HHS Public Access

Author manuscript

Proc SPIE Int Soc Opt Eng. Author manuscript; available in PMC 2022 February 01.

Published in final edited form as:

Proc SPIE Int Soc Opt Eng. 2021 February ; 11603: . doi:10.1117/12.2581789.

A Distributed System Improves Inter-Observer and AI Concordance in Annotating Interstitial Fibrosis and Tubular Atrophy

Avinash Kammardi Shashiprakash[1], **Brendon Lutnick**[2], **Brandon Ginley**[2], **Darshana Govind**[2], **Nicholas Lucarelli**[1], **Kuang-Yu Jen**[3], **Avi Z Rosenberg**[4], **Anatoly Urisman**[5], **Vighnesh Walavalkar**[5], **Jonathan E. Zuckerman**[6], **Marco Delsante**[7], **Mei Lin Z. Bissonnette**[8], **John E Tomaszewski**[1], **David Manthey**[9], **Pinaki Sarder**[2,*]

[1]Department of Biomedical Engineering, University at Buffalo – The State University of New York

[2]Department of Pathology and Anatomical Sciences, University at Buffalo – The State University of New York

[3]Department of Pathology, University of California at Davis

[4]Department of Pathology, Johns Hopkins University School of Medicine

[5]Department of Pathology, University of California San Francisco

[6]Department of Pathology and Laboratory Medicine, University of California Los Angeles

[7]Department of Medicine and Surgery, University of Parma, Parma, Italy

[8]Department of Pathology and Laboratory Medicine, University of British Columbia, Vancouver, BC, Canada

[9]Kitware Incorporated, Clifton Park, New York

Abstract

Histologic examination of interstitial fibrosis and tubular atrophy (IFTA) is critical to determine the extent of irreversible kidney injury in renal disease. The current clinical standard involves pathologist's visual assessment of IFTA, which is prone to inter-observer variability. To address this diagnostic variability, we designed two case studies (CSs), including seven pathologists, using HistomicsTK- a distributed system developed by Kitware Inc. (Clifton Park, NY). Twenty-five whole slide images (WSIs) were classified into a training set of 21 and a validation set of four. The training set was composed of seven unique subsets, each provided to an individual pathologist along with four common WSIs from the validation set. In CS 1, all pathologists individually annotated IFTA in their respective slides. These annotations were then used to train a deep learning algorithm to computationally segment IFTA. In CS 2, manual and computational annotations from CS 1 were first reviewed by the annotators to improve concordance of IFTA annotation. Both the manual and computational annotation processes were then repeated as in CS1. The inter-observer concordance in the validation set was measured by Krippendorff's alpha (KA). The KA for the seven pathologists in CS1 was 0.62 with CI [0.57, 0.67], and after reviewing

*Address all correspondence to: Pinaki Sarder, Tel: 716-829-2265; pinakisa@buffalo.edu.

each other's annotations in CS2, 0.66 with CI [0.60, 0.72]. The respective CS1 and CS2 KA were 0.58 with CI [0.52, 0.64] and 0.63 with CI [0.56, 0.69] when including the deep learner as an eighth annotator. These results suggest that our designed annotation framework refines agreement of spatial annotation of IFTA and demonstrates a human-AI approach to significantly improve the development of computational models.

I. INTRODUCTION

Histopathological assessment of interstitial fibrosis and tubular atrophy (IFTA) is important to understand the severity of chronic kidney disease related to kidney transplant. Pathologists' assessment based on visual analysis for the disease progression has a wide inter-observer variability due to the complex structure of IFTA[1]. To address this issue, we have come up with a two-step consensus process for a distributed study of annotation on IFTA. We employed HistomicsTK[2], a distributed method to achieve individual or collaborative annotation of IFTA on a whole slide image (WSI). Furthermore, we employed an iterative Human A.I Loop[3] (HAIL) pipeline, a machine learning tool for semantic segmentation to computationally annotate IFTA. This model was trained using pathologists' annotated ground truth, and compared the resulting prediction with pathologists' annotations.

II. RESULT

Distributed study using HistomicsTK and H-AI-L:

The goals of this study were 1) to learn the variability among pathologists from IFTA annotations, as IFTA does not have precise structure and definite boundary, and 2) to computationally segment IFTA using seven renal pathologists' annotations and HAIL. We used HistomicsTK[2], an online platform for distributed annotation, for the IFTA labeling by seven pathologists and used their annotated ground-truths to train HAIL. We used 25 WSIs, which were divided into 21 WSIs as a training set and four WSIs as a validation set. Fig. 1 shows the workflow. Each pathologist was assigned seven WSIs, of which unique WSIs were from the training set and four WSIs were from the validation set. Using HistomicsTK, pathologists annotated IFTA, using polygon and/or line tool (Fig. 1A). Next, all the marked IFTA regions were collected from each pathologist and a training set of 21 WSIs (three WSIs from each pathologist) (Fig. 1B) were used to train HAIL (Fig. 1C). The trained network was then tested for identifying IFTA on the validation set of four WSIs (Fig. 1D). For evaluation, annotations of pathologists and HAIL were compared, and the result was visualized using HistomicsTK. Unique color was assigned to each pathologist and network predicted annotations on the validation set (Fig. 1E).

Comparison between two studies for concordance analysis:

The Krippendorff's alpha[4] (KA), measuring the joint probability of agreement among the seven annotators and HAIL, was used to measure the overall inter-rater concordance for labeling IFTA. Calculation was done on the validation set among the seven annotators and HAIL as the eighth annotator. Two case studies were designed for the annotation. In Case Study (CS) 1, the deep learner was trained using annotations from all pathologists. In CS 2,

manual and computational annotations from CS 1 were first reviewed by the annotators to improve the concordance of annotation. KA showed improved concordance among the pathologists and HAIL in CS 2. Namely, the KA for the seven pathologists in CS1 was 0.62 with CI [0.57, 0.67], and after reviewing each other's annotations in CS2, 0.66 with CI [0.60, 0.72]. The respective CS1 and CS2 KA were 0.58 with CI [0.52, 0.64] and 0.63 with CI [0.56, 0.69] when including HAIL as an eighth annotator. To quantify the concordance between each annotator and the deep learner, we computed Cohen's kappa[5] measure for both case studies on the validation set (Tables 1 & 2). Namely, for each annotator and deep learner (DL), we compute Cohen's kappa measure with respect to each of the other annotators, and then compute average of the kappa measures for each annotator.

Notably, inter-annotator agreement measure was improved substantially in CS 2 compared to CS 1. Also, we found Annotator 3 and 6 converged with other annotators in CS 2 compared to CS 1.

III. METHOD

Image Acquisition:

The image data consisted of 25 WSIs from human transplant biopsies (21 training, 4 validation). Transplant cases were selected from the specimen archives to represent a spectrum of cortical IFTA amount that is typically encountered in clinic. The tissues sections were prepared at 2–3 μm thickness and stained with periodic acid-Schiff (PAS). Slides were scanned using a brightfield whole slide microscopy scanner at 40x magnification (0.25μm/pixel).

IFTA annotation by pathologist:

Seven renal pathologists annotated IFTA for this study.

Distributed annotation system – HistomicsTK:

HistomicsTK (Fig. 2A) is a web-based application with RESTful API developed by Kitware Inc. (Clifton Park, NY) for visualization and image analysis. This tool is supported by the OpenSlide library for handling proprietary digital pathology WSI formats. WSIs visualized in the browser are presented as a collection of images with different resolutions stored in a pyramid form using image compression.

HistomicsTK provides basic operations, namely, zoom and pan, and annotation tools, namely, point, line, and polygon. Annotations drawn on WSI can be organized in different layers based on classes. Additionally, HistomicsTK allows users to build a user-defined algorithm and/or plugin for seamless integration for dedicated image analysis task.

HistomicsTK provides an inbuilt data management system, called as Digital Slide Archive[6] (DSA) (Fig. 2B). Using DSA, we can create a user account and store WSI images at different levels of folder structure. It can also hold image metadata like pathology report, clinical data, and other associated files and even annotations marked by pathologists at different levels for a particular WSI. DSA allows to set the access level (permissions for each viewer) to WSI or the folder containing WSIs to be public or private.

Distributed study of pathological annotation on IFTA:

For our study, pathologists' were guided to annotate IFTA regions using polygon and line tools available in HistomicsTK. The definition of IFTA was followed based on definition provided in the work by Candice *et al*[7]. The 25 WSIs were divided into a training set of 21 and a validation set of four. Data from the validation set (four WSIs) were reviewed independently by all the pathologists, the remaining 21 WSIs from the training set were equally divided among the pathologists (3 each, Fig. 3). The study was established through two case studies (CS), and in each CS we made use of our developed HAIL pipeline for semantic segmentation. A deeplab V2 network with ResNet-50 encoder was used to train on the 21 WSIs from each case study. CNNs were configured for high resolution (40X magnification), learning rate of $2.5e^{-5}$ and batch size of 2 for 10 epochs. The validation set was used to test the model in both cases. The two case studies are described below.

a. *Individual annotations (CS 1):* Each pathologist was given seven WSIs and their annotations were held private, not disclosed with other raters. Performance metrics were computed for the validation set using all the seven pathologists' and HAIL annotated IFTA. We used Cohen's kappa[5] statistic to compare the concurrency between every pair of annotators as well as HAIL. Krippendorff's alpha[4] measure provides joint probability agreement among annotators and HAIL.

b. *Collaborative annotations (CS 2):* From CS 1, annotations from each pathologist and HAIL on the common training set were made available to all pathologists to review with a goal to improve their annotation concordance. To avoid bias, the validation set from the previous study was not shared. To test improvement in concordance between annotators, the same manual and computational annotation process was repeated for the originally assigned data set. The same performance metrics were used to compare the concurrency among pathologists and HAIL prediction. The performance for CS 1 ad CS2 were compared (Table 1 & 2).

IV. CONCLUSION

The problem of inter-observer variability on WSIs quantification can be refined by facilitating consensus using distributed systems like HistomicsTK. Our study suggests a substantial agreement among the pathologists in the identification of IFTA regions is feasible via distributed HistomicsTK annotation, which in turn improves the prediction of IFTA using deep learning algorithms.

REFERENCES

1. Farris AB and Colvin RB, Renal interstitial fibrosis: mechanisms and evaluation in: current opinion in nephrology and hypertension. Current opinion in nephrology and hypertension, 2012. 21(3): p. 289. [PubMed: 22449945]
2. Gutman DA, et al., The digital slide archive: A software platform for management, integration, and analysis of histology for cancer research. Cancer research, 2017. 77(21): p. e75–e78. [PubMed: 29092945]
3. Lutnick B, et al., Iterative annotation to ease neural network training: Specialized machine learning in medical image analysis. arXiv preprint arXiv:1812.07509, 2018.

4. Van Bockstal M, et al., Dichotomous histopathological assessment of ductal carcinoma in situ of the breast results in substantial interobserver concordance. Histopathology, 2018. 73(6): p. 923–932. [PubMed: 30168167]
5. Wei JW, et al., Pathologist-level classification of histologic patterns on resected lung adenocarcinoma slides with deep neural networks. Scientific reports, 2019. 9(1): p. 1–8. [PubMed: 30626917]
6. Amgad M, et al., Structured crowdsourcing enables convolutional segmentation of histology images. Bioinformatics, 2019. 35(18): p. 3461–3467. [PubMed: 30726865]
7. Roufosse C, et al., A 2018 reference guide to the Banff classification of renal allograft pathology. Transplantation, 2018. 102(11): p. 1795–1814. [PubMed: 30028786]

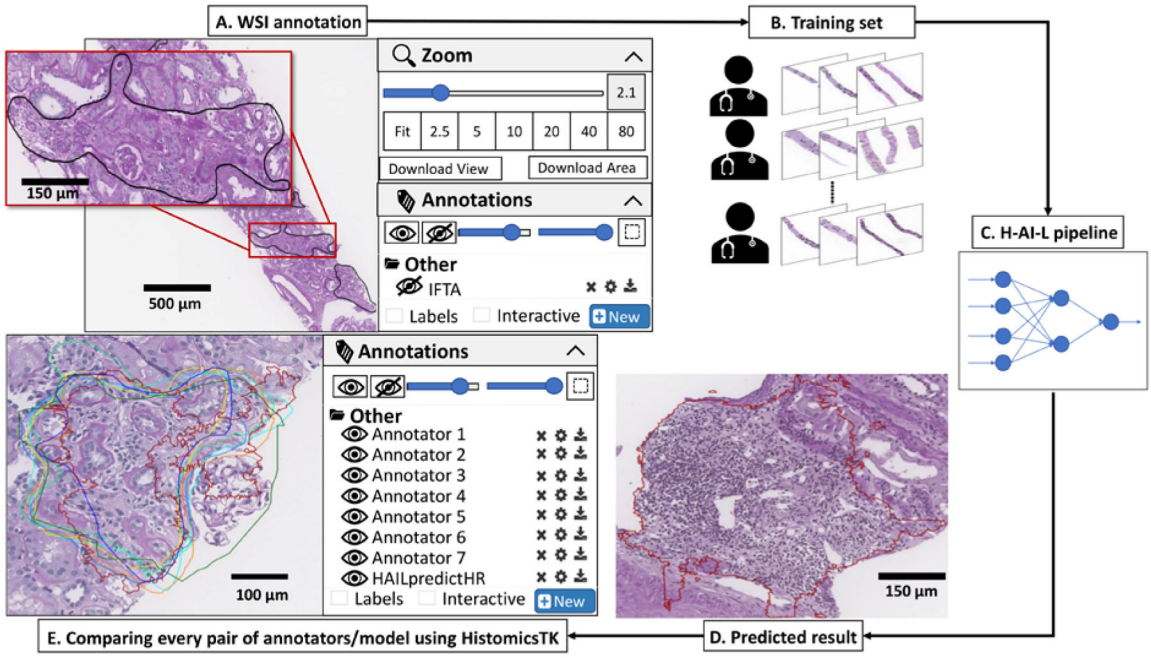

Figure 1. Overview of the distributed system:

(A) Using HistomicsTK to annotate IFTA via pathologists. **(B)** WSIs and associated annotations obtained from the pathologists. **(C)** Training set of WSIs and annotations were used to train HAIL. **(D)** Validation image and the corresponding predicted region. **(E)** Comparing validation image annotations from seven pathologists and HAIL by assigning different label colors in HistomicsTK.

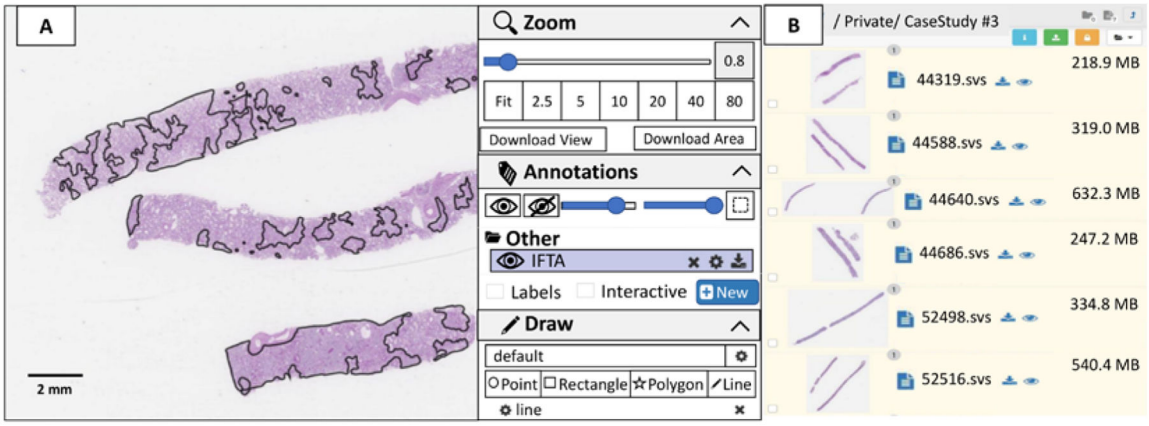

Figure 2. HistomicsTK and Digital Slide Archive (DSA) workflow.

(A) HistomicsTK viewer with basic operations, such as zoom, pan, and annotation. **(B)** Digital Slide Archive (DSA) interface which stores WSIs, annotation labels and meta data.

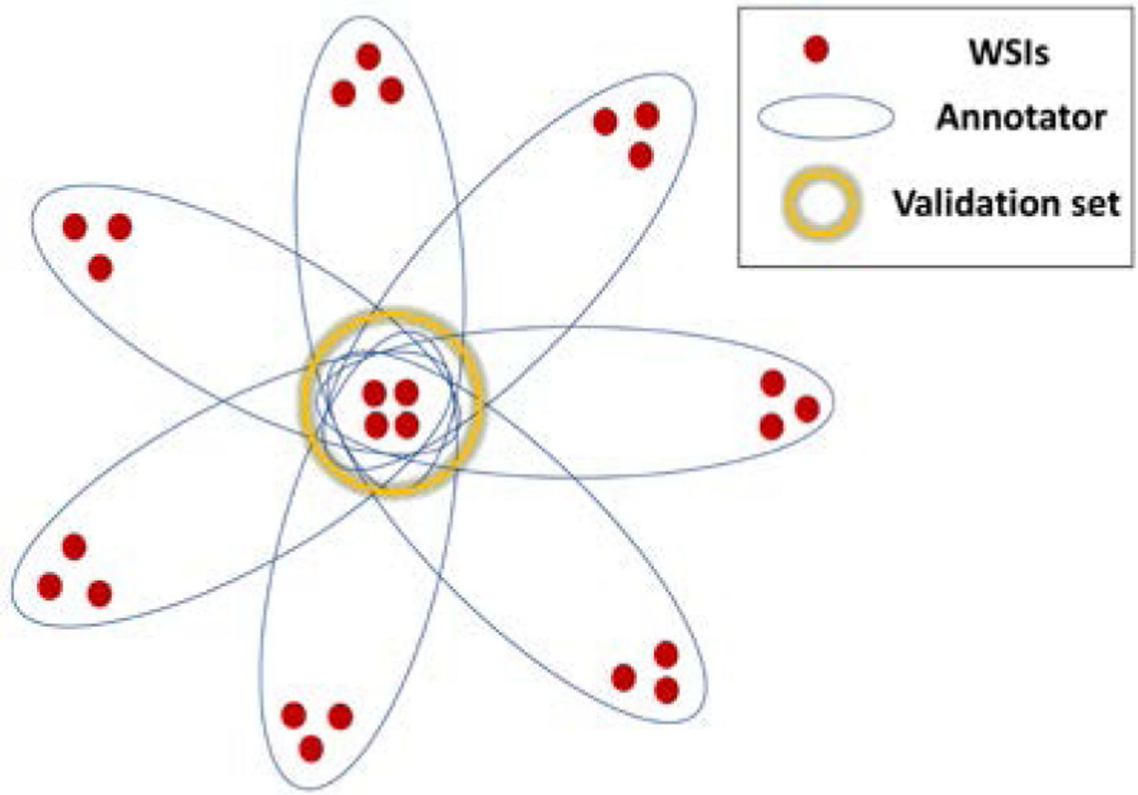

Figure 3. Overview of the project design to annotate IFTA on WSIs.

WSIs are allocated along different annotators to label IFTA. Inside yellow circle, WSIs are considered for validation set and rest of the WSIs are considered for training set for training HAIL.

Table 1.

Cohen’s kappa (κ) for CS 1.

Annotator	Others	
	Average (μ)	SD (σ)
A1	0.69	0.070
A2	0.66	0.065
A3	0.61	0.061
A4	0.67	0.101
A5	0.65	0.096
A6	0.56	0.073
A7	0.64	0.100
DL	0.53	0.039

Abbreviation: A-Annotator, DL-Deep learner

Table 2.

Cohen’s kappa (κ) for CS 2.

Annotator	Others	
	Average (μ)	SD (σ)
A1	0.69	0.075
A2	0.66	0.049
A3	0.68	0.044
A4	0.70	0.089
A5	0.69	0.076
A6	0.69	0.083
A7	0.65	0.068
DL	0.54	0.028

Abbreviation: A-Annotator, DL-Deep learner