AMERICAN ACADEMY OF OPHTHALMOLOGY®

Federated Learning for Microvasculature Segmentation and Diabetic Retinopathy Classification of OCT Data

Julian Lo, MASc,[1,] Timothy T. Yu, BASc,[1,*] Da Ma, PhD,[1] Pengxiao Zang, MEng,[2] Julia P. Owen, PhD,[3] Qinqin Zhang, PhD,[4] Ruikang K. Wang, PhD,[3,4] Mirza Faisal Beg, PhD,[1] Aaron Y. Lee, MD, MSc,[3] Yali Jia, PhD,[2] Marinko V. Sarunic, PhD, MBA[1]*

Purpose: To evaluate the performance of a federated learning framework for deep neural network-based retinal microvasculature segmentation and referable diabetic retinopathy (RDR) classification using OCT and OCT angiography (OCTA).

Design: Retrospective analysis of clinical OCT and OCTA scans of control participants and patients with diabetes.

Participants: The 153 OCTA en face images used for microvasculature segmentation were acquired from 4 OCT instruments with fields of view ranging from 2 × 2-mm to 6 × 6-mm. The 700 eyes used for RDR classification consisted of OCTA en face images and structural OCT projections acquired from 2 commercial OCT systems.

Methods: OCT angiography images used for microvasculature segmentation were delineated manually and verified by retina experts. Diabetic retinopathy (DR) severity was evaluated by retinal specialists and was condensed into 2 classes: non-RDR and RDR. The federated learning configuration was demonstrated via simulation using 4 clients for microvasculature segmentation and was compared with other collaborative training methods. Subsequently, federated learning was applied over multiple institutions for RDR classification and was compared with models trained and tested on data from the same institution (internal models) and different institutions (external models).

Main Outcome Measures: For microvasculature segmentation, we measured the accuracy and Dice similarity coefficient (DSC). For severity classification, we measured accuracy, area under the receiver operating characteristic curve (AUROC), area under the precision-recall curve, balanced accuracy, F1 score, sensitivity, and specificity.

Results: For both applications, federated learning achieved similar performance as internal models. Specifically, for microvasculature segmentation, the federated learning model achieved similar performance (mean DSC across all test sets, 0.793) as models trained on a fully centralized dataset (mean DSC, 0.807). For RDR classification, federated learning achieved a mean AUROC of 0.954 and 0.960; the internal models attained a mean AUROC of 0.956 and 0.973. Similar results are reflected in the other calculated evaluation metrics.

Conclusions: Federated learning showed similar results to traditional deep learning in both applications of segmentation and classification, while maintaining data privacy. Evaluation metrics highlight the potential of collaborative learning for increasing domain diversity and the generalizability of models used for the classification of OCT data. *Ophthalmology Science 2021;1:100069* © *2021 by the American Academy of Ophthalmology. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).*

Supplemental material available at www.ophthalmologyscience.org.

Diabetic retinopathy (DR) is a complication of diabetes mellitus, the most common cause of vision loss among people with diabetes, which affects 749 800 people in Canada[1] and is expected to affect 191.0 million people worldwide by 2030.[2] Diabetic retinopathy damages the structure of the blood vessels of the retina,[3] a light-sensitive tissue in the eye, leading to widespread areas of ischemia and loss of visual acuity.[4,5] Diabetic retinopathy is diagnosed and the severity is graded based on clinical examination findings, retinal funduscopic photography results, and fluorescein angiography results.[6] Direct, noninvasive, and quantitative analysis of the retinal microvasculature has significant potential to improve the clinical management of DR. One of the most promising methods for diagnostic imaging of the retinal microvasculature is OCT angiography (OCTA), which allows for volumetric imaging of the retinal vasculature with resolution down to the level of retinal capillaries.[7,8]

https://doi.org/10.1016/j.xops.2021.100069
ISSN 2666-9145/21

The use of artificial intelligence has been extended to numerous problems in the medical industry and is advancing rapidly in ophthalmic applications. Reviews of deep learning and artificial intelligence in medicine discuss the research and future directions,[9] including applications in glaucoma[10] and DR.[11–13] In this study, we expanded on previously published frameworks for microvasculature segmentation[14] and quantification,[15] as well as DR classification.[16,17] As deep learning applications increase in complexity, the amount of data required to train a robust and accurate deep neural network model increases significantly. However, for medical applications, data are often guarded behind privacy regulations regarding sensitive patient information. This presents a nearly insurmountable hurdle for collaborative data sharing between institutions. Additionally, a possibility exists that a model trained solely on medical data available in its own so-called data island is significantly overfitted, especially if all the data originate from 1 source.[18] This is the case for image processing algorithms, where images originating from 1 source may have distinct features that may lead to overfitting as training progresses.

Federated learning is a distributed machine learning approach that enables model training on a large corpus of decentralized data originating from different sources without directly accessing the sensitive data.[19] Cross-device federated learning, as originally described by Google,[20] consists of a framework in which a central server could distribute copies of a machine learning model to a set of clients for training. Each client locally could perform 1 or multiple steps of gradient descent ("learning") on local data and subsequently return its results to the central server to be averaged with the rest of the client base. Frameworks for developing federated learning algorithms such as Tensorflow Federated, NVIDIA Clara,[21] and PySyft[22] exist; however, for medical applications with a small number of participating (collaborating) institutions, this approach can be simplified. This is termed *cross-silo federated learning*.[23] Compared with cross-device federated learning, the small number of collaborative participants in a cross-silo setting simplifies execution by allowing for the training to be synchronous. The cross-silo setting also assumes that the participants are trusted and do not present an adversarial risk toward federated training, which can include white-box and black-box inference attacks[24] or exploiting the gradients to reconstruct the training data.[25] This approach has been explored by various groups in medical research, most notably for the coronavirus disease 2019 (COVID-19) pandemic. A collaborative federated learning platform for computed tomography scan-based COVID-19 diagnosis using a 3-dimensional dense convolutional neural network also was developed recently,[26] with additional work conducted for COVID-19 region segmentation in computed tomography scans.[27] In addition, a federated approach to both L1 regularization and multilayer perceptron models was applied to electronic health records to predict COVID-19 mortality, showing improvement over models trained locally.[28] A federated learning framework was also developed for functional magnetic resonance imaging analysis using domain adaptation.[29] In the field of OCT for ophthalmic imaging, to the best of our knowledge, no previous report has examined the use of federated learning. A related model-to-data approach was applied to intraretinal fluid segmentation in OCT volumes[30] with significant success. This report represents the next step in the progression of model to data to federated learning for ophthalmic applications.

Federated learning may have a large impact on niche research topics and rare diseases where datasets are locked within each institution and open-sourced datasets are limited. Clinical applications with large and open-sourced datasets contain many edge cases and may have reduced benefit from collaboration through federated learning. Using federated learning to improve the generalizability of the neural networks through the collaboration of multiple institutions with access to large datasets is important and should be investigated for the effects on neural network robustness. However, the focus of this study is toward facilitating collaboration between institutions investigating problems with small datasets. In this study, we investigated the federated learning approach to apply microvasculature OCTA segmentation to multiple datasets in a simulated cross-silo environment. The performance of the federated model was compared with that of models trained solely on local data, models trained on a fully centralized dataset, as well as alternative methods of collaborative deep learning. The framework subsequently was extended to a true collaboration between multiple institutions for referable DR (RDR)—eyes diagnosed with moderate DR or higher—classification using OCT and OCTA imaging. The performance of the federated RDR classification model was compared with that of models trained solely on local data across the institutions participating in this study.

Methods

Ethics Statement

Institutional review board (IRB) or ethics committee approval was obtained before implementation, and the experiments were conducted in accordance with the tenets of the Declaration of Helsinki independently at both Simon Fraser University (SFU) and Oregon Health & Science University (OHSU). A separate IRB approval was required for each institution for neural network model sharing, but the approval process was more efficient than the process required to share sensitive patient information and data. The requirement for informed consent was waived because of the retrospective nature of the study.

Framework Implementation

The main components of the federated learning framework involved the central server and the individual clients from which the private data were sourced. The central server served as the hub of the framework, coordinating training and defining hyperparameters for each client. The central server also aggregated each client's updates and performed averaging to compute a new global model for redistribution. Model updates from each client, as well as each iteration of the aggregated global model, were saved locally, and thus were not accessed by any

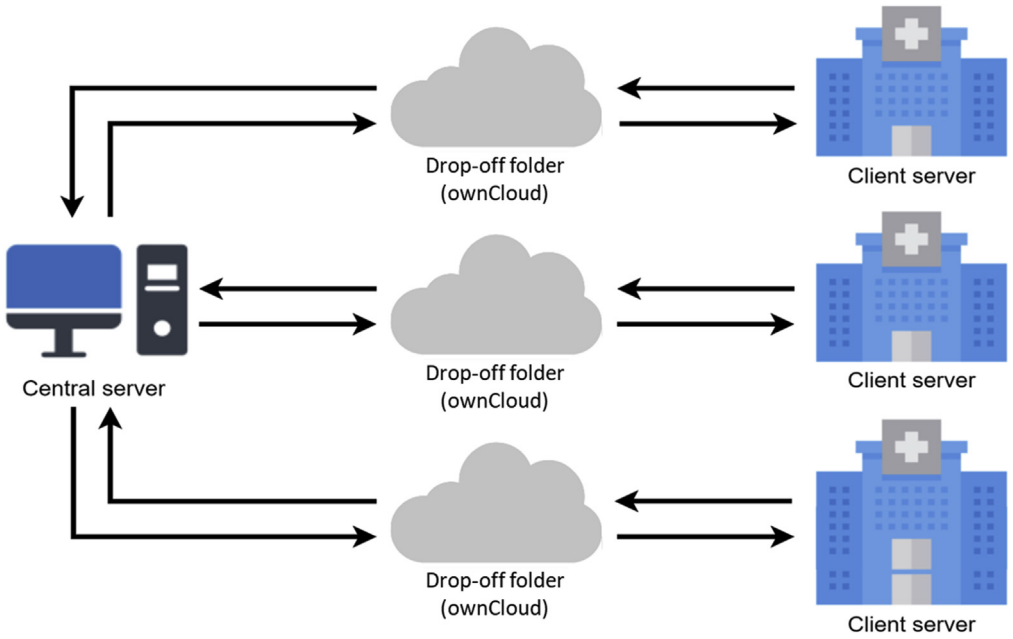

Figure 1. Diagram showing federated learning schematic.

participating client. Each client's training data remained on its own servers, preventing access by any outside party. Secure model transfer between each client and the central server was handled through the use of cloud-based drop-off folders implemented using the self-hosted open-source file synchronization and share server ownCloud (ownCloud GmbH) software, as shown in Figure 1. Each client was designated its own folder that was inaccessible by other clients. OwnCloud allows the model distribution to be under the control of a participating institution. In this study, we hosted the ownCloud folder on SFU servers, but this folder could be hosted at any site designated as the aggregator using their own ownCloud implementation.

For each training cycle, the central server distributed the aggregate model, alongside instructions in a configuration file unique to each client. This configuration file included information such as the current epoch, as well as the learning rate at which that epoch would be trained. The model and configuration file transfers to and from ownCloud folders were automated using the client module from the open-sourced ownCloud application programming interface (API). Using the ownCloud API, the clients and aggregator accessed the ownCloud folder using security keys specified in an institution-specific initialization configuration file. We implemented a robust system of checks to ensure that interruptions in connection at any stage between the client or aggregator and the ownCloud folder were handled adequately by the program. Continuous connection attempts were made if a client was unable to connect with the ownCloud server.

On receiving the model and configuration file, each client performed data augmentation and trained for a full epoch on its training set and then validated on its internal validation set. The resulting model and comma-separated values file containing the loss and accuracy for both training and validation were returned to the central server. As soon as the central server received the model and comma-separated values file from each client, it further validated each client model on a small validation set to isolate models that could disrupt the overall training process. Consequently, any client model that scored below a specified tolerance value would be omitted from the aggregated model. The tolerance value was minimal (0.3 or less) to remove potentially diverged models without inadvertently introducing bias to the training. The tolerance value of 0.3 that was selected as a safeguard was determined heuristically based on an SFU pilot dataset, which was a subset of the SFU training set. The tolerance was low enough to allow models from different domains that may perform poorly on the pilot dataset to be included in the aggregated model. It filtered out data contamination as a result of potential misconfiguration and improperly processed data and was never reached at any point in our study, as it should not be, if all participants were diligent. The paths to the saved models and training data were defined in a configuration file located locally on each client and not accessed by the aggregator. The open-source code can be found online on Github (https://github.com/borg-sfu/federated-learning-oct).

Federated Learning Applications

Microvasculature Segmentation. In the first experiment, we applied the federated learning framework to the application of microvasculature segmentation on OCTA en face images, expanding on a previously published study.[15] To examine the efficacy and clinical usefulness of a federated learning approach, the resulting model was compared with alternative collaborative approaches, which include pooling all the data into a centralized location. To allow for this, the available data were restricted to OCTA en face images that were locally acquired and sourced. The resulting 4 individual datasets used for training are described in detail elsewhere[14] and are summarized in Table 1. The dataset from each image source was split into training, validation, and testing partitions. The 6 × 6-mm images from the PLEX Elite instrument [Carl Zeiss Meditec] were divided into quadrants to ensure image size similarity. Four simulated clients, each hosting 1 of the 4 datasets, occupied 4 separate compute nodes on a supercomputer cluster

Table 1. Datasets Used for Microvasculature Segmentation

Image Source	Field of View (mm)	Included Capillary Complexes	No. of Available Images	Dataset Partitions (Quadrants)
SFU prototype swept-source OCTA	2 × 2	SVC	30	18 training 6 validation 6 testing
RTVue XR Avanti (OptoVue, Inc.)	3 × 3	SVC	26	16 training 5 validation 5 testing
Angioplex (Carl Zeiss Meditec)	3 × 3	SVC	24	14 training 5 validation 5 testing
PLEX Elite 9000 (Carl Zeiss Meditec)	6 × 6	SVC and DVC	73 (292 quadrants)	42 (168) training 15 (60) validation 16 (64) testing

DVC = deep vascular complex; SFU = Simon Fraser University; SVC = superficial vascular complex.
Overview of the 4 individual datasets used for the federated learning simulation of microvasculature segmentation. Images in the PLEX Elite 6 × 6-mm dataset were split after partitioning into training, validation, and test sets.

(Compute Canada). Another local computer was used as the central server.

Model Architecture and Training Parameters. The architecture used on each client was the residual U-Net,[31] which is shown in Figure S2. Neural network training was performed over a maximum of 1000 epochs on an NVIDIA Tesla P100 graphics processing unit, after which the model with the lowest validation loss was selected for evaluation. All clients were trained using stochastic gradient descent optimization with a cyclic learning rate schedule that used warm restarts. The learning rate would decrease from 0.1 to 0.001 with a decay factor of 0.1, and repeat. Each client used the same data augmentation steps, which included random pixel dropout (5% to 10%), linear (0.5 to 3), and γ (0.4 to 1.75) contrast adjustment; rotations (−15° to 15°); and horizontal and vertical flips using the ImgAug Python library.

Comparison with Other Collaborative Deep Learning Methods. To investigate the performance of the federated learning framework, we evaluated several simulated collaborative training scenarios. First, the model-to-data approach was investigated by training a model on each dataset individually, but without transfer learning, as was done previously.[30] Second, a combined dataset with all available images was used for training in the ideal case, where all data are available. The effects of the dataset sizes (and the resulting diversity in training examples) were explored by constructing 2 additional combined datasets with an equal number of images—1 with 14 images randomly sampled from each source (the maximum possible, because the Angioplex [Carl Zeiss Meditec] 3 × 3-mm training set contained 14 images), and 1 with only 4 images randomly sampled from each image source—to approximate the size of a smaller dataset. Finally, the federated model was compared with a model trained on all 4 datasets sequentially in the order shown in Table 1 to simulate a naïve collaborative deep learning approach.

Referable Diabetic Retinopathy Classification

In the second experiment, we applied the federated learning framework to RDR classification. Data collected from SFU and OHSU were used to investigate the relative performance of federated learning for the classification of RDR in OCT en face images. The image acquisition protocol, severity grading, and en face OCTA generation algorithm were as described in previous reports from our groups.[16,17] Images with a signal strength of more than 8 of 10 or with sufficient capillary network visibility through manual evaluation were included in the SFU dataset.[32] Images with signal strength index of more than 50 were included in the OHSU dataset.[33] Details of the 2 datasets, including the image acquisition systems, dataset stratification, and dataset allocation, are as described in Table 2. The 3-channel input for RDR classification was generated from a combination of OCTA en face images from the superficial vascular complex, OCTA en face images from the deep vascular complex, and a maximum intensity projection calculated from both of the OCT structural en face superficial vascular complex and deep vascular complex. The deep and superficial vascular complex boundary extraction algorithms were specific to the commercial OCT image acquisition system.[16,17] As with the microvasculature segmentation experiment, the federated learning performance was compared against models trained on 1 specific dataset, but fully collaborative approaches could not be explored to uphold data privacy.

The preprocessing and data augmentation pipeline were consistent across both institutions. Each dataset was preprocessed to have DR severity-stratified balance to enforce fairness in 4-fold cross-validation. Each client randomly allocated its data into 5 folds with an equal number of institution-specific DR severity eyes in each fold, i.e., each OHSU fold contained the following (numbers in the parentheses indicate the number of eyes in each category): normal (19−20), mild (2−3) moderate (4−5), severe (1−2), proliferative DR (36−37); each SFU fold contained the following: normal (31−32), mild (13−14), moderate (5−6), severe (12−13), proliferative DR (12−13). Oregon Health and Science University and SFU each allocated 1 fold for testing and used the remaining folds through 4-fold cross-validation to train and validate 4 models. In a post hoc review of the SFU data, 1 eye in the training dataset was identified as having been assigned incorrectly to the severe group instead of the proliferative DR group. A balanced distribution between RDR and non-RDR (NRDR) was created through random upsampling. The model training process was further augmented randomly throughout training. Augmentations included random dropout (5% to 10%), linear contrast changes (0.9 to 1.1), rotations (−10° to 10°), horizontal and vertical translations (−0.05 to 0.05), and horizontal and vertical flips using the ImgAug and Keras preprocessing Python libraries. Each image was resized to 512 × 512 pixels for cross-institution consistency. Channel-wise normalization into a range from 0 to 1 was performed to harmonize the different sites.

Model Architecture and Training Parameters. Transfer learning of a VGG19 architecture with ImageNet weights was used for feature extraction, and the classifier consisted of 2 fully connected layers, shown in Figure S3. The training hyperparameters were determined through 4-fold cross-validation. Models

Table 2. Datasets Used for Referable Diabetic Retinopathy Classification

Institution	Commercial OCT Systems (Field of View; mm)	Binary Stratification (No. of Images in Each)	Institution-Specific Stratification (No. of Images in Each)	Dataset Allocation
OHSU[17]	OptoVue, Avanti RTVue-XR SD OCT (3 × 3)	Non-RDR (n = 111), RDR (n = 212)	Normal (n = 99) Mild (n = 12) Moderate (n = 22) Severe (n = 7) Proliferative (n = 183)	20% testing 20% validation 60% training
SFU[16]	Zeiss, PLEX Elite SS OCT (3 × 3)	Non-RDR (n = 226); RDR (n = 151)	Normal (n = 157) Mild (n = 69) Moderate (n = 27) Severe (n = 61) Proliferative (n = 63)	20% testing 20% validation 60% training

OHSU = Oregon Health and Science University; RDR = referable diabetic retinopathy; SD = spectral-domain; SFU = Simon Fraser University; SS = swept-source.
Overview of the datasets from the collaborating institutions for the task of RDR classification.

that were trained on data from a single institution were trained for 100 epochs with a learning rate that decayed from 5×10^{-4} to 5×10^{-6}. The federated approach used a cyclic learning rate, as described in the segmentation application, decaying from 3×10^{-4} to 1×10^{-6} twice throughout 100 epochs.

Performance Evaluation. The following evaluation metrics were calculated for each of the models: accuracy, area under the receiver operating characteristic curve, area under the precision-recall curve, balanced accuracy, F1 score, sensitivity, specificity, and the severity-specific accuracies. Benjamini-Hochberg adjusted two-tailed *t* tests were performed to calculate a statistically significant ($P < 0.05$) difference in means of the evaluation metrics between federated learning and models trained on 1 institution's dataset. The means and standard deviations are reported for each of the evaluation metrics. The optimal threshold values for classification were calculated from the receiver operating characteristics curve performance of the respective validation set during model training.

The effect of thresholding was explored, and the binary classification of models was further evaluated on datasets stratified into the 5 stages of DR. The 4 models, 1 from each training fold, were ensembled using majority soft voting by averaging out the probabilities calculated from each model. The probabilities for the positive class (i.e., whether a given input image belongs to a patient with RDR) are graphically displayed through histogram plots for each training method. The output class of each eye is the ensembled probability thresholded using the average threshold value calculated during training. To further understand the range of severities most affected by thresholding, the allocated testing data were stratified further into their original 5 severities. As is the case with many smaller medical datasets, the class imbalance is an issue that exists not only between NRDR and RDR, but across all 5 severities. The participating institutions of this study have more eyes distributed toward the extremes (normal and proliferative DR).

Results

Microvasculature Segmentation

Table 3 shows the segmentation accuracy when all 9 training methods are evaluated on each dataset's test set. In this case, an internal model is a model trained and tested on data from 1 institution, and an external model is a model trained and tested on data from different institutions. The model trained with federated learning attained accuracy scores comparable with those of the internal models, except for the PLEX Elite 6 × 6-mm dataset, which resulted in a minor reduction in accuracy compared with the internal models. It also achieved similar performance as the models trained on combined datasets, suggesting that pooling all datasets into a centralized location may provide only a marginal benefit over federated learning. Additionally, the federated model outperformed the sequentially trained model for the 2 × 2-mm and 3 × 3-mm datasets, as the sequentially trained model biased toward the most recently seen PLEX Elite 6 × 6-mm dataset. A similar trend is seen when calculating the Dice similarity coefficient (DSC), as detailed in Table 4. The segmentation performance of the federated learning model exceeded the dataset-exclusive model for the 2 × 2-mm dataset, achieved comparable scores for the two 3 × 3-mm datasets, and resulted in a minor reduction in DSC for the 6 × 6-mm dataset. The federated learning model also achieved comparable scores as all combined datasets, similar to the results in Table 3.

Referable Diabetic Retinopathy Classification

Similar to the microvasculature segmentation experiment, for each test set (SFU and OHSU), the federated learning model was compared with both internal and external models, but, to preserve data privacy, combining datasets for centralized training was not possible. The corresponding thresholds for classification were calculated from their respective validation set during the federated training process. First, we evaluated the overall classification performance of the 3 experimental setups. Both federated learning and internal models significantly outperformed external models, as shown in Table 5. The federated learning approach was comparable with internal models when tested on the SFU (Table 5, top panel) and OHSU (Table 5, bottom panel) datasets. We further investigate the model performance on each stratified diagnostic group to gain an in-depth understanding of the relationship between the model and the diagnostic severity (Table 6). Figures 4 and 5 demonstrate the effect of thresholding on model performance because the method of acquiring a

Table 3. Accuracy of Federated Learning on Microvasculature Segmentation

Model Training Method (No. of Images)	Simon Fraser University Prototype 2 × 2-mm	OptoVue 3 × 3-mm	Angioplex 3 × 3-mm	PLEX Elite 6 × 6-mm
Federated learning	0.857 ± 0.031	0.815 ± 0.006	0.850 ± 0.030	0.784 ± 0.053
Sequential	0.835 ± 0.033	0.789 ± 0.006	0.821 ± 0.036	0.810 ± 0.039
Only SFU prototype (n = 18)	0.858 ± 0.033*	0.721 ± 0.017	0.808 ± 0.057	0.574 ± 0.061
Only Optovue (n = 16)	0.831 ± 0.049	0.829 ± 0.015*,†	0.801 ± 0.059	0.621 ± 0.073
Only Angioplex (n = 14)	0.818 ± 0.045	0.817 ± 0.009	0.858 ± 0.024*,†	0.713 ± 0.075
Only PLEX Elite (n = 168)	0.814 ± 0.034	0.800 ± 0.007	0.828 ± 0.024	0.817 ± 0.038*,†
Combined (all images)	0.855 ± 0.029	0.804 ± 0.005	0.842 ± 0.030	0.806 ± 0.036
Combined equally (n = 14 each)	0.861 ± 0.033†	0.829 ± 0.014†	0.857 ± 0.029	0.795 ± 0.043
Combined equally (n = 4 each)	0.855 ± 0.030	0.829 ± 0.014†	0.853 ± 0.026	0.785 ± 0.039

SFU = Simon Fraser University.
Data are presented as mean ± standard deviation. Accuracy for each training method when evaluated on each dataset's test set is shown.
*Internal model.
†Highest value(s) in each column.

threshold during training is optimized for the participating institutions. For each of the training approaches, we portray the classification across all 5 severities of DR using confusion matrices in Figures 6 and 7.

Discussion

As deep learning applications grow in complexity, the need for labeled ground-truth data increases significantly. In many cases, a single institution does not have enough resources to procure the data needed to train a robust model. Furthermore, medical images are guarded securely behind various privacy regulations, resulting in a significant barrier to collaborative data sharing between institutions. Also, a possibility exists that models trained mainly within 1 data island are significantly overfitted, which limits their eventual application on unseen data. Federated learning provides a path for collaboratively training a model while keeping the image data secure. The contributions of this study are as follows: (1) the design of an open-source robust federated learning framework to enable cross-silo training for hardware-agnostic applications, (2) a simulated test for microvasculature segmentation in OCTA en face images acquired from 4 imaging sources, and (3) collaboration with a separate institution for RDR classification on OCTA and OCT structural en face images. The framework involved individual clients training a model on its local training data and sending the weights to a central server. The central server aggregated the weights from all clients and redistributed the new global model. Secure file transfer was handled through a cloud-based drop-off folder hosted on SFU servers, eliminating the need for remote secure shell access between the central server and each participant.

Institutional review board approval was required across each institution for federated model transfer. However, the process was more efficient than the approval process for sharing sensitive patient data. The IRB approval was required by both the institution hosting the central server and the participating client(s). However, this may not apply

Table 4. Dice Similarity Coefficient of Federated Learning on Microvasculature Segmentation

Model Training Method (No. of Images)	Simon Fraser University Prototype 2 × 2-mm	OptoVue 3 × 3-mm	Angioplex 3 × 3-mm	PLEX Elite 6 × 6-mm
Federated learning	0.773 ± 0.060*	0.814 ± 0.011	0.824 ± 0.011	0.761 ± 0.078
Sequential	0.752 ± 0.060	0.782 ± 0.006	0.782 ± 0.016	0.798 ± 0.045
Only SFU prototype (n = 18)	0.756 ± 0.074†	0.663 ± 0.019	0.746 ± 0.051	0.265 ± 0.141
Only Optovue (n = 16)	0.671 ± 0.153	0.836 ± 0.020†	0.734 ± 0.061	0.439 ± 0.165
Only Angioplex (n = 14)	0.735 ± 0.079	0.818 ± 0.011	0.836 ± 0.009*,†	0.637 ± 0.141
Only PLEX Elite (n = 168)	0.738 ± 0.059	0.801 ± 0.010	0.801 ± 0.011	0.816 ± 0.038*,†
Combined (all images)	0.755 ± 0.059	0.797 ± 0.007	0.806 ± 0.010	0.797 ± 0.039
Combined equally (14 each)	0.772 ± 0.070	0.835 ± 0.018	0.835 ± 0.011	0.784 ± 0.049
Combined equally (4 each)	0.772 ± 0.058	0.838 ± 0.018*	0.831 ± 0.010	0.782 ± 0.042

SFU = Simon Fraser University.
Data are presented as mean ± standard deviation. Dice similarity coefficients for each training method, evaluated on each dataset's test set, are shown.
*Internal model.
†Highest value(s) in each column.

Table 5. Performance of Federated Learning for Referable Diabetic Retinopathy Classification

Testing Set	Training Model	Accuracy	Area under the Receiver Operating Characteristic Curve	Area under the Precision-Recall Curve	Balanced Accuracy	F1 Score	Sensitivity	Specificity
SFU	Internal (SFU)	0.869 ± 0.046	0.956 ± 0.011	0.927 ± 0.015	0.870 ± 0.035	0.847 ± 0.041	0.875 ± 0.102	0.864 ± 0.127
	External (OHSU)	0.676 ± 0.081*,†	0.852 ± 0.036*,†	0.814 ± 0.031*,†	0.611 ± 0.104*,†	0.341 ± 0.293*,†	0.250 ± 0.235*,†	0.973 ± 0.041
	Federated learning	0.875 ± 0.003	0.960 ± 0.011	0.936 ± 0.017	0.876 ± 0.037	0.851 ± 0.043	0.883 ± 0.103	0.870 ± 0.059
OHSU	Internal (OHSU)	0.884 ± 0.014	0.973 ± 0.008†	0.986 ± 0.004†	0.891 ± 0.016	0.908 ± 0.014	0.869 ± 0.050	0.913 ± 0.071
	External (SFU)	0.586 ± 0.131*,†	0.766 ± 0.137	0.864 ± 0.087	0.594 ± 0.032†	0.585 ± 0.252	0.568 ± 0.378	0.620 ± 0.349
	Federated learning	0.888 ± 0.019	0.954 ± 0.004*	0.972 ± 0.002*	0.897 ± 0.024	0.911 ± 0.015	0.869 ± 0.011	0.924 ± 0.042

OHSU = Oregon Health and Science University; SFU = Simon Fraser University.
Comparing federated learning with internal and external models for the calculated evaluation metrics: mean values are calculated with 1 standard deviation in parentheses.
*Benjamini-Hochberg–adjusted statistically significant ($P < 0.05$) difference in means when compared with internal learning models.
†Benjamini-Hochberg–adjusted statistically significant ($P < 0.05$) difference in means when compared with federated learning models.

to all potential future institutions, and further clarification regarding the criteria for IRB approval is needed for each participant institution. Because federated learning is relatively new, no precedent exists, and we expect that when federated learning frameworks become standardized, the IRB approval process will be more streamlined.

Our study explored the application of federated learning to increase the effective dataset size that often comes with the relatively niche explorations in emerging techniques like vessel segmentation or disease classification on OCTA and OCT structural en face images. Federated learning could facilitate collaboration between groups investigating rare diseases where open-sourced publicly available datasets are limited, and images are locked within each institution. Although value exists in using federated learning to improve the generalizability of tasks with widely available datasets such as fundus photography, we focused on the use-case of federated learning that facilitates multi-institutional collaborative studies toward more niche research topics.

To evaluate the performance of this framework for microvasculature segmentation, we simulated its performance on 4 datasets from 4 separate imaging devices, consisting of fields of view ranging from 2 × 2 to 6 × 6-mm. The resulting federated model achieved minor decreases in accuracy and DSC compared with an internal model (a model trained and tested on data from 1 source), showing that a model can converge despite training in a decentralized manner, although a tradeoff exists between accuracy and generalizability, as shown in Tables 3 and 4. The federated model also achieved similar performance as models trained on centralized datasets, where multiple configurations were tested: combining all training images naïvely and randomly sampling an equal number of images from each dataset. The federated model was also compared with the sequential training method, providing better performance on the first 3 individual datasets in the sequence because of its bias toward the most recently seen fourth dataset. It was also observed that the difference in DSC (in Table 4) for both models trained on randomly sampled combined datasets varied by a maximum of 0.4% for each test set. This suggests that increasing the diversity of examples by adding images from additional imaging sources (through combining the datasets or federated learning) benefits the overall model more than simply expanding a pre-existing dataset.

The federated learning framework was subsequently investigated for the application of RDR classification using OCTA and OCT data from different institutions, using different instruments. Federated learning performance was comparable with an internal model, with both obtaining significantly higher performance compared with external models when evaluated on the test sets from SFU and OHSU. Further analysis on eyes stratified into the original 5 severities was conducted to provide more transparency and insight into the true performance of the model tested on a small dataset. Notably, Table 6 suggests that the classification performance near the decision boundary is lower than the classification performance of eyes on the extreme ends of the DR severity spectrum. Figures 6 and 7 graphically show that in an ensembled method of

Table 6. Performance of Federated Learning on Stratified Diabetic Retinopathy Severities

Testing Set	Training Model	Normal	Mild	Moderate	Severe	Proliferative
SFU	Internal (SFU)	0.914 ± 0.090	0.750 ± 0.222	0.542 ± 0.315	0.942 ± 0.074	0.962 ± 0.077
	External (OHSU)	1.000 ± 0.000	0.911 ± 0.135	0.083 ± 0.096†	0.250 ± 0.202*,†	0.327 ± 0.335†
	Federated learning	0.922 ± 0.054	0.750 ± 0.071	0.750 ± 0.215	0.846 ± 0.126	0.981 ± 0.038
OHSU	Internal (OHSU)	0.900 ± 0.082	1.000 ± 0.000	0.600 ± 0.365	0.500 ± 0.000	0.926 ± 0.014
	External (SFU)	0.625 ± 0.328	0.583 ± 0.500	0.700 ± 0.383	0.375 ± 0.479	0.561 ± 0.376
	Federated learning	0.913 ± 0.048	1.000 ± 0.000	0.400 ± 0.000	0.500 ± 0.408	0.953 ± 0.014

OHSU = Oregon Health and Science University; SFU = Simon Fraser University.
Comparing federated learning to external and internal model accuracies at the institution-specific diabetic retinopathy severity stages: mean values are calculated with 1 standard deviation in parentheses.
*Benjamini-Hochberg–adjusted statistically significant ($P < 0.05$) difference in means when compared with federated learning models.
†Benjamini-Hochberg–adjusted statistically significant ($P < 0.05$) difference in means when compared with internal learning models.

aggregating the cross-validation folds, the federated learning model was able to correctly classify 2 additional eyes with moderate DR and 1 eye with mild DR in the SFU dataset when compared with a model trained with SFU data. However, the federated learning model misclassified 1 more eye with moderate DR in the OHSU dataset when compared with a model trained on OHSU data.

An interesting observation from this study is that the threshold for classification may retain biases of participating institutions and is an important hyperparameter to consider. Our methods allow us to obtain a threshold during training, but when testing on an external dataset from an institution unaffiliated with our training process, the optimal threshold becomes an unknown factor. Optimal thresholds of the federated learning model for the OHSU data ranged from 0.80 to 0.85, whereas classification on SFU data was optimized using thresholds ranging from 0.18 to 0.26. The optimal thresholds seemingly diverge, and the histograms in Figures 4 and 5 suggest that thresholding has a greater impact on eyes with mild and moderate DR that lie near the decision boundary. As shown in Figures 4 and 5, a portion of eyes would be misclassified if the threshold value was set to 0.5. This suggests that the federated learning model still retains the domain differences of the data from each institution. Threshold tuning from the receiver operating characteristic curve is effective when calculated during the training of data from a homogeneous domain; however, this is not possible in a scenario such as federated learning. Therefore, when considering future deployment of federated learning models, domain trends should be discussed, with site harmonization (the adaptation of domain information across participating institutions) potentially allowing for the calculation of a single optimized threshold value. In future work, the effect of site harmonization should be investigated to determine whether the increase in generalization of federated learning models is comparable with models trained on 1 large dataset in disease severity classification.

The results outlined for the segmentation and classification problems suggest that all participants benefit from federated learning when evaluated on data from the other domain and resulted in a model superior to those trained and tested at different institutions. We speculate that the largest benefit is to the participating institutions with fewer data (assuming that the data are similar). As demonstrated by the results of the microvasculature segmentation experiment shown in Tables 3 and 4, the performance of the models trained through the federated learning framework was relatively comparable with the internal models of the silos with fewer data (SFU prototype, Optovue, Angioplex). Conversely, on the 6 × 6-mm PLEX Elite dataset, which has a larger number of images compared with other silos, the federated learning models performed worse than the internal models. This warrants further investigations toward the effect that data distribution and imbalanced datasets have on federated learning. In both experiments, with the exception of microvasculature segmentation for the 6 × 6-mm PLEX Elite dataset, the federated learning framework facilitated the training of a model that is capable of generalizing to multiple datasets with performance comparable with that of internal models. We speculate that the federated models trained on a more diverse pool of data will result in better performance on images from an unseen source. However, one area that merits further exploration is how robust each model would perform on data acquired by instruments from the same manufacturer as in Tables 1 and 2, but at different institutions with different imaging protocols. If resources are available, this can be used in conjunction with a filter bank implementation, where individual deep neural network models are specifically trained to analyze images from only 1 image source and are shared among collaborators using this brand of instrument. This seems to be the most appropriate solution for a set of clients with a large variance in their respective datasets; however, we expect this approach to achieve lower performance on new and unseen data when compared with federated learning. An alternative approach is to use the federated learning model as a starting point for transfer learning, allowing each institution to tailor the model to their data, extending on the methods presented previously.[30]

Federated learning can be used to improve generalizability through the inclusion of diverse data from different institutions. We anticipate that a predetermined threshold could be used for an unseen dataset that is similar to one of the participating institutions. However, if the domain difference is large between the external data and the data in the

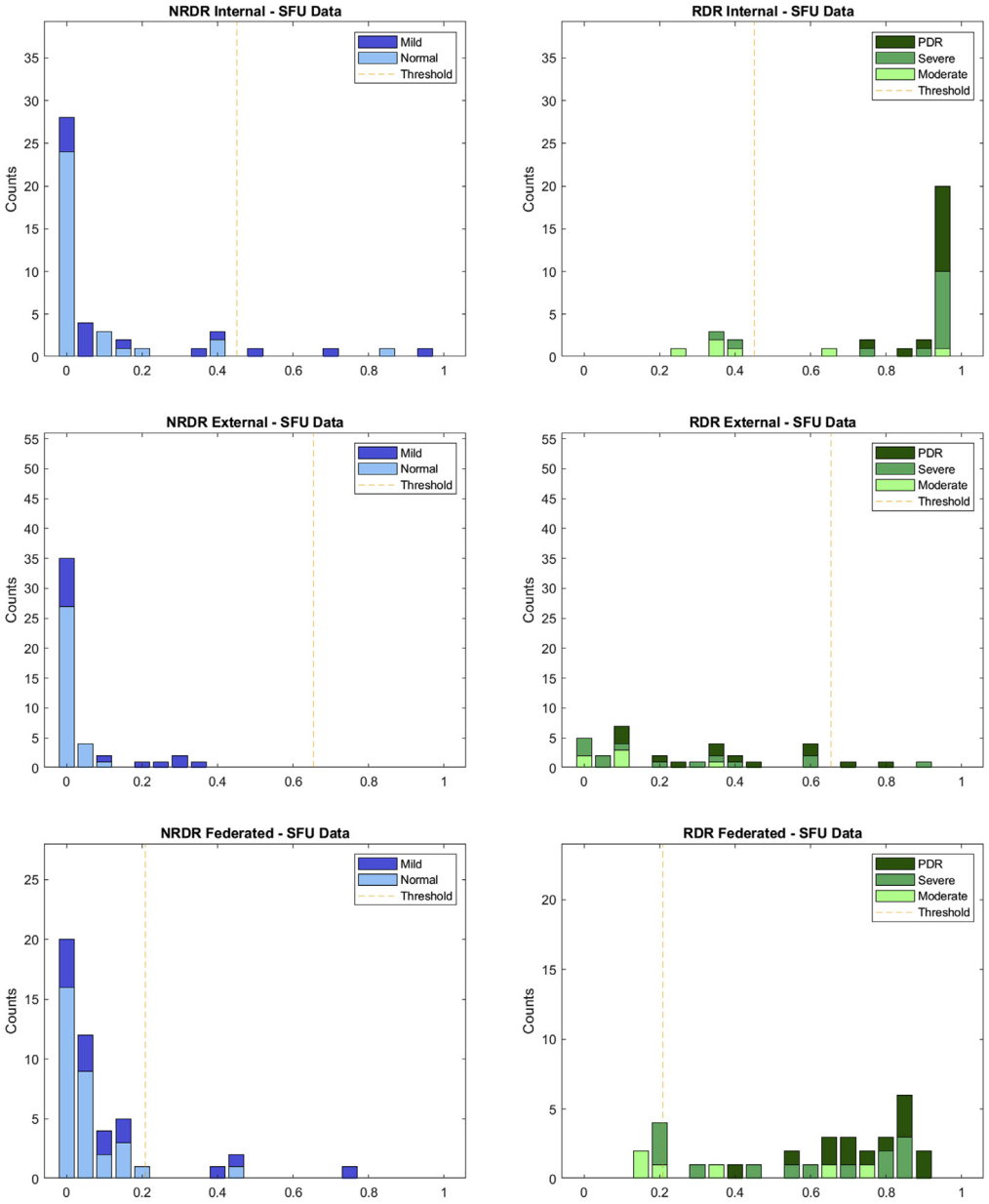

Figure 4. Histograms showing representations of model output probability distributions on Simon Fraser University (SFU) data showing the number of images (y-axes) with the corresponding probability score for referable diabetic retinopathy (RDR; x-axes): non-RDR images (left column) and RDR images (right column). Further 5-stage severity stratification is distinguished by the different shades within each subgroup. PDR = proliferative diabetic retinopathy.

federation, the issue becomes more challenging to address. Inference on a completely unseen dataset is challenging for any deep learning framework, and domain adaptation methods should be explored to improve the generalizability.

Similar to traditional deep learning, a limitation of federated learning is the quantity and diversity of the training data. For the microvasculature segmentation experiment, the PLEX Elite dataset (shown in Table 1) contained significantly more images than the other 3 datasets, resulting in 1 client iterating over more steps per epoch during training. However, when aggregating the client models, each was weighted equally during averaging. This was done to ensure that the federated model does not bias toward a single data source, despite the data imbalance. A potential solution to this is to perform additional augmentation on the smaller datasets; however, because each client model is aggregated into a global model, the benefits may be minimal. Similarly, for the RDR classification application, the overall class distributions in the data corpus directly correlate to model generalizability and should be discussed by all participants before beginning training. Upsampling classes with fewer images through additional augmentations provides a suitable solution; however, acquiring more

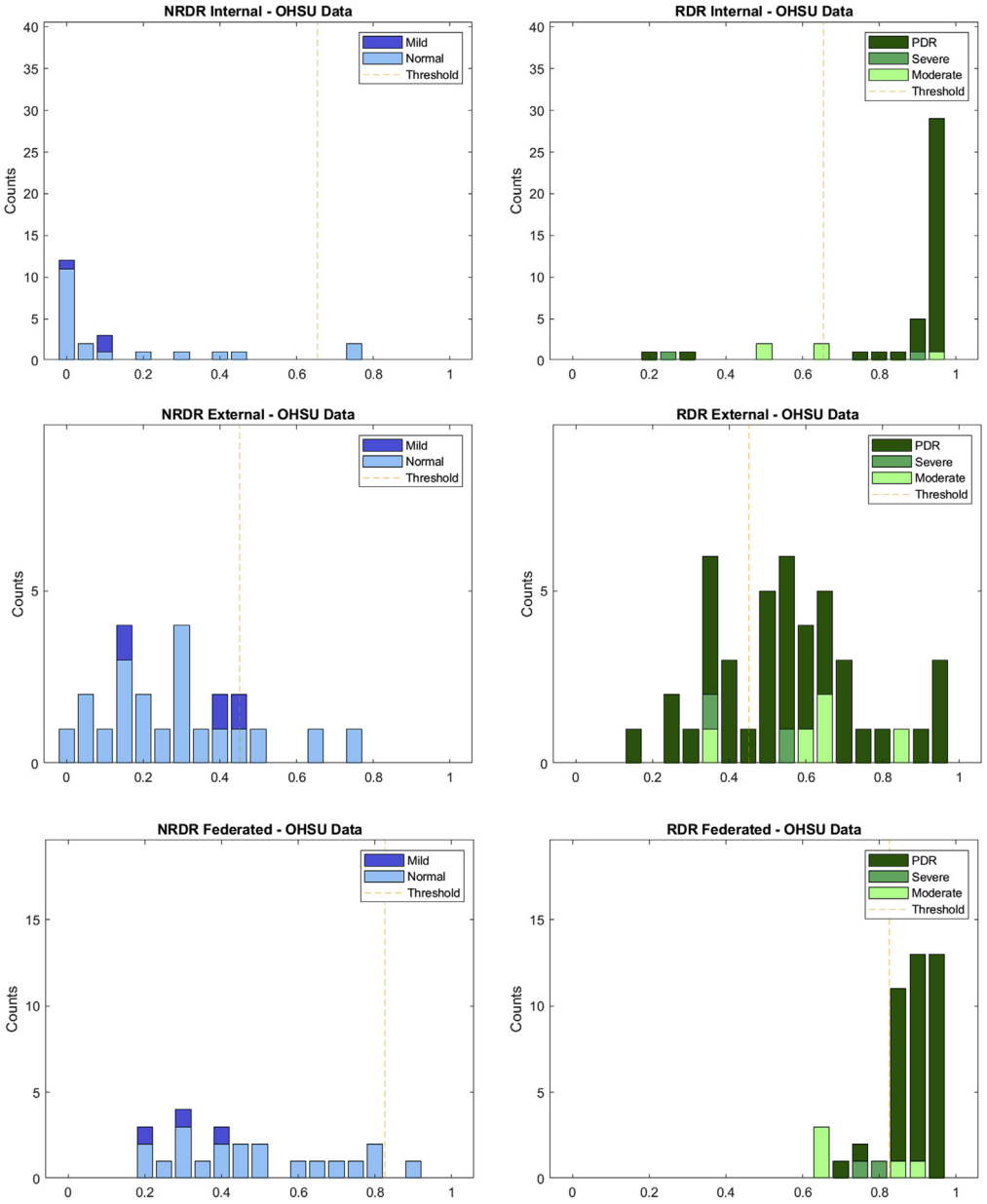

Figure 5. Histograms showing representations of model output probability distributions on Oregon Health and Science University (OHSU) data showing the number of images (y-axes) with the corresponding probability of referable diabetic retinopathy (RDR; x-axes): non-RDR images (left column) and RDR images (right images). Further 5-stage severity stratification is distinguished by the different shades within each subgroup. PDR = proliferative diabetic retinopathy.

training examples through labeled data would be significantly more beneficial. The collaborative and synchronous approach of cross-silo federated learning created issues throughout the training process. Errors resulting from network connection and other technical difficulties in connecting with the drop-off folder or training the model delays training until resolved. For example, 4-fold cross-validated training from an individual institution was completed in approximately 12 hours, whereas the same training using the federated learning framework required upward of 90 hours. As the number of participating institutions increase, these effects are magnified. The collaboration resulting from federated learning allows for the pooling of expertise and clinician knowledge, which is especially important in cases where data are especially limited, which include rare diseases or newest-generation technologies, including adaptive optics.

As with any attempts of deep learning, several risks exist if one has an adversarial motive. Specifically, with federated learning, a malicious "collaborator" may provide poor-quality images or may alter the processing pipeline to attack the training process. This study is grounded in

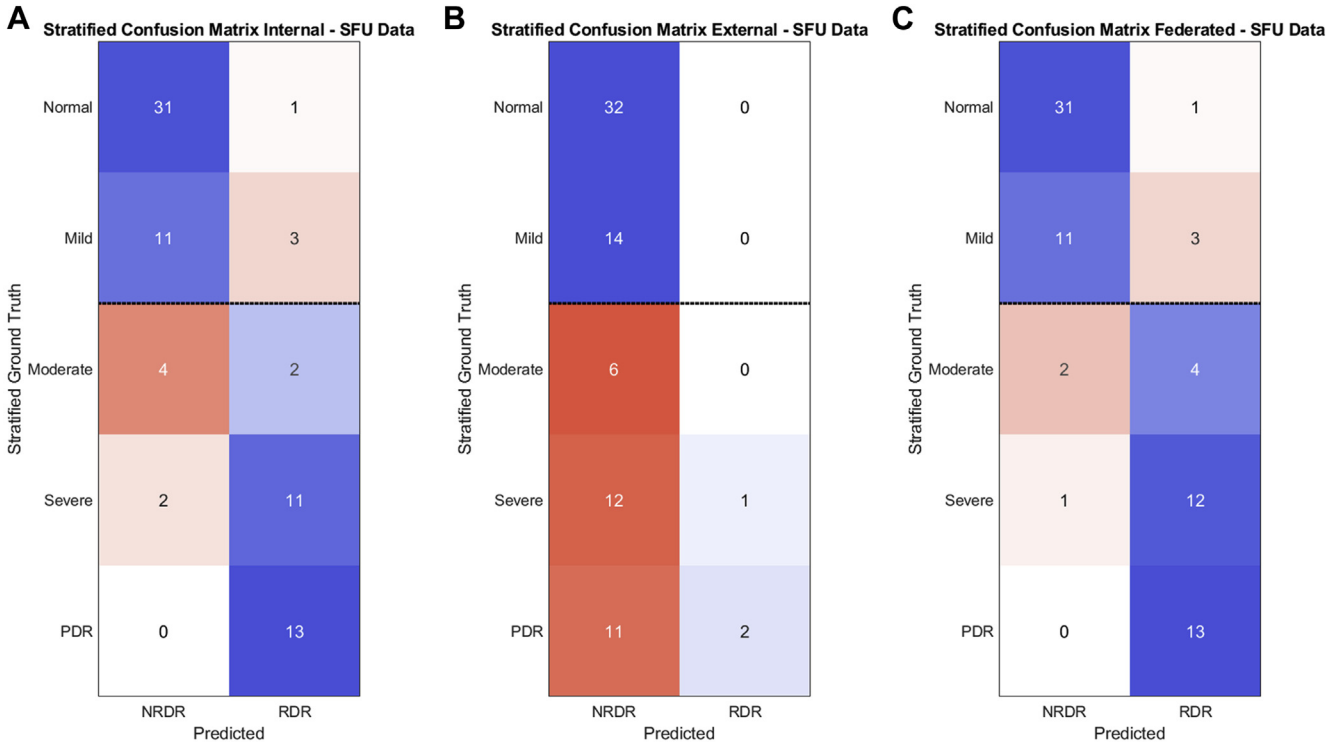

Figure 6. Confusion matrices investigating the referable diabetic retinopathy (RDR) classification performance for each of the 5 diabetic retinopathy severities on Simon Fraser University (SFU) data. Entries shaded blue and red represent correct and incorrect classification, respectively. PDR = proliferative diabetic retinopathy; NRDR = non-RDR.

collaboration and the attainment of a common goal within our federation. We believe that selecting quality collaborators is fundamental toward any use of frameworks where the training and testing data are not shared. Even under the assumption of no malintent, corrupted data have the potential to affect the federated learning model being trained negatively, which we mitigate through a framework-wide validation step before averaging. Ongoing research is addressing attackers with access to model parameters (white box) and with no access to models or parameters (black box),[24] where adversaries are motivated to disrupt the training process through white-box and black-box attacks. Aside from federated learning, attacks exist that attempt to gain insight on private training data where so-called dummy gradients can be introduced from a client to leak private training data.[25] Presently, this applies only to image classification problems with low batch sizes and low-resolution images (up to 64 × 64), which is unsuitable for the high-resolution images seen in medical data and the substantially higher number of parameters seen in image segmentation architectures. Whereas federated learning holds great potential for collaboration, the onus is still on the participants to ensure good-spirited collaboration and data quality. However, developing sophisticated safeguards against these attacks is not a part of the collaborative intentions of our research and is beyond the scope of this study.

Although many options exist for distributed training over multiple graphics processing units, those options provide only a method to speed up neural network convergence, without providing measures for preserving data privacy. At the time of writing this report, NVIDIA Clara[21] had presented another viable option for implementing federated learning that was developed concurrently with our framework. Additional frameworks were considered for our applications such as Tensorflow Federated, as well as PySyft.[22] However, the simplicity of our cross-silo federated learning framework allowed for independent development. Our federated learning framework was intended to facilitate collaboration and exploration of approaches for federated learning, but may differ from other open-sourced frameworks in areas such as scalability. Examples of potential further development of our federated learning framework include the automation of the folder creation through the ownCloud API and improving the efficiency of the aggregator by incorporating multithreading to allow the aggregator to upload and download to and from the ownCloud folders more efficiently.

In conclusion, we designed and developed a framework for multiple participants to conduct federated learning on a decentralized data corpus. The framework is implemented over an ownCloud instance; however, adaptation to other APIs can be added easily. Through our results, we showed that models trained with federated learning perform at a comparable level as internal models, presenting a viable method for increasing available data while maintaining patient privacy.

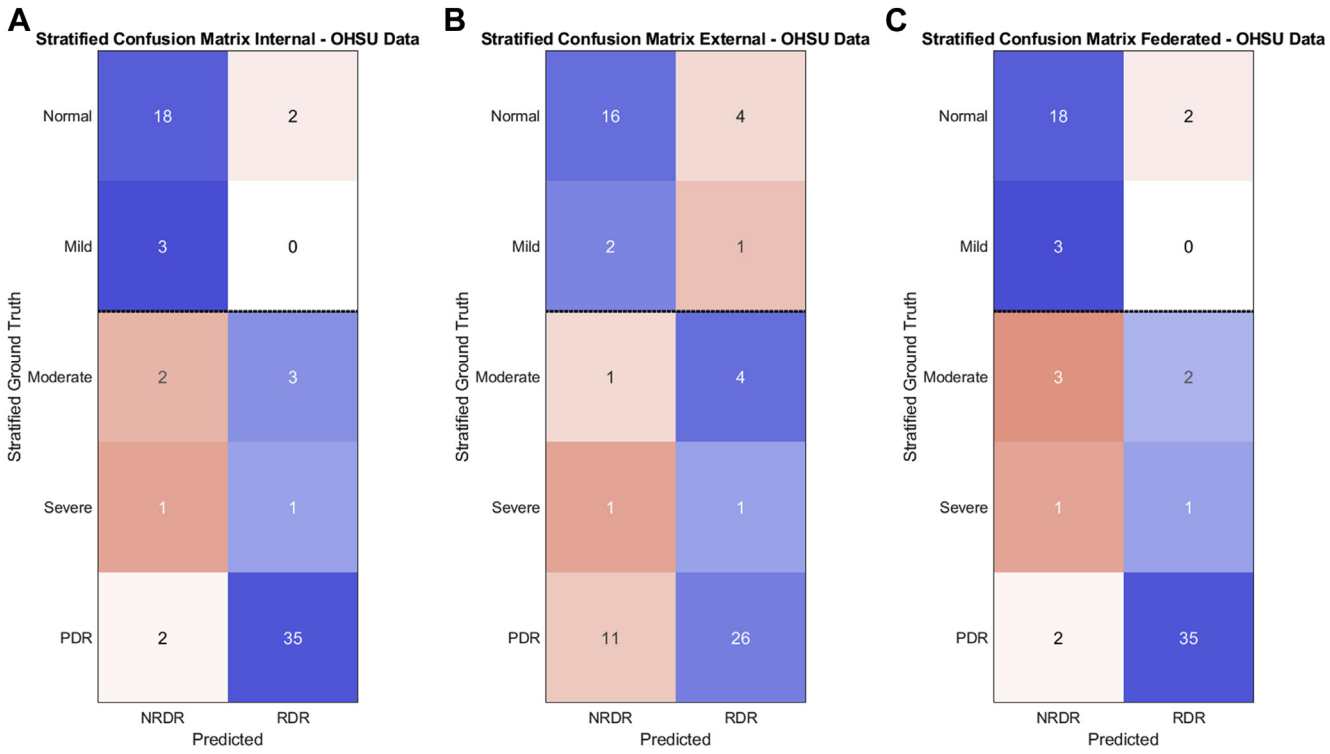

Figure 7. Confusion matrices investigating the referable diabetic retinopathy (RDR) classification performance for each of the 5 diabetic retinopathy severities on Oregon Health and Science University (OHSU) data. Entries shaded blue and red represent correct and incorrect classification, respectively. PDR = proliferative diabetic retinopathy; NRDR = non-RDR.

Footnotes and Disclosures

Originally received: May 12, 2021.
Final revision: September 1, 2021.
Accepted: September 28, 2021.
Available online: October 8, 2021. Manuscript no. D-21-00069.

[1] School of Engineering Science, Simon Fraser University, Burnaby, Canada.

[2] Casey Eye Institute, Oregon Health and Science University, Portland, Oregon.

[3] Department of Ophthalmology, University of Washington, Seattle, Washington.

[4] Department of Bioengineering, University of Washington, Seattle, Washington.

*Both authors contributed equally as first authors.

Disclosure(s):

All authors have completed and submitted the ICMJE disclosures form.

The author(s) have made the following disclosure(s): R.K.W.: Consultant — Carl Zeiss Meditec, Inc, Cyberdontics; Financial support — Carl Zeiss Meditec, Inc, Moptim, Inc, Facebook Technologies, Inc, Colgate-Palmolive Company, Shiseido, Inc, Estee Lauder

A.Y.L.: Financial support — Santen, Regeneron, Novartis, Microsoft, NVIDIA, Verana Health, Genentech/Roche, Gyroscope, Microsoft, Johnson and Johnson, United States Food and Drug Administration, Carl Zeiss Meditec, Topcon

Y.J.: Financial support — Optovue, Inc; Patent — Optos Plc

M.V.S.: Equity owner — Seymour Vision, Inc.

Supported by the National Sciences and Engineering Research Council of Canada (M.F.B., M.V.S.); the Canadian Institutes of Health Research (M.F.B., M.V.S.); the Michael Smith Foundation for Health Research (M.F.B., M.V.S.); Compute Canada (M.F.B., M.V.S.); the National Eye Institute, Bethesda, Maryland (grant no.: K23EY029246 [A.Y.L.]); National Institutes of Health, Bethesda, Maryland (P.Z.; grant nos.: R01EY028753 [R.K.W.], R01EY027833 [Y.J.], R01EY024544 [Y.J.], and P30EY010572 [Y.J.]); Research to Prevent Blindness, Inc., New York, New York (A.Y.L., R.K.W.); and the Lantham Vision Innovation Award (A.Y.L.). The sponsor or funding organizations had no role in the design or conduct of this research.

Aaron Y. Lee, an Associate Editor of this journal, was recused from the peer-review process of this article and had no access to information regarding its peer-review.

HUMAN SUBJECTS: Human subjects were not included in this study. The human ethics committees at Simon Fraser University and Oregon Health and Science University approved the study. All research adhered to the tenets of the Declaration of Helsinki. The requirement for informed consent was waived because of the retrospective nature of the study.

No animal subjects were included in this study.

Author Contributions:

Conception and design: Lo, Yu, Ma, Lee, Sarunic

Analysis and interpretation: Lo, Yu, Ma, Beg, Sarunic

Data collection: Lo, Yu, Ma, Zang, Owen, Zhang, Wang, Lee, Jia, Sarunic

Obtained funding: N/A; Study was performed as part of regular employment duties at Simon Fraser University, Oregon Health and Science University, and the University of Washington. No additional funding was provided.

Overall responsibility: Lo, Yu, Ma, Zang, Wang, Beg, Lee, Jia, Sarunic

Abbreviations and Acronyms:
API = application programming interface; **COVID-19** = coronavirus disease 2019; **DSC** = Dice similarity coefficient; **DR** = diabetic retinopathy; **IRB** = institutional review board; **NRDR** = non-referable diabetic retinopathy; **OCTA** = OCT angiography; **OHSU** = Oregon Health and Science University; **RDR** = referable diabetic retinopathy; **SFU** = Simon Fraser University.

Keywords:
Diabetic retinopathy, Federated learning, Machine learning, Neural network, OCT.

Correspondence:
Marinko V. Sarunic, PhD, MBA, School of Engineering Science, Simon Fraser University, 8888 University Drive, Burnaby, British Columbia V5A 1S6, Canada. E-mail: msarunic@sfu.ca.

References

1. Blindness in Canada, CNIB. Available at: https://cnib.ca/en/sight-loss-info/blindness/blindness-canada.
2. Zheng Y, He M, Congdon N. The worldwide epidemic of diabetic retinopathy. *Indian J. Ophthalmol*. 2012;60(5):428−431.
3. Nentwich MM, Ulbig MW. Diabetic retinopathy—ocular complications of diabetes mellitus. *World J. Diabetes*. 2015;6(3):489−499.
4. Arend O, Wolf S, Harris A, Reim M. The relationship of macular microcirculation to visual acuity in diabetic patients. *Arch Ophthalmol*. 1995;113(5):610−614.
5. Balaratnasingam C, Inoue M, Ahn S, et al. Visual acuity is correlated with the area of the foveal avascular zone in diabetic retinopathy and retinal vein occlusion. *Ophthalmology*. 2016;123(11):2352−2367.
6. Salz DA, Witkin AJ. Imaging in diabetic retinopathy. *Middle East Afr J Ophthalmol*. 2015;22(2):145−150.
7. Chen C-L, Wang RK. Optical coherence tomography based angiography. *Biomed Opt Express*. 2017;8(2):1056.
8. Kashani AH, Chen C-L, Gahm JK, et al. Optical coherence tomography angiography: a comprehensive review of current methods and clinical applications. *Prog Retin Eye Res*. 2017;60:66−100.
9. Rajkomar A, Dean J, Kohane I. Machine learning in medicine. *N Engl J Med*. 2019;380(14):1347−1358.
10. Ran AR, Tham CC, Chan PP, et al. Deep learning in glaucoma with optical coherence tomography: a review. *Eye*. 2021;35(1):188−201.
11. Ting DSW, Pasquale LR, Peng L, et al. Artificial intelligence and deep learning in ophthalmology. *Br J Ophthalmol*. 2019;103(2):167−175.
12. Wagner SK, Fu DJ, Faes L, et al. Insights into systemic disease through retinal imaging-based oculomics. *Transl Vis Sci Technol*. 2020;9(2):6.
13. Schmidt-Erfurth U, Sadeghipour A, Gerendas BS, et al. Artificial intelligence in retina. *Prog Retin Eye Res*. 2018;67:1−29.
14. Heisler M, Chan F, Mammo Z, et al. Deep learning vessel segmentation and quantification of the foveal avascular zone using commercial and prototype OCT-A platforms, arXiv: 1909.11289 [eess.IV], 2019.
15. Lo J, Heisler M, Vanzan V, et al. Microvasculature segmentation and inter-capillary area quantification of the deep vascular complex using transfer learning, *Transl Vis Sci Technol.* 2020;9(2):38.
16. Heisler M, Karst S, Lo J, et al. Ensemble deep learning for diabetic retinopathy detection using optical coherence tomography angiography. *Transl Vis Sci Technol*. 2020;9(2):20.
17. Zang P, Gao L, Hormel TT, et al. DcardNet: diabetic retinopathy classification at multiple levels based on structural and angiographic optical coherence tomography. *IEEE Trans Biomed Eng*. 2021;68(6):1859−1870.
18. Robinson C, Trivedi A, Blazes M, et al. Deep learning models for COVID-19 chest x-ray classification: preventing shortcut learning using feature disentanglement. *medRxiv*. 2021;(Preprint). https://doi.org/10.1101/2021.02.11.20196766.
19. Sheller M, Edwards B, Reina G, et al. Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data. *Scientific Reports*. 2020;10(1):12598.
20. McMahan HB, Moore E, Ramage D, et al. Communication-efficient learning of deep networks from decentralized data, arXiv:1602.05629 [cs.LG], 2016. Available at: https://arxiv.org/abs/1909.11289.
21. Rieke N, Hancox J, Li W, et al. The future of digital health with federated learning. *NPJ Digital Medicine*. 2020;3(1).
22. Ryffel T, Trask A, Dahl M, et al. A generic framework for privacy preserving deep learning, arXiv:1811.04017 [cs.LG], 2018.
23. Kairouz P, McMahan HB, Avent B, et al. Advances and open problems in federated learning, arXiv:1912.04977 [cs.LG], 2019.
24. Nasr M, Shokri R, Houmansadr A. Comprehensive privacy analysis of deep learning: passive and active white-box inference attacks against centralized and federated learning, arXiv: 1812.00910 [stat.ML], 2018.
25. Zhu L, Liu Z, Han S. Deep leakage from gradients, arXiv: 1906.08935 [cs.LG], 2019.
26. Xu Y, Ma L, Yang F, et al. A collaborative online AI engine for CT-based COVID-19 diagnosis. *medRxiv*. 2020. https://doi.org/10.1101/2020.05.10.20096073.
27. Yang D, Xu Z, Li W, et al. Federated semi-supervised learning for COVID region segmentation in chest CT using multi-national data from China, Italy, Japan. *Med Image Anal*. 2021;70:101992.
28. Vaid A, Jaladanki SK, Xu J, et al. Federated learning of electronic health records to improve mortality prediction in hospitalized patients with COVID-19: machine learning approach. *JMIR Med Informatics*. 2021;9(1):e24207.
29. Li X, Gu Y, Dvornek N, et al. Multi-site fMRI analysis using privacy-preserving federated learning and domain adaptation: ABIDE results. *Med Image Anal*. 2020;65:101765.
30. Mehta N, Lee CS, Mendonça LSM, et al. Model-to-data approach for deep learning in optical coherence tomography intraretinal fluid segmentation. *JAMA Ophthalmol*. 2020;138(10):1017−1024.
31. Zhang Z, Liu Q, Wang Y. Road extraction by deep residual U-Net, arXiv:1711.10684 [cs.CV], 2017.
32. Karst SG, Heisler M, Lo J, et al. Evaluating signs of microangiopathy secondary to diabetes in different areas of the retina with swept source OCTA. *Invest Ophthalmol Vis Sci*. 2020;61(5):8.
33. Guo Y, Hormel TT, Xiong H, et al. Development and validation of a deep learning algorithm for distinguishing the nonperfusion area from signal reduction artifacts on OCT angiography. *Biomed Opt Express*. 2019;10(7):3257−3268.