Segmentation of Glomeruli Within Trichrome Images Using Deep Learning

Shruti Kannan[1], Laura A. Morgan[1], McKenzie G. Cheung[1], Benjamin Liang[1], Christopher Q. Lin[1], Dan Mun[2], Jean M. Francis[3], Vipul C. Chitalia[3, 4], Joel M. Henderson[4], Vijaya B. Kolachalama[5]

[1]College of Engineering, Boston University, Boston, MA – 02215, USA
[2]College of Health & Rehabilitation Sciences, Sargent College, Boston University, Boston, MA – 02215, USA
[3]Renal Section, Department of Medicine, Boston University School of Medicine, Boston, MA – 02118
[4]Department of Pathology and Laboratory Medicine, Boston University School of Medicine, Boston, MA – 02118
[5]Section of Computational Biomedicine, Department of Medicine & Whitaker Cardiovascular Institute, Boston University School of Medicine, Boston, MA – 02118

Key words: Kidney biopsy, Glomerulus, Digital pathology, Image segmentation, Deep learning, Computational pathology

Corresponding author:
Vijaya B. Kolachalama, PhD
Boston University School of Medicine
72 E. Concord Street, Evans 636, Boston, MA, USA – 02118
Email: vkola@bu.edu
Phone: 617-358-7253

ABSTRACT

Background The number of glomeruli on a kidney biopsy slide followed by glomerular assessment constitute as standard components of a renal pathology report. The prevailing method for glomerular identification and assessment remains manual, labor intensive and non-standardized. In the era of digitized kidney biopsies, an automated method to identify, segment and count glomeruli is highly desirable.

Methods and results We developed an automated method to detect and segment the glomeruli within digitized kidney biopsy images by leveraging a deep learning architecture based on convolutional neural networks (CNN). A total of 275 trichrome-stained images (Average image size: 2560x1920x3 pixels, 1-2 unique images per patient, Scale: 0.85 μm/pixel) processed at 40x magnification from renal biopsies of 171 chronic kidney disease patients treated at the Boston Medical Center from 2009-2012 were analyzed. A sliding window operation was defined to crop each 40x image to smaller images of size 300x300x3 pixels. Each cropped image was then evaluated by clinical experts to identify the presence of a unique glomerulus, and each identified glomerulus was included in the training dataset (n = 751). About the same number of cropped images, containing the non-glomerular regions of the kidney biopsy, served as control cases. The CNN model was constructed as a binary classification problem to discriminate glomerular images from the non-glomerular ones (Performance on test data - Accuracy: 97.47±0.31%; Sensitivity: 96.43±1.89%; Specificity: 98.76±1.44%). Using the trained CNN model, another sliding window operator was developed to scan the digitized biopsies. A heatmap was generated to highlight regions of intensity that the CNN model classified as glomerular regions. Subsequently, two independent image processing strategies, one using steps such as image binarization and image erosion, and the other using image binarization, distance transform and watershed segmentation, were performed on the heatmaps to generate discriminatory signatures of the identified glomeruli. The final step involved automatically drawing a box around the higher intensity areas leading to output images with segmented glomerular regions (Performance on test data - Accuracy: 99.97±0.0086%; Sensitivity: 54.37±0.23%; Specificity: 99.99±0.009%).

Conclusion While used in the context of nephropathology, this study demonstrates the power of artificial intelligence to assess complex histologic structures and identify structural variations. Adoption of such methods of counting and classifying glomeruli using standard histological staining without disturbing the workflow can expedite the assessment of slides by the pathologists and serve as a first step toward more comprehensive automated analysis.

INTRODUCTION

Glomerular damage is a common manifestation in a spectrum of renal diseases that lead to kidney failure [1]. Morphological and ultrastructural alterations within the glomeruli provide valuable information on the mechanisms of renal impairment and facilitate accurate clinical diagnosis [1-3]. Identification and assessment of this highly relevant structure are therefore integral to histopathological analysis of kidney biopsies. A fundamental morphologic parameter in nephropathology is the quantification of normal and lesional glomeruli in the light microscopic material. Glomerular quantification is required for assessment of tissue sufficiency in kidney transplant pathology, for determination of severity of relatively common diseases such as lupus nephritis or IgA nephropathy, and for quantification of the extent of chronic damage in any kidney biopsy. Therefore, histological analysis of glomerular diseases involves careful examination of the entire kidney biopsy slide, and this includes in part, identification of all the glomeruli present, assessment of the state of each glomerulus, and integration of this data with other parameters to pinpoint the diagnosis of the glomerular disease [4-7]. While this seemingly tedious process of counting and assessing all the glomeruli can be handled efficiently at large medical centers under the supervision of an in-house nephropathologist, this expertise is not available at all locations across the globe. In addition, even for the institutions that have the expert nephropathologist, we need approaches that can automatically perform some of these tasks in order to assist clinical nephrology practice to maximize their efficiency.

Machine learning (ML) techniques have the ability to perform these tasks in an efficient fashion. ML approaches give computers the ability to integrate discrete datasets in an agnostic manner to detect previously indecipherable patterns and generate a disease-specific fingerprint. Especially in subspecialties that rely on imaging, these tools are being adopted quite rapidly, since ML can leverage many images as inputs and correlate patterns and features with clinical outcomes. Building on the advances of ML, scientists recently have developed so-called "deep learning" frameworks such as convolutional neural networks (CNN) for object recognition and classification [8]. CNN techniques are now being rapidly adopted as unbiased, self-learning approaches for pathologic assessment [9-23].

Using an established CNN that can perform image classification, we developed a framework to automatically identify and segment the glomeruli present within digitized images of human kidney biopsies. Trichrome stained kidney biopsies obtained from 171 patients treated at the Boston Medical Center (BMC) were digitized to generate 275 unique images at 40x magnification. These images were further processed to generate a training dataset containing unique images of the glomerular (n=751) and non-glomerular compartments (n=751) of the kidney biopsy. CNN model training on this dataset followed by cross-validation generated a highly accurate classifier with the ability to discriminate between images containing glomerular and non-glomerular aspects of the kidney. The validated model was then used in the form of a sliding window operator to further process the original 40x test images that were not used for CNN training, to identify and segment the glomeruli. Sensitivity analyses underscored the robust performance of the CNN model to classify the glomeruli and the consistent ability for it to segment glomeruli within the digitized trichrome images.

MATERIALS AND METHODS

Data collection

Anonymized kidney biopsies were obtained and digitized after approval by the Boston University Medical Campus' institutional review board. Biopsy procedures were performed on all patients treated at BMC between January 2009 and December 2012 (**Table 1**). More than 300 biopsy samples were processed, of which 171 biopsy slides were available for subsequent imaging. These biopsy samples were obtained from adult patients who had a native or an allograft biopsy, independent of the indication for the biopsy procedure [15]. The criterion for inclusion was the availability of pathological slides.

Imaging

Biopsy samples were obtained in the form of individual trichrome-stained slides prepared from formalin-fixed, paraffin-embedded core-needle biopsy tissue. A selected core visible on each slide was imaged at 40x magnification (indicating a 4x objective and a 10x eyepiece) using a Nikon Eclipse TE-2000 microscope (Melville, NY; http://www.bumc.bu.edu/busm/research/cores/). Images were generated with a special consideration to cover the entirety of the biopsy sample which resulted in multiple 40x images per patient. All the images were manually focused using the NIS-Elements AR software (Nikon, Tokyo, Japan) that was installed on the computer connected to the microscope. A total of 275 unique images (~2 40x images per patient) were used and the average size of each image was about 2560x1920x3 pixels, resulting in a length scale of 0.85 µm/pixel.

Glomerular dataset generation

A sliding window method was developed to automatically crop the trichrome images from their original size into several independent images of size 300x300x3 pixels (**Figure 1**). After manual review, the window size was determined as the minimum required to fit the largest glomerulus observed within all the original images. For each original image, the sliding window operator began at the top left corner of the 40x image and moved right with a stride of 20 pixels after cropping the 300x300x3 window. The stride size was empirically determined such that it resulted in each glomerulus being captured completely in at least one cropped image. In total, this process generated about 107525 unique images of size 300x300x3 from 275 unique 40x images of size 2560x1920x3 pixels. In order to expedite the cropping process, windows consisting of purely non-biopsy portions (i.e. background) were automatically ignored. For each cropped segment, we computed the median intensity of all the pixels and selected only those cases with a median bin frequency value lower than an empirically estimated threshold intensity of 150 (**Figure 2**). Our idea of thresholding was based on selecting a single cutoff value for the bin frequency (=150). The underlying assumption was that an image with only the background had a higher frequency of pixels with the same intensity values (**A1** in **Figure 2**), whereas a cropped image with tissue within it had a more distributed range of intensity values with lower frequencies (**A2** in **Figure 2**). In order to compute the median, the histogram bins (x-axis) were arranged in the ascending order of their corresponding frequencies (y-axis). In case of the background segment, since the frequencies were concentrated in a very narrow bin range when arranged in the ascending order, the median turned out to be lesser than the threshold value of frequency (=150) (**B1** in **Figure 2**). For the tissue segments, since the frequencies are distributed over a range of bins, when arranged in ascending order, the median frequency was found to be higher than the threshold value of frequency (=150) (**B2** in **Figure 2**). Following histogram-based thresholding, we manually selected a unique set of glomerular and non-glomerular images from them that were then used for CNN model development. This process selected about 44438 images from the 107525 images. We then manually examined all the non-background images and then selected images that included a unique glomerulus (n=751). An equal

number of images were selected from the remaining non-background images with non-glomerular tissue as control cases. Together, these images formed the training data, with an output label of '0' assigned to the non-glomerular images and '1' for the glomerular images. Note that the non-glomerular images were selected across different portions of the biopsy in order to capture variability within a patient and to include several unique aspects of the biopsy (such as tubular elements, interstitial spaces, vascular regions, etc.).

Model training

We used Google's Inception v3 CNN architecture, which was pre-trained on millions of images with 1000 object classes [24], incorporated minor changes to fine-tune the framework and trained it to predict the presence or absence of a glomerulus within the cropped trichrome images (**Figure 3**). Specifically, we removed the final classification layer from the network and retrained it with our dataset using the 2 output labels. We then performed fine-tuning of the parameters at all layers. This procedure, known as transfer learning, is optimal, given the amount of data available. See **Supplemental material** for more details.

The cropped image dataset was randomly split on a patient-by-patient basis. Specifically, in order to capture intra- and inter-patient variabilities, and to verify whether the CNN model works well on images and image characteristics which it has not been trained on, the patient list was randomly split into 2 parts in a 7:3 ratio (70% training, 30% testing). This resulted in 120 patients in the train set, and 51 patients in the test set. Cropped images belonging to each patient in the list were included in the corresponding dataset (training vs testing). Also, for consistency, we repeated the process of random splitting 3 times. CNN model training and testing were performed on each split, and average performance values were recorded.

Data augmentation

Some of the glomeruli on the biopsy images were observed on the edges of the tissue sample. When cropping was performed to capture these cases, a portion of the cropped region had only the background pixels. All these images were used as part of the training data, but they were not in sufficient number to be able to generate a model that could accurately identify the glomeruli present on the edges of the biopsy. We therefore augmented the training data by creating copies (n=5) of each image by randomly whitening a small fraction (=0.2) of the total pixels in the images, resulting in 6 total images per original cropped image (**Figure 4**). See **Supplemental material** for more details.

Image segmentation

Using the CNN model with the best test performance, we tested 2 different image processing routines to scan the test images to identify and segment the glomeruli (**Figure 5**). The sliding window operation was used again to scan the entire test image of size ~2560x1920x3 pixels in increments of 300x300x3 pixels. Each cropped image was then processed through the trained CNN model that predicted if there was a glomerulus. An output of '0' indicated that the CNN model determined that no glomerulus was present whereas an output of '1' indicated a glomerulus was detected within that cropped image. When a glomerulus was detected, the pixel coordinates of the four corners of that image were stored in an array. This process was repeated as the sliding window operation swept from one end of the corner to the other, which resulted in bright patches that corresponded to the areas that were predicted to contain a glomerulus. A heatmap was generated using these corners. The brightness of the patch in the heatmap was found to be directly proportional to how confident the model was in terms of detecting the presence of a glomerulus in that area. Every non-bright region (i.e., area with pixel intensity close to 0) on the heatmap then represents all the non-glomerulus regions.

Generated heatmaps were processed further to segment the identified glomeruli using a simple annotation defined as a 'green box' surrounding the glomerular region. We performed this task using 2 different approaches. In the first approach, we first performed an erosion operation on the heatmap. Since most glomeruli possess a round shape, erosion helped in the separation of overlapping regions of heatmap intensities. We then binarized the heatmap with eroded objects to create 'blobs' representing identified glomeruli. Note that the threshold value for binarization was empirically determined (=20), after examining several images. This was the lowest value that was able to efficiently highlight the regions where the model predicted the presence of a glomerulus. Values lower than this resulted in missing some glomeruli, whereas values higher than this resulted in false detections. Image erosion operation was performed again to remove any overlapping boundaries from the blobs to finally generate the image objects identified as unique glomeruli. In the second approach, we first binarized the image using the Otsu's method [25]. Subsequently, a distance transform was applied on the heatmaps, which simply calculated the distance of each foreground pixel from the nearest background pixel. We then performed watershed segmentation to separate the identified 'blobs' in the image. The watershed transformation treats the image it operates upon like a topographic map, with the brightness of each pixel representing its height, and finds the lines that run along the tops of ridges [26]. Finally, for both approaches, a green box was automatically placed by the segmentation algorithm to highlight the identified glomerulus.

Performance metrics

CNN and sliding window-based segmentation model performances were evaluated by computing overall mean accuracy, mean sensitivity and mean specificity on the test data for each train-test split that was generated. We also computed F1-score as a measure of model accuracy that considers both the precision and recall of a test. We also computed Matthews correlation coefficient (MCC), which is a balanced measure of quality for dataset classes of different sizes of a binary classifier. Mean receiver operating characteristic (ROC) curves were plotted along with the standard deviation observed across the runs, followed by estimation of area under curve (AUC) for these cases.

RESULTS

Patient population

Baseline characteristics describe a patient cohort representative of Boston's inner-city population comprising 46% African American population. About 60% of the patients were male, about 84% had hypertension, about 75% cardiovascular disease and about 43% had diabetes. About 82% of patients had chronic kidney disease (CKD) stage 3 to 5; 6% had stage 2 CKD, and the rest had stage 1 CKD. About 35% of patients had nephrotic-range proteinuria (>3.5 g/day). On the basis of varied genetic background and several co-morbidities (described above), it is worth noting that the dataset that we generated provides a wide range of glomerular morphologies including few cases of normal glomeruli.

Glomerular classification model

Our first goal was to develop a classifier that could accurately detect the presence of a glomerulus in an image. Light microscopy images of Masson trichrome–stained sections were captured at 40x magnification and processed in NIS-Elements AR software (Nikon). We selected 40x as the level of magnification and generated 275 unique images from 171 patient biopsies, where each image size was about 2560x1920x3 pixels, corresponding to a field of 2.176 x 1.632 mm^2. These images were then converted to 8-bit red–green–blue (RGB) color images in TIFF format. Each of these images represent a large portion of the digitized biopsy (**Figure 2**), and the information contained within them had to be filtered in order to train a glomerular image classifier. Therefore, we created a dataset that was more amenable for CNN model training using the sliding window operator. This operation allowed us to systematically crop the original images into smaller ones of size 300x300x3 pixels. The size of the cropped images was empirically chosen such that any glomerulus observed within the original 40x images was able to fit well within the bounds of the cropped image size.

The sliding window operator with a small stride (20 pixels) generated a large number of cropped images, and histogram-based thresholding selected many of the images containing the kidney tissue from the ones that contained only the background (**Figure 2**). This thresholding method saved significant amount of time as it was able to filter more than half of the cropped images. Using the selected data, binary classification models, constructed by fine-tuning a well-known pre-trained CNN architecture (Inception V3) [24], identified images with a glomerulus with high accuracy across 3 different train-test splits (**Table 2**). Combining random whitening and other data augmentation strategies resulted in superior CNN model performance on the testing data as exemplified by model accuracy, sensitivity, specificity, F1-score, MCC and AUC (**Table 2 & Figure 5**). While F1-score can be viewed as a weighted average of precision and recall, MCC can be considered as a more robust measure of model performance.

Glomerular segmentation model

We developed two different methods to segment the identified glomeruli (**Table 3, Figures 7 & 8**). The threshold method involved binarization and using a fixed threshold value and the other method used the watershed algorithm (**Approach 1 in Table 3**). When the threshold method was tested, it was found to have high overall accuracy and specificity, but its sensitivity was low (**Table 3**). The watershed implementation of the segmentation program had also high accuracy and specificity, but with a slightly higher sensitivity than the previous method (**Approach 2**; **Table 3**). Although the threshold method yielded fewer false positives, the threshold method could not detect clumps of glomeruli. Also, the watershed method detected a higher percentage of glomeruli. In our test data there were a total of 364 glomeruli and 131 were detected by Approach 1 and 197 by Approach 2. This shows that the approach

based on watershed segmentation was able to detect multiple glomeruli within a clump better than the threshold method.

Another value that was compared between the two methods were the number of false positives detected. False positives in this case would be green rectangles that were drawn around areas that did not contain a glomerulus. The number of false positives detected using Approach 1 were 45, and 201 for Approach 2. Even though the number of false positives was higher and the precision was lower with Approach 2, it still had an accuracy of 99.97%. With an overall accuracy this high, the other aspects of the watershed method made it more ideal than the threshold method. These aspects include being able to detect multiple glomeruli that are closely associated and the fact that it is generalizable and can be used on different data sets.

DISCUSSION

Deep learning algorithms are transforming the way by which medical images and other forms of data are analyzed to uncover hidden patterns and facilitate patient diagnosis as well as to improve the delivery and effectiveness of patient care [27]. This is especially the case in the field of digital pathology where several researchers are employing these powerful techniques to address specific questions in a spectrum of disease scenarios [12, 15, 28-31]. For many of these cases, the clinical workflow is quite similar, i.e., a biopsy procedure is performed to extract a tiny portion of the organ, which is then subjected to a series of histological staining processes. Tissue slides that result from this effort are then digitized to generate image representations of the diseased organ. These computerized images then serve as the input data of interest, and the deep learning algorithms read and process these digital signatures to extract relevant quantitative information or associate them with corresponding outputs of interest. Once trained on sufficient number of cases, these models can have the ability to predict on new test cases that the model has never seen before with remarkable accuracy. Note that the process of biopsy digitization is not fully integrated within all clinical practices as of today. However, there is a growing interest in terms of moving in this direction as the community is realizing the enormous potential deep learning and other machine learning frameworks can have on quantitative assessment of biopsies and to assist the pathologist.

Assessment of renal pathology slides has several features worthy of consideration. The biopsy report methodically deals with all the components of the slide with different staining and clinical correlation is pursued to eventually arrive at a diagnosis. Some of the features are objective (number of glomeruli) and some are descriptive (type of sclerosis). While the latter item calls upon the expertise of pathologist, the former can be automated. Also, the location (cellular or compartmental) and distribution (focal or diffuse or segmental or global) of the intra-glomerular damage determine the type of glomerular disease (glomerulonephritis vs glomerulosclerosis) [32-42]. Availability of digitized images provide an immense resource which has opened up opportunities to leverage different tools to improve the analysis of some of the features in an automated manner.

The goal of this work was to develop a method by which to automate a seemingly straightforward, but rather important task of automatically identifying and segmenting glomeruli within digitized kidney biopsies. Several steps had to be undertaken to correctly perform glomerular segmentation. The first step was data generation for training the CNN model. The sliding window operator scanned the original 40x images of all the 120 patients that were used for model training to generate over 100,000 cropped images. The task at this stage was for trained technicians to manually examine each cropped image further to determine whether it had a glomerulus or a non-glomerular compartment of the kidney biopsy. Note that the technicians had prior knowledge and skill needed to perform this task. Subsequently, a histogram-based thresholding approach filtered a large majority of the cropped images that were part of the background (**Figure 2**), and the remaining images were manually selected to ensure that no duplicate images were included as part of the training and testing data sets. Care was taken to not select cropped images containing the same glomerulus again. Similarly, no two cropped images containing the same non-glomerular part of the kidney biopsy were selected. For the glomeruli segments, we checked the segment against the complete image to ensure each glomerulus came up only once. For the non-glomerular segments, the segments were taken from different parts of the biopsy image, thus ensuring there was no overlap. This entire process was fairly time consuming and by no means trivial but was critical to train an accurate CNN model that detected glomerular images from those that did not have a glomerulus.

Our CNN model identified the presence of a glomerulus with high accuracy on the test data (**Table 2**). The transfer learning approach turned out to be very effective, given the number of cropped images (n=1502) used for model training. Moreover, techniques such as random whitening and other known data augmentation techniques performed only on the training data enhanced the overall model performance on the testing data. Note that the datasets for training and testing were divided in such a fashion that none of the images belonging to patients in the testing set were available to the model while training. This implies that the training-testing split was done at the patient level as opposed to the image level and the variability in the test images was not part of the images which the model was trained on. This helped us evaluate how well the model performed on completely new patient data. Also, the process of introducing additional noise/variability using random whitening and other data augmentation strategies over than what could possibly be contained within the cropped images has shown to limit model overfitting and increase model generalizability (**Table 2**).

Even though the CNN model demonstrated a remarkable performance in terms of identifying the presence of a glomerulus, our sliding window strategy segmented the glomeruli with high accuracy and high specificity but with low sensitivity. Several factors may have contributed to this outcome. First, the sliding window operator begins scanning an image from the top left corner and moves forward with a pre-defined stride. At each instant, it scans the region within the sliding window, and the trained CNN model predicts whether there is a glomerulus present in that portion, and if so, it generates a bright patch on the heatmap. This process continues until the last portion at the bottom right corner of the image is scanned and processed by the CNN model. During this process, the stride length of the sliding window plays a major role; choosing a small stride can create redundant heatmaps whereas choosing a large stride can simply not capture all the glomeruli. After several experiments, we selected the stride length as 20 pixels as this gave us the best compromise between precision and recall of the sliding window operator in terms of identifying a glomerulus.

The segmentation process did not end at this stage as the heatmaps had to be further processed to represent the identified glomeruli. Our two image analysis pipelines processed the heatmaps to complete the segmentation process that resulted in output images with highlighted areas of the selected glomeruli. While one pipeline used a series of standard image erosion and binarization processes, the other followed well-known techniques such as Otsu binarization and computing a distance transform followed by watershed segmentation. Both methods had strengths and weaknesses as some images when processed through both the pipelines gave a similar output (**Figure 6**), whereas one image processing routine outperformed the other in few cases (**Figures 7 & 8**). Note that while simple binarization involves thresholding an image based on a pre-selected value, Otsu's binarization assumes that an image contains two classes of pixels, and then searches for a threshold that minimizes intra-class variance between the classes [25]. For the erosion process, a local minimum gets computed over the area of a kernel with a pre-determined size and replaces the image pixel under the anchor point. On the other hand, the watershed transformation treats the image it operates upon like a topographic map, with the brightness of each point representing its height, and finds the lines that run along the tops of ridges [26]. Ultimately, the nature of the locations of the glomeruli dictated the performance of each of the image processing pipelines.

We selected trichrome-stained kidney biopsy images as our data set under consideration primarily due to its use as a standard staining for kidney biopsies at most centers, but our paradigm can be easily extended to images generated using other staining protocols. Recent studies have also shown interesting results in terms of identifying glomeruli from digitized kidney biopsies [43, 44]. For all these cases, a carefully generated dataset of images with glomerular and non-glomerular compartments is

needed in order to train an accurate CNN model. The subsequent steps associated with the image segmentation pipeline may be different than what were used for this paper, and it would depend on the type of images generated using different staining protocols (i.e. H&E, Periodic-acid Schiff, Jones' silver stain, Congo red, etc.). Another limitation in our study is that we relied on manually observing cropped images and classifying them as the ones containing glomerular or non-glomerular aspects of the biopsy. A "gold standard" for glomerular identification can help minimize the bias associated with manual selection.

Ultimately, we envision that the glomerular segmentation strategy proposed here can be integrated within a software application (or "App") to derive predictions based on a digitized kidney biopsy image. The input for this App can simply be a digitized kidney biopsy in a common image format, which will then be first processed by the CNN model in conjunction with the sliding window operator to generate heatmaps highlighting the areas that are predicted to contain the glomeruli. The App would then automatically process these heatmap signatures using customized image segmentation steps to derive a final representation of the digitized kidney biopsy with segmented regions of glomeruli highlighted on them (**Figures 6-8**). Such a tool could save time and improve accuracy in the routine but complex evaluation of a kidney biopsy. Ultimately, this tool could serve as one among a set of automated morphologic assessments available within a more comprehensive software application. Such an App might even include AI-based tools which reveal entirely new diagnostic information, which the pathologist may incorporate into their decision-making process.

CONCLUSION

We demonstrated the effectiveness of using a deep learning strategy combined with a series of image processing operations to accurately identify and segment glomerular regions from trichrome-stained histologic images obtained at the time of kidney biopsy. This rapid, scalable method can be utilized in the form of a software tool at the point-of-care to assist nephropathologists. This framework can also be adapted to other images obtained via different histological staining protocols. Further validation of the deep learning framework along with the image processing operations across different clinical practices and image datasets is necessary to validate this technique across the full distribution and spectrum of lesions encountered in a typical nephropathology service.

DISCLOSURE

All the authors declared no competing interests.

ACKNOWLEDGMENTS

This project was supported in part by the National Center for Advancing Translational Sciences, National Institutes of Health, through BU-CTSI Grant (1UL1TR001430), and a Scientist Development Grant (17SDG33670323) from the American Heart Association to VBK, Boston University's Undergraduate Research Opportunities Program funding to LAM, and NIH grants (R01-HL132325 and R01-CA175382) to VCC.

FIGURE CAPTIONS

Table 1. Patient and digitized kidney biopsy characteristics used for this study. All the patients underwent treatment for chronic kidney disease at the Boston Medical Center between 2009 and 2012.

Figure 1. Cropped images. The sliding window operator was used to generate different sets of images to train the CNN model. The first row contains images with a single glomerulus in each image, and the second row contains images with non-glomerular aspects of the tissue in each image. Each cropped image is of size 300x300x3 pixels.

Figure 2. Histogram-based thresholding. A sliding window operator scanned the entire original image of size 2560x1920x3 pixels and generated cropped images of size 300x300x3 pixels. For each cropped image, a histogram based on pixel intensity was generated (**A1** for a cropped image representing the background and **A2** representing a portion of the kidney biopsy). These histograms were then reordered according to the bin frequency. A threshold value of 150 was empirically selected as a cutoff and median value for the bin frequency was computed. Images with a median value below the cutoff were selected as the ones representing the background (**B1**), and the ones with a median value above the cutoff were selected as part of the kidney biopsy (**B2**).

Figure 3. Deep neural network model. Our classification technique is based on using a transfer learning approach on Google Inception V3 convolutional neural network (CNN) architecture pre-trained on the ImageNet dataset (1.28 million images over 1000 generic object classes) and fine-tuned on our dataset (see Methods). Inception v3 CNN architecture reprinted with permission from the Google blog "Train Your Own Image Classifier With Inception in TensorFlow" (https://research.googleblog.com/2016/03/train-your-own-image-classifier-with.html).

Figure 4. Whitening transformation used for data augmentation. On each cropped image with a single glomerulus (**A**, **B**) or a non-glomerular aspect (**C** & **D**) of the kidney biopsy, about 20% of the pixels were randomly selected and a whitening transform was applied on them. This process generated images that still contained a major portion of the original content that represented either the glomerular or non-glomerular aspects of the kidney biopsy (**E**, **F**, **G** & **H**).

Table 2. CNN model performance. Four different models were developed to understand the effect of random whitening as well as other data augmentation strategies on the CNN model performance. Model performance is shown on test data that was not used for model training.

Figure 5. ROC curves for the CNN models. Four different model performances are shown: (A) No whitening, no augmentation, (B) No whitening, augmentation, (C) Whitening, no augmentation, and (D) Whitening and augmentation.

Figure 6. Segmentation pipeline. The trained CNN model was used in conjunction with the sliding window operator to scan a test image (A) that was not used in model training. (B) A heat map was generated based on how the CNN model detected the presence of glomeruli. Two different segmentation routines were developed to further segment the heatmap. In one case, an erosion operation (C_1), followed by image binarization (D_1), followed by another erosion operation (E_1). In the other case, an Otsu binarization operation was attempted on the heatmap (C_2), followed by a distance transform (D_2) and then watershed segmentation (E_2). For this test image (A), both image processing pipelines resulted in segmentation of 4 distinct glomeruli (F).

Table 3. Segmentation model performance. Two different approaches were developed to segment the glomeruli. Approach 1, in this case, took the generated heatmap and performed an erosion operation followed by image binarization followed by another erosion operation. Approach 2 used the heatmap and performed an Otsu binarization followed by a distance transform and then a Watershed segmentation. The two approaches performance is shown on the test data that was not used for model training.

Figure 7. Segmentation pipeline. The trained CNN model was used in conjunction with the sliding window operator to scan a test image (A) that was not used in model training. (B) A heat map was generated based on how the CNN model detected the presence of glomeruli. Two different segmentation routines were developed to further segment the heatmap. In one case, an erosion operation (C_1), followed by image binarization (D_1), followed by another erosion operation (E_1). In the other case, an Otsu binarization operations was attempted on the heatmap (C_2), followed by a distance transform (D_2) and then watershed segmentation (E_2). For this test image (A), the erosion operation resulted in the segmentation of 3 distinct glomeruli (F_1) while the watershed operation resulted in the same segmentation of 3 distinct glomeruli but also drew an extra, incorrect box (label F_{2a}) around non-glomerular tissue (F_2).

Figure 8. Segmentation pipeline. The trained CNN model was used in conjunction with the sliding window operator to scan a test image (A) that was not used in model training. (B) A heat map was generated based on how the CNN model detected the presence of glomeruli. Two different segmentation routines were developed to further segment the heatmap. In one case, an erosion operation (C_1), followed by image binarization (D_1), followed by another erosion operation (E_1). In the other case, an Otsu binarization operations was attempted on the heatmap (C_2), followed by a distance transform (D_2) and then watershed segmentation (E_2). For this test image (A), the erosion operation resulted in the segmentation of 1 distinct glomerulus but was unable to correctly identify the second glomerulus (F_1). The watershed operation resulted in the segmentation of the 2 distinct glomeruli (F_2).

Supplemental table 1. Train and test data used for training the models. Each model, incorporating different combinations of whitening and augmentation, was trained using different training data therefore resulting in different test data for each model.

REFERENCES

1. Greenberg, A. and A.K. Cheung, *Primer on kidney diseases*. 5th ed. 2009, Philadelphia, PA: Saunders/Elsevier : National Kidney Foundation. xvii, 594 p.
2. Lees, G.E., R.E. Cianciolo, and F.J. Clubb, Jr., *Renal biopsy and pathologic evaluation of glomerular disease.* Top Companion Anim Med, 2011. **26**(3): p. 143-53.
3. Rayat, C.S., et al., *Glomerular morphometry in biopsy evaluation of minimal change disease, membranous glomerulonephritis, thin basement membrane disease and Alport's syndrome.* Anal Quant Cytol Histol, 2007. **29**(3): p. 173-82.
4. Puelles, V.G. and J.F. Bertram, *Counting glomeruli and podocytes: rationale and methodologies.* Curr Opin Nephrol Hypertens, 2015. **24**(3): p. 224-30.
5. Basgen, J.M., et al., *Estimating glomerular number in situ using magnetic resonance imaging and biopsy.* Kidney Int, 1994. **45**(6): p. 1668-72.
6. Bertram, J.F., *Estimating glomerular number: why we do it and how.* Clin Exp Pharmacol Physiol, 2013. **40**(11): p. 785-8.
7. Cullen-McEwen, L.A., et al., *A design-based method for estimating glomerular number in the developing kidney.* Am J Physiol Renal Physiol, 2011. **300**(6): p. F1448-53.
8. LeCun, Y., Y. Bengio, and G. Hinton, *Deep learning.* Nature, 2015. **521**(7553): p. 436-44.
9. Arevalo, J., et al., *An unsupervised feature learning framework for basal cell carcinoma image analysis.* Artif Intell Med, 2015. **64**(2): p. 131-45.
10. Cruz-Roa, A., et al., *Accurate and reproducible invasive breast cancer detection in whole-slide images: A Deep Learning approach for quantifying tumor extent.* Sci Rep, 2017. **7**: p. 46450.
11. Durant, T.J.S., et al., *Very Deep Convolutional Neural Networks for Morphologic Classification of Erythrocytes.* Clin Chem, 2017. **63**(12): p. 1847-1855.
12. Ertosun, M.G. and D.L. Rubin, *Automated Grading of Gliomas using Deep Learning in Digital Pathology Images: A modular approach with ensemble of convolutional neural networks.* AMIA Annu Symp Proc, 2015. **2015**: p. 1899-908.
13. Kainz, P., M. Pfeiffer, and M. Urschler, *Segmentation and classification of colon glands with deep convolutional neural networks and total variation regularization.* PeerJ, 2017. **5**: p. e3874.
14. Khosravi, P., et al., *Deep Convolutional Neural Networks Enable Discrimination of Heterogeneous Digital Pathology Images.* EBioMedicine, 2018. **27**: p. 317-328.
15. Kolachalama, V.B., et al., *Association of Pathological Fibrosis With Renal Survival Using Deep Neural Networks.* Kidney Int Rep, 2018. **3**(2): p. 464-475.
16. Malon, C.D. and E. Cosatto, *Classification of mitotic figures with convolutional neural networks and seeded blob features.* J Pathol Inform, 2013. **4**: p. 9.
17. Rachmadi, M.F., et al., *Segmentation of white matter hyperintensities using convolutional neural networks with global spatial information in routine clinical brain MRI with none or mild vascular pathology.* Comput Med Imaging Graph, 2018. **66**: p. 28-43.
18. Schaumberg, A.J., et al., *DeepScope: Nonintrusive Whole Slide Saliency Annotation and Prediction from Pathologists at the Microscope.* Comput Intell Methods Bioinform Biostat (2016), 2017. **10477**: p. 42-58.
19. Sharma, H., et al., *Deep convolutional neural networks for automatic classification of gastric carcinoma using whole slide images in digital histopathology.* Comput Med Imaging Graph, 2017. **61**: p. 2-13.
20. Turkki, R., et al., *Antibody-supervised deep learning for quantification of tumor-infiltrating immune cells in hematoxylin and eosin stained breast cancer samples.* J Pathol Inform, 2016. **7**: p. 38.

21. Wang, H., et al., *Mitosis detection in breast cancer pathology images by combining handcrafted and convolutional neural network features.* J Med Imaging (Bellingham), 2014. **1**(3): p. 034003.
22. Xu, Y., et al., *Large scale tissue histopathology image classification, segmentation, and visualization via deep convolutional activation features.* BMC Bioinformatics, 2017. **18**(1): p. 281.
23. Yi, F., et al., *Microvessel prediction in H&E Stained Pathology Images using fully convolutional neural networks.* BMC Bioinformatics, 2018. **19**(1): p. 64.
24. Szegedy, C., et al., *Rethinking the inception architecture for computer vision*, in *https://arxiv.org/abs/1512.00567*. 2015, Cornell University.
25. Otsu, N., *A threshold selection method from gray-level histograms.* IEEE Trans. Sys., Man., Cyber., 1979. **9**(1): p. 62-66.
26. Beucher, S., *Watershed, hierarchical segmentation and waterfall algorithm.* Mathematical Morphology and Its Applications to Image Processing, 1994. **2**: p. 69-76.
27. Miotto, R., et al., *Deep learning for healthcare: review, opportunities and challenges.* Brief Bioinform, 2017.
28. Janowczyk, A. and A. Madabhushi, *Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases.* J Pathol Inform, 2016. **7**: p. 29.
29. Qiu, J.X., et al., *Deep Learning for Automated Extraction of Primary Sites From Cancer Pathology Reports.* IEEE J Biomed Health Inform, 2018. **22**(1): p. 244-251.
30. Saltz, J., et al., *Spatial Organization and Molecular Correlation of Tumor-Infiltrating Lymphocytes Using Deep Learning on Pathology Images.* Cell Rep, 2018. **23**(1): p. 181-193 e7.
31. Esteva, A., et al., *Dermatologist-level classification of skin cancer with deep neural networks.* Nature, 2017. **542**(7639): p. 115-118.
32. Brill, G. and H. Mendelow, *Intercapillary glomerulosclerosis; a clinico-pathologic study.* J Mt Sinai Hosp N Y, 1956. **23**(5): p. 663-70.
33. Conlon, P.J., et al., *Clinical and pathologic features of familial focal segmental glomerulosclerosis.* Am J Kidney Dis, 1995. **26**(1): p. 34-40.
34. D'Agati, V., *Pathologic classification of focal segmental glomerulosclerosis.* Semin Nephrol, 2003. **23**(2): p. 117-34.
35. D'Agati, V.D., et al., *Pathologic classification of focal segmental glomerulosclerosis: a working proposal.* Am J Kidney Dis, 2004. **43**(2): p. 368-82.
36. Gupta, R., et al., *Focal and segmental glomerulosclerosis in renal allograft recipients: a clinico-pathologic study of 37 cases.* Saudi J Kidney Dis Transpl, 2013. **24**(1): p. 8-14.
37. Iskandar, S.S., R.J. Falk, and J.C. Jennette, *Clinical and pathologic features of fibrillary glomerulonephritis.* Kidney Int, 1992. **42**(6): p. 1401-7.
38. Robbins, S.L., J. Rogers, and O.J. Wollenman, Jr., *Intercapillary glomerulosclerosis a clinical and pathologic study. III. A pathologic study of 100 cases.* Am J Med, 1952. **12**(6): p. 700-5.
39. Rogers, J. and S.L. Robbins, *Intercapillary glomerulosclerosis: a clinical and pathologic study. I. Specificity of the clinical syndrome.* Am J Med, 1952. **12**(6): p. 688-91.
40. Rogers, J., S.L. Robbins, and H. Jeghers, *Intercapillary glomerulosclerosis: a clinical and pathologic study. II. A clinical study of 100 anatomically proven cases.* Am J Med, 1952. **12**(6): p. 692-9.
41. Stokes, M.B., et al., *Cellular focal segmental glomerulosclerosis: Clinical and pathologic features.* Kidney Int, 2006. **70**(10): p. 1783-92.
42. Thomas, D.B., et al., *Clinical and pathologic characteristics of focal segmental glomerulosclerosis pathologic variants.* Kidney Int, 2006. **69**(5): p. 920-6.
43. Kato, T., et al., *Segmental HOG: new descriptor for glomerulus detection in kidney microscopy image.* BMC Bioinformatics, 2015. **16**: p. 316.

44. Simon, O., et al., *Multi-radial LBP Features as a Tool for Rapid Glomerular Detection and Assessment in Whole Slide Histopathology Images.* Sci Rep, 2018. **8**(1): p. 2032.

Table 1

Characteristic	Value
Number of patients	171
Magnification	40x
Full biopsy image size	2560x1920x3 pixels
Cropped image size	300x300x3 pixels
Number of non-glomerular images	751
Number of glomerular images	751

Figure 1

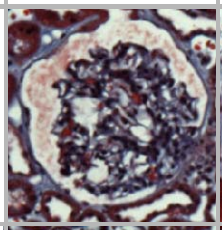
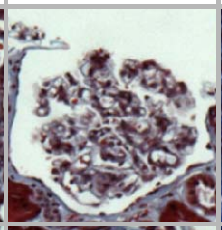
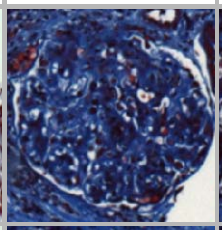
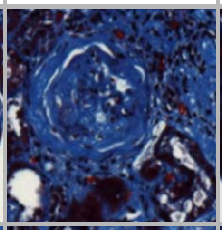
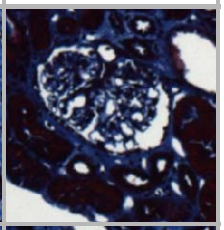
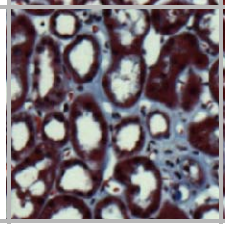
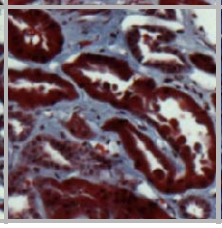
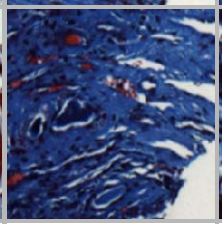
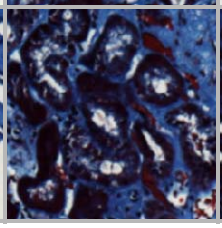
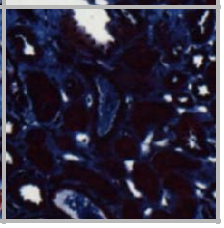
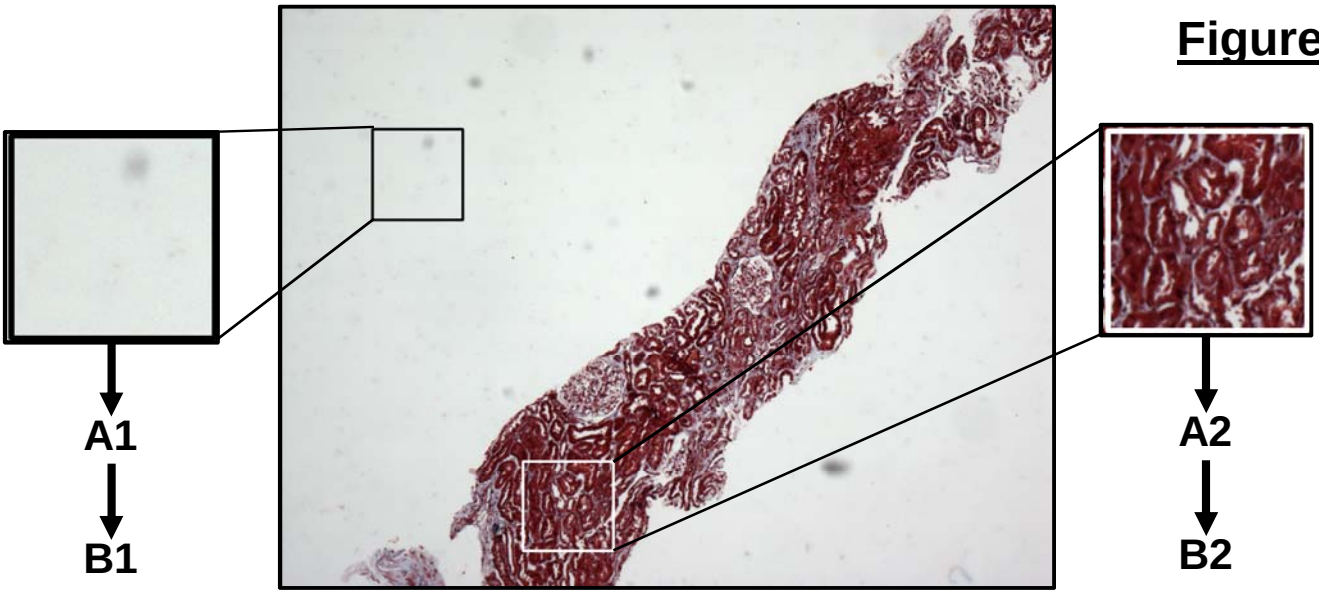
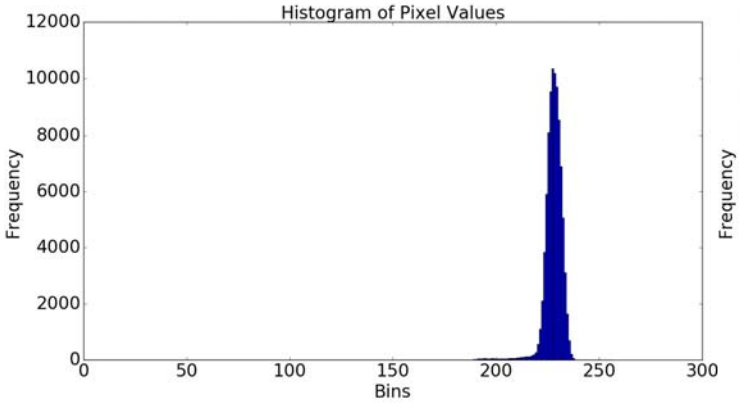
	Patient 1	Patient 2	Patient 3	Patient 4	Patient 5
Glomerular region					
Non-glomerular region					

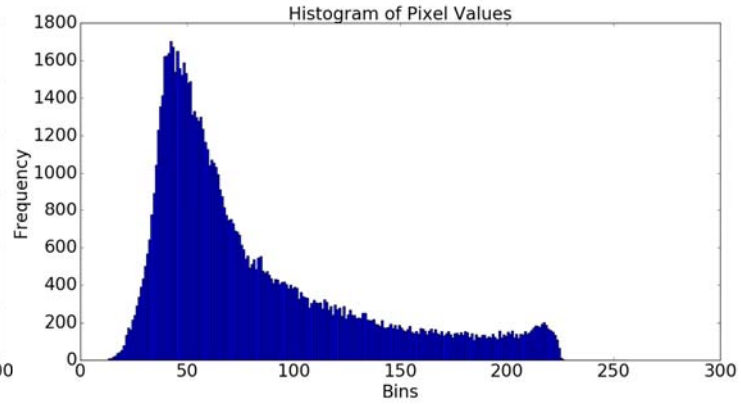
Figure 2



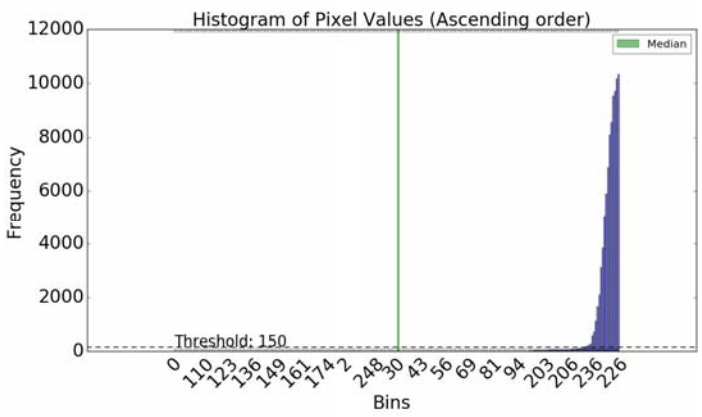
A1



A2



B1



B2

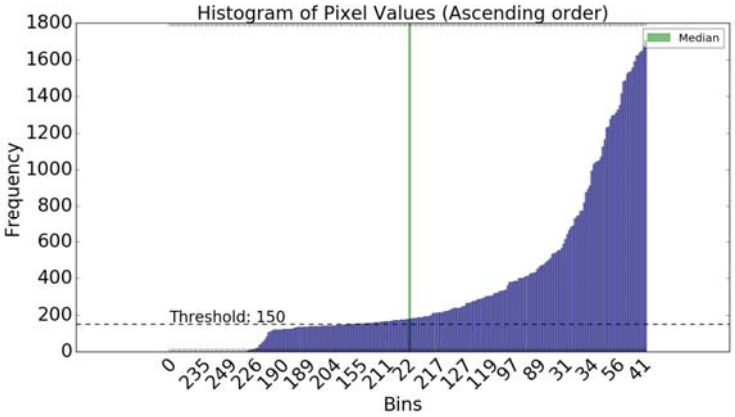


Figure 3

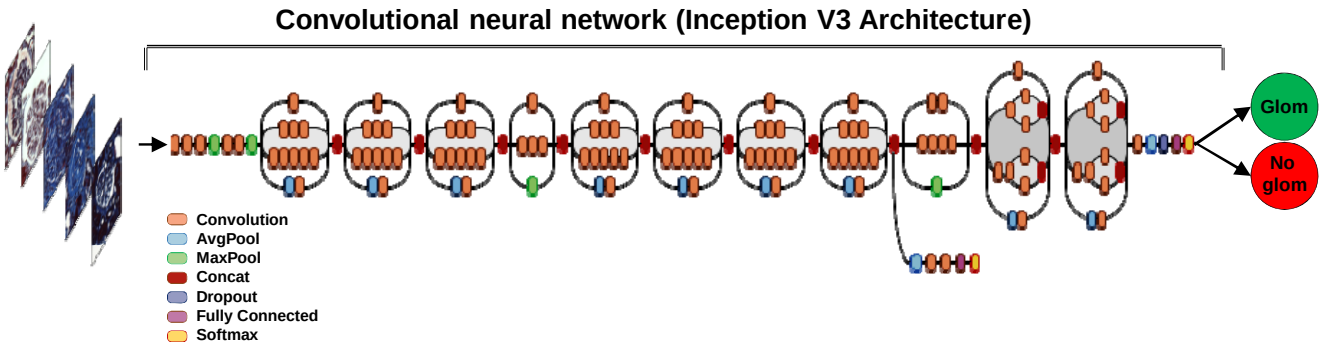


Figure 4

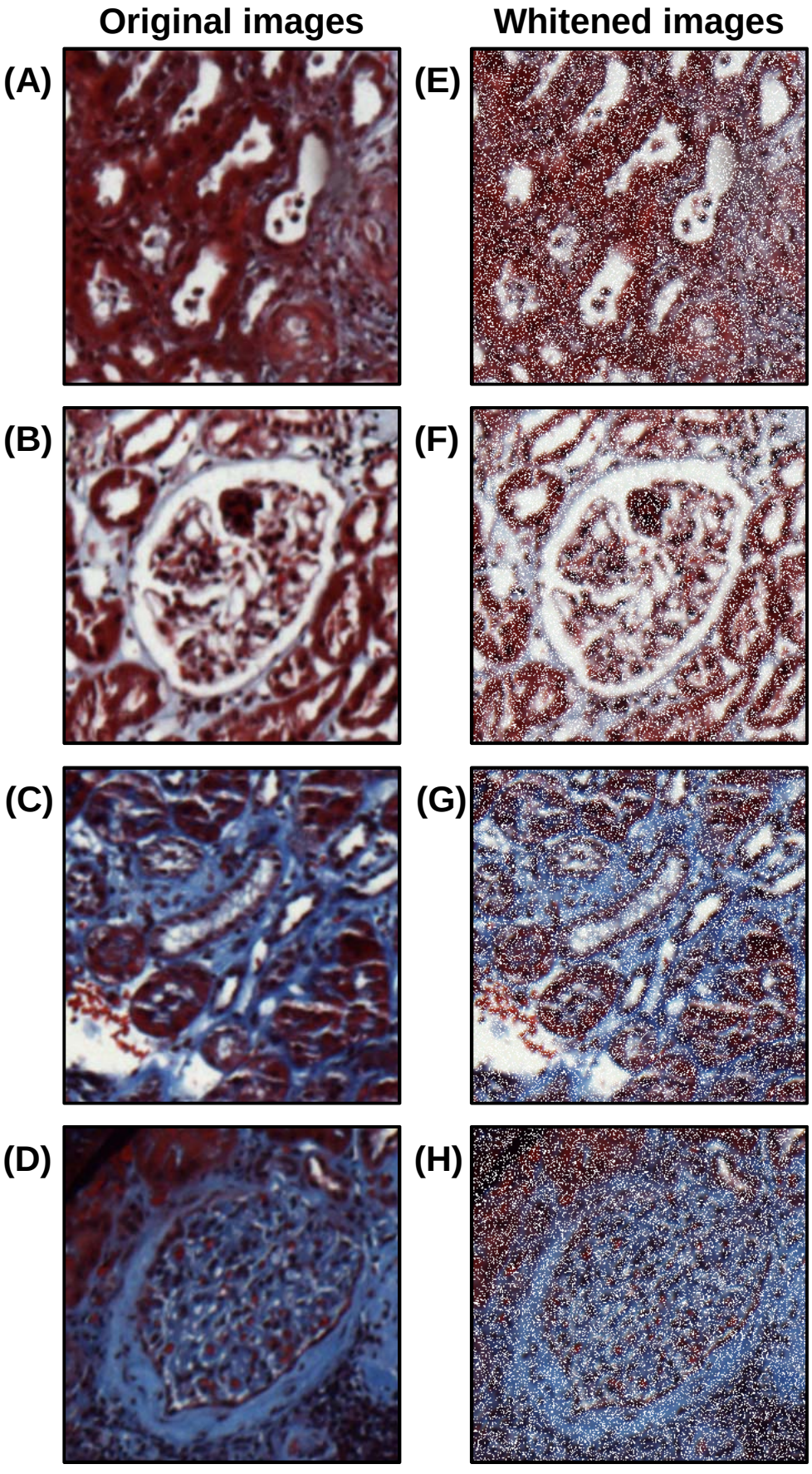


Table 2

Whitening factor	Augmentation factor	Accuracy	Specificity	Sensitivity	F1 score	MCC
0	0	0.944±0.2750	0.966±0.0631	0.922±0.0248	0.942± 0.0262	0.889±0.0532
0	10	0.927±0.0051	0.919±0.0069	0.935±0.0101	0.926±0.0032	0.856±0.0092
5	0	0.965±0.0031	0.994±0.0144	0.937±0.0189	0.965±0.0025	0.931±0.0102
5	10	0.975±0.0031	0.988±0.0144	0.964±0.0189	0.976±0.0025	0.953±0.0102

Figure 5

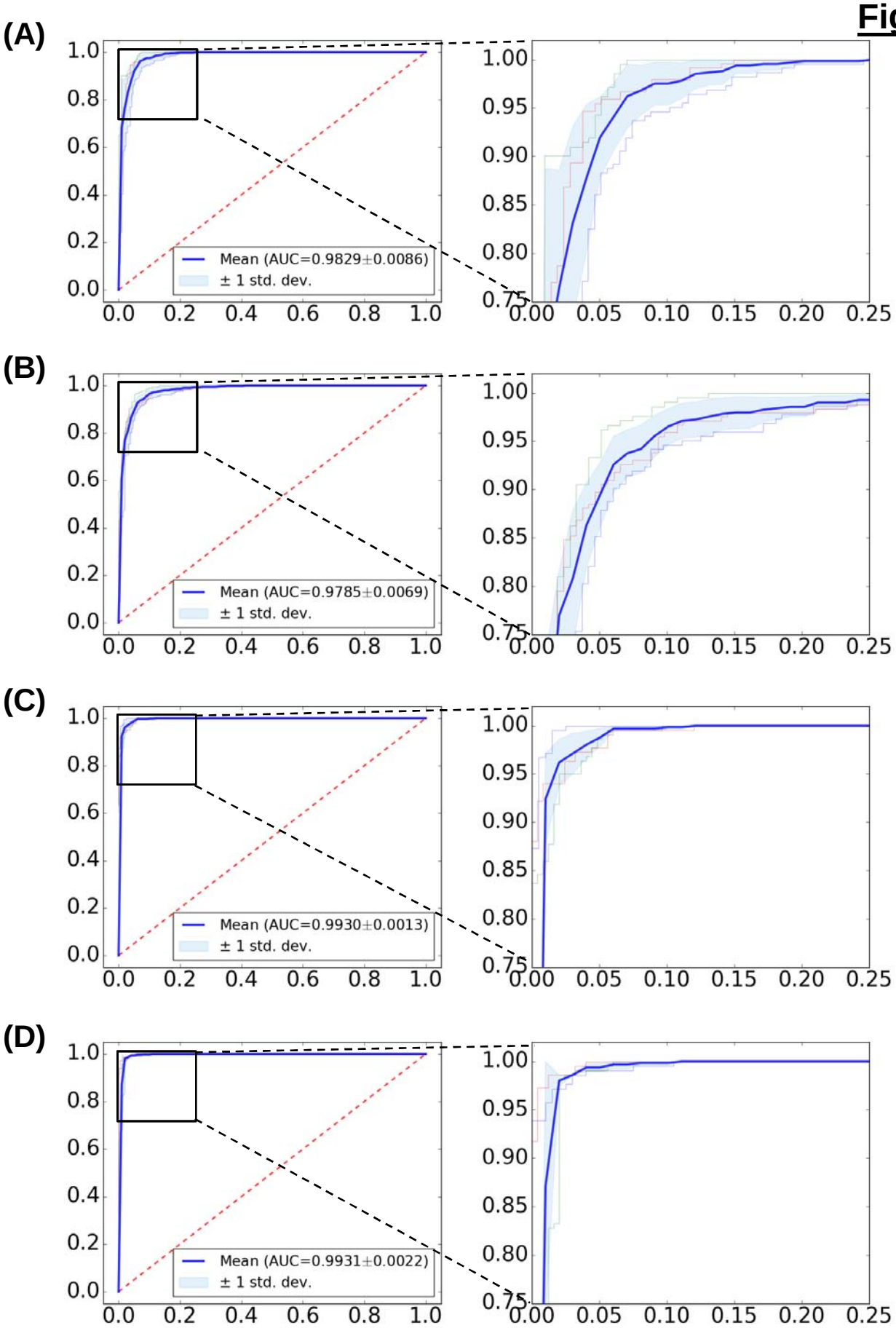


Figure 6

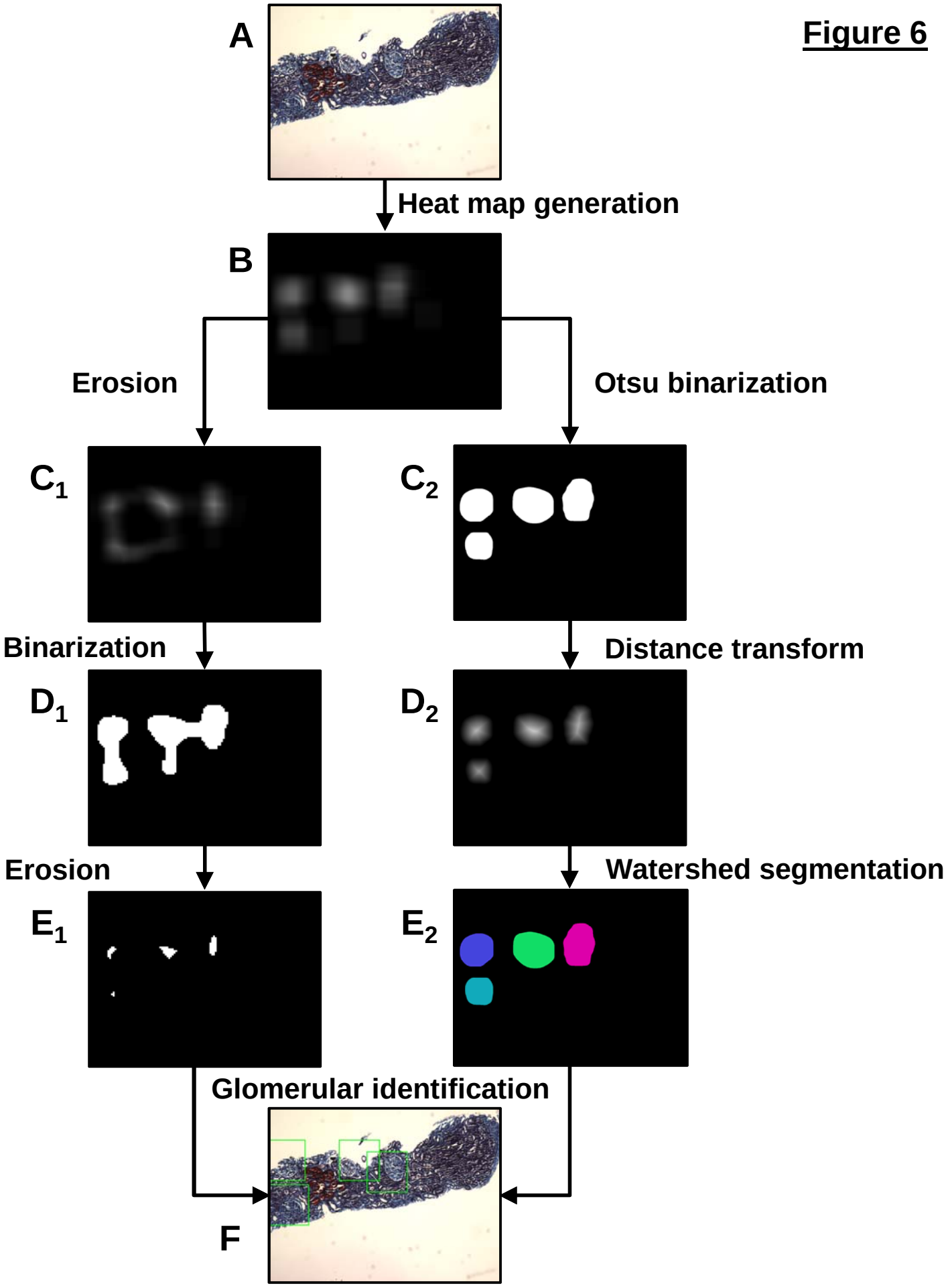


Table 3

	Accuracy	Specificity	Sensitivity	F1 Score	MCC
Approach 1	0.9998±0.0001	0.9999±0.0001	0.3601±0.0776	0.4844±0.0788	0.5235±0.0672
Approach 2	0.9997±0.0001	0.9999±0.0001	0.5437±0.0023	0.5248±0.0873	0.5291±0.0837

Figure 7

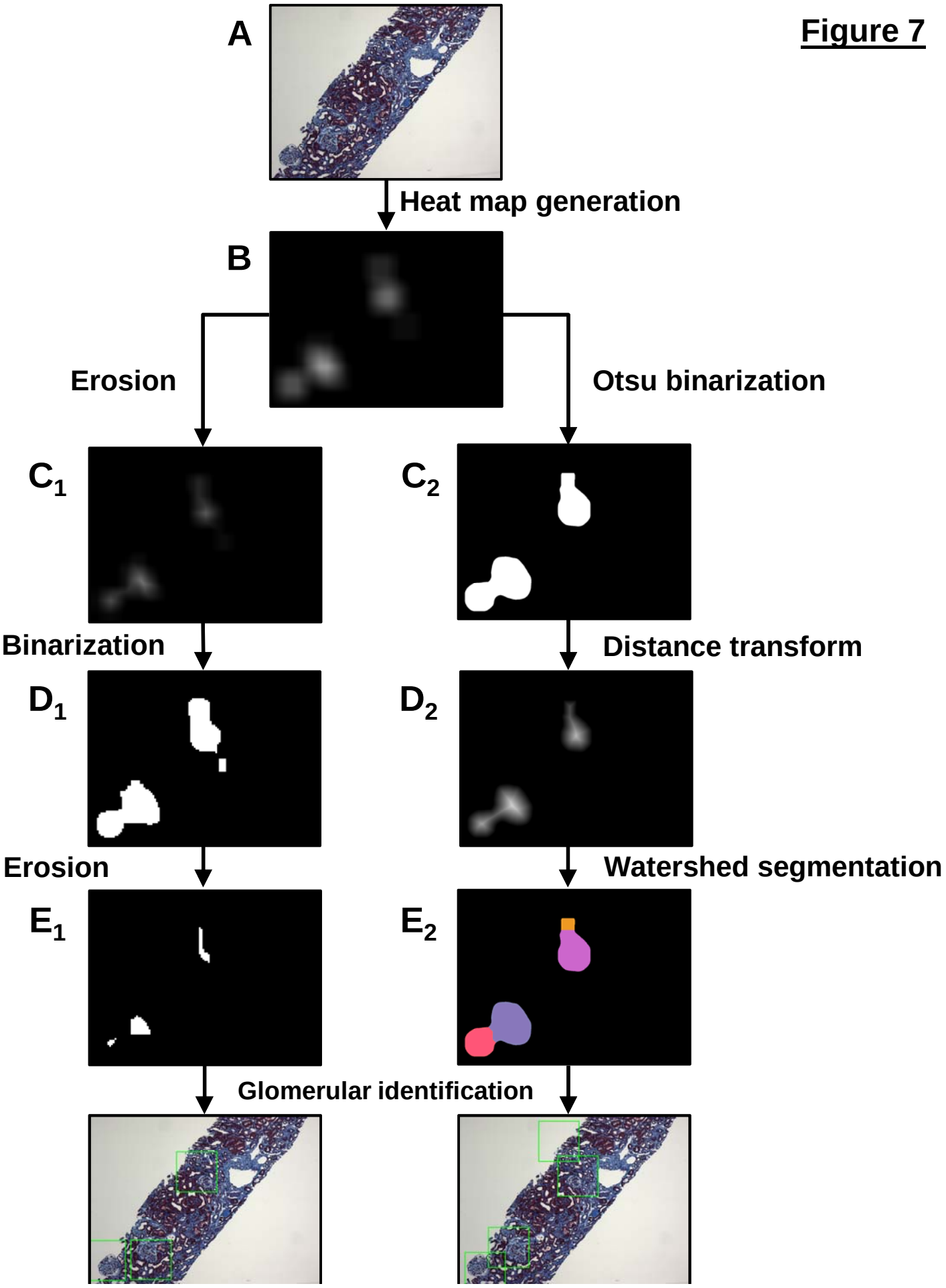
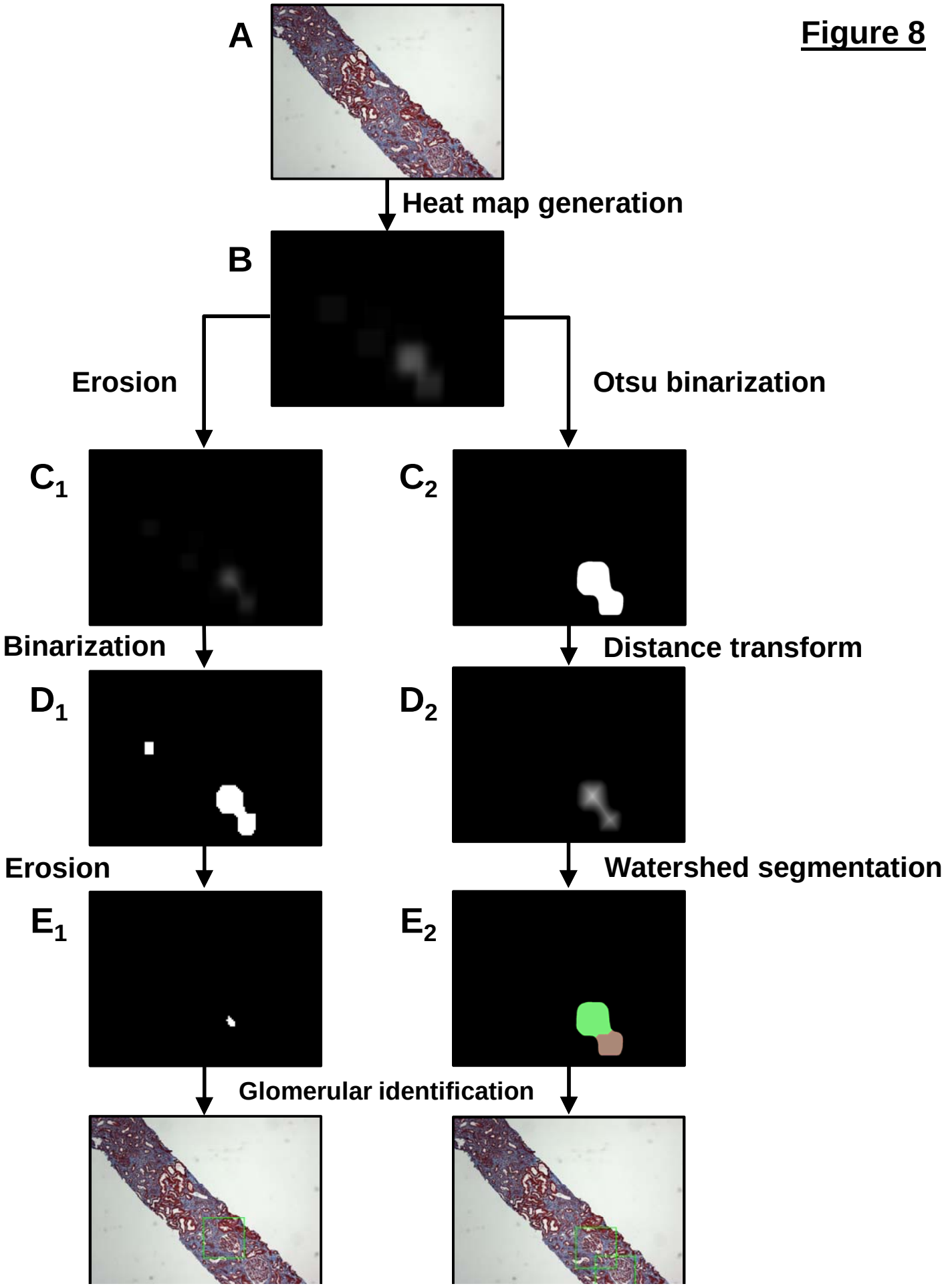


Figure 8



Supplemental Table 1

Dataset	Whitening	Aug.	Train Images (Total)	Test Images (Total)	Model
Round 1					
Case 11	No	No	1063	439	Model 11
Case 12	No	Yes	1063	439	Model 12
Case 13	Yes	No	6534	413	Model 13
Case 14	Yes	Yes	6534	413	Model 14
Round 2					
Case 21	No	No	1040	462	Model 21
Case 22	No	Yes	1040	462	Model 22
Case 23	Yes	No	6192	470	Model 23
Case 24	Yes	Yes	6192	470	Model 24
Round 3					
Case 31	No	No	1044	458	Model 31
Case 32	No	Yes	1044	458	Model 32
Case 33	Yes	No	6204	468	Model 33
Case 34	Yes	Yes	6204	468	Model 34