applied sciences

Article

Federated Learning for Computer-Aided Diagnosis of Glaucoma Using Retinal Fundus Images

Telmo Baptista [1,*], Carlos Soares [1,2], Tiago Oliveira [3] and Filipe Soares [1,*]

[1] Fraunhofer Portugal AICOS, Rua Alfredo Allen 455/461, 4200-135 Porto, Portugal; carlos.soares@aicos.fraunhofer.pt
[2] Faculdade de Engenharia da Universidade do Porto, Rua Dr. Roberto Frias, s/n, 4200-465 Porto, Portugal; csoares@fe.up.pt
[3] First Solutions—Sistemas de Informação S.A., Rua Conselheiro Costa Braga, 502F, 4450-102 Matosinhos, Portugal; tiago.oliveira@first-global.com
* Correspondence: telmo.baptista@aicos.fraunhofer.pt (T.B.); filipe.soares@aicos.fraunhofer.pt (F.S.)

Abstract: Deep learning approaches require a large amount of data to be transferred to centralized entities. However, this is often not a feasible option in healthcare, as it raises privacy concerns over sharing sensitive information. Federated Learning (FL) aims to address this issue by allowing machine learning without transferring the data to a centralized entity. FL has shown great potential to ensure privacy in digital healthcare while maintaining performance. Despite this, there is a lack of research on the impact of different types of data heterogeneity on the results. In this study, we research the robustness of various FL strategies on different data distributions and data quality for glaucoma diagnosis using retinal fundus images. We use RetinaQualEvaluator to generate quality labels for the datasets and then a data distributor to achieve our desired distributions. Finally, we evaluate the performance of the different strategies on local data and an independent test dataset. We observe that federated learning shows the potential to enable high-performance models without compromising sensitive data. Furthermore, we infer that FedProx is more suitable to scenarios where the distributions and quality of the data of the participating clients is diverse with less communication cost.

Keywords: federated learning; image processing; fundus images; retina quality; glaucoma

check for updates

Citation: Baptista, T.; Soares, C.; Oliveira, T.; Soares, F. Federated Learning for Computer-Aided Diagnosis of Glaucoma Using Retinal Fundus Images. *Appl. Sci.* **2023**, *13*, 11620. https://doi.org/10.3390/app132111620

Academic Editor: Hongming Shan

Received: 25 September 2023
Revised: 18 October 2023
Accepted: 23 October 2023
Published: 24 October 2023

1. Introduction

In digital healthcare, machine learning and, more specifically, deep learning (DL) models have seen growth in their descriptive and predictive capabilities. DL models, particularly convolutional neural networks, have been widely explored for medical imaging to process and analyze various imaging techniques, such as MRI [1], CT scans [2], X-rays [3] and ultrasounds [4]. Physicians use these images to detect and diagnose numerous diseases like cancer and cardiovascular and eye diseases. In ophthalmology, retinal fundus photography (RFP) and, more recently, optical coherence tomography (OCT) are used to screen for the main eye diseases—diabetic retinopathy, glaucoma and age-related macular degeneration [5]. The vast amount of data and the diversity of the diseases make it difficult for physicians to analyse the images manually. Therefore, developing these ML models has become increasingly important for assisting physicians and improving the diagnosis process.

However, DL models rely on access to a large amount of data, often centralized, which is not always feasible. Hospitals and institutions must comply with the General Data Protection Regulation (GDPR) and other privacy regulations, which restrict the sharing of patient data [6]. Additionally, approval by institutional review boards is required to determine to what extent data can be shared with peers, and anonymization of data is often required. Data anonymization is a complex and time-consuming task and needs to ensure that patient identification cannot be obtained from the data. Nevertheless, studies have already shown the possibility of reconstruction of identification from anonymized data [7], resulting in a critical liability regarding the privacy of sensitive information. In this context, there is a need for solutions that enable the development of ML models in a distributed, privacy-preserving manner.

Federated Learning (FL) is a machine learning paradigm that enables the development of ML models in a distributed manner without data sharing. To achieve this, each federation participant, often called a client, trains a model on its local data. Afterwards, each client sends the model weights to a central server, which aggregates the various local models received from the clients to generate a global model. For this reason, FL has been gaining traction in digital healthcare as one of the privacy-preserving methods for artificial intelligence (AI) in medical applications, as it enables the possibility of training a model that learns on data from various data centres without requiring any of the data to leave each data centre's network. Furthermore, FL can be further complemented with other privacy-preserving techniques, such as differential privacy [8], which enable a higher degree of privacy of the sensitive data.

The server uses the FL algorithm or federated strategy to generate the global model, which combines the clients' model weights. For example, one possible algorithm is to perform a weighted average of the clients' model weights, which denotes the Federated Average (FedAvg) algorithm [9]. One of the biggest challenges in FL stands on the heterogeneity of the clients, which can be caused by differences in the types of data, amounts of data, quality of data, etc. In this context, FL algorithms must adapt to the heterogeneity of the clients, and multiple algorithms and variations have been proposed to tackle this challenge. As such, choosing the FL algorithm is a crucial step that needs to be considered and depends on the type of heterogeneity present. However, there is a gap in the literature regarding the extent to which these algorithms are affected by the heterogeneity of the clients, whether in terms of data distribution, data quality or other factors.

This paper aims to study and analyse (a) the federated approach compared to the centralized approach; (b) the impact of heterogeneous data, such as different label distributions and varying data quality, on the performance of federated approaches; and (c) which FL is the most suitable given the nature of the data and the objective of the model. Various experiments are conducted in varied settings regarding (i) label distribution on the client's local datasets, (ii) data quality of said datasets and (iii) the federated strategy used by the server to aggregate the models received from the clients. We also evaluate non-federated approaches to serve as comparison points for the federated approach in order to validate the results obtained and support the contributions of this work.

The rest of the paper is organized as follows: Section 2 outlines the relevant work present in the literature; Section 3 provides an overview of the methodology, a description of all the tools and datasets used, the implementation details of the federated setup and the experiments that are performed; in Section 4, the results obtained are analysed; this is followed by a discussion in Section 5. Finally, Section 7 summarizes what was discussed in this paper and our main conclusions.

2. Related Work

In a setting where data sharing is heavily regulated for ethical and legal reasons, centralising patient data to develop medically significant machine learning models is often impossible or has several limitations. FL is one of the approaches to address these issues. Several studies on FL applied to the medical field have been done. Roth et al. [10] developed a federated approach to breast density classification using a weighted averaging strategy (FedAvg) and compared results against locally trained models at each institution. The federated model obtained better agreement on classifications between institutions than the models trained on local data. However, the model's performance is still not in an acceptable state to be used due to the heterogeneity of the data across the institutions. One of the possible reasons is the strategy used, as FedAvg assumes uniformly distributed data across the institutions. Thus, this research could be complemented with other strategies to address this issue. Ho et al. [11] apply FL for COVID-19 detection using X-ray images and information about the symptoms. The authors study the performance of the FL model in various settings, such as IID and non-IID settings, with varying numbers of participants and underlying DL models. The authors also combine FL with the differential privacy method to achieve a more secure system while evaluating model performance trade-offs. FL has also been applied to brain tumour segmentation [12,13]. Sheller et al. [13] evaluate the performance of FL comparatively to a data-sharing centralized approach and other collaborative ML approaches. In their work, the authors show the potential of FL over other collaborative ML approaches in a real-world setting. Similarly, Li et al. [12] also evaluated the performance of FL comparatively to a centralized approach. Furthermore, the authors study the impact of the percentage of the model shared on the performance.

In ophthalmology, Lo et al. [14] evaluate the performance of FL for microvasculature segmentation and diabetic retinopathy detection using OCT and OCTA en face images. The authors study and validate the potential of FL to adapt to settings with a small amount of data available across clients. FL has also been applied to RFP [15–18]. Lu et al. [17] evaluate the performance of FL compared to centralized learning for detecting retinopathy of prematurity (ROP) using fundus images. In the same domain, Hanif et al. [18] propose an FL approach that generates the diagnosis's vascular severity score (VSS). The authors studied the variations in the VSS for the different categories of ROP. In this study, the authors found significant differences in the severity scores between the early stage and the advanced stage of ROP despite not showing any variance when comparing the intermediate stage with the advanced stage.

Nielsen et al. [16] investigate the vulnerability of FL to gradient inversion attacks. These attacks exploit the fact that gradients carry some information about the data that influenced their calculation, and manipulation of the model's input can enable the extraction of information regarding the training data. In the paper, the authors evaluate the possibility of such attacks for reconstructing RFP images from an FL model used to classify diabetic retinopathy. The authors show the possibility of reconstructing clinical features to identify an individual from gradient information transmitted during the federation.

While these studies provide a great foundation for the potential of FL in heterogeneous environments for medical imaging, several challenges still need to be addressed. First, the performance of DL models trained on local data is often not inferior to that of FL models. This presents a drawback to adopting FL, as institutions may not benefit from the collaboration, namely participants with large amounts of high-quality data, especially regarding the potential leakage of sensitive information via the FL model. Furthermore, there is a gap in the literature on the impact of bad-quality participants and how to restrict them from the FL process.

3. Materials and Methods

3.1. Overview

In this study, we developed a horizontal cross-silo federated learning approach applied to glaucoma screening from fundus images with three independent clients. This is an approach where the clients have the same feature space but with different data samples (horizontal) and where each client has greater computational power and larger amounts of data (cross-silo). This approach can be applied to scenarios that aim to represent real-world institutions with varying data distributions (here, this term refers to how the data are distributed across the clients, not statistical distribution). The methodology can be divided into two main modules: (i) Data Preparation and Distribution: information about the data from a collection of datasets is gathered and then distributed to different clients for the federated setup and (ii) Federated Setup: where multiple strategies for model aggregation are evaluated. Figure 1 represents, in a compact form, our methodology process.

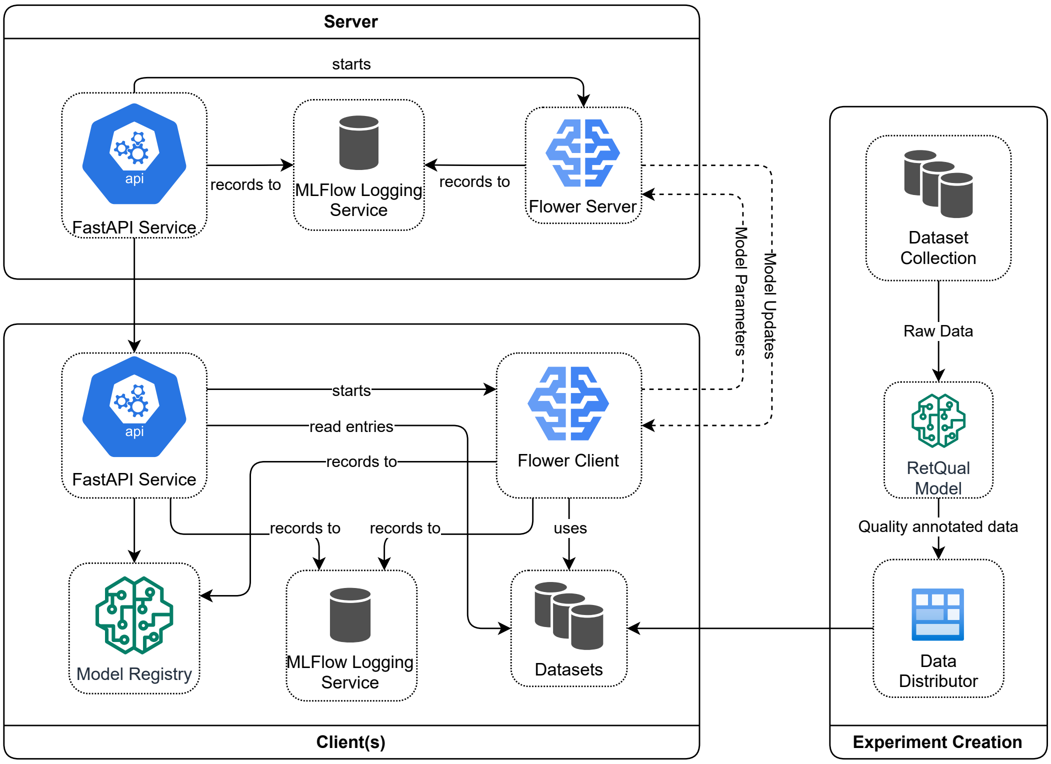

Figure 1. Architecture diagram for the methodology proposed.

In the following sections, we provide more details of each module.

3.2. Data Preparation and Distribution

For our study, we used a collection of public datasets with varying regions of collection, equipment used, image quality and position of eye structures: ORIGA [19], REFUGE [20], GAMMA [21], Drishti [22] and EyePACS AIROGS [23]. The image quality labels for the datasets were then inferred using the RetinaQualEvaluator model Leonardo et al. [24], which attributes a label of "Reject", "Usable" or "Good" to each sample.

Due to the size of AIROGS compared to the remaining datasets (containing 98.4% of the data), in this study, this dataset is only used to complement the other datasets to form the different experiments. After distribution, the samples that remain from the AIROGS dataset are then used to compose an additional dataset to serve as an independent test dataset for the federated setting. As such, a data distributor component was developed to combine one or more base datasets (excluding AIROGS). Afterwards, the distributor uses the AIROGS dataset to complement the base datasets to achieve a target label and quality distribution, noting that data from the complementary dataset cannot be repeated across clients.

Using this distributor, we formed two data distributions for each client in the following manner:

1. In the first distribution, which we label "Dataset Heavily Diverse (Dataset-HD)", we aim to study the effect of heavily heterogeneous clients on the performance of the federated approach. To achieve such, all clients have different label and quality distributions: (a) Client 1's dataset comprises mostly negative cases (80%) and high-quality images; (b) Client 2's dataset has an even label distribution (50%) and medium-quality images; (c) Client 3's dataset serves as the polar opposite of Client 1, mainly comprising positive cases (80%) and lower-quality images.
2. For the second distribution, which we label "Dataset Quality Diverse (Dataset-QD)", we aim to study the effect of different data quality distributions across the clients while following a real-world distribution of glaucomatous cases. For such, we follow

the glaucoma incidence observed in screening clinics [25], where around 20% of the adults suffer from glaucoma or are suspects. As such, all clients' distributions contain 80% of the cases being negative, while the data quality distribution remains similar to that of the first distribution.

Client 1's dataset is the same in both distributions and is used as a common comparison point between the two experiments. Table 1 presents the distributions and the datasets forming each client's data.

Table 1. Data distribution parameters across the different clients.

Target Dataset	Dataset	Base Dataset	Complementary Dataset	Target Samples	Target Negative Labels	Target Positive Labels
Client 1	Dataset-HD	REFUGE	EyePACS AIROGS	2000	80%	20%
Dataset	Dataset-QD	REFUGE	EyePACS AIROGS	2000	80%	20%
Client 2	Dataset-HD	ORIGA + GAMMA	EyePACS AIROGS	2000	50%	50%
Dataset	Dataset-QD	ORIGA + GAMMA	EyePACS AIROGS	2000	80%	20%
Client 3	Dataset-HD	Drishti	EyePACS AIROGS	2000	20%	80%
Dataset	Dataset-QD	Drishti	EyePACS AIROGS	2000	80%	20%
Outsider	Dataset-HD	EyePACS AIROGS	-	-	50%	50%
Dataset	Dataset-QD	EyePACS AIROGS	-	-	50%	50%

3.3. Federated Setup

The federated setup comprises three identical Jetson AGX Xavier setups representing the three independent institutions in the federated network. The server of the federated network is also an institution independent from the three clients above. However, due to restrictions on the network, one of the Jetsons had to act as both a server and a client. Despite the restriction, this does not hinder any performance of the machine, and the server and client running on the same machine are still treated as independent processes.

Each client only has access to its dataset located on the Jetson and has no access or information regarding any other clients' datasets. The server, meanwhile, has no access to any information on any of the clients' datasets other than any information communicated during federation, such as the number of samples trained. The model trained on each client is also only available to the client itself, and only the model weights, either totally or partially depending on the federated strategy, are shared with the server.

For the model training, each dataset was divided into three splits: training (60%), validation (20%) and test (20%). The model developed by Leonardo et al. [24] serves as the baseline and initiation model for the federated setup. The federation was implemented using Flower (https://flower.dev/ accessed on 22 October 2023) as the backend framework responsible for the communication protocols between the server and the clients, which is started by our API. In this study, we experimented with various federated strategies: FedAvg [9], FedSGD [9], FedProx [26], FedBN [27] and FedYogi [28]. The first two strategies are part of one of the initial studies done on FL and serve as the standard baseline for FL strategies. The remaining three were chosen as they have often been mentioned and used as benchmarks for multiple studies [29–32]. FedAvg aggregates the clients' weights by performing a weighted average to generate the new global model. FedSGD works in the same way as FedAvg but only performs a single local epoch (step) in each federated round. FedProx introduces one key difference compared to FedAvg by adding a proximal term to the model optimizer in order to limit the divergence of the local model from the global model in each federated round. Similarly, FedBN introduces one key difference to FedAvg by excluding the batch normalization layers from being sent to the server for aggregation to alleviate feature shifts. Finally, FedYogi adapts the Yogi optimizer into a federated setting. With regards to the strategies' hyperparameters, we varied FedProx's proximal term (μ) between 1.0, 0.01 and 0.1, following the values from the authors. FedYogi also had its server learning rate parameter (η) varied between 0.01 and 0.001. Alongside the federated approach, we also conducted experiments with non-federated approaches, for which the hyperparameters were not changed: (i) local-only, where we train a local model on each client using solely its own local dataset; and (ii) centralized, where we combine all the clients' data into a single data centre and use them for training the model.

Furthermore, to evaluate the impact of image quality on the training of a federated model, each strategy was trained twice: (i) with all the images from the dataset available and (ii) where images labelled "Reject" were excluded from the training process.

The local training process was implemented with Tensorflow (https://www.tensorflow.org/ accessed on 22 October 2023). The model was trained with binary cross-entropy loss and the Adam optimizer with a learning rate of 2.5×10^{-5}, decay of 0.0 and first and second momentum estimates of 0.9 and 0.999, respectively. Each strategy was trained for five server rounds with 25 local epochs. To compensate for the single epoch per round required for FedSGD, this strategy was trained for 125 server rounds instead. On the same note, local-only and centralized models were trained for 125 local epochs due to the nonexistence of server rounds. To aid in tracking all the experiment parameters, models and metrics, we report all the results to an MLFlow (https://mlflow.org/ accessed on 22 October 2023) service that we use to analyse the results obtained.

4. Results

To evaluate the performance of the federated strategies, we first analyse the performance of each strategy on a specific client in detail, along with the performance on the outsider test dataset. Afterwards, we analyse the performance of the federated strategies on a global overview across the various clients.

Table 2 shows the strategies' performance on Client 3's local and outsider test datasets. We highlight in orange the best performance between the centralized and the local-only models—representing the non-federated scenario—and the best-performing strategies overall are marked with a green highlight. All federated strategies, except FedYogi, performed similarly, with only marginal differences. Despite FedYogi presenting similar convergence to those of the remaining strategies, the strategy diverged in one of the variations, labelling every image as a negative case. This divergence was likely caused by a loss explosion due to a higher server learning rate parameter, $\eta = 0.01$, as we observed non-divergence with a lower server learning rate, $\eta = 0.001$, albeit with lower performance than the other strategies. The variations to the hyperparameter of FedProx (μ) did not result in significant differences in performance, and the best-performing value for the proximal term varied depending on what client and what dataset were to be used. Nevertheless, based on our experiments, this hyperparameter optimization for FedProx can improve the performance by 1%, and in other scenarios, it would require another search for the optimal value.

Despite faster convergence on the training and validation datasets, the local-only model's performance is lower than that of all of federated strategies. The performance gain from the federation ranges from a marginal difference for clients with high-quality data, such as Client 1, to a boost of 2–6% as observed in Client 3, which contains lower-quality images, on both experiments. These gains can be seen in Table 3, where the best-performing strategy is marked with a green highlight for each client and experiment.

Another noteworthy metric to analyse is the performance of the different strategies on the local test dataset and the outsider test dataset. From this, we can identify the strategies that fit better into a more heterogeneous environment from those that better fit into a more controlled environment. In Tables 2 and 3, we can notice two trends:

1. FedBN tends to achieve higher performance on the local test dataset than other strategies. This is noticeable in the second experiment, where the distributions across the clients are similar and the heterogeneity of the datasets comes from the varying quality (i.e., due to different capturing equipment). This scenario aligns with the objective of FedBN, which uses the batch normalization layers kept as local-only information to alleviate feature shifts. However, this strategy falls behind in the experiment with multiple types of heterogeneity, where the data's distribution and quality varied across every client.

2. FedSGD and FedProx achieve the most generalizable models in both experiments. FedSGD achieves this by aggregating the weights more often and thus preventing the models from diverging too much from the global model. This comes at the cost of larger training time due to the increased time spent communicating the model. FedProx achieves the same result by adding the proximal term, which prevents the model from diverging too much from the global model. However, the FedProx approach does not result in increased communication costs.

Table 2. Performance of the different strategies on the local and outsider test datasets. In orange, the best non-federated approach. In green, the best federated approach.

		Local Performance					**Outsider Performance**				
Strategy	**Fit Time (h)**	**Accuracy**	**Precision**	**Recall**	**F1 Score**	**Specificity**	**Accuracy**	**Precision**	**Recall**	**F1 Score**	**Specificity**
Baseline	-	0.407	0.781	0.382	0.513	0.521	0.459	0.444	0.326	0.376	0.591
Local-Only	**1.8**	0.850	0.899	**0.920**	0.909	0.534	0.754	**0.712**	0.854	0.777	**0.655**
Centralized	3.9	**0.870**	**0.931**	0.908	**0.920**	**0.699**	**0.779**	0.711	**0.940**	**0.810**	0.618
FedAvg	2.2	0.860	**0.928**	0.908	0.913	**0.685**	0.761	0.693	0.934	0.796	0.587
FedSGD	3.3	0.860	0.925	0.902	0.913	0.671	0.763	0.695	**0.939**	**0.799**	0.588
FedBN	2.2	**0.868**	**0.928**	0.908	**0.918**	**0.685**	0.760	0.692	0.936	0.796	0.584
FedProx x.0 $\mu = 1.0$	2.3	0.863	**0.928**	0.902	0.915	**0.685**	0.760	0.693	0.933	0.795	0.587
FedProx x.1 $\mu = 0.01$	2.3	0.858	0.925	0.899	0.912	0.671	0.764	0.697	0.933	0.798	0.595
FedProx x.2 $\mu = 0.1$	2.3	0.858	0.925	0.899	0.912	0.671	0.758	0.692	0.930	0.793	0.585
FedYogi x.0 $\eta = 0.01$	2.4			Diverged					-		
FedYogi x.1 $\eta = 0.001$	2.4	0.752	0.837	0.865	0.851	0.247	0.534	0.52	0.894	0.658	0.174

Performance was measured for the third client using Dataset-HD and no filtering during training.

Table 3. Accuracy achieved on the test datasets for each federated strategy. In orange, the best non-federated approach. In green, the best federated approach.

		Dataset-HD				**Dataset-QD**			
Strategy	**Filter**	**Client 1**	**Client 2**	**Client 3**	**Outsider**	**Client 1**	**Client 2**	**Client 3**	**Outsider**
Baseline	-	0.368	0.522	0.530	0.468	0.368	0.472	0.407	0.459
Local-Only	✗	0.825	0.752	0.738	0.776 [2]	0.827	0.817	0.850	0.754 [2]
	✓	0.820	0.738	0.760	0.783 [2]	0.822	0.813	0.822	0.728 [2]
Centralized	✗	**0.842**	**0.777**	**0.803**	**0.839**	**0.873**	0.832	0.870	**0.779**
	✓	0.832	0.760	**0.803**	0.837	0.868	**0.842**	**0.873**	0.771
FedAvg	✗	**0.827**	0.762	0.772	0.810	0.868	0.827	0.860	0.761
	✓	0.815	0.748	0.762	0.809	0.884	0.820	0.863	0.759
FedSGD	✗	0.825	0.762	**0.788**	**0.817**	0.860	**0.830**	0.860	0.763
	✓	**0.827**	0.748	0.765	**0.817**	0.848	0.825	0.863	**0.764**
FedBN	✗	0.820	0.762	0.772	0.810	0.868	**0.830**	**0.868**	0.760
	✓	0.820	0.750	0.760	0.810	0.845	0.817	0.863	0.760
FedProx [1]	✗	**0.827**	**0.772**	0.777	**0.817**	**0.873**	0.827	0.863	**0.764**
	✓	0.817	0.755	0.767	0.807	0.850	0.825	0.863	0.758
FedYogi [1]	✗	0.705	0.613	0.485	0.580	0.772	0.740	0.752	0.534
	✓	0.700	0.610	0.468	0.569	0.767	0.735	0.740	0.528

[1] For federated strategies with multiple experiments, only the best is represented in the table. [2] The local-only performance on the outsider was obtained by evaluating each client's local-only model on the outsider dataset. The model that obtained the best results is represented here.

Despite not achieving the best performance, FedAvg remained close to every strategy in all scenarios, making it a viable choice in scenarios where there is a lack of information regarding the data without sacrificing performance or increasing training costs. Additionally, we also observed that adding a quality filter for training did not significantly affect the performance of the models in either direction and can be useful to cut down training times, as we observed up to 20% reduction in the total training time. However, this addition may require additional information regarding the amount of data the client has and its quality. Thus, depending on the amount of data available on a client after filtering, the filter can negatively affect the performance, as observed for Client 3 on Dataset-HD, which had a reduction of 24% in the number of training samples and resulted in a drop of 3% on average on the local test dataset. The global model can also suffer a negative impact by adding this filter. However, we only noticed significant changes in the FedProx approach.

We can also analyse the impact of the heterogeneity of the clients on the performance with Client 1's test dataset. As mentioned in the experimental setup, Client 1's dataset is the same across both experiments. We can observe a significant increase of 6% in the model performance when the distributions on the target label were similar across the clients, as shown in Table 3. Clients in both experiments had similar quality distributions: Client 1 presented higher-quality data and Client 3 had lower-quality data. In the first experiment, where the clients had diverse distributions on the target labels, FedProx achieved an accuracy of 0.827. In contrast, FedProx achieved an accuracy of 0.873 in the second experiment, where the clients had similar distributions on the target label. This difference can be attributed to the variance introduced by the clients' heterogeneous data distributions on the local models aggregated to generate the global model.

At last, we can also compare the federated approach to the centralized one. The centralized approach achieved higher or equal performance in all scenarios, with gains ranging from under 1% to around 6%. However, this requires data sharing among the clients to create a centralized data source, which is often not possible due to restrictions discussed earlier in this document. The federated approach essentially trains on the same amount of data. Still, due to its distributed nature, the variance between the local models causes this performance drop—but with the advantage of maintaining data integrity regarding privacy and security.

5. Discussion

In this section, we summarise and address the results obtained, considering the objectives of this paper. Firstly, we analysed the convergence of the various strategies and observed a similar pattern for all strategies except for FedYogi, which diverged in one experiment and had a significantly slower convergence in the other. We also observed that the centralized and local-only models converge faster than all the federated models and that early stopping does not negatively affect the performance of the models.

Despite this, we later observed that the performance of the local-only model on the test datasets was lower than that of most federated strategies, especially on clients whose local data were comprised of lower-quality images. From our results, we can answer our first objective (a) in Section 1, concluding that employing a federated approach does pose a benefit to the model performance and its generalizability compared to training a model with only the client's local data, as it essentially allows the model to learn from a larger pool of data while not compromising the privacy of such data. This approach can also benefit clients with higher-quality datasets, albeit the differences in those cases are more marginal. And while the centralized approach still had the overall best performance, federated approaches could reach performances close to that of the centralized method while maintaining data security and isolation in their own client's network.

Furthermore, we also analysed the performance differences of each strategy for each of the experiments, which allows us to answer our second objective (b) in Section 1. We observed that the model's performance is lower if trained in a more heterogeneous environment, as seen in Client 1's tests, the data for which remained identical across experiments. However, eliminating lower-quality images from the training process to achieve a more consistent quality distribution across clients does not always benefit the model's performance, as we observed in our experiments. With this, the impact caused by having clients with diverse distributions during training can degrade the model performance, as observed with Client 1, with some strategies being more affected than others.

Indeed, the last observation from our experiments is the adaptability of specific strategies to each context, which answers our final objective of the paper (c) in Section 1. We observed that FedBN generally performs the best given a scenario wherein the federated model's goal is to optimise the performance on the client's local data given that those clients follow similar distributions with their data. On the other hand, FedSGD and FedProx exhibited higher performance under scenarios wherein the clients' data are more diverse. These strategies also showed higher generalizability, as we can observe by the performances on the outsider dataset. Based on the results, we conclude that FedSGD and FedProx are the most suitable strategies in scenarios for which the clients' data are more diverse or if said distributions are unknown. This choice also applies if the objective is to obtain a more generalizable model. On the other hand, we opt to choose FedBN to optimise the model for the client's local data. FedAvg poses a solid choice in both scenarios, albeit not optimal.

6. Limitations

The experiments consist of a limited set of federated strategies and require further evaluation using more recent approaches, such as incorporating reinforcement learning to learn the aggregation of the clients' models [30] and incorporation of domain generalization and personalization approaches into the federation [32,33].

The incorporation of cases from the AIROGS dataset into the clients can introduce a common ground for the clients' data characteristics, and an evaluation with the client's data consisting of a single data source, regardless of its size, should be conducted to validate the potential of federated learning further.

This study was conducted under the area of application of glaucoma diagnosis using retinal fundus images with a single underlying DL model, and for future work, we also intend to evaluate this approach using a different image modality, such as OCT scans, and for different tasks.

The API developed lacks SSL (Secure Sockets Layer) certification and encryption support, which is necessary when deploying this system with real-world institutions and data.

7. Conclusions

FL approaches show great potential in digital healthcare as they ensure a more secure environment when working with sensitive data. Research on this topic has been rapidly growing in the past years, and we expect it to grow even further with frameworks such as NVIDIA Flare [34], which aims to provide a scalable framework for federated learning, along with new algorithms to generate more optimal global models.

In this paper, we aimed to study various FL algorithms and analyse the impact of heterogeneous clients on the performance of a federated approach to glaucoma diagnosis. For this, we conducted various experiments using public datasets of retinal fundus images from different sources to create clients with different characteristics. The model developed by Leonardo et al. [24] was used as the base model for the federated setup. We evaluated different FL algorithms in these settings and analysed the results obtained.

We concluded that the use of federated learning in this area of application resulted in an improvement in the model performance over the local-only approach while also achieving comparable performance to the centralized model without requiring the sharing of sensitive data, supporting federated learning as a viable method for enabling privacy-preserving machine learning without negatively impacting the capabilities of the trained models. Additionally, we extracted some scenarios where a specific FL algorithm is more suitable to be deployed, given the objective of the model and the nature of the data of the clients in the federation. In a scenario where the model's generalisation is crucial, FedProx would be a suitable choice based on our experiments and results achieved on the outsider datasets.

Author Contributions: Conceptualization, F.S.; methodology, T.B., F.S. and C.S.; software, T.B.; validation, T.B. and F.S.; formal analysis, T.B., F.S. and C.S.; investigation, T.B. and F.S.; data curation, T.B. and F.S.; writing—review and editing, T.B., F.S., C.S. and T.O.; supervision, F.S. and C.S.; project administration, F.S.; funding acquisition, T.O. and F.S. All authors have read and agreed to the published version of the manuscript.

Funding: This work was financially supported by the project Transparent Artificial Medical Intelligence (TAMI) co-funded by Portugal 2020 framed under the Operational Programme for Competitiveness and Internationalisation (COMPETE 2020), Fundação para a Ciência and Technology (FCT), Carnegie Mellon University, and the European Regional Development Fund under grant NORTE-01-0247-FEDER-045905.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: All retinal fundus images used in this study belong to public datasets. This data can be found in the aforementioned references.

Acknowledgments: The authors would like to thank João Gonçalves for helping with the hardware setup.

Conflicts of Interest: The authors declare that they have no competing interests.

Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial Intelligence
FL	Federated Learning
FedAvg	Federated Averaging
FedBN	Federated Batch Normalization
FedSDG	Federated Stochastic Gradient Descent
OCT	Optical Coherence Tomography
RFP	Retinal Fundus Photograph
SSL	Secure Sockets Layer

References

1. Magadza, T.; Viriri, S. Deep Learning for Brain Tumor Segmentation: A Survey of State-of-the-Art. *J. Imaging* **2021**, *7*, 19. [CrossRef]
2. Serte, S.; Demirel, H. Deep learning for diagnosis of COVID-19 using 3D CT scans. *Comput. Biol. Med.* **2021**, *132*, 104306. [CrossRef]
3. Lehman, C.D.; Yala, A.; Schuster, T.; Dontchos, B.; Bahl, M.; Swanson, K.; Barzilay, R. Mammographic Breast Density Assessment Using Deep Learning: Clinical Implementation. *Radiology* **2019**, *290*, 52–58. [CrossRef] [PubMed]
4. Liu, S.; Wang, Y.; Yang, X.; Lei, B.; Liu, L.; Li, S.X.; Ni, D.; Wang, T. Deep Learning in Medical Ultrasound Analysis: A Review. *Engineering* **2019**, *5*, 261–275. [CrossRef]
5. Prathiba, V.; Rema, M. Teleophthalmology: A Model for Eye Care Delivery in Rural and Underserved Areas of India. *Int. J. Fam. Med.* **2011**, *2011*, 683267. [CrossRef] [PubMed]
6. Yuan, B.; Li, J. The Policy Effect of the General Data Protection Regulation (GDPR) on the Digital Public Health Sector in the European Union: An Empirical Investigation. *Int. J. Environ. Res. Public Health* **2019**, *16*, 1070. [CrossRef] [PubMed]
7. Rocher, L.; Hendrickx, J.M.; de Montjoye, Y.A. Estimating the success of re-identifications in incomplete datasets using generative models. *Nat. Commun.* **2019**, *10*, 3069. [CrossRef] [PubMed]
8. McMahan, H.B.; Ramage, D.; Talwar, K.; Zhang, L. Learning Differentially Private Recurrent Language Models. *arXiv* **2018**, arXiv:1710.06963.
9. McMahan, H.B.; Moore, E.; Ramage, D.; Hampson, S.; Arcas, B.A.Y. Communication-Efficient Learning of Deep Networks from Decentralized Data. *arXiv* **2017**, arXiv:1602.05629.
10. Roth, H.R.; Chang, K.; Singh, P.; Neumark, N.; Li, W.; Gupta, V.; Gupta, S.; Qu, L.; Ihsani, A.; Bizzo, B.C.; et al. Federated Learning for Breast Density Classification: A Real-World Implementation. In *Domain Adaptation and Representation Transfer, and Distributed and Collaborative Learning* ; Springer: Cham, Switzerland, 2020; Volume 12444, pp. 181–191. [CrossRef]

11. Ho, T.T.; Tran, K.D.; Huang, Y. FedSGDCOVID: Federated SGD COVID-19 Detection under Local Differential Privacy Using Chest X-ray Images and Symptom Information. *Sensors* **2022**, *22*, 3728. [CrossRef]
12. Li, W.; Milletarì, F.; Xu, D.; Rieke, N.; Hancox, J.; Zhu, W.; Baust, M.; Cheng, Y.; Ourselin, S.; Cardoso, M.J.; et al. Privacy-preserving Federated Brain Tumour Segmentation. *arXiv* **2019**, arXiv:1910.00962.
13. Sheller, M.J.; Reina, G.A.; Edwards, B.; Martin, J.; Bakas, S. Multi-Institutional Deep Learning Modeling Without Sharing Patient Data: A Feasibility Study on Brain Tumor Segmentation. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke Trauma and Brain Injuries. BrainLes (Workshop)*; Springer: Cham, Switzerland, 2019; Volume 11383, pp. 92–104. [CrossRef]
14. Lo, J.; Yu, T.T.; Ma, D.; Zang, P.; Owen, J.P.; Zhang, Q.; Wang, R.K.; Beg, M.F.; Lee, A.Y.; Jia, Y.; et al. Federated Learning for Microvasculature Segmentation and Diabetic Retinopathy Classification of OCT Data. *Ophthalmol. Sci.* **2021**, *1*, 100069. [CrossRef]
15. Soni, M.; Singh, N.K.; Das, P.; Shabaz, M.; Shukla, P.K.; Sarkar, P.; Singh, S.; Keshta, I.; Rizwan, A. IoT-Based Federated Learning Model for Hypertensive Retinopathy Lesions Classification. *IEEE Trans. Comput. Soc. Syst.* **2022**, *10*, 1722–1731. [CrossRef]
16. Nielsen, C.; Tuladhar, A.; Forkert, N.D. Investigating the Vulnerability of Federated Learning-Based Diabetic Retinopathy Grade Classification to Gradient Inversion Attacks. In *Ophthalmic Medical Image Analysis*; Antony, B., Fu, H., Lee, C.S., MacGillivray, T., Xu, Y., Zheng, Y., Eds.; Lecture Notes in Computer Science; Springer: Cham, Switzerland, 2022; pp. 183–192. [CrossRef]
17. Lu, C.; Hanif, A.; Singh, P.; Chang, K.; Coyner, A.S.; Brown, J.M.; Ostmo, S.; Chan, R.V.P.; Rubin, D.; Chiang, M.F.; et al. Federated Learning for Multicenter Collaboration in Ophthalmology: Improving Classification Performance in Retinopathy of Prematurity. *Ophthalmol. Retin.* **2022**, *6*, 657–663. [CrossRef]
18. Hanif, A.; Lu, C.; Chang, K.; Singh, P.; Coyner, A.S.; Brown, J.M.; Ostmo, S.; Chan, R.V.P.; Rubin, D.; Chiang, M.F.; et al. Federated Learning for Multicenter Collaboration in Ophthalmology: Implications for Clinical Diagnosis and Disease Epidemiology. *Ophthalmol. Retin.* **2022**, *6*, 650–656. [CrossRef] [PubMed]
19. Zhang, Z.; Yin, F.S.; Liu, J.; Wong, W.K.; Tan, N.M.; Lee, B.H.; Cheng, J.; Wong, T.Y. ORIGA-light: An online retinal fundus image database for glaucoma analysis and research. In Proceedings of the 2010 Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBC'10, Buenos Aires, Argentina, 31 August–4 September 2010; pp. 3065–3068. [CrossRef]
20. Orlando, J.I.; Fu, H.; Barbosa Breda, J.; van Keer, K.; Bathula, D.R.; Diaz-Pinto, A.; Fang, R.; Heng, P.A.; Kim, J.; Lee, J.; et al. REFUGE Challenge: A unified framework for evaluating automated methods for glaucoma assessment from fundus photographs. *Med. Image Anal.* **2020**, *59*, 101570. [CrossRef] [PubMed]
21. Wu, J.; Fang, H.; Li, F.; Fu, H.; Lin, F.; Li, J.; Huang, L.; Yu, Q.; Song, S.; Xu, X.; et al. GAMMA Challenge:Glaucoma grAding from Multi-Modality imAges. *arXiv* **2022**, arXiv:2202.06511.
22. Sivaswamy, J.; Krishnadas, S.R.; Datt Joshi, G.; Jain, M.; Syed Tabish, A.U. Drishti-GS: Retinal image dataset for optic nerve head(ONH) segmentation. In Proceedings of the 2014 IEEE 11th International Symposium on Biomedical Imaging (ISBI), Beijing, China, 29 April–2 May 2014; pp. 53–56. [CrossRef]
23. de Vente, C.; Vermeer, K.A.; Jaccard, N.; Wang, H.; Sun, H.; Khader, F.; Truhn, D.; Aimyshev, T.; Zhanibekuly, Y.; Le, T.D.; et al. AIROGS: Artificial Intelligence for RObust Glaucoma Screening Challenge. *arXiv* **2023**, arXiv:2302.01738.
24. Leonardo, R.; Gonçalves, J.; Carreiro, A.; Simões, B.; Oliveira, T.; Soares, F. Impact of Generative Modeling for Fundus Image Augmentation With Improved and Degraded Quality in the Classification of Glaucoma. *IEEE Access* **2022**, *10*, 111636–111649. [CrossRef]
25. Daba, K.T.; Gessesse, G.W.; Sori, S.B. Proportion of Glaucoma among Voluntary People Coming for Glaucoma Screening Program at Jimma University Department of Ophthalmology, Jimma, Ethiopia. *Ethiop. J. Health Sci.* **2020**, *30*, 13–22. [CrossRef]
26. Li, T.; Sahu, A.K.; Zaheer, M.; Sanjabi, M.; Talwalkar, A.; Smith, V. Federated Optimization in Heterogeneous Networks. *arXiv* **2020**, arXiv:1812.06127.
27. Li, X.; Jiang, M.; Zhang, X.; Kamp, M.; Dou, Q. FedBN: Federated Learning on Non-IID Features via Local Batch Normalization. *arXiv* **2021**, arXiv:2102.07623.
28. Reddi, S.; Charles, Z.; Zaheer, M.; Garrett, Z.; Rush, K.; Konečný, J.; Kumar, S.; McMahan, H.B. Adaptive Federated Optimization. *arXiv* **2021**, arXiv:2003.00295.
29. Xia, Y.; Yang, D.; Li, W.; Myronenko, A.; Xu, D.; Obinata, H.; Mori, H.; An, P.; Harmon, S.; Turkbey, E.; et al. Auto-FedAvg: Learnable Federated Averaging for Multi-Institutional Medical Image Segmentation. *arXiv* **2021**, arXiv:2104.10195. https://doi.org/10.48550/arXiv.2104.10195.
30. Guo, P.; Yang, D.; Hatamizadeh, A.; Xu, A.; Xu, Z.; Li, W.; Zhao, C.; Xu, D.; Harmon, S.; Turkbey, E.; et al. Auto-FedRL: Federated Hyperparameter Optimization for Multi-institutional Medical Image Segmentation. *arXiv* **2022**, arXiv:2203.06338.
31. Gunesli, G.N.; Bilal, M.; Raza, S.E.A.; Rajpoot, N.M. FedDropoutAvg: Generalizable federated learning for histopathology image classification. *arXiv* **2021**, arXiv:2111.13230.
32. Jiang, M.; Yang, H.; Cheng, C.; Dou, Q. IOP-FL: Inside-Outside Personalization for Federated Medical Image Segmentation. *arXiv* **2022**, arXiv:2204.08467.

33. Liu, Q.; Chen, C.; Qin, J.; Dou, Q.; Heng, P.A. FedDG: Federated Domain Generalization on Medical Image Segmentation via Episodic Learning in Continuous Frequency Space. *arXiv* **2021**, arXiv:2103.06030.
34. Roth, H.R.; Cheng, Y.; Wen, Y.; Yang, I.; Xu, Z.; Hsieh, Y.T.; Kersten, K.; Harouni, A.; Zhao, C.; Lu, K.; et al. NVIDIA FLARE: Federated Learning from Simulation to Real-World. *arXiv* **2022**, arXiv:2210.13291.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.