Federated learning for diagnosis of age-related macular degeneration

Sina Gholami[1], Jennifer I. Lim[2], Theodore Leng[3], Sally Shin Yee Ong[4], and Minhaj Nur Alam[1]

[1]Department of Electrical Engineering, University of North Carolina at Charlotte, Charlotte, NC, USA
[2]Department of Ophthalmology and Visual Science, University of Illinois at Chicago, Chicago, IL, USA
[3]Department of Ophthalmology Stanford University School of Medicine, Stanford, CA, USA
[4]Department of Ophthalmology, Wake Forest University, Winston-Salem, NC, USA

July 6, 2023

Abstract

This paper presents a thorough comparison of local, centralized, and federated learning (FL) approaches in training deep learning (DL) models for Optical Coherence Tomography (OCT) image classification. We employ the use of ResNet18 and vision transformer (ViT) encoders for the binary classification of OCT images into Normal and Age-related Macular Degeneration (AMD) categories, integrating three unique domain adaptation (DA) techniques to address domain shift issues. Experimental results indicate that despite longer training times, FL strategies can achieve competitive performance rates, similar to centralized ResNet18 and ViT models. Notably, the FedMRI strategy stood out in our FL evaluations, consistently delivering high performance across all tests due to its contrastive loss function. Furthermore, ViTs showed improved performance on the OCT-500 set compared to ResNet18, emphasizing their capability to focus on global image features. The study provides valuable insights into the efficacy of simpler architectures in image classification tasks, particularly in scenarios where data privacy and decentralization are critical. It suggests future exploration into deeper models and other FL strategies for a more nuanced understanding of these models' performance.

1 Introduction

AMD is a common eye condition and a leading cause of vision loss among people aged 50 and older [Wong et al., 2014]. This causes damage to the macula, the part of the eye that provides sharp, central vision, which is located near the center of the retina. As a result, everyday activities such as reading and driving may be difficult to perform. In order to prevent severe vision impairment and preserve vision, early intervention and detection of AMD in its early stages are crucial to implementing appropriate treatments, such as medications or procedures.

Artificial intelligence (AI) plays a pivotal role in the preliminary identification and classification of AMD [De Fauw et al., 2018, Burlina et al., 2017a, Kaymak and Serener, 2018, Russakoff et al., 2019, Lee et al., 2017, Yim et al., 2020, Treder et al., 2018]. Its proficiency in discerning the disparate stages of both wet and dry AMD results in substantial enhancement of the prognosis of treatment outcomes. DL models significantly refine the precision and accuracy of AMD diagnosis, capable of detecting subtle ocular changes that might elude human scrutiny [Burlina

et al., 2017b]. The remarkable capacity of AI for rapid analysis of imaging data facilitates more expeditious and efficient diagnosis, a critical factor in timely disease management [Brown et al., 2018].

The broad applications of AI include large-scale AMD screening within populations, a critical feature, particularly in areas where accessibility to ophthalmologists is restricted [Rajalakshmi et al., 2018]. Beyond these clinical uses, AI's potential to discern patterns and correlations in expansive data sets could yield innovative perspectives into the origins and evolution of AMD, potentially influencing future research trajectories [Balyen and Peto, 2019].

AI models employed in AMD diagnosis predominantly utilize centralized learning. This traditional method accumulates data from diverse sources, collating them in a centralized server or location for the training of a machine learning (ML) model [Xu et al., 2021]. In the medical sphere, this approach encounters hurdles due to data privacy and security concerns. Patient data's delicate and regulated characteristics complicate the transfer, sharing, and storage of data centrally. Adherence to data protection regulations such as the Health Insurance Portability and Accountability Act (HIPAA) is paramount in healthcare environments [Tom et al., 2020].

The introduction of FL allows model training without the dissemination of raw patient data, thereby circumventing privacy issues as data remains local. The possibilities proffered by FL involve enhancing diagnostic accuracy, prediction capability, and personalized treatment within ophthalmology, whilst harnessing large, diverse data sets. However, for successful FL integration, it is necessary to address challenges linked with data heterogeneity, along with assuring the reliability and security of the learning process.

Inconsistencies in optical coherence tomography (OCT) image acquisition parameters and scanning protocols can induce variations in image quality [De Carlo et al., 2015]. Clinical and technical hurdles including differing standards and regulations among various institutional review boards, and limited training data sets for rare diseases can exacerbate the complexities of constructing and implementing DL techniques [Ting et al., 2019]. Such variations can impact the efficacy and generalizability of DL models [Coyner et al., 2018].

The problem of domain shift poses a significant challenge in the context of FL [McMahan et al., 2017, Smith et al., 2017]. Domain shift arises when there is a substantial difference in data distributions across various devices or nodes, also termed as clients, within the FL system. The non-identically distributed nature of decentralized data, a key characteristic of FL, can potentially compromise model learning performance [Li et al., 2020b]. Rectifying this issue necessitates strategic and robust methodologies.

In the research of [Li et al., 2020b], domain adaptation techniques are outlined for optimizing learning algorithms irrespective of disparities in data distribution. Employing domain-invariant features or transfer learning methodologies, these techniques endeavor to lessen the impact of varied data distributions. Additionally, data augmentation can be leveraged to artificially enhance data representation, thereby diminishing the effects of domain shift [Perez and Wang, 2017].

An array of other methods can also be utilized to counter this challenge, encompassing client selection and sampling strategies, model aggregation procedures, and proactive domain exploration [Yang et al., 2019]. By effectively tackling domain shift, FL can bolster the model's generalization capacity and augment performance across disparate domains.

This manuscript delineates the practicality of employing DA FL in the diagnosis of AMD. Using data from three distinct data sets, this study examines various FL strategies to address the domain shift issue for retinal OCT binary classification for AMD.

The performance of these FL strategies is juxtaposed with a baseline centralized approach to evaluate the prospective advantages of deploying various FL techniques to alleviate the domain shift. The FL methodologies put into practice encompass the baseline approach, Federated Averaging (FedAvg) [McMahan et al., 2017], along with domain shift-specific techniques such as Federated Proximal (FedProx) [Li et al., 2020c], FedSR [Nguyen et al., 2022], and a modified

FedMRI [Feng et al., 2022].

2 Methods

We leveraged OCT data derived from three distinct researches for our study [Kermany et al. [2018], Srinivasan et al. [2014], Li et al. [2020a]], hereinafter referred to as Kermany, Srinivasan, and OCT-500. The utilization of these distinct data sets facilitated the simulation of three disparate institutions (FL nodes) intent on training a DL model for binary image classification (Normal Vs AMD). Hence, each node is allocated its own training, validation, and testing set.

2.1 Data

Kermany data set encompasses a total of 84,484 OCT retinal images which are classified into four categories: normal, Choroidal Neovascularization (CNV), Diabetic Macular Edema (DME), and Drusen. These images are compartmentalized into three distinct folders: training, validation, and testing. However, it was observed that some images were duplicated across the validation and testing folders as well as the training folder. To eliminate redundancy, we amalgamated the validation and testing folders and compared each image with those in the training set using the mean square error (MSE) technique. An MSE score of zero signified the presence of identical images, leading to the identification of 8,520 duplicates within the dataset. These issues stemmed from 34 images that were marked as both normal and diseased retina. We discarded these images and exclusively used Drusen and Normal retinal images for this binary classification task. In the end, around 3% of the patient samples were chosen as the test set, which contained varying numbers of scans per patient.

The srinivasan data set comprises retinal images from 45 subjects, including 15 individuals each with normal retinas, AMD, and DME. The initial 11 patients' data was designated for training, the 12th for validation, and the remaining data were used for testing the model.

The OCT-500 data set includes OCT imaging of 500 subjects captured under two fields of view (FOVs): 3-mm and 6-mm. Each 3-mm file contains 304 scans from a single patient, while each 6-mm file accommodates 400 scans. Given the lack of a substantial role of peripheral retinal parts in classification, we focused only on fovea images (image numbers 100 to 180 for 3-mm and 160 to 240 for 6-mm).

The distribution of the data across the three data sets is visually represented in Figure 1 and tabulated in Table 1.

Table 1: The data distribution of the three data sets for the training and test sets, including the number of normal and ADM retinas for each set.

Data set	train		test		total	
	Normal	AMD	Normal	AMD	Normal	AMD
Kermany	23794	7214	1237	434	25031	7648
Srinivasan	1006	530	291	147	1297	677
OCT-500	17320	3480	2268	243	19588	3723

2.2 Centralized and Local Models

Our hypothesis is that an FL model can be trained to deliver performance on par with a centralized model that has access to all data and surpass the performance of local models trained solely on locally available data. We established baseline models, both centralized and local, and evaluated their performance against test sets. Local models were trained exclusively

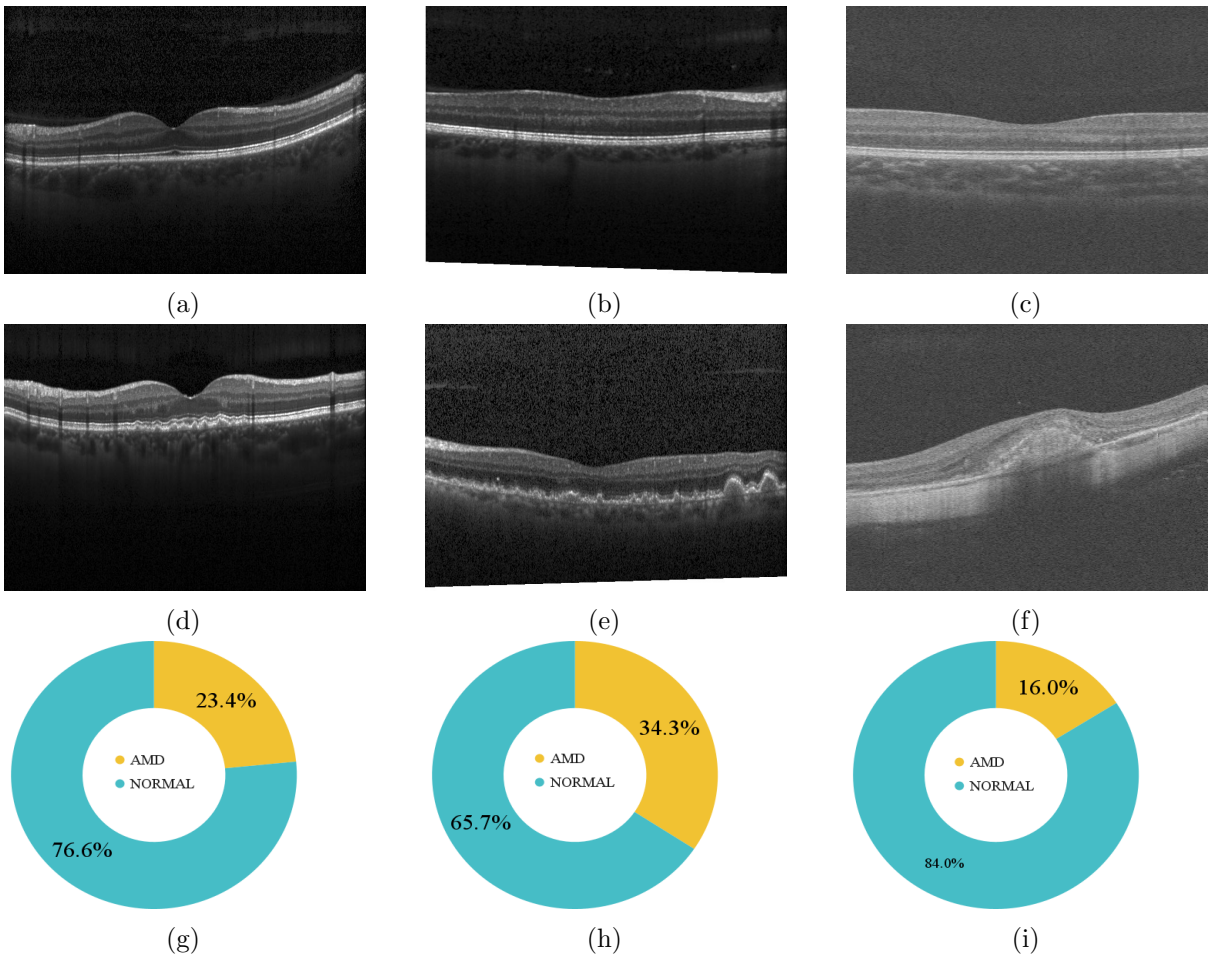


Figure 1: Illustration of various OCTs from different data sets. The first column (**a**, **d**, **g**) shows samples from the Kermany data set and its data distribution, the second column (**b**, **e**, **h**) from the Srinivasan data set, and the third column (**c**, **f**, **i**) from the OCT-500 data set.

on their respective data and then subjected to performance assessments using all three test sets. In contrast, the centralized model underwent training that utilized all three data sets before its evaluation. This rigorous testing methodology provided us with a robust comparative analysis of the performance metrics of these models. Our model's structure is designed with two main components: an encoder and a classification head. After evaluating various options such as Residual Networks (ResNet), ViT, VGG16, InceptionV3, and EfficientNet, we settled on ResNet18 with 11.2 million parameters and ViT with 4.8 million parameters as the encoding mechanisms for our models, conducting thorough comparisons of their performances across diverse benchmarks. The ViT encoders consist of six transformer blocks and eight heads in the multi-head attention layer.

We utilized the Area Under the Receiver Operating Characteristics (ROC) Curve (AUC) as our metric for evaluation. Initial findings indicate that the Adaptive Momentum Estimation with Weight Decay (AdamW) surpasses the Stochastic Gradient Descent (SGD) in terms of performance at both the local and centralized levels, after hyperparameter optimization through grid search. Detailed descriptions of the final hyperparameters for ResNet18 and ViT models are provided in the appendix. The optimal hyperparameter combination was determined through the maximization of the AUC on the validation set, taking care to prevent data leakage from the test set. To examine the impact of the number of epochs (E) on the models, we trained the models with $E = 10$ and $E = 100$, implementing early stopping based on the AUC of the validation set and patience of ten epochs when $E = 100$. Kermany data set was processed on a setup equipped with 2× Nvidia RTX A6000 graphics cards (node 1), while Srinivasan data set was allocated to a machine furnished with 2× Nvidia Titan V units (node 2), and OCT-500 data set was trained on a node with 8× Nvidia GTX1080Ti cards (node 3). Centralized benchmarks were executed on node 3. The summarized results can be found in Table 2.

2.3 FL Framework

Traditional FL algorithms involve a central server that oversees model updates and circulates the global model to all participating nodes. Local models are trained on the respective data and subsequently transmitted back to the server, where they are integrated into the global model [Li et al., 2020b]. Notably, the main FL algorithms utilized include FedAvg [McMahan et al., 2017] and Federated Stochastic Gradient Descent [Smith et al., 2017], along with their variants.

However, the decentralized character of FL introduces substantial challenges, especially in terms of data heterogeneity and distribution shifts. For instance, in ophthalmology, considerable variations in retinal images across different institutions can be attributable to factors such as the use of distinct imaging devices [Chen et al., 2012], heterogeneous patient populations [Vickers, 2017], and inconsistencies in image acquisition protocols [Khanifar et al., 2008].

Addressing these challenges necessitates domain alignment, also referred to as DA. This essential process modifies an ML model trained on one domain to perform proficiently on a related domain. Numerous techniques have been proposed to mitigate the domain shift problem, making it crucial to implement these methods for successful DA. In our FL framework, we have compared four DA strategies: FedProx, FedSR, and FedMRI.

FedProx: FedProx [Li et al., 2020c], is specifically designed to counter the data heterogeneity challenge in FL. It utilizes proximal regularization to incorporate a penalty term into the loss function and avoid overfitting. By maintaining local updates close to the initial global model parameters, FedProx is particularly useful when dealing with not independent and identically distributed data. This ensures each local model does not veer too far from the global model during training, yielding a more resilient global model that performs well across a broader spectrum of data distributions.

FedSR: FedSR [Nguyen et al., 2022] simplifies the model's representation and encourages it to extract only essential information. This method employs two regularizers: an L-2 norm regularizer on the representation and conditional mutual information between the data and

the representation given by the label. These regularizers limit the quantity of information the representation can contain. By enforcing these regularizers, FedSR facilitates learning data representations that generalize well across diverse domains, all while maintaining data privacy between nodes—a crucial advantage in an FL context.

FedMRI: FedMRI [Feng et al., 2022] addresses the issue of domain shift that might surface during local node optimization. It does so through the implementation of a weighted contrastive regularization, which helps guide the update direction of the network parameters, thus directly rectifying any discrepancies between the local nodes and the server during optimization. This approach contrasts with traditional contrastive learning, which relies on identifying positive and negative pairs from data. In experiments involving multi-institutional data, FedMRI has demonstrated superior performance in image reconstruction tasks compared to state-of-the-art FL methods. As our task lies within the realm of binary image classification, we customized the FedMRI approach. Specifically, we excluded the decoder component and employed the weighted contrastive loss as an auxiliary loss exclusively.

Our FL structure integrates three FL nodes with a central server, and it is developed based on the Flower framework. Before running the local training on these nodes, the server needs to be operational, necessitating the selection of a certain FL strategy, FL, and training configurations. The strategy oversees several elements of the training and evaluation protocol, such as weight initialization and aggregation. FL configuration outlines necessary parameters for FL training, encompassing the minimum number of FL nodes needed for training and subsequent evaluation. Furthermore, the training configuration encapsulates requisite parameters for DL model training, including the number of epochs, learning rate, and weight decay.

The procedure to train the FL model generally follows these steps: Initially, the FL strategy (options include FedAvg, FedSR, FedProx, and FedMRI) will be designated, as well as the FL and training configurations. Subsequently, the server waits for the necessary minimum number of FL nodes to establish a connection, following which it dispatches the training configuration and the initial weights (based on the selected FL strategy) to each node. Each node then updates its local model using the weights received from the server, instigating the training process. Upon completion of the training, each node transmits its local model weights back to the server. Finally, the server aggregates these weights using the designated strategy (such as FedAvg) and reciprocates by sending the updated weights back to each client, marking the conclusion of one round (R).

In this framework, there is an optional step called evaluation, where each local node assesses its performance after receiving the global FL model and evaluation configuration over local test sets. The evaluation configuration is similar to the training configuration and may contain various hyperparameters for model evaluation, such as early stopping or batch size. After evaluation, the performance of each node is sent to the server to demonstrate the FL model's overall performance across all node test sets. Data heterogeneity can be handled by varying the number of local training epochs. This way, each round of training can be more productive, reducing convergence time and communication costs [Mendieta et al., 2022]. To assess the training productivity of each strategy, we examined its AUC with three different allotted local training epochs per round in Table 3.

During each benchmarking session, one of the nodes played a dual role by serving as both a server and an FL node. The other two resources solely functioned as FL nodes and communicated with the server. Whenever $E = 10$, node 1 assumed the role of both server and FL node. When $E = 5$, node 2 became the server, and when $E = 1$, node 3 took on the role of server. The DL models at each local node were trained using the hyperparameters detailed in the preceding section. Note that, the value of R in all FL benchmarks is 10.

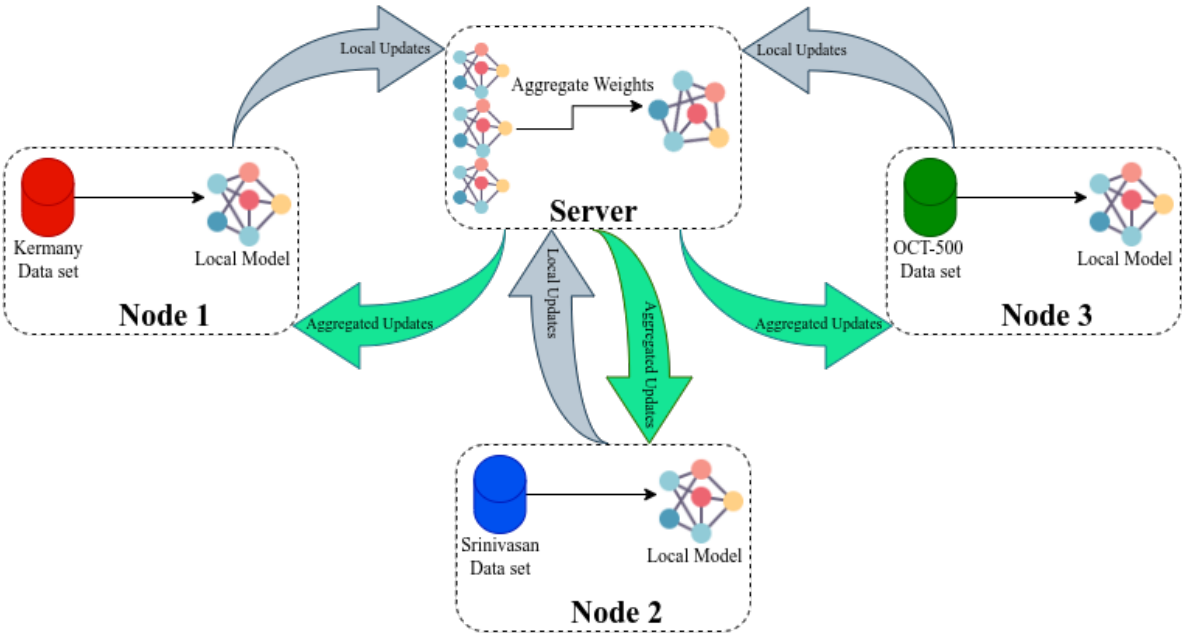


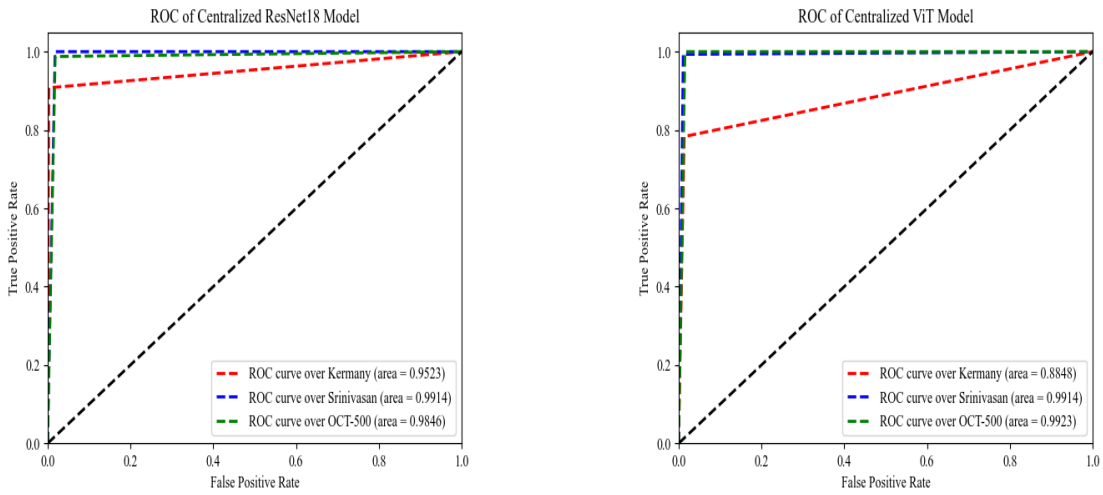
Figure 2: General Overview of FL Framework.

2.4 Results

Table 2: The AUC of the Centralized and Local Models using $E = 10$ and $E = 100$ with early stopping on the validation AUC.

Method	Test Set		
	Kermany	Srinivasan	OCT-500
	E = 10		
Centralized (ResNet18)	92.78 ± 1.83	98.9 ± 0.89	97.72 ± 2.09
Centralized (ViT)	85.64 ± 1.87	98.19 ± 0.82	99.22 ± 0.53
Local Kermany (ResNet18)	93.7 ± 1.31	98.83 ± 0.88	54.14 ± 3.75
Local Kermany (ViT)	85.23 ± 1.21	96.76 ± 0.92	88.08 ± 3.17
Local Srinivasan (ResNet18)	55.6 ± 2.6	80.40 ± 10.07	50 ± 0.002
Local Srinivasan (ViT)	51.07 ± 1.08	75.03 ± 5.54	53.63 ± 2.73
Local OCT-500 (ResNet18)	49.84 ± 0.78	54.65 ± 6.31	94.5 ± 3.74
Local OCT-500 (ViT)	48.39 ± 2.27	56.73 ± 9.13	86.3 ± 7.24
	E = 100		
Centralized (ResNet18)	**94.58 ± 0.87**	99.06 ± 0.57	**98.91 ± 0.93**
Centralized (ViT)	87.85 ± 1.67	**99.18 ± 0.76**	99.11 ± 0.54
Local Kermany (ResNet18)	94.04 ± 0.97	98.26 ± 1.21	51.74 ± 0.93
Local Kermany (ViT)	88.3 ± 2.09	97.7 ± 0.83	88.8 ± 3.82
Local Srinivasan (ResNet18)	57.03 ± 3.39	84.47 ± 11.33	50 ± 0
Local Srinivasan (ViT)	52.61 ± 1.94	83.16 ± 3.97	56.6 ± 4.45
Local OCT-500 (ResNet18)	49.99 ± 0.02	50 ± 0	91.98 ± 5.62
Local OCT-500 (ViT)	48.45 ± 3.95	55.49 ± 10.54	84.59 ± 5.85

Table 2 summarizes the results derived from training a range of local and centralized models, each evaluated against three different test sets upon completion of the training process. Both ResNet18 and ViT encoders were employed during the training phase, with their respective performance metrics detailed in the table. The latter segment of Table 2 specifically showcases the outcomes associated with training models at $E = 100$. Among these models, the centralized models record the highest performance, ResNet18 achieving a maximum of %95.45, and %99.94 on the Kermany, and Srinivasan, while ViT on OCT-500 test sets with %99.84. Figure 3 offers valuable insights into the operational efficiency of these two models, centralized ResNet18 and ViT, across the three test sets when $E = 100$ and the patient parameter is set to ten.



(a) (b)

Figure 3: AUC of Centralized model over the three data sets using (**a**) ResNet18 and (**b**) ViT.

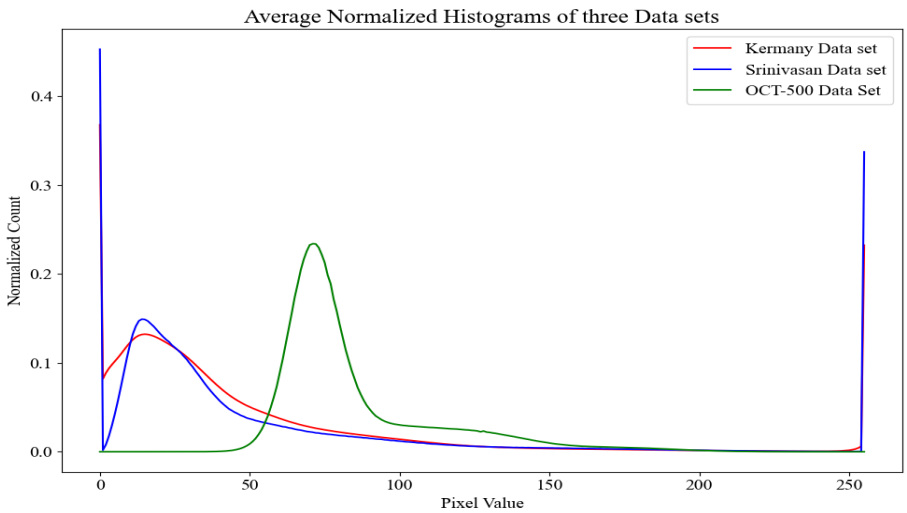


Figure 4: This figure displays the average normalized histograms of all OCT images in three data sets.

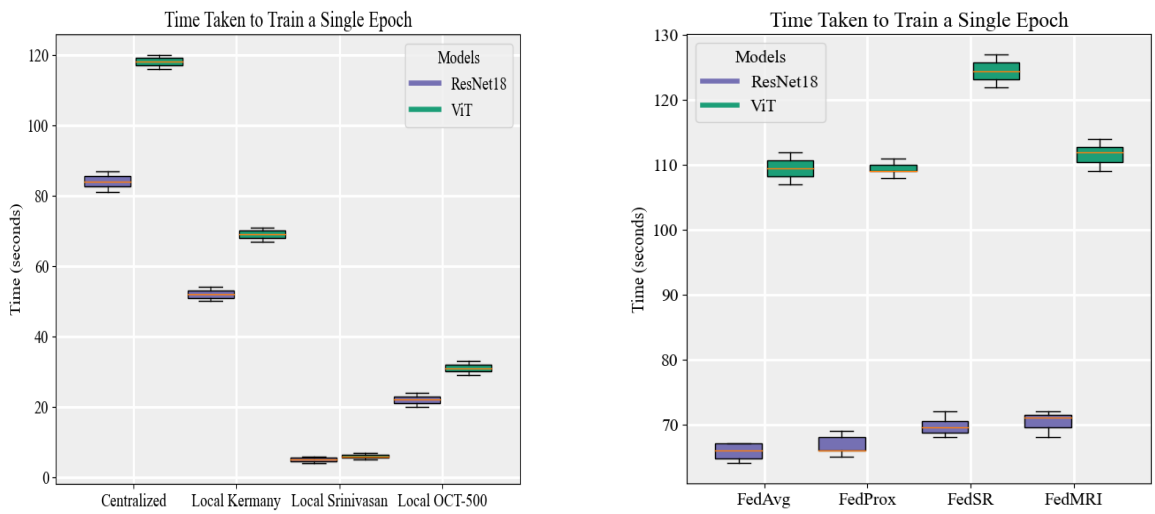
Moreover, we have presented the results of applying the FL strategies, including FedAvg, FedProx, FedSR, and FedMRI, in Table 3. We evaluated the performance of ResNet18 and ViT models using these strategies at $E = 1$, $E = 5$, and $E = 10$. The most successful model for each E value is highlighted in bold for easier identification. It is noticeable that the highest-performing

model on the Kermany test set changes depending on the value of E. Specifically, FedAvg leads when $E = 1$, whereas FedSR and FedMRI perform superiorly when $E = 5$ and 10, respectively. However, for the Srinivasan and OCT-500 test sets, FedSR and FedMRI respectively demonstrate the highest performance compared to the other implemented FL strategies. As mentioned before, $R = 10$ is consistent across all benchmarks.

Table 3: The AUCs of the different FL methods on the three test sets with $E = 1$, $E = 5$, $E = 10$, and $R = 10$.

Method	Test Set, $R = 10$		
	Kermany	Srinivasan	OCT-500
	E = 1		
FedAvg (ResNet18)	**90.47 ± 1.55**	99.01 ± 0.38	54.26 ± 1.96
FedAvg (ViT)	82.4 ± 0.46	97.25 ± 0.26	93.95 ± 0.6
FedProx (ResNet18)	85.46 ± 2.94	91.75 ± 5.08	53.67 ± 4.22
FedProx (ViT)	82.4 ± 0.53	97.33 ± 0.32	94.03 ± 0.66
FedSR (ResNet18)	89.29 ± 1.78	**99.45 ± 0.45**	59.12 ± 3.38
FedSR (ViT)	81.5 ± 0.33	95.01 ± 0.1	95.59 ± 0.81
FedMRI (ResNet18)	89.17 ± 1.39	95.46 ± 1.05	**98.32 ± 0.46**
FedMRI (ViT)	50 ± 0	50 ± 0.3	50 ± 0
	E = 5		
FedAvg (ResNet18)	92.48 ± 1.02	99.42 ± 0.44	55.87 ± 5.98
FedAvg (ViT)	81.61 ± 14.78	98.47 ± 0.47	92.16 ± 14
FedProx (ResNet18)	92.11 ± 1.5	97.22 ± 1.53	65.1 ± 5.17
FedProx (ViT)	88.16 ± 0.51	98.32 ± 0.45	96.53 ± 0.63
FedSR (ResNet18)	**92.99 ± 0.69**	**99.79 ± 0.2**	56.91 ± 5.03
FedSR (ViT)	85.58 ± 0.77	98.22 ± 0.53	97.44 ± 0.64
FedMRI (ResNet18)	91.9 ± 1.42	93.95 ± 5	**98.01 ± 0.53**
FedMRI (ViT)	73.09 ± 0.16	92.62 ± 0.15	86.31 ± 0.72
	E = 10		
FedAvg (ResNet18)	89.88 ± 0.93	99.62 ± 0.18	54.54 ± 1.18
FedAvg (ViT)	88.55 ± 0.59	99.18 ± 0.37	98.07 ± 0.29
FedProx (ResNet18)	92.89 ± 0.88	99.54 ± 0.25	58.57 ± 5.97
FedProx (ViT)	88.86 ± 0.48	99.02 ± 0.32	98.17 ± 0.36
FedSR (ResNet18)	90.87 ± 0.95	**99.74 ± 0.25**	54.07 ± 2.81
FedSR (ViT)	88.61 ± 0.63	99.12 ± 0.26	98.23 ± 0.42
FedMRI (ResNet18)	**93.16 ± 1.3**	97.72 ± 1.07	**97.49 ± 1.58**
FedMRI (ViT)	81.84 ± 0.38	94.75 ± 0.59	94.09 ± 0.25

Figure 5 depicts the training duration for each model, with a noticeable pattern of longer training times for ViT models in comparison to the ResNet18 equivalents. This trend is consistently apparent across local, centralized, and various FL models, even persisting through FL training iterations at $E = 5$ and $E = 10$. Notably, FedSR stands out as having the lengthiest training time among all the benchmarks. The time difference is minimal when training the local model using the Srinivasan dataset - ResNet18 takes about 4-6 seconds, while ViT requires around 5-7 seconds. However, this difference grows when it comes to centralized and FL models, extending up to approximately 40 seconds for training one epoch. Keep in mind that the duration to train an FL model for one epoch is timed from the instant the server dispatches the initial weights to all nodes until it receives and aggregates all the parameters (FL training time).



(a) (b)

Figure 5: A comparison of the time taken to train an epoch for (**a**) local and centralized models, and (**b**) different FL strategies when local $E = 1$.

This calculation does not include the time spent on initializing the server, starting the nodes, connecting them to the server, and the evaluation stages. Due to this reason, FL strategies, with the exception of FedSR, tend to take less time to train than centralized models at $E = 1$.

3 Discussion

This study presented a comprehensive series of experiments exploring the comparative effectiveness of deploying DL models using local, centralized, and FL methodologies across three distinct data sets. The primary focus was the classification of OCT images into Normal and AMD binary categories, for which we utilized ResNet18 and ViT encoders. We also integrated three unique DA methods into our FL strategy to tackle the prevalent issue of domain shift. As our results show, DA FL strategies demonstrated impressive proficiency in training a global model well-suited to this specific problem, achieving competitive performance rates in comparison to centralized ResNet18 and ViT models.

ResNet architectures were selected due to their well-documented efficacy in image classification tasks, the depth of their architectural design enabling complex data pattern learning, and the convenience of existing pre-trained models. ViT were chosen for their ability to incorporate global image context, an invaluable trait for enhancing image classification performance, and their architecture, which eliminates the need for task-specific designs, permitting complex pattern discernment without requiring specialized configurations.

Our experimental process for both local and centralized DL models involved two distinct training scenarios: a short-duration training over 10 epochs and an extended training over 100 epochs. Our intent was to pinpoint the model offering optimal performance when trained over 100 epochs with matching training data, using the validation AUC as the stopping criteria. We also scrutinized the effect of altering the number of local epochs on the training efficiency of the FL strategies, setting E values at 1, 5, and 10.

The comparison demonstrated the expected superiority of centralized models over local models, owing to their full access to all data during training. An interesting observation was the performance discrepancy of the local Kermany ResNet18 on different test sets. While it displayed robust results on its own and the Srinivasan test sets, it faltered with the OCT-500 set. This issue arose from the variance in brightness distribution between the Kermany and

Srinivasan sets, and the OCT-500 set, which can be seen in figure 4. A review of the local Kermany ViT's performance highlighted the encoder's focus on global features, an aspect that may have contributed to its satisfactory performance with the OCT-500 set. However, the local OCT-500 and Srinivasan models performed subpar on tests outside their native environment. Potential causes include model generalization incapacity (for OCT-500) and inadequate training data (for Srinivasan). Additionally, local ResNet18 models demonstrated better performance than ViTs on their corresponding test sets, a likely result of their deeper and parameter-rich architecture.

Examining the FL training duration, the theoretical expectation would be for parallel training (inherent to FL models) to take less time than sequential training. While we observed a slight reduction in the training time for a single epoch of an FL model, the varying sizes of the data sets (with the Kermany set being larger) prevented significant outpacing of the centralized models. Training discrepancies between nodes also created bottlenecks, with Nodes 2 and 3 having to wait for Node 1's completion. This problem was exacerbated as the E value increased, resulting in extended idle periods for quicker nodes.

In our investigations, FedMRI distinguished itself as a superior FL strategy, consistently delivering high performance across all tests. This success can be attributed to its contrastive loss function, which guarantees alignment between local training and global parameters and underscores often-neglected features. As a result, FedMRI demonstrated both robustness and reliable performance. Interestingly, within FL scenarios, ResNet models underperformed on the OCT-500 set, a likely consequence of the averaging methodology that skewed towards the Kermany and Srinivasan data sets and led to suboptimal paths when three models displayed different gradient directions. Despite FL ViT models exhibiting improved performance on the OCT-500 compared to ResNet18, none managed to outperform the FedMRI ResNet18. The superior performance of FL ViT models on OCT-500 aligns with observations from the Local Kermany ViT and underscores the benefits of using ViTs over ResNet18 due to their focus on global features.

However, this study is not without its limitations. In the training phase, we opted for a relatively straightforward DL architecture and an aggregation policy based solely on a weighted average. Despite these limitations, our findings offer valuable insights into the comparative effectiveness of training simpler architectures for image classification tasks and contribute to our broader understanding of Federated Learning strategies.

In conclusion, our research underscores the potential of FL strategies, particularly when equipped with a contrastive loss function, in training robust models capable of consistent performance across varying data sets. This holds promise for future applications in scenarios where data privacy and decentralization are paramount.

As we look toward future research, we believe investigating other FL strategies would enrich our understanding of these models' performance nuances. We also consider the separate classification head as a prospective area of focus, with normalization of amplitude possibly enhancing FL network performance. Finally, the exploration of deeper models like ResNet50 or ResNet101, or ViTs with more transformer blocks and deeper multi-layer perceptron layers, may alter the performance dynamics and offer fresh insights.

References

Lokman Balyen and Tunde Peto. Promising artificial intelligence-machine learning-deep learning algorithms in ophthalmology. *The Asia-Pacific Journal of Ophthalmology*, 8(3):264–272, 2019.

James M Brown, J Peter Campbell, Andrew Beers, Ken Chang, Susan Ostmo, RV Paul Chan, Jennifer Dy, Deniz Erdogmus, Stratis Ioannidis, Jayashree Kalpathy-Cramer, et al. Automated

diagnosis of plus disease in retinopathy of prematurity using deep convolutional neural networks. *JAMA ophthalmology*, 136(7):803–810, 2018.

Philippe M. Burlina, Neil Joshi, Michael Pekala, Katia D. Pacheco, David E. Freund, and Neil M. Bressler. Automated Grading of Age-Related Macular Degeneration From Color Fundus Images Using Deep Convolutional Neural Networks. *JAMA Ophthalmology*, 135 (11):1170–1176, 11 2017a. ISSN 2168-6165. doi: 10.1001/jamaophthalmol.2017.3782. URL `https://doi.org/10.1001/jamaophthalmol.2017.3782`.

Philippe M Burlina, Neil Joshi, Michael Pekala, Katia D Pacheco, David E Freund, and Neil M Bressler. Automated grading of age-related macular degeneration from color fundus images using deep convolutional neural networks. *JAMA ophthalmology*, 135(11):1170–1176, 2017b.

Xinjian Chen, Meindert Niemeijer, Li Zhang, Kyungmoo Lee, Michael D. Abramoff, and Milan Sonka. Three-dimensional segmentation of fluid-associated abnormalities in retinal oct: Probability constrained graph-search-graph-cut. *IEEE Transactions on Medical Imaging*, 31 (8):1521–1531, 2012. doi: 10.1109/TMI.2012.2191302.

Aaron S Coyner, Ryan Swan, James M Brown, Jayashree Kalpathy-Cramer, Sang Jin Kim, J Peter Campbell, Karyn Jonas, RV Paul Chan, Susan Ostmo, and Michael F Chiang. Deep learning for image quality assessment of fundus images in retinopathy of prematurity. *Investigative Ophthalmology & Visual Science*, 59(9):2762–2762, 2018.

Talisa E De Carlo, Andre Romano, Nadia K Waheed, and Jay S Duker. A review of optical coherence tomography angiography (octa). *International journal of retina and vitreous*, 1: 1–15, 2015.

Jeffrey De Fauw, Joseph R Ledsam, Bernardino Romera-Paredes, Stanislav Nikolov, Nenad Tomasev, Sam Blackwell, Harry Askham, Xavier Glorot, Brendan O'Donoghue, Daniel Visentin, et al. Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nature medicine*, 24(9):1342–1350, 2018.

Chun-Mei Feng, Yunlu Yan, Shanshan Wang, Yong Xu, Ling Shao, and Huazhu Fu. Specificity-preserving federated learning for mr image reconstruction. *IEEE Transactions on Medical Imaging*, 2022.

Sertan Kaymak and Ali Serener. Automated age-related macular degeneration and diabetic macular edema detection on oct images using deep learning. In *2018 IEEE 14th International Conference on Intelligent Computer Communication and Processing (ICCP)*, pages 265–269, 2018. doi: 10.1109/ICCP.2018.8516635.

Daniel S Kermany, Michael Goldbaum, Wenjia Cai, Carolina CS Valentim, Huiying Liang, Sally L Baxter, Alex McKeown, Ge Yang, Xiaokang Wu, Fangbing Yan, et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. *cell*, 172(5):1122–1131, 2018.

Aziz A Khanifar, Anjum F Koreishi, Joseph A Izatt, and Cynthia A Toth. Drusen ultrastructure imaging with spectral domain optical coherence tomography in age-related macular degeneration. *Ophthalmology*, 115(11):1883–1890, 2008.

Cecilia S Lee, Doug M Baughman, and Aaron Y Lee. Deep learning is effective for classifying normal versus age-related macular degeneration oct images. *Ophthalmology Retina*, 1(4): 322–327, 2017.

Mingchao Li, Yuhan Zhang, Zexuan Ji, Keren Xie, Songtao Yuan, Qinghuai Liu, and Qiang Chen. Ipn-v2 and octa-500: Methodology and dataset for retinal image segmentation. *arXiv preprint arXiv:2012.07261*, 2020a.

Tian Li, Anit Kumar Sahu, Ameet Talwalkar, and Virginia Smith. Federated learning: Challenges, methods, and future directions. *IEEE signal processing magazine*, 37(3):50–60, 2020b.

Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. *Proceedings of Machine learning and systems*, 2:429–450, 2020c.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR, 2017.

Matias Mendieta, Taojiannan Yang, Pu Wang, Minwoo Lee, Zhengming Ding, and Chen Chen. Local learning matters: Rethinking data heterogeneity in federated learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8397–8406, 2022.

A Tuan Nguyen, Philip Torr, and Ser Nam Lim. Fedsr: A simple and effective domain generalization method for federated learning. *Advances in Neural Information Processing Systems*, 35:38831–38843, 2022.

Luis Perez and Jason Wang. The effectiveness of data augmentation in image classification using deep learning. *arXiv preprint arXiv:1712.04621*, 2017.

Ramachandran Rajalakshmi, Radhakrishnan Subashini, Ranjit Mohan Anjana, and Viswanathan Mohan. Automated diabetic retinopathy detection in smartphone-based fundus photography using artificial intelligence. *Eye*, 32(6):1138–1144, 2018.

Daniel B Russakoff, Ali Lamin, Jonathan D Oakley, Adam M Dubis, and Sobha Sivaprasad. Deep learning for prediction of amd progression: a pilot study. *Investigative ophthalmology & visual science*, 60(2):712–722, 2019.

Virginia Smith, Chao-Kai Chiang, Maziar Sanjabi, and Ameet S Talwalkar. Federated multi-task learning. *Advances in neural information processing systems*, 30, 2017.

Pratul P Srinivasan, Leo A Kim, Priyatham S Mettu, Scott W Cousins, Grant M Comer, Joseph A Izatt, and Sina Farsiu. Fully automated detection of diabetic macular edema and dry age-related macular degeneration from optical coherence tomography images. *Biomedical optics express*, 5(10):3568–3577, 2014.

Daniel Shu Wei Ting, Louis R Pasquale, Lily Peng, John Peter Campbell, Aaron Y Lee, Rajiv Raman, Gavin Siew Wei Tan, Leopold Schmetterer, Pearse A Keane, and Tien Yin Wong. Artificial intelligence and deep learning in ophthalmology. *British Journal of Ophthalmology*, 103(2):167–175, 2019. ISSN 0007-1161. doi: 10.1136/bjophthalmol-2018-313173. URL `https://bjo.bmj.com/content/103/2/167`.

Elysse Tom, Pearse A. Keane, Marian Blazes, Louis R. Pasquale, Michael F. Chiang, Aaron Y. Lee, Cecilia S. Lee, and AAO Artificial Intelligence Task Force. Protecting Data Privacy in the Age of AI-Enabled Ophthalmology. *Translational Vision Science & Technology*, 9(2):36–36, 07 2020. ISSN 2164-2591. doi: 10.1167/tvst.9.2.36. URL `https://doi.org/10.1167/tvst.9.2.36`.

Maximilian Treder, Jost Lennart Lauermann, and Nicole Eter. Automated detection of exudative age-related macular degeneration in spectral domain optical coherence tomography using deep learning. *Graefe's Archive for Clinical and Experimental Ophthalmology*, 256:259–265, 2018.

Neil J. Vickers. Animal communication: When i'm calling you, will you answer too? *Current Biology*, 27(14):R713–R715, 2017. ISSN 0960-9822. doi: https://doi.org/10.1016/j.cub.2017.05.064. URL `https://www.sciencedirect.com/science/article/pii/S0960982217306309`.

Wan Ling Wong, Xinyi Su, Xiang Li, Chui Ming G Cheung, Ronald Klein, Ching-Yu Cheng, and Tien Yin Wong. Global prevalence of age-related macular degeneration and disease burden projection for 2020 and 2040: a systematic review and meta-analysis. *The Lancet Global Health*, 2(2):e106–e116, 2014. ISSN 2214-109X. doi: https://doi.org/10.1016/S2214-109X(13)70145-1. URL `https://www.sciencedirect.com/science/article/pii/S2214109X13701451`.

Jie Xu, Benjamin S Glicksberg, Chang Su, Peter Walker, Jiang Bian, and Fei Wang. Federated learning for healthcare informatics. *Journal of Healthcare Informatics Research*, 5:1–19, 2021.

Qiang Yang, Yang Liu, Tianjian Chen, and Yongxin Tong. Federated machine learning: Concept and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 10(2): 1–19, 2019.

Jason Yim, Reena Chopra, Terry Spitz, Jim Winkens, Annette Obika, Christopher Kelly, Harry Askham, Marko Lukic, Josef Huemer, Katrin Fasler, et al. Predicting conversion to wet age-related macular degeneration using deep learning. *Nature Medicine*, 26(6):892–899, 2020.