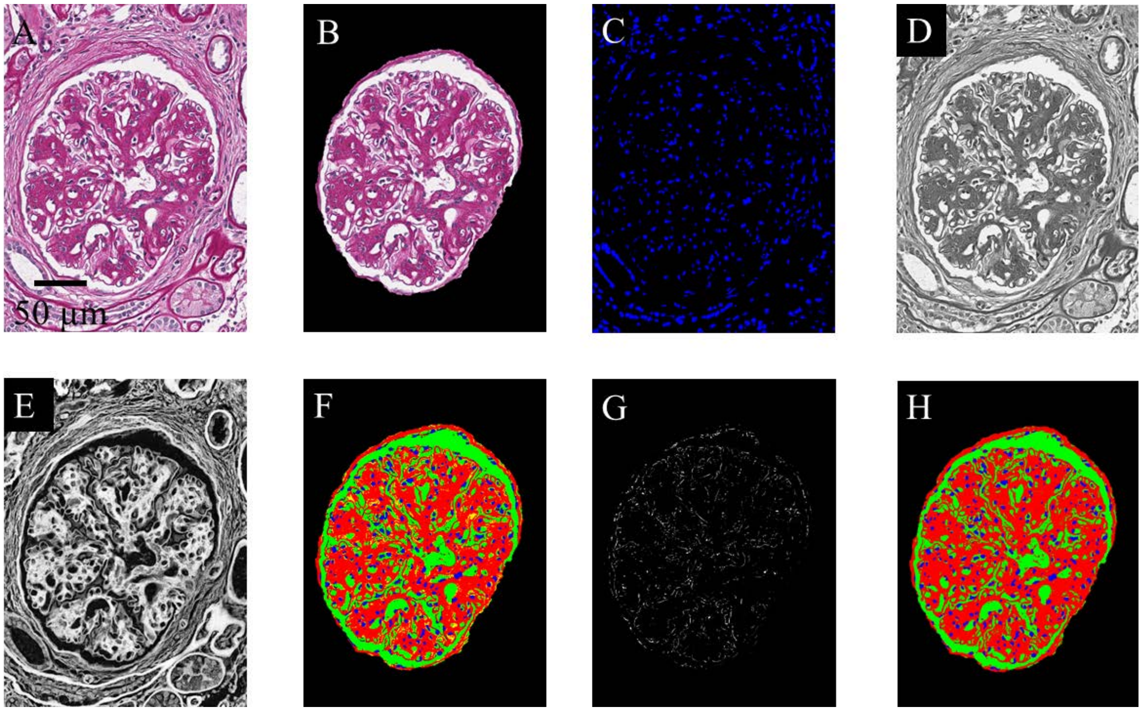

Figure 3.

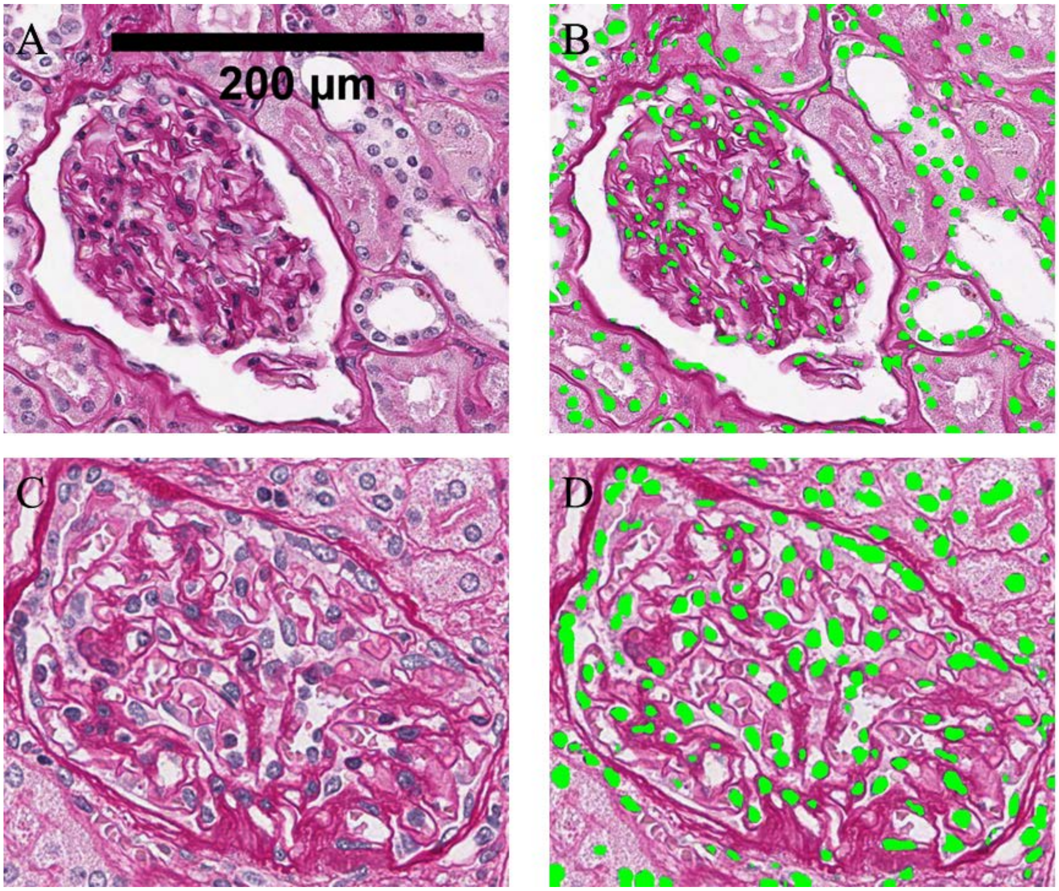

Figure captions

Figure 1. Boundary segmentation by deep CNN on DN images. 1A. Human glomerulus from control data. 1B-1F. Human glomeruli with respective stage annotations of I, IIa, IIb, III, and IV. 1G. Control mouse glomerulus. 1H. Mouse glomerulus from STZ treated group. 1I. PAS stained rat glomerulus. 1J. Trichrome stained rat glomerulus.

Figure 2. Pipeline for glomerular compartment segmentation. 2A. Example image of PAS stained human glomerulus. 2B. Segmented glomerular boundary. 2C. CNN segmentation of nuclei. 2D. Grayscale image depicting the lightness component of *L*a*b** color space, delineating luminal spaces. 2E. Grayscale image depicting stain deconvolution for PAS components, delineating mesangium and basement membranes. 2F. Preliminary compartment segmentation generated by CNN segmentation of nuclei and global thresholding of 2D and 2E. A Naïve Bayes classifier will be trained using these pixels. 2G. Pixels from 2F which do not yet have a label. The Naïve Bayes classifier will predict these labels. 2H. Final, complete segmentation of all three compartments after Naïve Bayes segmentation correction.

Figure 3. Nuclear segmentation by deep CNN. 3A. Example of a human glomerulus. 3B. Segmentation of nuclei from image in 3A by deep CNN. 3C. Glomerulus image which was prepared in a separate institute than glomerulus from image 3A. 3D. Segmentation of nuclei from 3C.

Supplementary Table of Contents