PROCEEDINGS OF SPIE

SPIEDigitalLibrary.org/conference-proceedings-of-spie

A cloud-based tool for federated segmentation of whole slide images

Brendon Lutnick, David Manthey, Jan Becker, Jonathan Zuckerman, Luis Rodrigues, et al.

Brendon Lutnick, David Manthey, Jan U. Becker, Jonathan E. Zuckerman, Luis Rodrigues, Kuang Yu Jen, Pinaki Sarder, "A cloud-based tool for federated segmentation of whole slide images," Proc. SPIE 12039, Medical Imaging 2022: Digital and Computational Pathology, 120391J (4 April 2022); doi: 10.1117/12.2613502

SPIE.

Event: SPIE Medical Imaging, 2022, San Diego, California, United States

A cloud-based tool for federated segmentation of whole slide images.

Brendon Lutnick[1], David Manthey[2], Jan U. Becker[3], Jonathan E. Zuckerman[4], Luis Rodrigues[5], Kuang Yu. Jen[6], and Pinaki Sarder[1,*]

[1]Department of Pathology and Anatomical Sciences, SUNY Buffalo, NY,
[2]Kitware Incorporated, Clifton Park, NY,
[3]Institute of Pathology, University Hospital Cologne, Germany,
[4]Department of Pathology and Laboratory Medicine, University of California at Los Angeles, CA,
[5]University Clinic of Nephrology, Faculty of Medicine, University of Coimbra, Portugal,
[6]Department of Pathology and Laboratory Medicine, University of California at Davis, CA.

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

ABSTRACT

It is commonly known that diverse datasets of WSIs are beneficial when training convolutional neural networks, however sharing medical data between institutions is often hindered by regulatory concerns. We have developed a cloud-based tool for federated WSI segmentation, allowing collaboration between institutions without the need to directly share data. To show the feasibility of federated learning on pathology data in the real world, We demonstrate this tool by segmenting IFTA from three institutions and show that keeping the three datasets separate does not hinder segmentation performance. This pipeline is deployed in the cloud for easy access for data viewing and annotation by each site's respective constituents.

Keywords: WSI segmentation, Federated learning, cloud based analysis, IFTA

I. INTRODUCTION

As the practice of digitizing histological slides has become common practice[1], the field of computational pathology has exploded. Modern image analysis technologies (such as deep learning[2]) are increasingly being applied to examine digitized whole slide images (WSIs). Training these networks is enhanced by access to diverse WSI datasets, as greater data variability is known to enhance model robustness[3]. For histological tissue, the institution where data is prepared often has a large effect on the quality and appearance of the tissue[4]. Practically this means gathering training data from multiple institutions. However sharing medical data across institutions can be complicated by regulatory challenges[5], limiting the scope of collaboration and therefore the generalizability of computational pathology tools.

Federated learning was recently proposed as an efficient solution for decentralized training of models without sharing data[6,7]. At the core of federated learning is federated averaging (FedAvg)[8], which is simply a weighted average of the network weights across training sites, performed at pre-selected intervals. FedAvg has been practically shown to achieve convergence in a reasonable amount of time with proper hyperparameter tuning[9]. Computational pathology datasets are a perfect candidate for

Medical Imaging 2022: Digital and Computational Pathology, edited by John E. Tomaszewski, Aaron D. Ward, Proc. of SPIE Vol. 12039, 120391J · © 2022 SPIE
1605-7422 · doi: 10.1117/12.2613502

federated learning where both file sizes of WSIs (gigapixels) and regulatory limitations hinder data sharing.

II. RESULTS

To show the feasibility of federated learning on pathology data in the real world, we have created a pipeline for federated segmentation on WSIs capable of deployment across multiple institutions. This pipeline is deployed in the cloud for easy access for data viewing and annotation by each site's respective constituents.

To test this system we designed an experiment for federated segmentation of interstitial fibrosis and tubular atrophy (IFTA) from renal biopsies. Three pathologists from different institutions provided 20, 48, and 22 PAS stained slides respectively. A holdout dataset was randomly selected by pooling 1/3rd of the slides from each institution (29 slides total). We trained 5 models using this dataset: The first model was trained across three federated servers, split by institution of origin. For a baseline performance, a second model was trained centrally using traditional gradient decent by pooling all the training data on a single server. Finally, to compare the performance in a data restricted setting, three additional models were trained using data from a single institution alone.

We note that IFTA boundaries are poorly defined, and subject to disagreement between pathologists[10], receiver operating characteristic (ROC) curves were used to better capture the performance characteristics of our trained models. These were generated by applying a varying threshold to the network logits for the prediction of IFTA regions. To measure performance, we calculate the area under the curve (AUC) which is a common metric for measuring performance when a ROC curve is available.

Testing these models on the holdout set, we observed that central training and federated training of the IFTA model performed similarly both with *AUC = 0.95*. Performance fell when testing the models trained using a single institutions data, giving *AUC = 0.92, 0.87, & 0.91* respectively. ROC plots of the performance of the five models is highlighted in Fig. 2a. An example of IFTA segmentation on a holdout slide using the federated model is shown in Fig. 2c. Here we use the network logits to display the predictions as a probabilistic heatmap which we believe is better for the display of structures with poorly defined boundaries such as IFTA.

A fourth pathologist from a different institution provided an additional 17 slides to be used as an independent testing dataset. When we applied the trained IFTA models to this independent set we observed a similar trend as the holdout set. Here the federated model performed best with *AUC = 0.90* and the central model also performed well with *AUC = 0.88*. Like the holdout set, performance of the models trained on a single institution was lower than federated or central models, with *AUC = 0.85, 0.81, & 0.84* respectively. ROC plots of the performance of the five models is highlighted in Fig. 2b.

III. METHODS

This work is heavily based upon our previously published Histo-Cloud tool[11], where we modified the DeepLab V3+ architecture[12] to work natively on WSIs and developed a series of plugins for running segmentation training and prediction in the cloud. This work was based on the Digital Slide Archive (DSA)[13] an open source slide viewer and repository developed by Kitware Inc.

Server coordination: In this pipeline, each site / institution has a worker node server with the DSA installed where training data is uploaded and annotated. A central (master) server manages the training cycle, uploading the global model parameters to each worker via the DSA REST API before requesting each run local training, an example schematic of this setup is depicted in Fig. 1. Training Jobs are submitted to each worker (training site) using the Histo-Cloud training plugin[11]. The training job scheduling is handled by the DSA internally using *slicer_cli_web*, which uses Celery[14] for task queue management, and RabbitMQ[15] as a message broker. The job status is monitored by the master server until completion. Upon job completion the master server requests and downloads the resultant saved local model parameters from each worker node. These parameters are averaged by the master server and the global model is updated accordingly. The next round of training is then initiated: the global model is uploaded to each worker and is trained further before being downloaded and averaged. If training fails on one of the participating workers, then it is excluded from the rest of the training round, but participates in future training rounds.

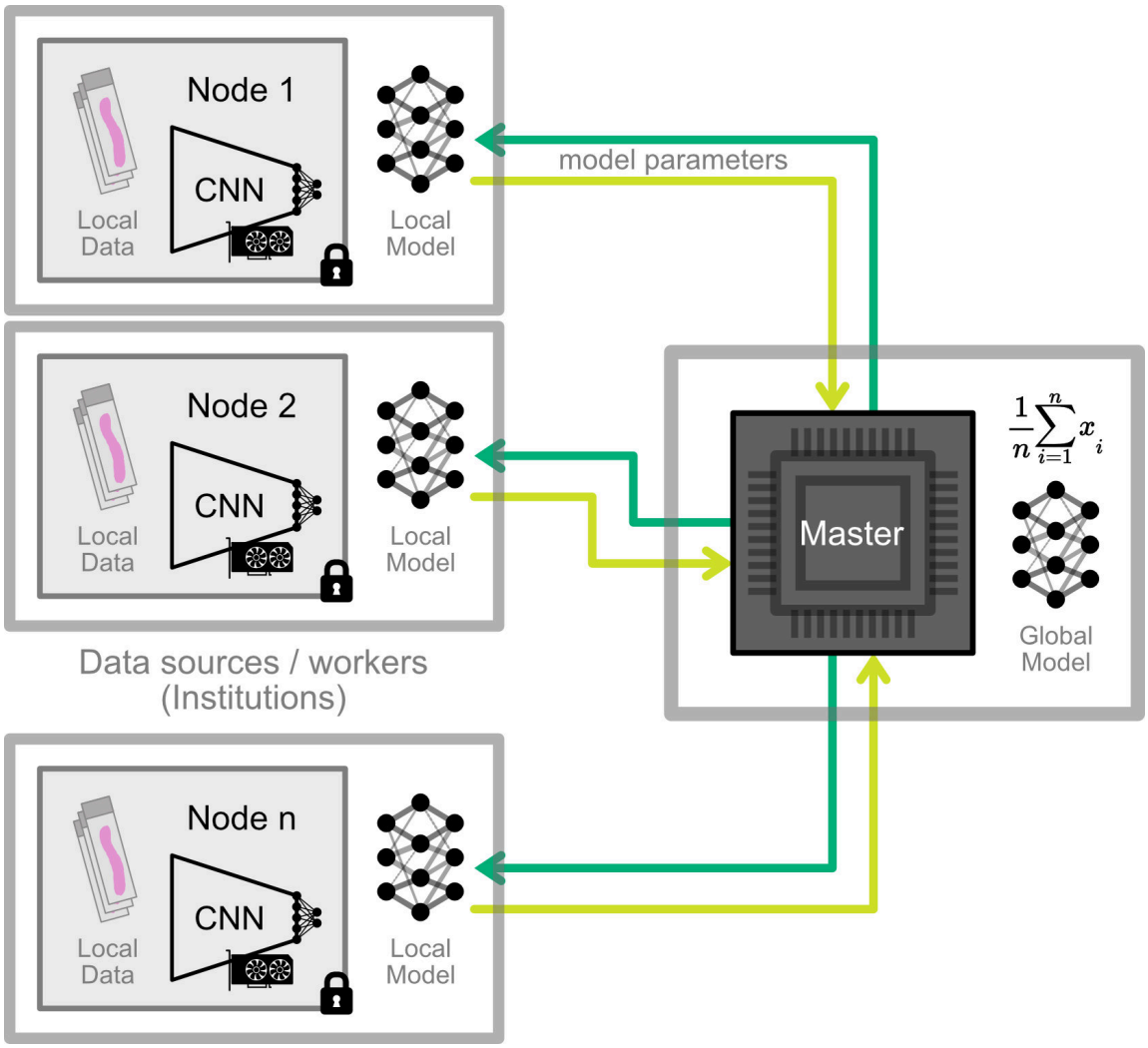

Fig. 1. Federated learning schematic. A schematic example of federated learning. Multiple worker nodes store data and model parameters locally at the institution of origin. The data stored on these worker nodes is never shared, and the nodes perform local training using this data upon the request of the master server. The local models are then shared with the

master server who performs parameter averaging, before sending the updated global model back to the worker nodes for further local training. This process is repeated iteratively throughout the training process, until model convergence.

Data management: The training WSI data is uploaded to the DSA worker servers, where it was annotated by expert pathologists. Training data is placed in a folder created on each worker for easy access by the Histo-Cloud training plugin. A separate folder was created for the models produced by training and uploaded after federated averaging. The ID of these folders is known by the master server so it can submit training jobs specifying the data and models to be used for training.

Training setup: For training we used three physically distinct Linux servers running Ubuntu 18.04.5 LTS, with the DSA installed. All computers had 2 GPUs that were produced by the Nvidia corporation and included:

1) Titan X Pascale (12GB VRAM) & GeForce GTX 1080 (8GB VRAM) – batch size 4
2) GeForce RTX 2080 Ti (11GB VRAM) & GeForce GTX 1080 (8GB VRAM) – batch size 4
3) 2X Quadro RTX 5000 (16 GB VRAM) – batch size 12

For training we used both available GPUs on each server and adjusted the batch size for each server to accommodate the individual VRAM (GPU memory) capacity of each.

Training hyperparameters: The goal of federated averaging is to speed up training by removing the overhead of frequent communication between training sites. This is done by training locally for multiple steps before updating the central model parameters using FedAvg. Practically when optimizing the hyperparameters of our training loop, we found that using 1000 training steps between FedAvg achieved repeatable convergence. We trained for a total of 40 rounds (40,000 steps), using the momentum optimizer[16] with an initial learning rate of $7e^{-3}$, using polynomial decay with a learning power of 0.9 and final learning rate of 0. To achieve stability at the start of training, we set the learning rate to $1e^{-4}$ for the first 750 steps. Finally, the gradients on the last layers of the network were scaled up by a factor of 10 to achieve faster convergence. These layers included the ASPP pooling layers and the layers in the decoder as defined by the DeepLab network architecture[12]. All the models were trained using transfer learning with parameters inherited from a model pre-trained on the ImageNet dataset.

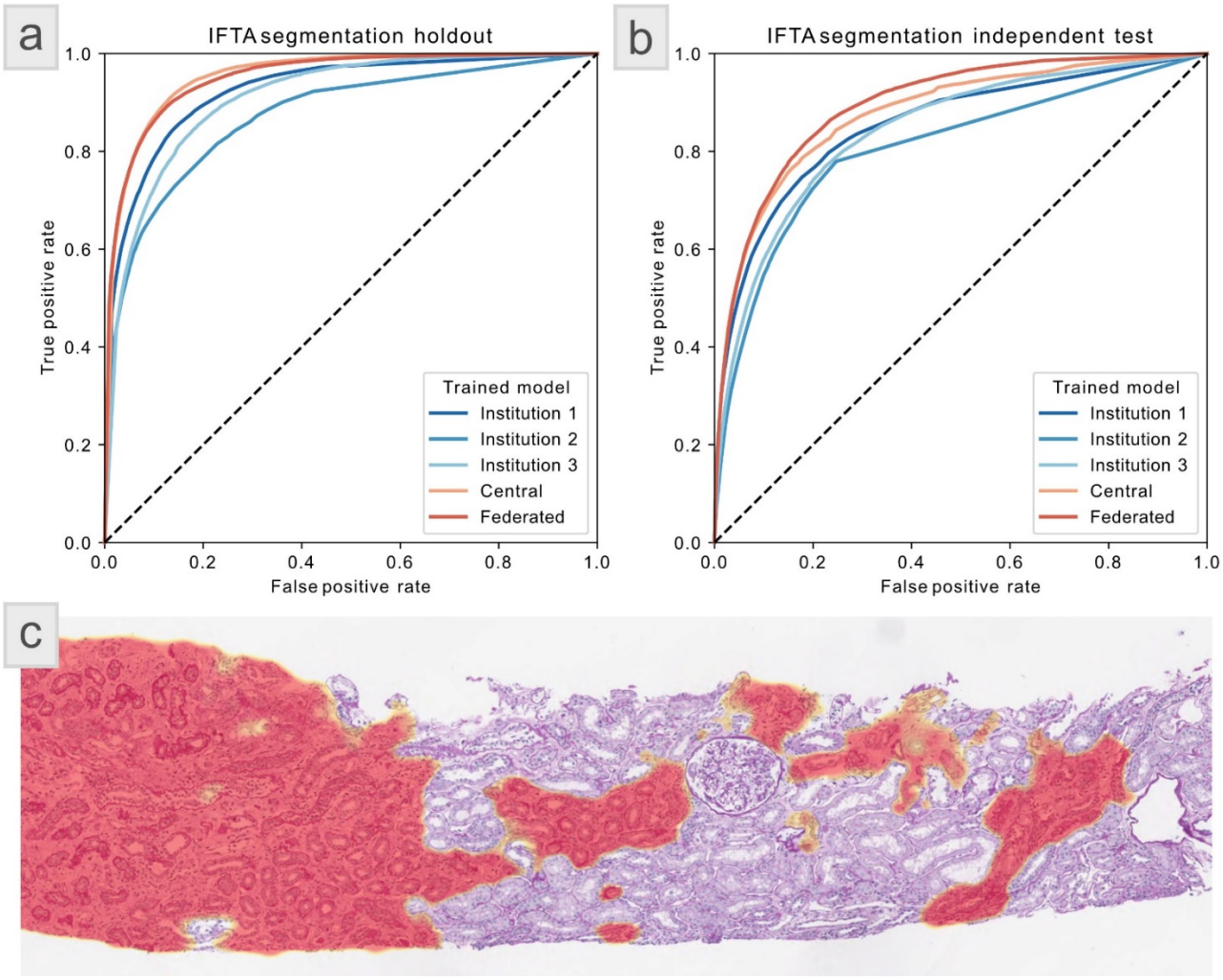

Fig. 2. Federated IFTA segmentation performance. [a] ROC curves showing each models performance on a dataset of 29 holdout WSIs which were randomly selected from the same data as the training set. We observed that central training and federated training of the IFTA model performed similarly both with *AUC = 0.95*. Performance fell when testing the models trained using a single institutions data, giving *AUC = 0.92, 0.87, & 0.91* respectively. **[b]** ROC curves showing each models performance on an independent test set of data containing 17 WSIs. This dataset was from an institution which did not provide any training data, and was annotated by an independent pathologist. Similar to the holdout set, the central and federated models outperformed the models trained on a single institution's data. Interestingly the federated model performed best with *AUC = 0.90* and the central model also performed well with *AUC = 0.88*. The institutions 1, 2, & 3 had *AUC = 0.85, 0.81, & 0.84* respectively. **[c]** an example of IFTA segmentation using the federated model on a slide from the holdout dataset. The prediction of IFTA is shown here using a heatmap, which reflects the network confidence in IFTA segmentation.

IV. CONCLUSION AND FUTURE WORK

The work presented in this write-up utilizes our recently developed Histo-Cloud tool[11] for segmentation of WSIs in the cloud. We leverage Histo-Cloud's deployment in the cloud to coordinate several instances of the tool for federated training of segmentation models on WSIs. Our experiment on IFTA segmentation shows that not only does federating training for WSI segmentation converge, but the resultant model outperforms training done with a single institutions data. Furthermore, the federated model performs on par with a model trained traditionally with multiple datasets gathered at a central location. Most importantly, these experiments demonstrate the feasibility of training and coordinating federated segmentation models, managing datasets distributed across physically separate servers, and training in reasonable time.

This experiment used the Digital Slide Archive (DSA) to handle data ingestion, annotation, and the transfer of parameters between federated nodes. During training, pathologists at each institution have access to the most current model and can evaluate its performance on any local testing data stored on their DSA instance. This could help identify new slides where the model struggles for inclusion in the training dataset for future rounds of training. Visualization of the models on testing data is done by using Histo-Cloud tool[11] to segment the slides. The output of this model can be displayed as a series of contours or a heatmap depicting the probability of a structure (as shown in Fig. 2), directly on the slide in the HistomicsUI viewer which is internal to the DSA.

In the future we will test this pipeline on more segmentation tests and develop further plugins for training models for WSI classification and multi-instance learning.

ACKNOWLEDGEMENT

This project was supported by NIH-NIDDK grant R01 DK114485 (PS), NIH-OD grant R01 DK114485 03S1 (PS), a glue grant (PS) from the NIH-NIDDK Kidney Precision Medicine Project grant U2C DK114886 (Contact: Dr. Jonathan Himmelfarb), a multi-disciplinary small team grant RSG201047.2 (PS) from the State University of New York, a pilot grant (PS) from the University of Buffalo's Clinical and Translational Science Institute (CTSI) grant 3UL1TR00141206 S1 (Contact: Dr. Timothy Murphy), a DiaComp Pilot & Feasibility Project 21AU4180 (PS) with support from NIDDK Diabetic Complications Consortium grants U24 DK076169 and U24 DK115255 (Contact: Dr. Richard A. McIndoe), NIH-OD grant U54 HL145608 (PS; Contact: Dr. Kun Zhang and Dr. Sanjay Jain), and NIDDK grant R01 DK131189 (PS; Contact: Dr. Farzad Fereidouni).

REFERENCES

1 Farahani, N., Parwani, A. V. & Pantanowitz, L. Whole slide imaging in pathology: advantages, limitations, and emerging perspectives. *Pathology and Laboratory Medicine International* **7**, 23-33 (2015).

2 LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *nature* **521**, 436-444 (2015).

3 Abels, E. *et al.* Computational pathology definitions, best practices, and recommendations for regulatory guidance: a white paper from the Digital Pathology Association. *The Journal of pathology* **249**, 286-294 (2019).

4 Dimitriou, N., Arandjelović, O. & Caie, P. D. Deep learning for whole slide image analysis: an overview. *Frontiers in medicine* **6**, 264 (2019).
5 Scheibner, J. *et al.* Data protection and ethics requirements for multisite research with health data: a comparative examination of legislative governance frameworks and the role of data protection technologies. *Journal of Law and the Biosciences* **7**, lsaa010 (2020).
6 Konečný, J. *et al.* Federated learning: Strategies for improving communication efficiency. *arXiv preprint arXiv:1610.05492* (2016).
7 Yang, Q., Liu, Y., Chen, T. & Tong, Y. Federated machine learning: Concept and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)* **10**, 1-19 (2019).
8 McMahan, B., Moore, E., Ramage, D., Hampson, S. & y Arcas, B. A. in *Artificial Intelligence and Statistics.* 1273-1282 (PMLR).
9 Li, X., Huang, K., Yang, W., Wang, S. & Zhang, Z. On the convergence of fedavg on non-iid data. *arXiv preprint arXiv:1907.02189* (2019).
10 Ginley, B. *et al.* Automated Computational Detection of Interstitial Fibrosis, Tubular Atrophy, and Glomerulosclerosis. *Journal of the American Society of Nephrology* (2021).
11 Lutnick, B. *et al.* A user-friendly tool for cloud-based whole slide image segmentation, with examples from renal histopathology. *bioRxiv*, 2021.2008.2016.456524, doi:10.1101/2021.08.16.456524 (2021).
12 Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F. & Adam, H. in *Proceedings of the European conference on computer vision (ECCV).* 801-818.
13 Gutman, D. A. *et al.* The digital slide archive: A software platform for management, integration, and analysis of histology for cancer research. *Cancer research* **77**, e75-e78 (2017).
14 Solem, A. *Celery - Distributed Task Queue*, <https://docs.celeryproject.org/en/stable/> (2021).
15 VMware. *RabbitMQ*, <https://www.rabbitmq.com/> (2021).
16 Sutskever, I., Martens, J., Dahl, G. & Hinton, G. in *International conference on machine learning.* 1139-1147 (PMLR).