Journal of Pathology Informatics 15 (2024) 100385

Contents lists available at ScienceDirect

Journal of Pathology Informatics

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

Short Communication

Smartphone-based machine learning model for real-time assessment of medical kidney biopsy

Odianosen J. Eigbire-Molen [a,*], Clarissa A. Cassol [a], Daniel J. Kenan [a], Johnathan O.H. Napier [a], Lyle J. Burdine [b], Shana M. Coley [a], Shree G. Sharma [a]

[a] Arkana Laboratories, 10810 Executive Center Dr. Suite 100, Little Rock, AR 72211, USA
[b] Department of Surgery, University of Arkansas for Medical Sciences, Little Rock, AR 72205, USA

ARTICLE INFO

Keywords:
Kidney biopsy
Deep learning
Smartphone imaging
Adequacy assessment

ABSTRACT

Background: Kidney biopsy is the gold-standard for diagnosing medical renal diseases, but the accuracy of the diagnosis greatly depends on the quality of the biopsy specimen, particularly the amount of renal cortex obtained. Inadequate biopsies, characterized by insufficient cortex or predominant medulla, can lead to inconclusive or incorrect diagnoses, and repeat biopsy. Unfortunately, there has been a concerning increase in the rate of inadequate kidney biopsies, and not all medical centers have access to trained professionals who can assess biopsy adequacy in real time. In response to this challenge, we aimed to develop a machine learning model capable of assessing the percentage cortex of each biopsy pass using smartphone images of the kidney biopsy tissue at the time of biopsy.

Methods: 747 kidney biopsy cores and corresponding smartphone macro images were collected from five unused deceased donor kidneys. Each core was imaged, formalin-fixed, sectioned, and stained with Periodic acid–Schiff (PAS) to determine cortex percentage. The fresh unfixed core images were captured using the macro camera on an iPhone 13 Pro. Two experienced renal pathologists independently reviewed the PAS-stained sections to determine the cortex percentage. For the purpose of this study, the biopsies with less than 30% cortex were labeled as inadequate, while those with 30% or more cortex were classified as adequate. The dataset was divided into training ($n=643$), validation ($n=30$), and test ($n=74$) sets. Preprocessing steps involved converting High-Efficiency Image Container iPhone format images to JPEG, normalization, and renal tissue segmentation using a U-Net deep learning model. Subsequently, a classification deep learning model was trained on the renal tissue region of interest and corresponding class label.

Results: The deep learning model achieved an accuracy of 85% on the training data. On the independent test dataset, the model exhibited an accuracy of 81%. For inadequate samples in the test dataset, the model showed a sensitivity of 71%, suggesting its capability to identify cases with inadequate cortical representation. The area under the receiver-operating curve (AUC-ROC) on the test dataset was 0.80.

Conclusion: We successfully developed and tested a machine learning model for classifying smartphone images of kidney biopsies as either adequate or inadequate, based on the amount of cortex determined by expert renal pathologists. The model's promising results suggest its potential as a smartphone application to assist real-time assessment of kidney biopsy tissue, particularly in settings with limited access to trained personnel. Further refinements and validations are warranted to optimize the model's performance.

Introduction

Kidney biopsy plays an essential role in the diagnosis and management of various medical renal diseases, providing critical information for treatment decisions and prognostication. The histopathological evaluation of renal tissue allows the identification of specific renal pathologies, such as glomerulonephritis and tubulointerstitial diseases, which guide appropriate therapeutic interventions[1,2]. One of the key determinants of biopsy quality is the representation of renal cortex within the biopsy core. Adequate cortex representation ensures a comprehensive assessment of glomeruli, tubules, interstitium, and blood vessels, while insufficient cortex may limit the pathologist's ability to arrive at a precise diagnosis[2,3] or preclude the estimate of chronic changes that may inform prognosis and help guide clinical management.

Over the last 15 years, the rate of kidney biopsies that are inadequate for complete diagnosis has significantly increased, representing a significant financial drain on the healthcare system as well as a direct risk to the patients who must be re-biopsied or go without an accurate diagnosis[3,4]. Ideally,

[*] Corresponding author.
E-mail address: odia.molen@arkanalabs.com (O.J. Eigbire-Molen).

http://dx.doi.org/10.1016/j.jpi.2024.100385
Received 25 March 2024; Received in revised form 8 May 2024; Accepted 27 May 2024
Available online 31 May 2024

sample adequacy should be assessed at the time of biopsy by examination of the obtained cores for presence of cortex under a microscope by a pathologist or trained personnel. Such real-time assessment enables the biopsy physician to obtain additional tissue cores if existing cores contain insufficient amounts of cortex. However, resource constraints have markedly limited the availability of real-time onsite adequacy assessment within many biopsy suites. When available, onsite assessment is often performed by a general pathologist, or pathology assistant, who might not have experience in cortex identification from fresh renal biopsies as nephropathologists. A recent study of 123,372 native kidney biopsies found that the miss rate increased markedly from 2% in 2005 to 14% in 2020[3]. This increase in kidney biopsy miss rate highlights the need for improved tools to assist assessment of the tissue collected before the patient leaves the biopsy suite.

Recent advances in artificial intelligence and machine learning have demonstrated their potential to assist pathologists in various diagnostic tasks, including lesion detection and classification[5,6]. Machine learning models, particularly deep learning algorithms, have shown promising results in automating image analysis, reducing inter-observer variability, and improving diagnostic accuracy across various medical specialties[7,8].

In this study, we sought to address the challenge of inadequate kidney biopsies by developing a machine learning model capable of identifying renal cortex from smartphone images of kidney needle-core biopsy tissue. The model can be deployed as a smartphone application to assist the biopsy clinician by providing information on the percentage cortex obtained on each biopsy pass. This will enable the biopsy clinician to make an informed decision on the number of passes required during percutaneous renal biopsy procedures, potentially improving the quality and accuracy of renal biopsy evaluations, especially in the settings with limited access to trained personnel.

Materials and methods

Patient selection and data acquisition

For this study, we prospectively collected 747 needle-core biopsies from five fresh unused deceased donor kidneys using a Bard Monopty 16-gauge biopsy gun under direct visualization. The utilization of deceased donor kidneys allowed us to obtain a standardized sample set with minimal confounding factors. Each biopsy core was imaged using the macro camera on an iPhone 13 Pro. The phone was handheld close to the biopsy core at approximately 2 in. to obtain an image in clear focus. The images were captured with ambient room lighting. No external lighting or photographic enhancement was used. The biopsy cores were photographed on standard lab absorbing paper. Then, each core was processed according to routine histological procedures including formalin fixation, paraffin embedding, sectioning, and staining with Periodic acid–Schiff (PAS) to facilitate the determination of cortex percentage. All computational analysis and implementation of the machine learning models were performed with R Statistical Software, “R version 4.2.1 (2022-06-23 ucrt)”, using Keras 2.11.0 and TensorFlow version 2.9.3[9–11].

Determination of cortex percentage

To determine the ground truth of percentage cortex in each biopsy core, two experienced renal pathologists independently reviewed the PAS-stained sections. Each pathologist analyzed the PAS image separately and provided their estimate of percentage of cortex. After that, the scores were aggregated. Any significant disagreement in scoring was resolved by a third pathologist. For the purpose of this study, biopsies with less than 30% cortex were labeled as inadequate, while those with 30% or more cortex were labeled as adequate. We selected a threshold of 30% based on logistic regression analysis of the glomeruli distribution in the 747 biopsy cores, which showed that cores with approximately 40% cortex had a 50% probability of containing more than 10 glomeruli (Supplementary Fig. 1). The 30% threshold was selected as it provides a 10% buffer to increase sensitivity for cores with insufficient cortex. During the biopsy procedure, the overall percentage of cortex obtained from all the passes determines the biopsy's adequacy. Therefore, this approach should assist the physician performing the biopsy in obtaining a satisfactory tissue sample from each pass to make sure the entire biopsy sample is adequate for the diagnosis.

Data split and preprocessing

The dataset, consisting of smartphone images and corresponding class labels (adequate/inadequate), was randomly divided into three subsets: a training set of 643 samples, a validation set of 30 samples, and a test set of 74 samples. The training set was used to train the machine learning model, while the validation set was utilized for hyperparameter tuning and model selection. The test set remained untouched during model development/training and served as an independent dataset to evaluate the model's performance. The breakdown of classes (adequate/inadequate) within each subset is shown in Table 1. To balance the training dataset size, the number of images in the inadequate class was up sampled from 40 to 600 with further data augmentation performed during model training.

Before feeding the images into the machine learning model, preprocessing steps were performed to ensure data compatibility and optimize image analysis. First, the iPhone's High-Efficiency Image Container images were converted to the widely used JPEG format using the magick package in R[12] and resized to 320 × 320 pixels. The machine learning framework consisted of two models employed in series: a U-Net deep learning model to segment the renal tissue from the background[13], followed by a deep learning model to classify the renal tissue (Fig. 1). The U-Net renal tissue segmentation model was trained on a random subset of 114 images from the training dataset and corresponding masks of the renal tissue region of interest. The masks were created in ImageJ[14] by selecting the renal tissue and creating a binary mask. After training for 350 epochs, the U-Net segmentation model achieved a Dice coefficient of 0.99 and AUC-ROC of 0.99. Of note, the renal tissue pixels occupy a small percentage (1–3%) of the total image pixels, and the presence of background noise in the images, such as handwritten sample identification numbers, sharp background transitions, and fiber textures, posed additional challenges for accurate segmentation. Addressing these complexities required extended training epochs to ensure the segmentation robustness. The U-Net model's performance was evaluated by testing on the remaining 633 images by visual inspection. Each JPEG image was input to the U-Net segmentation model, generating a segmentation mask of the renal tissue. This region of interest mask was applied to its corresponding image at the original image resolution, which was then cropped to the mask boundaries, enabling the classification model to focus on the renal tissue present in the image.

Classification model

A deep neural network was constructed to classify the smartphone images as either adequate or inadequate. The model consists of:

1. The input layer: The model input is the renal tissue region of interest obtained from the U-Net segmentation, resized to 320 × 320 pixels.

2. Data augmentation layers: A random horizontal and vertical flip, a random contrast of 0.1, a random rotation of 0.3, pixel normalization to values between 0 and 1, and addition of up to 10% Gaussian noise was performed on each input image.

Table 1
Breakdown of the dataset with the number of samples in each group.

	Adequate class	Inadequate class
Training	603	40
Validation	15	15
Test	60	14
Total	678	69

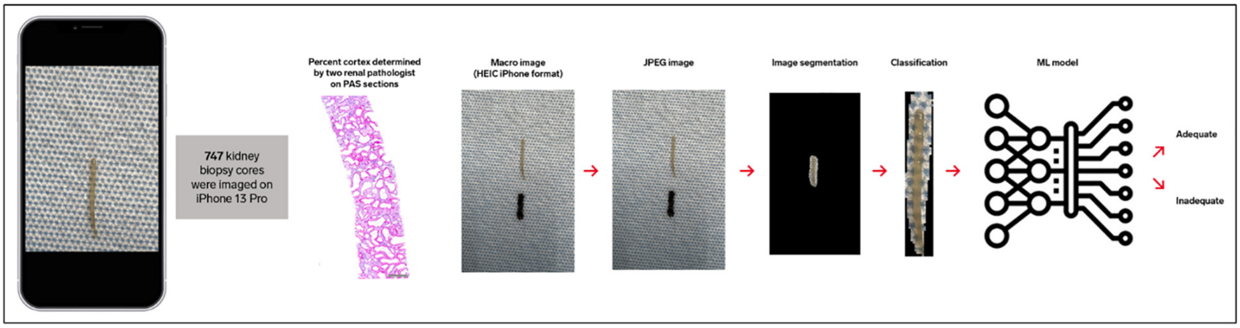
747 kidney biopsy cores were imaged on iPhone 13 Pro
Percent cortex determined by two renal pathologist on PAS sections
Macro image (HEIC iPhone format)
JPEG image
Image segmentation
Classification
ML model
Adequate
Inadequate

Fig. 1. Diagram of methodology indicating image input, data preprocessing and sequential model approach.

3. Feature extraction with attention mechanism: The image features were extracted using two fully connected layers, dense (1) and dense (32), separated by a flatten layer and a dropout layer (0.3). An attention mechanism was applied to the extracted features using a tanh activated layer and softmax weights.

4. A classification head: This consisted of two fully connected layers, dense (64) and dense (32), separated by a dropout layer (0.3). The output layer resulted the binary probability using a sigmoid activation function.

A schematic of the model architecture is shown in Fig. 2.

The classification model had 3,282,117 trainable parameters. Training was performed using root mean square propagation (rmsprop) as the optimization algorithm, at a learning rate of 0.001. The model's hyperparameters (layer unit size, batch size, learning rate, and dropout) were fine-tuned on the validation dataset to optimize the model's performance. The training process was iterated over 100 epochs with a batch size of eight. Using model checkpoint and callback, the best model was selected based on the best-combined training and validation accuracy.

Evaluation metrics

Multiple metrics were employed to evaluate the model's performance: accuracy, sensitivity, specificity, and the area under the receiver-operating curve (AUC-ROC). Accuracy represents the proportion of correctly classified samples out of the total number of samples, while sensitivity reflects the model's ability to correctly identify inadequate samples among all inadequate samples. Specificity, on the other hand, denotes the model's ability to correctly identify adequate samples among all adequate samples. AUC-ROC provides a comprehensive measure of the model's discrimination power, with higher values indicating better performance in distinguishing between the two classes.

Results

The developed deep learning model demonstrated an accuracy of 85% and AUC-ROC of 0.92 on the training data. When evaluated on the independent test dataset, the model achieved an accuracy of 81% and AUC-ROC of 0.80. For the subset of inadequate biopsy samples in the test dataset, the model exhibited a sensitivity of 71%. This result implies that the model could correctly identify 71% of inadequate biopsy samples, providing valuable support in flagging cases requiring further evaluation. For the subset of adequate biopsy samples in the test dataset, the model had a specificity of 83%. This implies that the model's performance did not come at a significant cost of falsely rejecting adequate biopsy samples.

A simple android application was made to demonstrate how the model could be deployed on a smartphone (Supplementary Video 1).

Discussion

Our study presents a novel approach to address the issue of inadequate kidney biopsies by leveraging smartphone-based imaging and machine learning technology. The rise in inadequate kidney biopsy rates poses significant challenges to accurate diagnosis and patient management, making it imperative to develop innovative solutions to improve the quality of biopsy evaluations in real time, when it is still possible to obtain additional tissue, if needed. The utilization of smartphones as imaging tools offers several advantages, including ease of use, portability, and availability[4]. Smartphone-based imaging allows for real-time capture and processing of high-quality images, enabling rapid assessment of the tissue obtained. By leveraging machine learning algorithms, particularly deep learning models, we aimed to automate the classification of smartphone images as adequate or inadequate, based on the percentage of renal cortex determined by expert renal pathologists. The deep learning model uses only renal pathologist-reported percent cortex as ground-truth labels for training, thereby avoiding expensive and time-consuming pixel-wise manual annotations. Additionally, using expert renal pathologist ground-truth labels of the biopsy cores is a marked improvement from current practice, where assessments of fresh tissue are conducted by individuals who are either untrained or lack specific training in medical renal disease. The encouraging performance of our deep learning model, with an accuracy of 81% and a sensitivity of 71% for inadequate samples on the independent test dataset, indicates its potential utility for assisting real-time biopsy assessment. The deep learning model can be deployed as a smartphone application, utilizing the smartphone camera to obtain high-quality images of fresh biopsy tissue, and determining whether the tissue is likely to contain sufficient cortex for diagnosis, thereby informing the biopsy physician whether additional cores should be obtained and decreasing the patient's risk of re-biopsy. In settings with limited access to trained personnel, this technology could serve as a valuable tool to facilitate onsite assessment of kidney biopsies.

Although the model's performance is promising, certain limitations warrant consideration. First, the dataset used in this study was collected from

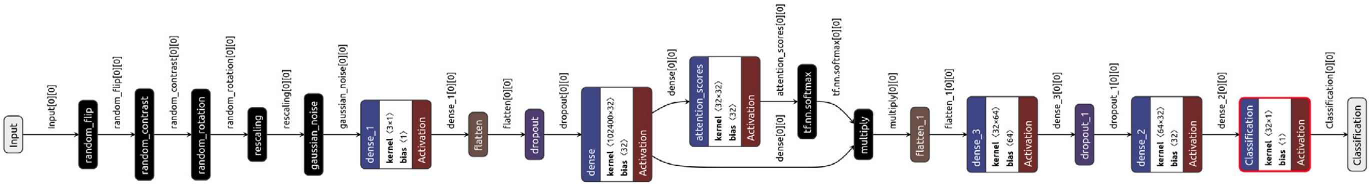

Fig. 2. Classification model visualized using Netron[15]. Input(320x320x3) → random flip → random contrast → random rotation → rescaling → Gaussian noise → Dense (1) → Dropout(0.3) → Dense(32) → Tanh(32)$_a$ → Softmax(32)$_b$ → Multiply(a,b) → Dense(64) → Dropout(0.3) → Dense(32) → Sigmoid (1).

five unused deceased donor kidneys, representing a specific subset of renal biopsies, which are different in appearance and imaging qualities from biopsies performed at the bedside. The model's generalization to biopsies from patients with various renal diseases needs to be further investigated. Second, a kidney biopsy core typically comprises renal cortex and non-cortex tissues, including renal medulla, perirenal fat, and fibrous capsule. Various visual characteristics such as color, size, and texture offer clues to distinguish renal cortex from non-cortex tissue. Renal cortex exhibits a smooth reddish appearance attributed to glomerular capillaries and other blood vessels, often revealing circular structures resembling glomeruli. Conversely, renal medulla appears pale with reddish streaks and texture. Perirenal fat and connective tissue tend to appear transparent, yellowish to brown-tan. Cores with ample cortical sampling are generally larger and more intact. In our study, biopsy cores were sourced from deceased donor kidneys, which lack perfusion, resulting in a brown-tan appearance for the renal cortex and white-tan for the renal medulla. These visual disparities between cortex and non-cortex tissues present potential targets for the classification deep learning model to differentiate cortex-rich and cortex-poor biopsy cores. Nevertheless, further investigations are warranted to elucidate how different regions of the biopsy core image influence the model's class output. Lastly, the dataset sample size is small with significant class imbalance. We mitigated the effects of class imbalance by upsampling the minority class and adding multiple data augmentation operations[16]. A small model size and utilization of dropout layers also helped reduce overfitting[17]. However, we acknowledge the potential limitations on generalizability due to the small dataset of this study. Additionally, the study was limited to a single center, and the model's performance may vary in different clinical settings with varying biopsy practices and sample preparation techniques. Multi-center validations on diverse patient populations would be important to assess the model's robustness and reliability in different contexts. The model's performance may be influenced by image quality variations, such as lighting conditions and imaging artifacts. Further improvements in preprocessing techniques and the incorporation of advanced image enhancement methods could enhance the model's resilience to such variations. Finally, refinement of the model's accuracy and ensuring compatibility with different smartphone models and operating systems are needed. In future work, we plan to collaborate with other institutions with hopes of addressing these issues.

Conclusion

This study demonstrates the successful development of a machine learning model capable of classifying smartphone images of kidney needle-core biopsy tissue, based on the percentage cortex determined by expert renal pathologists. The model's promising performance suggests its potential utility as a smartphone application for assisting real-time assessment of renal biopsy tissue, particularly in settings where access to expert onsite assessment is limited. Further refinements and validations are warranted to optimize the model's performance and facilitate its integration into routine clinical practice, ultimately benefiting patients. By providing the biopsy physician with additional information on the amount of cortex in each biopsy pass, this innovative approach could reduce inter-observer variability and improve kidney biopsy yield.

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jpi.2024.100385.

Funding

This work was supported by the NIH SBIR phase 1 grant (R43-DK134273).

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Shree Sharma reports financial support was provided by National Institutes of Health. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors would like to thank Chassity Jordan for creating the electronic artwork.

References

1. Colvin RB, Chang A, Dressler R, et al. Kidney biopsy as a guide to therapeutic decisions in renal disease. Kidney Int 2014;85(2):246–252.
2. Walker PD. The renal biopsy. Arch Pathol Lab Med 2009;133(2):181–188.
3. Nissen CJ, Moreno V, Davis VG, Walker PD. Increasing incidence of inadequate kidney biopsy samples over time: a 16-year retrospective analysis from a large national renal biopsy laboratory. Kidney Int Rep 2021 Dec 8;7(2):251–258.
4. Singh G, Massak M, Czaplicki M, et al. Use of a smartphone camera at the bedside to assess adequacy of kidney biopsies. J Am Soc Nephrol 2021;32(12):3024.
5. Zheng Y, Cassol CA, Jung S, et al. Deep-learning–driven quantification of interstitial fibrosis in digitized kidney biopsies. Am J Pathol 2021;191(8):1442–1453.
6. Ba W, Wang S, Shang M, et al. Assessment of deep learning assistance for the pathological diagnosis of gastric cancer. Mod Pathol 2022;35(9):1262–1268.
7. Campanella G, Hanna MG, Geneslaw L, et al. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. Nat Med 2019;25(8): 1301–1309.
8. Xiong Z, He J, Valkema P, et al. Advances in kidney biopsy structural assessment through dense instance segmentation. arXiv:230917166 [cs]. Published online September 29, 2023.
9. R Core Team. *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing. 20, 22, https://www.R-project.org/.(Accessed 2/20/2024).
10. Kalinowski T, Falbel D, Allaire JJ, et al. R Interface to Keras. R package vers ion 2.11.0. https://CRAN.R-project.org/package=keras 2023.
11. Kalinowski T, Falbel D, Allaire JJ, et al. R Interface to TensorFlow. R package vers ion 2.9.3. https://CRAN.R-project.org/package=tensorflow 2023.
12. Ooms J. magick: Advanced Graphics and Image-Process ing in R. R package version 2.7.5. https://CRAN.R-project.org/package=magick 2023.
13. Ronneberger O, Fischer P, Brox T. U-Net: convolutional networks for biomedical image segmentation. In: *Navab N, Hornegger J, Wells WM, Frangi AF, eds. Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015.* Lecture Notes in Computer ScienceCham: Springer; 2015. p. 234–241.
14. Schneider CA, Rasband WS, Eliceiri KW. NIH image to ImageJ: 25 years of image analysis. Nat Methods 2012;9(7):671–675.
15. Netron: Visualizer for neural network, deep learning and machine learning models. https://netron.app/. (Accessed: 2/16/2024).
16. Shorten C, Khoshgoftaar TM. A survey on image data augmentation for deep learning. J Big Data 2019;6(1):60.
17. Chollet F, Allaire JJ. *Deep Learning with R.* Manning Publications Co. 2018:95-102.