

FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

Collaborative Machine Learning Without Centralized Data for Computer-Aided Diagnosis of Glaucoma using Retinal Fundus Images

Telmo Baptista



FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

Mestrado em Engenharia Informática e Computação

Supervisor: Carlos Soares

Co-supervisor: Filipe Soares

July 27, 2023

Collaborative Machine Learning Without Centralized Data for Computer-Aided Diagnosis of Glaucoma using Retinal Fundus Images

Telmo Baptista

Mestrado em Engenharia Informática e Computação

Approved by:

President: Rui Camacho

Referee: João Vinagre

Referee: Carlos Soares

Referee: Filipe Soares

July 27, 2023

Resumo

Machine learning mostra grande potencial no auxílio ao diagnóstico de inúmeras doenças, bem como outras tarefas. Estes sistemas são, geralmente, baseados em abordagens de deep learning que necessitam que uma grande quantidade de dados seja transferido para uma entidade central. No entanto, isso geralmente não é uma opção viável, pois levanta questões quanto à privacidade e partilha de informação sensível. Federated Learning (FL) é uma abordagem recente que visa solucionar esse problema, permitindo o treino de modelos deep learning em dados de varias entidades, sem a necessidade da partilha desses dados com uma entidade central, garantindo assim uma maior privacidade e segurança. Isto é notório na área da saúde, uma vez que permite desenvolver modelos sem sacrificar a precisão destes. Apesar disso, as abordagens FL ainda estão na suas fases iniciais e há uma falta de pesquisa sobre o impacto de diferentes tipos de heterogeneidade de dados no desempenho dos algoritmos FL. Nesta dissertação, pretendemos estudar e avaliar a robustez dos algoritmos FL para diferentes distribuições de dados e qualidade de dados no diagnóstico de glaucoma usando imagens da retina. Para alcançar esse objetivo, realizamos várias experiências com distribuições de dados variados e qualidade de imagem variável nos participantes da federação. De modo a controlar as distribuições, utilizamos o RetQualEvaluator para categorizar os datasets em termos de qualidade, e em seguida, utilizamos um distribuidor para obter as distribuições pretendidas. Por fim, avaliamos o desempenho da abordagem centralizada e das diferentes estratégias federadas nos dados dos clientes e em um conjunto de dados de teste independentes da federação. A partir dos resultados das experiências, observamos que federated learning mostra potencial em permitir a cooperação entre instituições, sem comprometer dados confidenciais, de modo a atingir um modelo com desempenho superior. Além disso, também observamos que determinadas estratégias eram mais adequadas a certos cenários, dependendo da distribuição e qualidade dos dados dos participantes e do objetivo pretendido para o modelo desenvolvido. Em geral, estas experiências permitiram-nos compreender em mais detalhe o comportamento de várias estratégias federadas em determinados cenários, o que poderá ser crucial no desenvolvimento de modelos com maior performance para o diagnóstico de glaucoma ou outras aplicações médicas.

Abstract

Machine learning has shown great potential in aiding the diagnosis of a multitude of diseases, as well as other tasks. These systems are often based on deep learning approaches that require a large amount of data to be transferred to a centralized entity. However, this is often not a feasible option, as it raises privacy concerns over the sharing of sensitive information. Federated Learning (FL) is a recent approach that aims to address this issue by allowing the training of deep learning models on the data of multiple entities without the need to transfer the data to a centralized entity. FL has shown great potential to ensure privacy in digital healthcare while maintaining the accuracy of the models. Despite that, FL approaches are still in their infancy, and there is a lack of research on the impact of different types of data heterogeneity on the performance of FL algorithms. In this thesis, we aim to study and evaluate the robustness of FL algorithms to varying data distributions and data quality in glaucoma diagnosis using retinal fundus images. To achieve this, we conduct experiments with different data distributions and varying image quality on the participating clients. To control these distributions, we use RetQualEvaluator to generate quality labels for the datasets and then a data distributor to achieve our desired distributions. Finally, we evaluate the performance of the centralized approach and the different federated strategies on the clients' data and an independent test dataset. From our experiment results, we observed that federated learning shows the potential to enable higher-performance models without compromising sensitive data. Furthermore, we inferred that certain federated strategies were more suitable to certain scenarios based on the data distribution of the participating clients and the model objective. Overall these experiments allowed us to have a deeper understanding of the behaviours of various federated strategies under certain scenarios, which can be crucial in developing high-performance models for glaucoma diagnosis or other medical applications.

Keywords: Federated Learning, Image Processing, Fundus images, Glaucoma

Contents

1	Introduction	1
1.1	Objectives	2
1.2	Document Structure	2
2	Background	4
2.1	Glaucoma	4
2.2	Federated Learning	5
3	Related Work	10
3.1	Computer-aided glaucoma diagnosis	10
3.2	Federated Strategies	11
3.3	Federated Learning in the medical field	13
4	Methodology and Experimental Setup	15
4.1	Methodology	15
4.2	Retina Quality Evaluator	16
4.3	Datasets	18
4.3.1	ORIGA	18
4.3.2	REFUGE Challenge	18
4.3.3	GAMMA	19
4.3.4	Drishti	20
4.3.5	EyePACS AIROGS	22
4.4	Glaucoma Classification Model	25
4.5	Federated Setup	25
4.5.1	Hardware Setup	25
4.5.2	Data Distributor	25
4.5.3	Federation Implementation	29
5	Results and Analysis	36
5.1	Convergence Analysis	36
5.2	Performance Analysis	38
5.2.1	Analysis on the local-only approach	38
5.2.2	Analysis of the federated approach	39
5.3	Final Remarks	42
6	Conclusions	44
	References	46

List of Figures

2.1	Fundus image with the key regions of interest annotated.	5
2.2	Centralized machine learning paradigm	7
2.3	Federated machine learning paradigm	8
3.1	Number of publications	11
4.1	Architecture diagram for the methodology proposed.	16
4.2	Samples from AIROGS dataset and their quality label inferred by RetQualEvaluator.	17
4.3	Sample images from ORIGA dataset.	19
4.4	Sample images from REFUGE Challenge dataset.	21
4.5	Sample images from GAMMA dataset.	22
4.6	Sample images from Drishti dataset.	23
4.7	Sample images from AIROGS dataset.	24
4.8	Client 1’s dataset 1.0 and 2.0.	28
4.9	Client 2’s dataset 1.0.	29
4.10	Client 3’s dataset 1.0.	30
4.11	Outsider’s dataset 1.0.	31
4.12	Client 2’s dataset 2.0.	32
4.13	Client 3’s dataset 2.0.	33
4.14	Outsider’s dataset 2.0.	34
5.1	Evolution of the training loss curve and accuracy of each strategy on experiment 3.0 in client 1.	37
5.2	Evolution of the validation loss curve and accuracy of each strategy on experiment 3.0 in client 1.	37

List of Tables

4.1	ORIGA dataset overview.	18
4.2	REFUGE Challenge dataset overview - training and validation splits.	20
4.3	REFUGE Challenge dataset overview - combined splits.	20
4.4	GAMMA dataset overview.	21
4.5	Drishti dataset overview.	23
4.6	EyePACS AIROGS dataset overview.	24
4.7	Data distribution parameters across the different clients.	27
4.8	Experimental setups details.	31
4.9	Values for FedProx’s proximal term (μ) on each experiment.	33
4.10	Values for FedYogi’s hyperparameters on each experiment.	35
5.1	Effect of early-stopping on the performance of the local-only model.	38
5.2	Performance of the clients’ local-only model on its local test dataset and the out- sider test dataset.	39
5.3	Performance of the clients’ local-only model on its local test dataset and the out- sider test dataset.	39
5.4	Performance of the different strategies on the local test dataset and the outsider test dataset.	41
5.5	Accuracy achieved on the test datasets for each federated strategy.	42

Abbreviations

AI	Artificial Intelligence
CNN	Convolutional Neural Network
DL	Deep Learning
FL	Federated Learning
FedAvg	Federated Averaging
FedBN	Federated Batch Normalization
FedSDG	Federated Stochastic Gradient Descent
GDPR	General Data Protection Regulation
IOP	Intraocular Pressure
ML	Machine Learning
OCT	Optical Coherence Tomography
RFP	Retinal Fundus Photograph

Chapter 1

Introduction

In digital healthcare, machine learning (ML) and, specifically, deep learning (DL) models have seen growth in their descriptive and predictive capabilities. DL models, particularly convolutional neural networks (CNN), have been widely explored for medical imaging to process and analyze various imaging techniques, such as MRI [38], CT scans [55], X-rays [26] and ultrasounds [33]. Physicians use these images to detect and diagnose numerous diseases, such as cancer, cardiovascular diseases and eye diseases. In ophthalmology, retinal fundus photography (RFP) and, more recently, optical coherence tomography (OCT) are used to screen the main eye diseases—diabetic retinopathy, glaucoma and age-related macular degeneration [48]. 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 in 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 a feasible option. Hospitals and institutions must comply with the General Data Protection Regulation (GDPR) and other privacy regulations, which restrict the sharing of patient data [66]. Additionally, it requires approval by the institutional review board to determine to what extent data can be shared with peers and often requires anonymization of data. Data anonymization is a complex and time-consuming task and needs to ensure that patient identification can not be obtained from the data. Nevertheless, studies have already shown the possibility of reconstruction of identification from anonymized data [51], resulting in a critical liability regarding the privacy of sensitive information. In this context, there's 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 will aggregate 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 of 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 [41], which enable a higher degree of privacy of the sensitive data.

To generate the global model, the server uses the FL algorithm or federated strategy, 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 [40]. One of the biggest challenges in FL stands on the heterogeneity of the clients, whether caused by differences in the type of data, amount of data, quality of data, etc. In this context, FL algorithms must be able to adapt to the heterogeneity of the clients, and multiple different 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 depending 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.

1.1 Objectives

In this thesis, we aim to study and analyse the federated approach compared to the centralized approach, the impact of heterogeneous data, such as different label distribution and varying data quality, on the performance of federated approaches, and which FL is the most suitable given the nature of the data and the objective of the model. In this context, the following research questions are raised:

- Does the federated approach have performance gains over non-federated approaches?
- What is the impact of the data distribution and data quality of the clients on the performance of the FL algorithms?
- Given the nature of the data, such as data distribution or data quality, and the model's objective, such as optimising local performance or generalized performance, what is the most suitable FL algorithm?

With this thesis, we aim to contribute to the state-of-the-art with a deeper understanding of FL algorithms under heterogeneous data across clients and the impact of data distribution and quality on the performance of FL algorithms. Another contribution involves empirically validating various FL algorithms on classifying suspicious glaucomatous optic nerves using retinal fundus images.

1.2 Document Structure

This document is divided into six chapters. Chapter 2 introduces the necessary context and background knowledge for this thesis. We provide an overview of what is glaucoma and its diagnosis

methods, followed by an introduction to machine learning, deep learning and federated learning essential concepts for the understanding of this thesis.

In chapter 3, we review the previous work developed in the relevant related areas. We start by analysing the existing work in the field of glaucoma diagnosis using ML models. In the following section, we review several existing FL algorithms. Finally, we look into applications of FL in the medical domain.

Chapter 4 provides an overview of the methodology, a detailed description of all the tools and datasets used, the implementation details of the federated setup and the experiments that will be performed.

In chapter 5, we analyze the results obtained from the experiments and answer the research questions proposed. Finally, chapter 6 summarises what was discussed in this document, our main conclusions and proposed future work that can be done to expand our research.

Chapter 2

Background

This chapter aims to introduce and explain the necessary concepts and knowledge to read this work. Section 2.1 explains what glaucoma is and what screening techniques are used for diagnosis. Section 2.2 explains briefly how standard machine learning and deep learning techniques work, as well as their limitations.

2.1 Glaucoma

Glaucoma is a group of eye diseases that damage the optic nerve due to high intraocular pressure (IOP). It is the second leading cause of blindness [9] worldwide, and if untreated, it usually is irreversible and results in reduced quality of life. The number of people afflicted has increased in the last two years, with over 80 million cases, and is expected to grow even further [54]. Early detection and treatment are crucial to control and prevent irreversible damage to the optic nerve. Risk factors for glaucoma include family history (genetic disease), age, sex, race, cornea thickness, diabetes, high blood pressure, and mainly an increased intraocular pressure (IOP).

To accurately and safely diagnose glaucoma, a doctor will examine the IOP using a tonometry test, which relies on flattening the cornea slightly using a tonometer to measure the pressure in the eye [4]. IOP readings can be influenced by corneal thickness and lead to an improper diagnosis. Patients with thinner corneal thickness may present artificially low IOP and be falsely labelled as a negative case of glaucoma. On the other hand, patients with higher corneal thickness may present IOP levels above average and lead to false positive cases of glaucoma [42]. A pachymetry test is performed to measure corneal thickness and is a vital procedure to prevent the issues above. There are two types of techniques for performing a pachymetry test. Ultrasound pachymetry uses a device that produces sound waves to measure the thickness. However, it is prone to measurement errors, as any probe displacement can cause inaccurate readings [6]. The second technique, optical pachymetry, has various types of devices to make the measurement, with optical coherence tomographies (OCT) gaining popularity. OCT is a non-invasive imaging technique used to generate a high-resolution picture of the retina by measuring the amount of low-coherence light reflected by

the retina and optic nerve. The image, then, can be used to measure the thickness of the retinal layers and optic nerve and to diagnose various eye diseases, including glaucoma [5, 1].

Another common image modality in glaucoma screening is retinal fundus photography (RFP) [15, 23]. RFP is a non-invasive process of taking photographs of the interior of the eye through the pupil. Fundus imaging is performed using a special colour camera, a fundus camera. There are many types of fundus cameras, from stationary to portable, and the capturing of these images can be performed by any medical professional, making it a more affordable and flexible technique for screening various eye diseases. The ophthalmologist can then examine these images to analyze damage to the optic nerve. The two principal eye structures analyzed in retinal images are the optic disc and the optic cup, demonstrated in Fig. 2.1. Specifically, the cup-to-disc ratio (CDR) is the ratio of the vertical diameter of the optic cup to the vertical diameter of the optic disc. A healthy eye has a CDR of approximately 0.3. The increase in the IOP causes the cup to enlarge, and thus, a higher CDR than normal is a potential indicator of the presence of glaucoma.

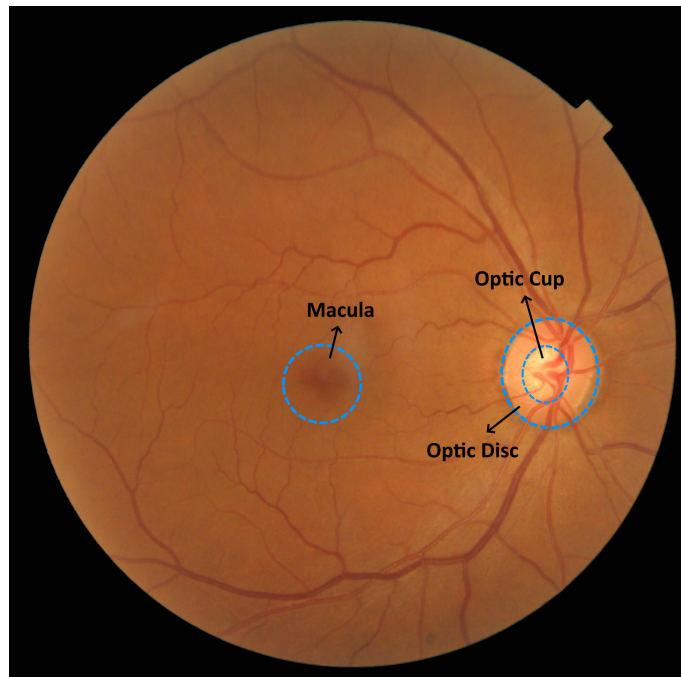


Figure 2.1: Fundus image with the key regions of interest annotated.

2.2 Federated Learning

Machine learning (ML) is a subfield of AI that focuses on developing algorithms and models that learn from data without being explicitly programmed. This involves designing and constructing mathematical models, such as linear regressions [44], capable of learning patterns and relationships in the data to make predictions or decisions. These algorithms can be categorised into various types. However, in our study, we will only study a supervised learning approach. In supervised learning, the algorithm is trained with labelled data, in which the input data is paired with its target

label. With this, the algorithm aims to learn how to map the input data to the correct target label by identifying patterns and relationships between the two. To aid in this mapping, the relevant characteristics and properties of the input data need to be extracted so it can be relayed to the machine learning algorithm. These properties are called features and can take various forms depending on the type of input data and the domain where the algorithm will be applied. For example, for developing an algorithm with the task of predicting if a patient has glaucoma or not, a possible set of features could be comprised of the following: the age of the patient, the patient's CDR, blood pressure, etc. The selection of these features is called feature engineering. This complex and time-consuming task involves engineers and domain-specific specialists to derive the most representative features of the input data. An improper selection can lead to a negative impact on the algorithm's performance and may fail to converge. In ML, an algorithm that has converged refers to the state where the algorithm has reached a stable or optimal solution; this is, the algorithm has learnt the underlying patterns in the data, and further iterations on the learning process may not improve the algorithm's performance. Generally, the monitoring of the algorithm's convergence is essential in ensuring the algorithm is learning and avoiding possible underfitting and overfitting (lack of learning, and excessive learning, respectively).

Deep learning is a subset of machine learning that focuses on designing and training deep neural networks, which are artificial neural networks with multiple layers. Artificial neural networks are models composed of nodes, called artificial neurons, interconnected and organized in layers. The connection between neurons is represented by weights which influence the strength of the information flowing through that connection. These weights are then adjusted during the training process to optimise the network's performance on the given task. These networks can learn complex patterns from the data, including learning a suitable set of features to represent the input data, removing the need for explicit feature engineering.

Deep learning has demonstrated success in various domains, namely in image processing tasks, such as image classification, object detection and object segmentation. In image classification, a model assigns a label to the entire image. For example, a model designed to diagnose glaucoma based on a retinal fundus image is an image classification task. In object detection, a model aims to identify and localise various objects using bounding boxes. An example object detection task would be identifying and locating the optic disc on a retinal fundus image. Finally, in object segmentation, the model aims to accurately segment objects by assigning them a label on a pixel level. An example object segmentation task would be accurately labelling which pixels compose the optic disc and which compose the optic cup. In digital healthcare, deep learning models have been developed for a variety of tasks, including brain lesion segmentation [38], skin lesion segmentation [7], and various ocular diseases [68, 18, 49].

However, these approaches have a requirement of vast amount of data to achieve convergence and high performance. This generally relies on aggregating all training data into a single data centre, as illustrated in Fig. 2.2. In digital healthcare, however, this aggregation raises legal, ethical and economic complications, as sharing patient data is heavily regulated and valuable for the institutions [2, 25] and can hinder the availability of data required for training these models.

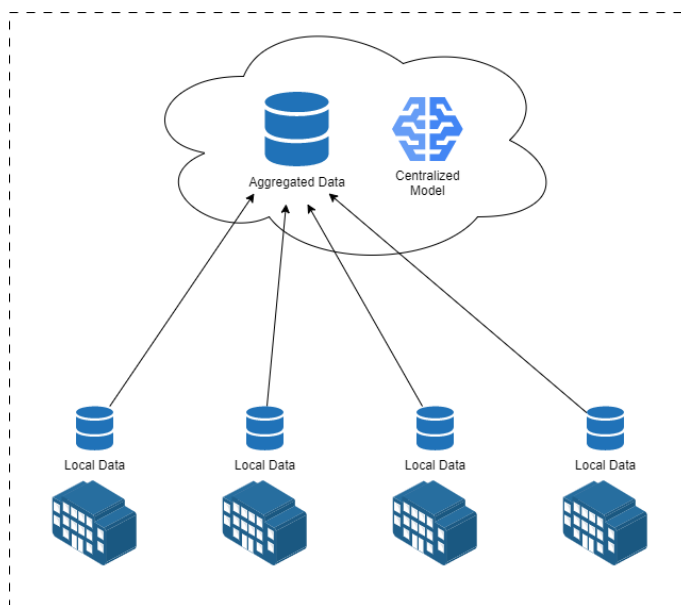


Figure 2.2: Centralized machine learning paradigm

Federated learning (FL) aims to solve this problem by removing the need for data centralization by training a model in a decentralized collaborative fashion. Instead of sharing data, each institution (i.e., client) trains a model locally on their data and submits the model updates to a central institution (i.e., server). The server then aggregates all the client's updates and generates a new model that is sent back to begin another federated round. The aggregation process can be executed using different algorithms called federated strategies. The strategy is responsible for how each client's model contributes to the global model and what information about the client's model is utilised to generate the global model. For example, a strategy can generate the global model by performing a weighted average of the clients' model [40]. The federated paradigm is illustrated in Fig. 2.3.

FL can be classified based on the data distribution and the type of participants. The data can be classified into two definitions: (i) independent and identically distributed data (IID data) and (ii) non-IID data. In statistics, an identically distributed variable is a random variable that has the same probability distribution as another random variable. Additionally, the samples being independent means there's no correlation between them, i.e. are independent events. Larger amounts of IID data during the training process are important to ensure that the training data is representative of the data distribution. Without it, the model may become biased towards certain patterns in the data and lead to worse performance. In FL, it's unrealistic to assume that the local data of each participant is IID. For instance, the data from a participant may be ordered in some way, such as, by time, by disease, and thus introduce a correlation between data points, violating data independence. Participants may also have different data distributions caused by different data collection methods, different data sources, or different data collection periods. For example, image quality can vary due to the equipment used, thus making the data non-identically distributed. Zhao et al. [69] has shown the accuracy of an FL model can be significantly affected by non-IID data,

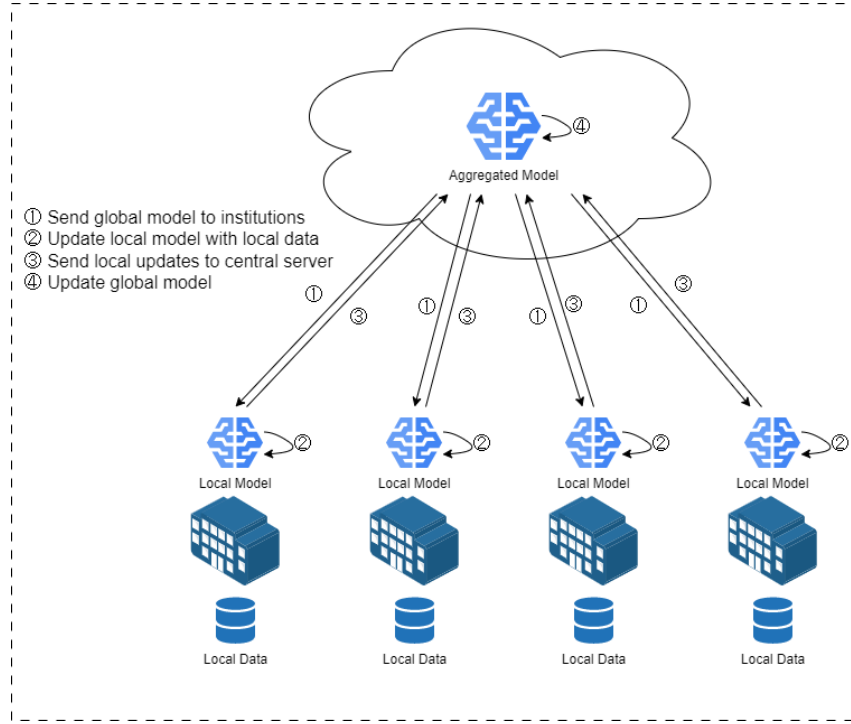


Figure 2.3: Federated machine learning paradigm

where the performance of the model was 50% lower than expected.

Furthermore, depending on how the data is split, FL implementation can be classified into: (i) horizontal FL, (ii) vertical FL and (iii) federated transfer learning [35]. Horizontal FL (HFL), also known as sample-based FL, refers to scenarios where the data has a similar set of features but from different samples. For example, in a healthcare scenario, multiple hospitals may share a similar set of features (age, gender, symptoms, etc.), but each hospital has its own set of patient data. Vertical FL (VFL), or feature-based FL, refers to scenarios where the samples are shared but present a different set of features. For example, suppose two hospitals hold information on the same population, but one has information regarding the patient's ophthalmology tests and the other regarding blood tests. In that case, VFL allows both hospitals to cooperate in generating a model that combines both data types. Finally, federated transfer learning (FTL) applies to scenarios where datasets vary in samples and feature space. The goal is to transform feature spaces from different sources into a common feature space. FTL allows the transfer of knowledge across domains and the building of application-specific models with scarce data and without compromising user privacy [34].

Finally, depending on the type of participants, FL can be classified into two groups: (i) cross-device FL and (ii) cross-silo FL. Cross-device FL refers to an implementation where the learning occurs on various devices, such as smartphones, edge devices, apps, etc. This implementation is usually deployed at scale, requiring millions of participants with small amounts of data, as in the case of Google Board's FL [40, 3]. Challenges in this implementation are the heterogeneity of the devices, the limited computational power, and the limited communication bandwidth, not

to limit the user experience. On cross-silo FL, the number of participants is reduced, but with a larger amount of data per participant [21]. This implementation is usually deployed in a controlled environment, such as hospitals and banks, which provide more powerful and scalable computing power and communication bandwidth. With this, we can classify the implementation in our study as horizontal cross-silo FL.

Chapter 3

Related Work

In this chapter, we analyse and introduce the related work done in the areas relevant to this thesis. This includes computer-aided glaucoma diagnosis, proposed federated strategies and applications of federated learning in the medical field.

3.1 Computer-aided glaucoma diagnosis

Computer-aided diagnosis (CAD) are systems that provide support to doctors and clinicians in the diagnosis of diseases and interpretation of medical images. CAD systems exist for over 30 years. However, with recent advances in deep learning, traction for developing CAD systems has increased [46]. Searching on PubMed¹'s database for the queries: (i) "computer-aided diagnosis", (ii) "deep learning" AND "medical imaging" and (iii) "deep learning" AND "glaucoma", we can observe an exponential trend on the last 5 years of publications, as observed in Fig. 3.1.

Several studies and reviews have assessed the current state-of-the-art artificial intelligence applied to medical imaging, including glaucoma diagnosis [18, 49, 10].

Liu et al. [31] proposes a multiple kernel learning-based classifier that combines image features extracted from retinal fundus images with patient personal data, such as demographic data (i.e., age, gender, height, etc.), examination data (i.e., IOP, corneal thickness) and historical medical data, and finally genome data. The authors utilize a previous method [62] to obtain the candidate patch, which is later used to extract the optic disc utilizing a variational level-set approach. The image features to be used on the model are obtained from the colour and texture descriptors of each grid on the optic disc image. The combination of data allowed the model to achieve a higher AUC compared to the state-of-the-art. The study also shows that the genome information results in the highest AUC, both when evaluated as a single source and when paired with another data source. This approach, however, has not been validated on a larger and more heterogeneous dataset, and the collection of such data for a larger scale system can raise concerns regarding the data privacy of the patients as personal data. Single-nucleotide polymorphisms (SNP) can be used to identify a person uniquely [65]. Since then, several approaches and improvements have been

¹<https://pubmed.ncbi.nlm.nih.gov/>

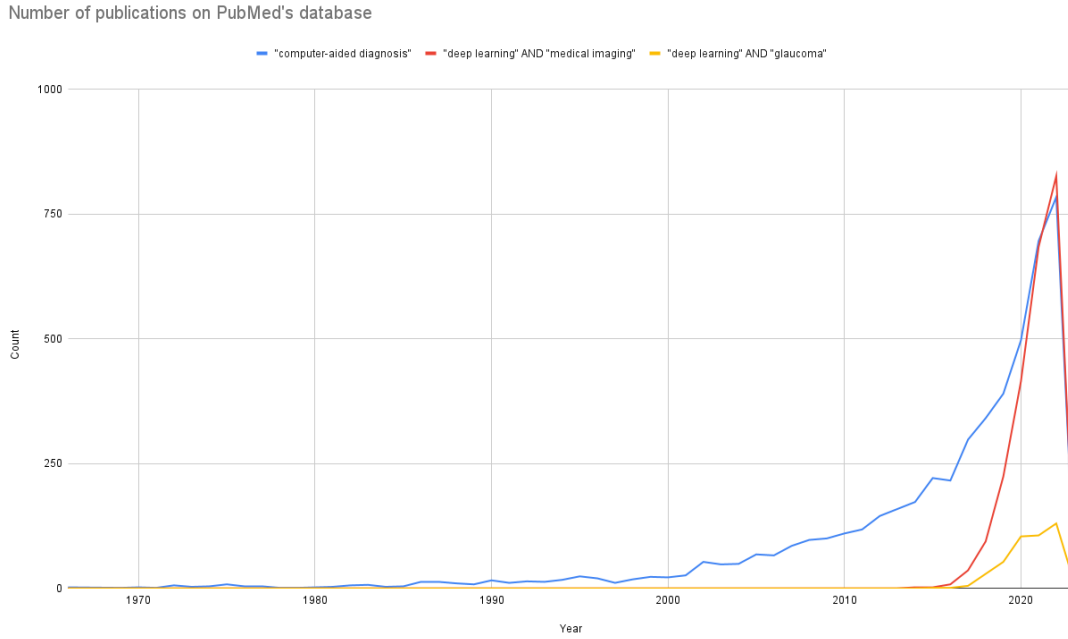


Figure 3.1: Number of publications on (i) "computer-aided diagnosis", (ii) "deep learning" AND "medical imaging", and (iii) "deep learning" AND "glaucoma" over the years on PubMed's database.

achieved, especially with the advances in deep learning. Christopher et al. [12] assessed various architectures and evaluated them in a racially and ethnically diverse dataset. Martins et al. [39] proposed a pipeline utilizing a two-phase approach capable of running on mobile devices. The optic disc and optic cup are first segmented. The results of the segmentation are then used to extract various morphological features and to train the classification model for glaucoma detection. The authors merge several public datasets, performing data augmentation to increase the number of training samples. The proposed pipeline achieved performances higher than other state-of-the-art architectures while also exploring the interpretability of the model using the morphological features extracted and other techniques. Nawaz et al. [43] proposed an efficient two-phase for optic disc and optic cup localization and glaucoma diagnosis. Using a bidirectional feature pyramid network allows the authors to utilize multi-scale key points with equal contribution of each scale, resulting in a more robust model to the diversity of lesion behaviours. The proposed solution managed to achieve an AUC of 0.98. Leonardo et al. [27] improves the performance of the model proposed by Martins et al. [39] by introducing artificially generated images using a Generative Adversarial Network (GAN) supported by a Retinal Quality Evaluator model.

3.2 Federated Strategies

The application of Stochastic Gradient Descent (SGD) in a distributed manner [58, 11] is one of the earliest applications of federated learning. The server starts by broadcasting the global model

weights to each client. Afterwards, the server selects a random fraction of the clients for training. Each client then computes the gradient of the loss on its local data. The server, finally, aggregates the gradients to update the global model weights. This algorithm, denoted as FedSGD [40], is simple and efficient. Still, it has a high communication cost, requiring the clients to communicate with the server after each local update. To reduce the communication cost, Federated Averaging (FedAvg) [40] is built on the same principle as FedSGD but performs several local SGD steps in each client on each round before aggregating the global model using a weighted average. Due to the weighted average, the contribution each client has to the global model is dependent on the amount of local data available. In an IID setting, this leads to uniformly distributed weights across the clients and, thus, an evenly balanced global model, and it is commonly used as the baseline for FL algorithms. However, as discussed earlier, it's not realistic to assume IID data distributions in real-world scenarios, which can lead to severe drops in the accuracy of both the global model and the local models. Several approaches have been proposed to mitigate this issue.

To tackle the problem of non-IID, various researchers introduce regularization terms to prevent model drifts, which normally refers to preventing the locally trained models from diverging too much from the global model or reduce the impact of model variance during aggregation process. FedProx [28] improves on FedAvg by adding a proximal term to restrict the local updates from diverging too much from the global model. FedCurv [57] adapts the Elastic Weight Consolidation (EWC) [24] algorithm to FL. Ahmad et al. [8] proposes a hybrid approach using Hessian coefficients and weighted average which improves on FedAvg at increased communication cost. Reddi et al. [50] adapts various optimizers, such as Adagrad, Yogi, etc., to a federated setting with the aim of improving on FedAvg. Gunesli et al. [16] proposes FedDropoutAvg, where a subset of weights from each local model is randomly dropped. Li et al. [30] proposes FedBN, where the batch normalization layers of the model are not sent to the server, to alleviate feature shifts caused by non-iid data.

Approaches based on learning the aggregation weights and other hyperparameters have also been proposed. Auto-FedAvg[64] aims to learn the aggregation weights as a stochastic model. The algorithm starts by sending the global model to all the clients and then gathering the local models, like FedAvg. However, instead of using a weighted average, the authors introduce a set of learnable weights to generate the global model. Afterwards, the server sends all the local models to each client to compute the loss of the aggregation weights. Due to the communication burden introduced by this step, the authors only execute this learning step every few rounds, configurable by a hyperparameter introduced by the method. The algorithm was then validated on real-world scenarios, such as COVID-19 lesion segmentation, and showed improvements over FedAvg and FedProx. Auto-FedRL [17] also enables the learning of the aggregation weights, but instead of using a stochastic model, the authors introduce a Reinforcement Learning (RL) agent.

The approaches mentioned above all aim to improve the global model and the models of internal clients. However, the model's generalizability to new clients is also an important aspect of FL. To address this issue, Liu et al. [32] propose a method based on domain generalization, FedDG, that shares extra frequency information of the data across clients. This extra information is

then used to learn generalized parameters based on episodic learning by simulating domain shifts. Jiang et al. [22] proposes a personalization strategy to optimize both clients inside and outside the federation, to obtain a more generalizable model without sacrificing local performance. The authors keep track of both global and local gradients to form a personalized model that is optimized for each client. Furthermore, the authors also use the global model and the local, personalized models to create a routing space to personalize the models for outside clients via a learnable routing matrix. This routing matrix is obtained using an unsupervised learning approach utilizing a set of losses designed by the authors. The method is then validated on two medical imaging tasks, on which the authors show an increase in performance compared to FedAvg and other methods.

3.3 Federated Learning in the medical field

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. [52] developed a federated approach to breast density classification using a weighted averaging strategy (FedAvg) and compared results against locally trained models on each institution. The federated model obtained a better agreement on classification between institutions than the model 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 that try to address this issue. Ho et al. [20] applies FL for COVID-19 detection using X-ray images and information on the symptoms. The authors study the performance of the FL model in various settings, such as IID and non-IID settings, varying numbers of participants and the underlying DL model. Furthermore, the authors combine FL with the differential privacy method to achieve a more secure system while evaluating the trade-off in model performance. FL has also been applied to brain tumour segmentation [29, 56]. Sheller et al. [56] evaluates 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. [29] 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. [36] evaluates 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 [60, 45, 37, 19]. Lu et al. [37] evaluates the performance of FL compared to centralized learning for detecting retinopathy of prematurity (ROP) using fundus images. On the same domain, Hanif et al. [19] proposes an FL approach that generates a vascular severity score (VSS) for the diagnosis. The authors studied the variations of the

VSS for the different categories of ROP. In this study, the authors found significant differences in the severity score between the early stage and advanced stage of ROP, despite not showing any variance when comparing the intermediate stage with the advanced stage.

Nielsen et al. [45] investigates the vulnerability of FL to gradient inversion attacks. These attacks exploit the fact that gradients carry a bit of information about the data that influenced their calculation and manipulate the model's input in a manner that enables 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, there are still several challenges that need to be addressed. First, the performance of DL models trained on local data is often not inferior to the FL model. 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's a gap in the literature on the impact of bad-quality participants and how to restrict them from the FL process.

Chapter 4

Methodology and Experimental Setup

In this chapter, we first introduce the methodology used for this study in the first section. Afterwards, the experimental setup will be discussed in detail. This includes an exploratory analysis of the datasets, the models used, their function, and a detailed explanation of implementing the federated setup and the experiments.

4.1 Methodology

Our research questions introduced in Chapter 1 concern the performance of federated approaches compared to non-federated approaches. It also involves the impact of the nature of the client’s data, such as label distribution, data quality, etc., on the performance of federated strategies. And finally, the differences in the performance of the various federated strategies under different settings and use-case of the federated model.

To answer these questions, in this thesis, we will develop a horizontal cross-silo federated learning approach applied to glaucoma screening from fundus images. This approach will be applied to scenarios that aim to represent real-world institutions with varying data distributions (this term here refers to how the data is distributed across the clients, and not statistical distribution). A collection of public datasets will be used and treated as independent data sources to create these scenarios. To gather information regarding the quality of the images in the various datasets, we will use a retinal fundus image quality evaluator model [27] to infer quality labels for each dataset. Furthermore, a data distributor will be developed to aid the creation of these scenarios based on target distributions and the quality of data wanted.

Afterwards, for each scenario, we will vary the federated strategy used to train the federated model to evaluate the differences between them. We will also train local models on each client using solely their local data and a centralized model where we combine all the client’s data. These two compose the non-federated approaches that will be evaluated against the federated approach. And to further study the impact of data quality on the performance of the various approaches, we will train each on a non-filtered scenario and a filtered scenario, i.e., where all low-quality data is removed from the training data.

Finally, the different approaches and the various federated strategies will be evaluated methodically using adequate metrics for the task, such as accuracy, precision, recall, and specificity. We will also evaluate the convergence rate and the computational cost of the various approaches and federated strategies. Fig. 4.1 represents, in a compact form, our methodology process.

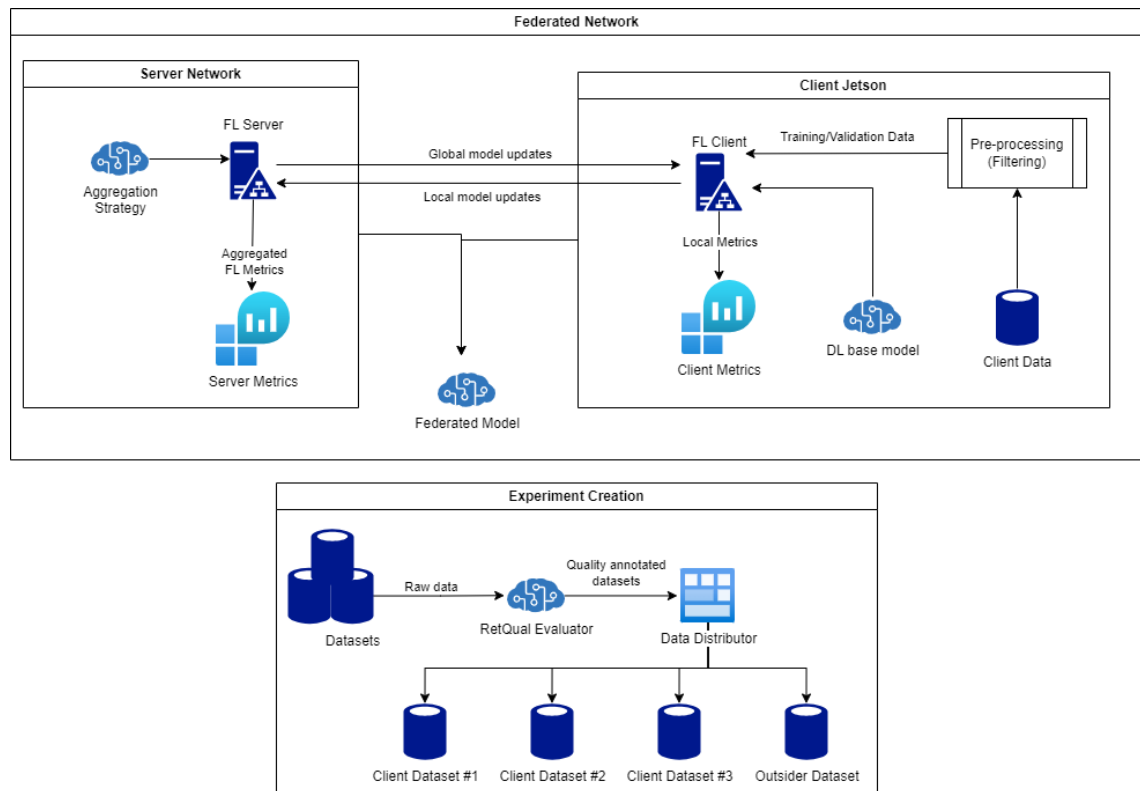


Figure 4.1: Architecture diagram for the methodology proposed.

4.2 Retina Quality Evaluator

The RetinaQualEvaluator model developed by Leonardo et al. [27] was created to distinguish between three image qualities, i.e., 0 for "Good" quality, 1 for "Usable", and 2 for "Reject". The quality of the image is related to not only the properties of the image but also the presence of distinguishable eye structures required to make a proper diagnosis, such as the presence of the optic disc and cup that can be easily identified. Fig. 2.1 shows the location of the important structures of a retina fundus image. The exact positioning of these structures varies depending on the equipment used to obtain the images, with the common locations being: (i) centred on the optic disc; (ii) centred on the midpoint of the macula and the optic disc (posterior pole); and (iii) centred on the macula. In this study, this model is used to infer quality labels for the used datasets to perform a study on the impact of image quality in training a federated setup. Fig. 4.2 presents some examples of images labelled by RetQualEvaluator.

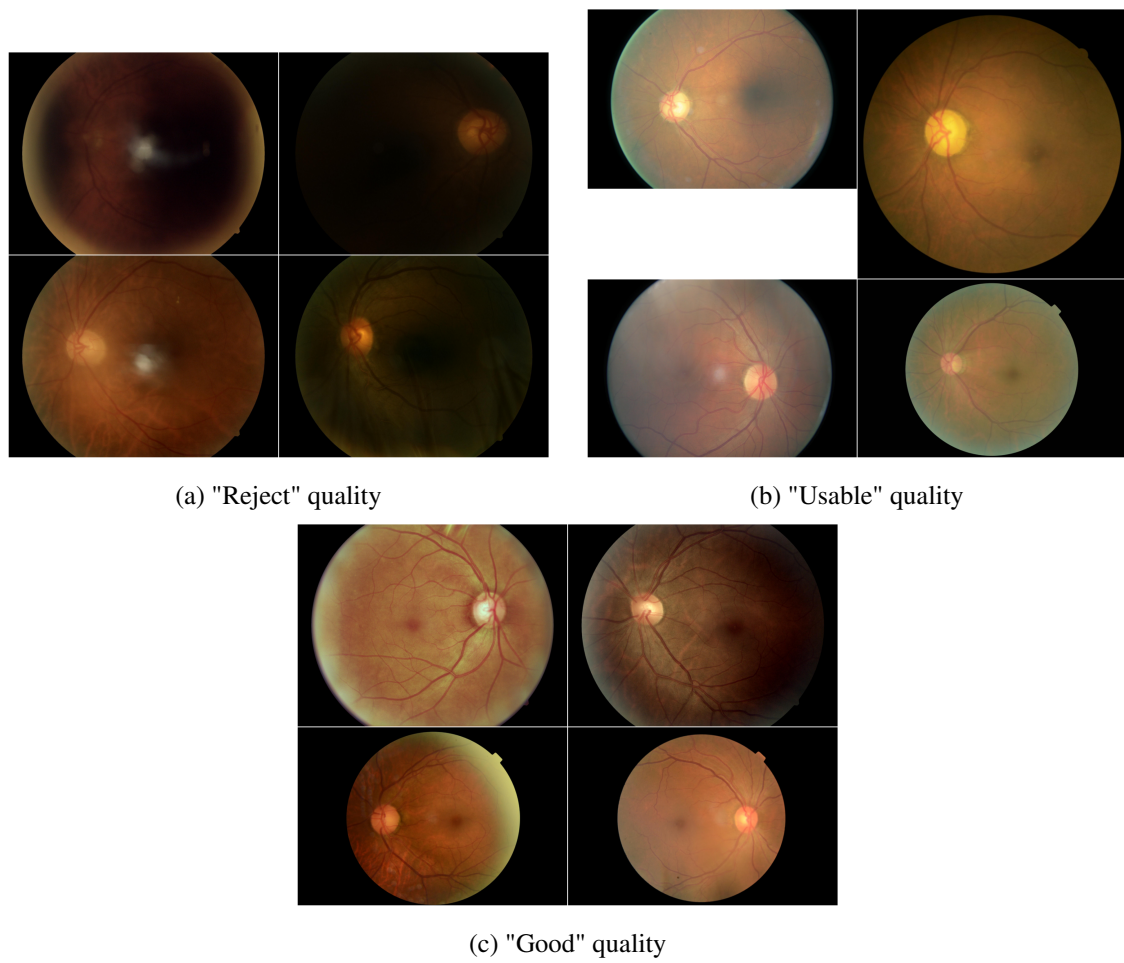


Figure 4.2: Samples from AIROGS dataset and their quality label inferred by RetQualEvaluator.

4.3 Datasets

Various public datasets were used to create the federated setup. To study the image quality of each dataset, the RetinaQualEvaluator model presented in the section above was used to obtain quality classifications for all the images in these datasets. In this section, each dataset will be introduced and discussed.

4.3.1 ORIGA

The ORIGA [67] dataset contains retinal images collected by the Singapore Eye Research Institute for the Singapore Malay Eye Study (SiMES). It contains 650 cases labelled by trained professionals from Malay individuals ranging from 40 to 80 years old. All images in this dataset are full FOV, uncropped, and contain scans from both the left and right eye. The optic disc is usually positioned off-centred on the image to either side, depending on which eye the scan is from, with some cases having the optic disc centred.

The quality labels inferred by the RetQualEvaluator aided in gaining insight into the quality distribution of the dataset and the potential heterogeneity stemming from such distribution. Of the 650 cases, 451 (69.4%) were of "Good" quality, while only 37 (5.7%) were of "Reject" quality. Regarding the glaucoma incidence on the dataset, 482 (74.1%) cases are negative, i.e., non-glaucomatous. This is similar to the glaucoma incidence observed in glaucoma screening clinics, where 10% of the adults screened have glaucoma and an additional 13% were suspects of glaucoma [13]. Thus, the dataset poses a realistic scenario of a dataset provided by a screening clinic. A more detailed breakdown of the dataset is shown in Table 4.1, and some example samples from the dataset are shown in Fig. 4.3.

Table 4.1: ORIGA dataset overview.

Glaucoma Classification	Samples	Quality		
		Good	Usable	Reject
Glaucoma	168 (25.8%)	88 (13.5%)	58 (8.9%)	22 (3.3%)
Non-Glaucoma	482 (74.1%)	363 (55.8%)	104 (16%)	15 (2.3%)
Total	650	451 (69.4%)	162 (24.9%)	37 (5.7%)

4.3.2 REFUGE Challenge

The REFUGE Orlando et al. [47] dataset is part of a challenge held during MICCAI 2018. The data consists of 800 cases split into training and validation splits, each containing 400 cases. The data was collected from multiple hospitals and clinics with Chinese patients, with a close

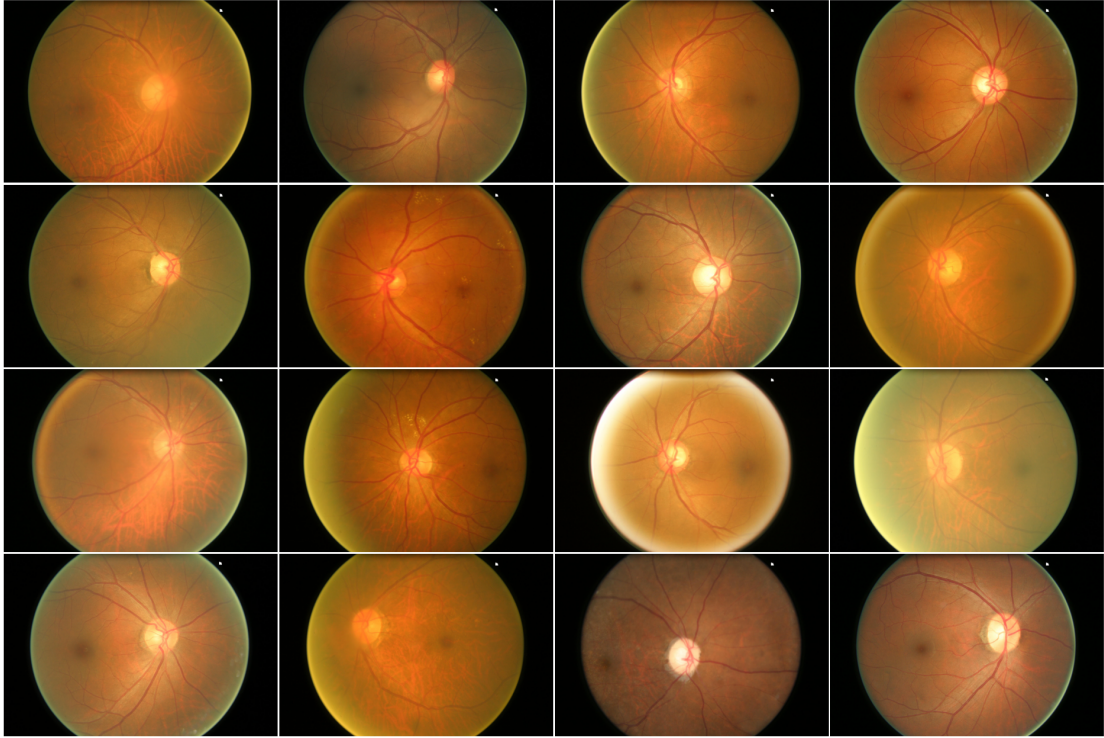


Figure 4.3: Sample images from ORIGA dataset.

to even distribution of gender. The non-glaucomatous cases comprise healthy eye patients and patients afflicted by other eye diseases, such as diabetic retinopathy, contributing to a more realistic representation of clinic data. The training split images were captured using a Zeiss Visucam 500 fundus camera with a resolution of 2124×2056 pixels, while the validation split images were captured using a Canon CR-2 with a resolution of 1634×1634 pixels. In both splits, the images are centred at the posterior pole, having both the optic disc and macula always visible.

Regarding the quality of the images, the dataset has already been curated only to contain high-quality images, which was reflected in the quality labels inferred. Of the 800 cases, only 6 (0.7%) were "Reject". However, a difference between each split is present, as 58.5% of the cases in the training split are of "Good" quality compared to the 98% of the cases in the validation split. The distribution of glaucoma is the same on both splits, with 90% of the cases being non-glaucomatous. For this study, both splits will be merged into a single pool of data that will be used to compose the federated setups. Table 4.2 and Table 4.3 show a more detailed breakdown of the dataset, and some example images are shown in Fig. 4.4.

4.3.3 GAMMA

The GAMMA [63] dataset was made available as part of the Glaucoma grAding from Multi-Modality imAges (GAMMA) Challenge. The data consists of 100 cases labelled for three stages of glaucoma (non-glaucomatous, early glaucoma and mid-advanced glaucoma). The early and

Table 4.2: REFUGE Challenge dataset overview - training and validation splits.

Glaucoma Classification	Training				Validation			
	Samples	Quality			Samples	Quality		
		Good	Usable	Reject		Good	Usable	Reject
Glaucoma	40 (10%)	16 (4%)	20 (5%)	4 (1%)	40 (10%)	37 (9.2%)	3 (0.8%)	0 (0%)
Non-Glaucoma	360 (90%)	218 (54.5%)	140 (35%)	2 (0.5%)	360 (90%)	355 (88.8%)	5 (1.2%)	0 (0%)
Total	400	234 (58.5%)	160 (40%)	6 (1.5%)	400	392 (98%)	8 (2%)	0 (0%)

Table 4.3: REFUGE Challenge dataset overview - combined splits.

Glaucoma Classification	Quality			
	Samples	Good	Usable	Reject
Glaucoma	80 (10%)	53 (6.6%)	23 (2.9%)	4 (0.5%)
Non-Glaucoma	720 (90%)	573 (71.6%)	145 (18.1%)	2 (0.3%)
Total	800	626 (78.3%)	168 (21%)	6 (0.7%)

mid-advanced glaucoma labels were combined into a single label for this study. Similarly to the previous dataset, the images collected for this dataset are mainly from Chinese patients ranging from 19 to 77 years old, with an average of 40.64 ± 14.53 years old. Two different fundus cameras were used to collect the images: KOWA camera with a resolution of 2000 x 2992 pixels and Topcon TRC-NW400 camera with a resolution of 1934 x 1956 pixels. The images are centred on the macula or the posterior pole; both structures are always visible.

The quality of the images was manually verified by the authors of the dataset, which was reflected on the quality labels inferred by the RetQualEvaluator. Only eight were "Reject" by the model. Contrary to the previous datasets, the glaucoma incidence for this dataset was evenly split, with 50% of the cases being of healthy patients and the remaining of glaucomatous patients. A more detailed breakdown of the dataset is shown in Table 4.4 and some example images in Fig. 4.5.

4.3.4 Drishti

The Drishti [59] dataset consists of 101 cases acquired from a single hospital in India. All the images were collected using a 30° FOV camera with a 2896 x 1944 pixels resolution centred on the optic disc. The population has an evenly split gender distribution from 40 to 80 years old.

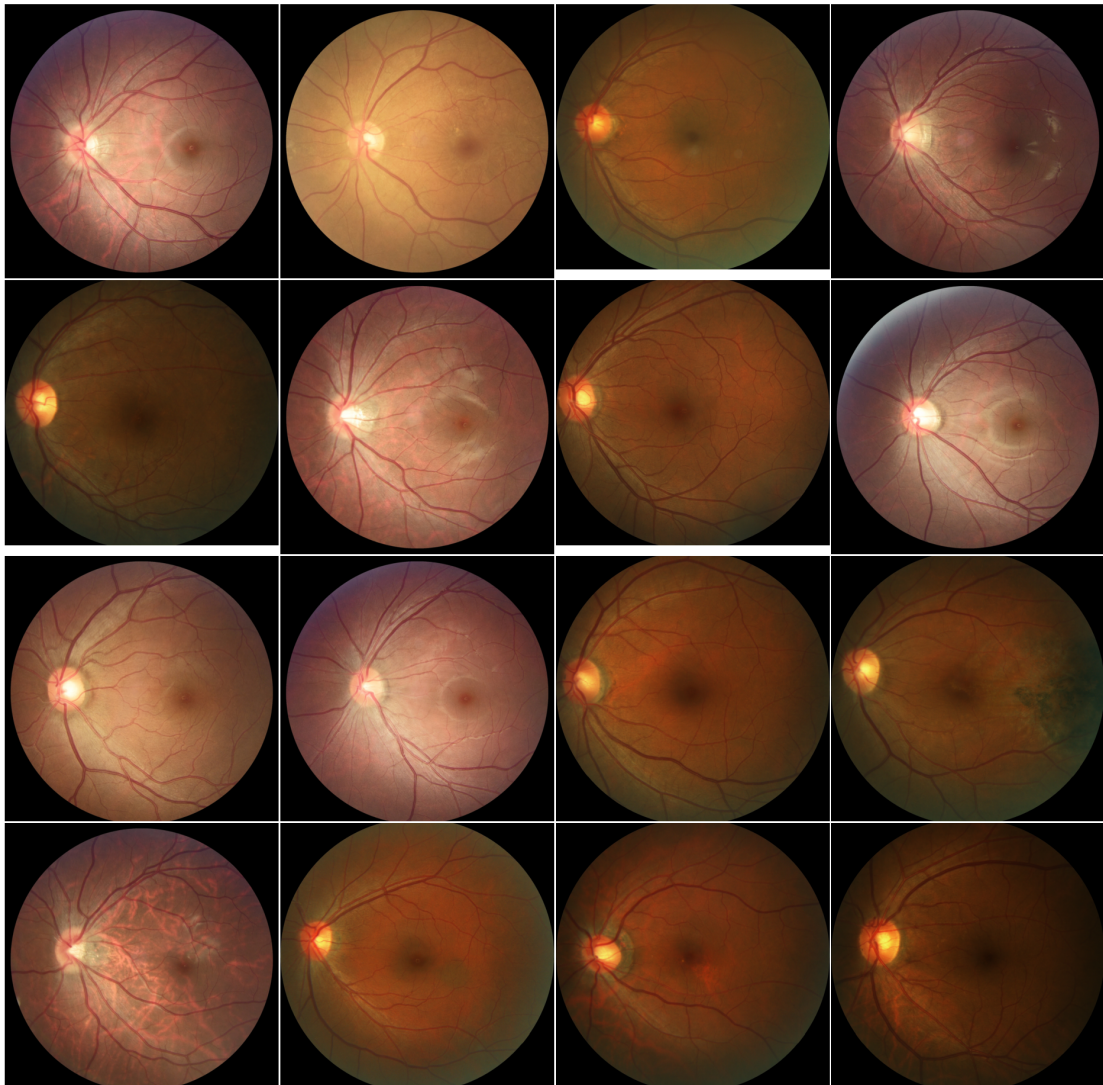


Figure 4.4: Sample images from REFUGE Challenge dataset.

Table 4.4: GAMMA dataset overview.

Glaucoma Classification	Samples	Quality		
		Good	Usable	Reject
Glaucoma	50 (50%)	32 (32%)	10 (10%)	8 (8%)
Non-Glaucoma	50 (50%)	48 (48%)	2 (2%)	0 (0%)
Total	100	80 (80%)	12 (12%)	8 (8%)

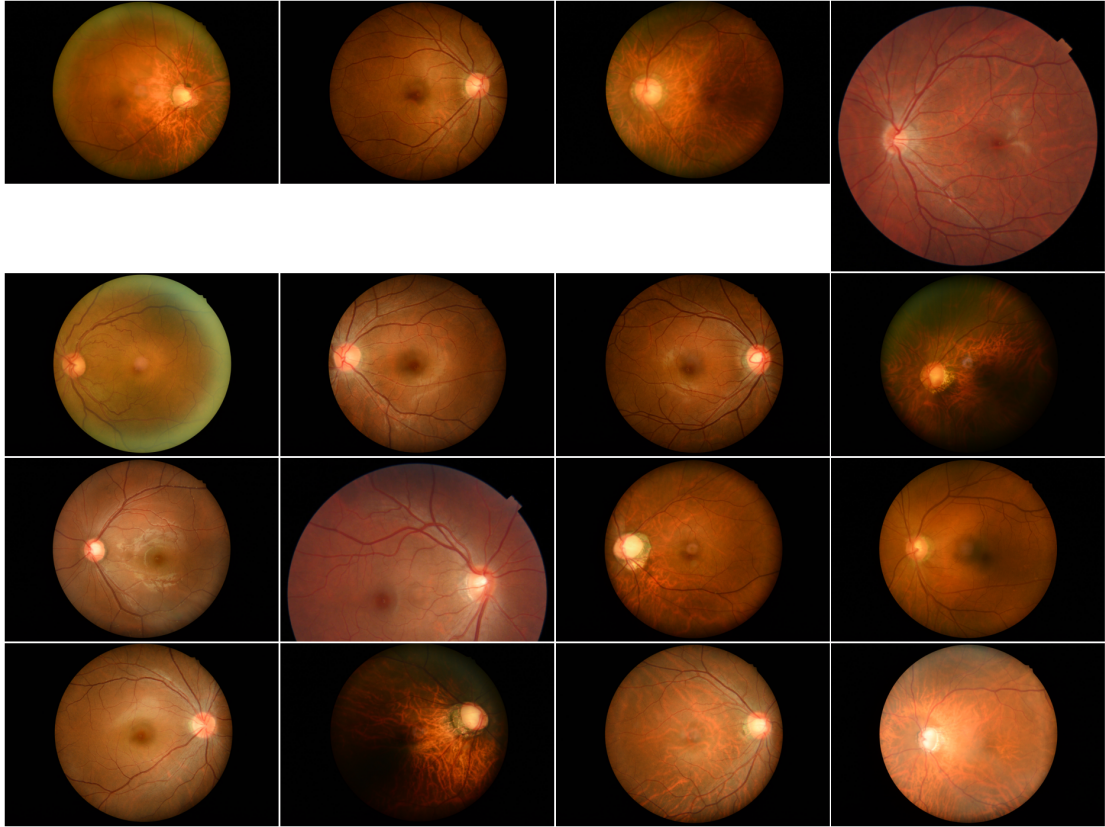


Figure 4.5: Sample images from GAMMA dataset.

Four different glaucoma specialists did the labelling process with experience ranging from 3 to 20 years.

The authors verified the quality of the images and discarded those deemed of bad quality, either by having poor contrast or incorrect positioning of the optic disc and others. Despite that, when analysing the labels inferred by the RetQualEvaluator model, this dataset was composed majorly of "Reject" images, with only 8.91% of the images being of "Good" quality. Compared to other datasets and the global incidence of glaucoma, the dataset is also majorly formed with glaucomatous cases, where 69.3% of the cases are of glaucomatous patients. A more detailed breakdown is shown in Table 4.5 and some example images in Fig. 4.6.

4.3.5 EyePACS AIROGS

The EyePACS AIROGS [14] data was available for the Artificial Intelligence for Robust Glaucoma Screening (AIROGS) challenge. This dataset is the largest and most diverse dataset used in this study. It consists of 101442 cases collected from 54274 patients with an average age of 56.7 ± 10.2 years old across 500 screening centres across the United States of America. On top of the heterogeneous ethnicity, the dataset also comprises images taken from a diverse set of equipment, with more than nine different cameras used. This results in realistic heterogeneous data with

Table 4.5: Drishti dataset overview.

Glaucoma Classification	Samples	Quality		
		Good	Usable	Reject
Glaucoma	70 (69.3%)	4 (4%)	11 (10.9%)	55 (54.4%)
Non-Glaucoma	31 (30.7%)	5 (5%)	5 (5%)	21 (20.8%)
Total	101	9 (8.91%)	16 (15.8%)	76 (75.2%)

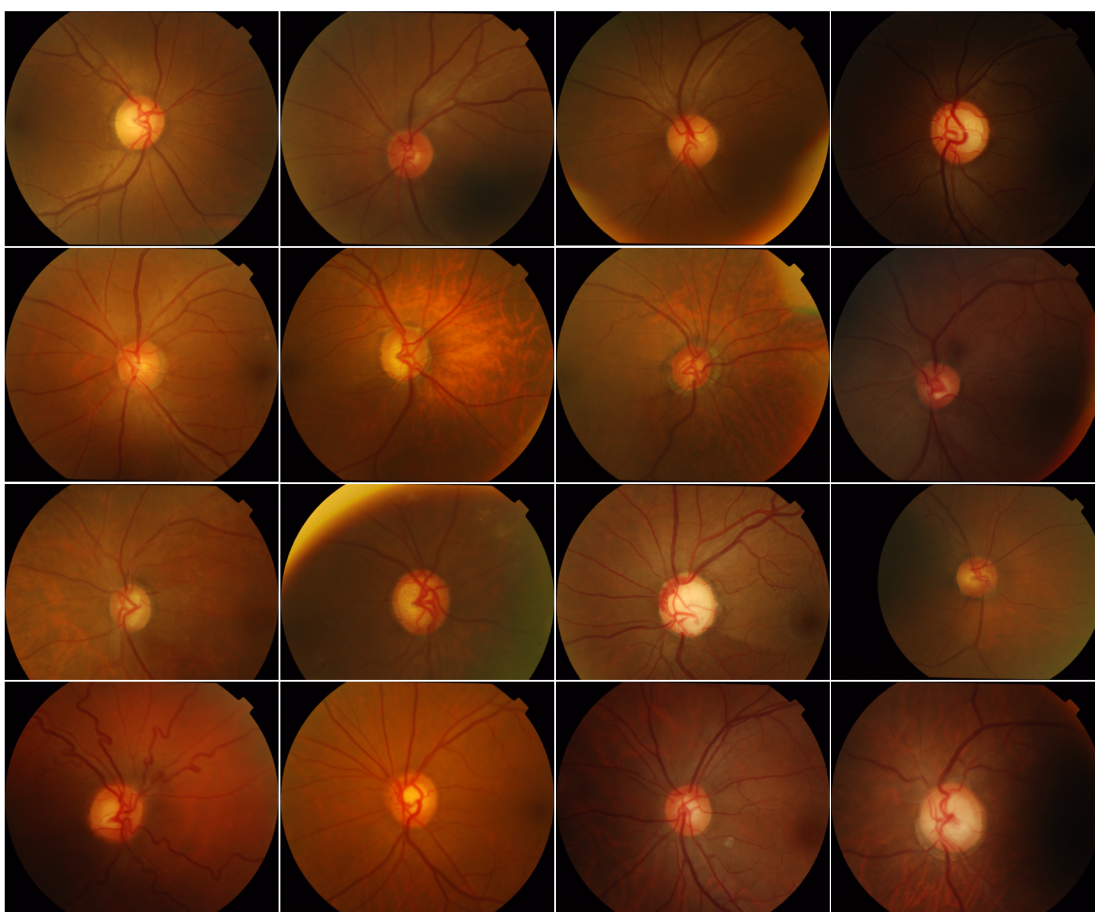


Figure 4.6: Sample images from Drishti dataset.

varied patterns and structure (i.e. optic disc location, image contrast, etc.). The labelling process was done by 30 specialists from a pool of 90, which resulted in a more reliable and consistent ground truth for the diagnosis.

The dataset also poses a great diversity in image quality, for which models in real-world applications would need to be prepared. 56.3% of the images were labelled as "Good" quality by

the evaluator, with 15.5% being of "Reject" quality. This distribution allows us to study image quality's impact on federated setups' performance in a realistic setting. The glaucoma incidence on the dataset is closer to the global prevalence of glaucoma [61], with 3.2% of the cases being of glaucomatous patients, as opposed to the distribution seen on screening centres [13]. Table 4.6 and some example images in Fig. 4.7 show a more detailed breakdown of the dataset.

Table 4.6: EyePACS AIROGS dataset overview.

Glaucoma Classification	Samples	Quality		
		Good	Usable	Reject
Glaucoma	3270 (3.2%)	1199 (1.2%)	1185 (1.2%)	886 (0.9%)
Non-Glaucoma	98172 (96.8%)	55882 (55.1%)	27411 (27%)	14879 (14.7%)
Total	101442	57081 (56.3%)	28596 (28.2%)	15765 (15.5%)

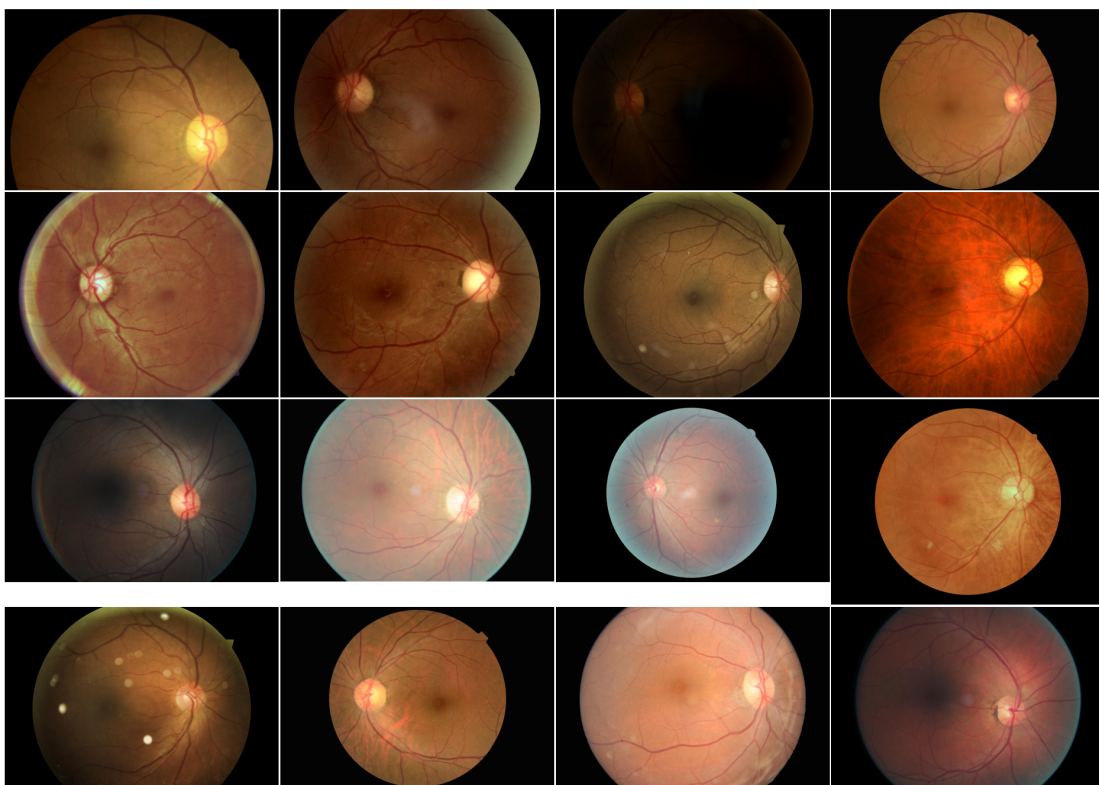


Figure 4.7: Sample images from AIROGS dataset.

4.4 Glaucoma Classification Model

The model developed by Leonardo et al. [27] will serve as the baseline model for this study and the initiation model in the federated setup. The model was trained on a collection of datasets, including ORIGA and Drishti datasets used in this study, with a synthetically augmented data generation module to enhance and degrade the quality of the images used for training. The collection of datasets amounted to 2549 cases on which the model trained for 200 epochs, with early stopping to avoid overfitting. With this, our initial model has already been trained on the same task with good performance and will serve as the underlying model for FL. Other than serving as an initialization model for the federated setup, this model will serve as the baseline for all the strategies (i.e. tested on the datasets without any additional training) and also be trained on the data used in each experiment to evaluate the performance of the centralized approach over the federated approach.

4.5 Federated Setup

In this section, we will discuss in detail how the federation was set up, including the hardware and its networking, as well as the data distribution across the clients. Furthermore, we also explain in detail what federated strategies we implement and what variations we test for each.

4.5.1 Hardware Setup

Three identical Jetsons AGX Xavier were used to run the federated setup to represent independent institutions representing clients of the federated network. Each Jetson was configured in the same manner, with the most up-to-date Jetpack ¹ version (5.1.1). The server of the federated network is also an independent institution 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' dataset. 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 and label distributions. 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.

4.5.2 Data Distributor

To create the various experiments for the federated setup, a data distributor was developed to distribute the datasets mentioned above to the different clients. Due to the size difference between the EyePACS AIROGS dataset and the rest, this dataset will complement the remaining datasets when

¹ <https://developer.nvidia.com/embedded/jetpack>

distributing the data for each client. We used the distributor for each client to pick one or more datasets from above to serve as the base data, excluding EyePACS AIROGS. Then we used the latter to fill in data to achieve a target label and quality distribution, noting that data from the complementary dataset can not be repeated across clients. Lastly, the remaining glaucomatous cases left on the EyePACS AIRGOS after distribution were used to generate a fourth dataset containing a fifty-fifty split of glaucoma and non-glaucoma cases to act as a test dataset representing an institution outside the federation. In table 4.7, we present a detailed breakdown of the distributions generated and the parameters used for the distributor².

Once distributed, the datasets of each client were prepared into three splits: training split (60%), validation split (20%) and test split (20%). This distribution was done randomly, with the only restriction of maintaining the class balance across each split.

²Client 1 presents the same data across both distributions to have a ground of comparison between experiments.

Table 4.7: Data distribution parameters across the different clients.

Target Dataset	Dataset ID	Base Dataset	Complementary Dataset	Target Samples	Target Negative Labels	Target Positive Labels	"Good" Quality Bias	"Usable" Quality Bias	"Reject" Quality Bias
Client 1 Dataset	1.0	REFUGE	EyePACS AIROGS	2000	80%	20%	1.0	1.0	1.0
	2.0	REFUGE	EyePACS AIROGS	2000	80%	20%	1.0	1.0	1.0
Client 2 Dataset	1.0	ORIGA + GAMMA	EyePACS AIROGS	2000	50%	50%	1.0	1.0	1.0
	2.0	ORIGA + GAMMA	EyePACS AIROGS	2000	80%	20%	1.0	1.0	1.0
Client 3 Dataset	1.0	Drishti	EyePACS AIROGS	2000	20%	80%	1.0	1.0	1.0
	2.0	Drishti	EyePACS AIROGS	2000	80%	20%	1.0	1.0	1.0
Outsider Dataset	1.0	EyePACS AIROGS	-	-	50%	50%	-	-	-
	2.0	EyePACS AIROGS	-	-	50%	50%	-	-	-

For the first experiment, which we will call "Dataset 1.0", we aim to study the effect of heavily heterogeneous clients on the performance of the federated approach, with both different glaucoma distributions and different quality distributions. For such, we distributed the data in the following manner:

1. Client 1's dataset comprises mostly negative cases (80%) and higher-quality images. Fig. 4.8 shows the resulting distribution on the dataset.
2. Client 2's dataset serves as the middle ground for the other two clients, with an even distribution on both glaucoma labels (50%) and quality labels, as shown in Fig. 4.9.
3. Client 3's dataset serves as the polar opposite of client 1, comprising mostly positive cases (80%) with lower-quality images. This can be seen in Fig. 4.10.
4. The remaining cases left on the AIROGS dataset was used to build an additional test dataset containing an even distribution of the target label and no cases of "Reject" quality. This dataset is shown in Fig. 4.11

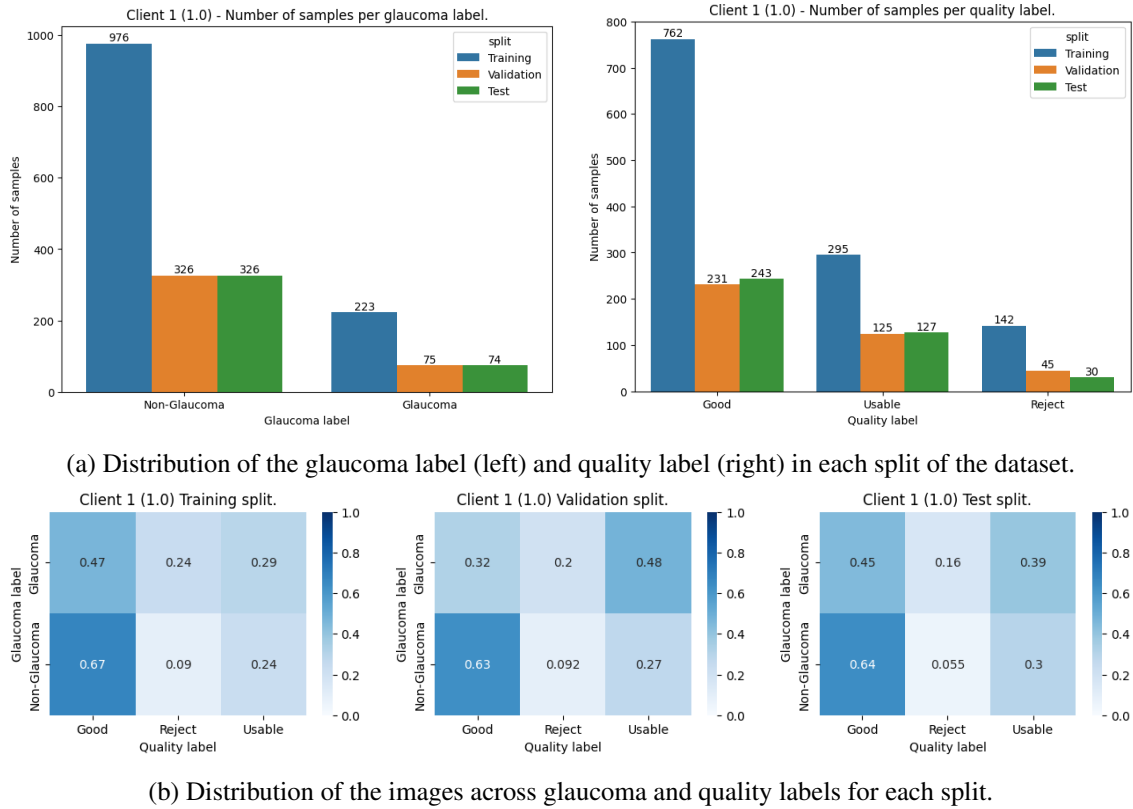


Figure 4.8: Client 1's dataset 1.0 and 2.0.

For the second experiment, which we will call "Dataset 2.0", we aim to study the effect of different distributions of data quality across clients while following a real-world distribution of

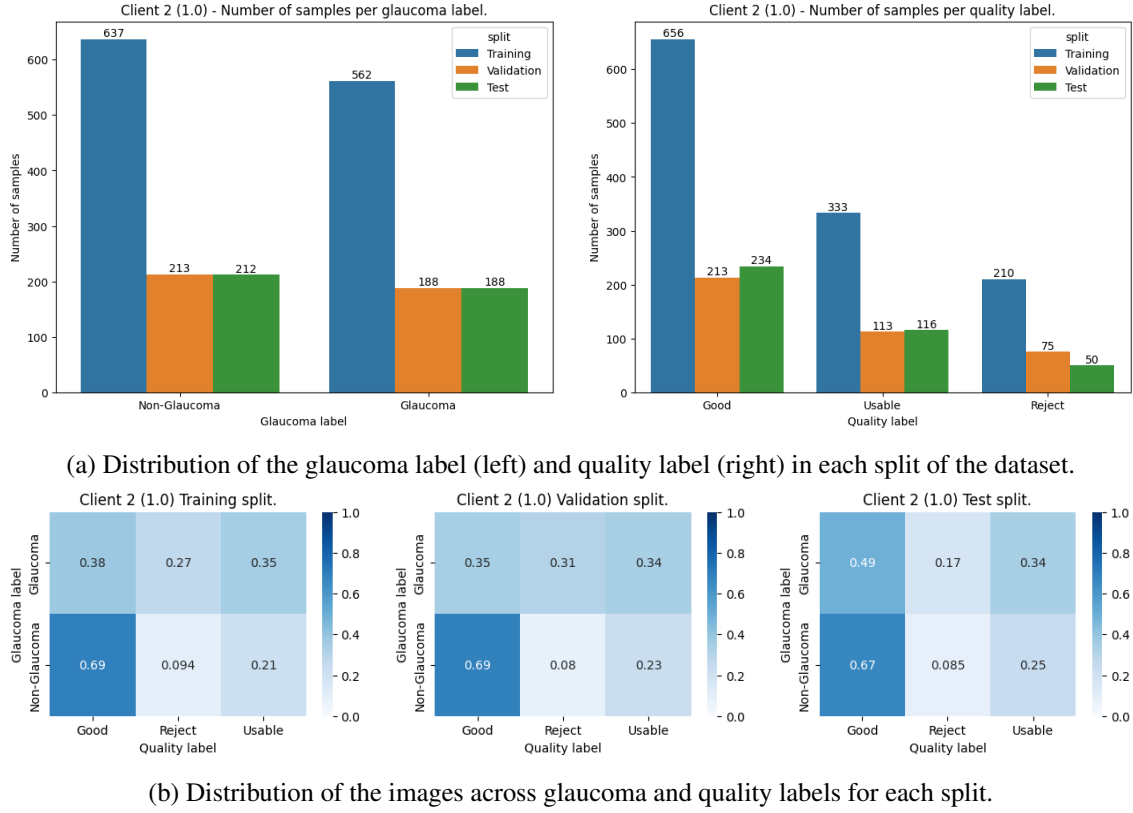


Figure 4.9: Client 2's dataset 1.0.

glaucomatous cases. For this, we follow the glaucoma incidence found in screening clinics [13] where around 20% of the adults suffer from glaucoma or are suspects. Therefore, we distributed all clients with 80% of the cases being negative (i.e. non-glaucomatous) and maintained the quality distribution as above, with client 1 presenting the highest quality and client 3 the lowest. The outsider dataset was built in the same manner, with the remaining cases of the AIROGS dataset being used to compose an even dataset in terms of the glaucoma label filtering out all the images with "Reject" quality. Note that client 1's dataset is the same in the two experiments. This was done to have a point of comparison for the two experiments. The details of the datasets for client 1, client 2, client 3 and the outsider in the second experiment can be seen in Fig. 4.8, Fig. 4.12, Fig. 4.13 and Fig. 4.14, respectively.

4.5.3 Federation Implementation

The federation was implemented using Flower³ as the backend framework responsible for the communication protocols between the server and the clients. This framework was chosen as it allowed us to implement an approach agnostic to the tensor library. Despite only utilising Tensorflow in this study, this choice allows us to expand this approach to other libraries in the future,

³<https://flower.dev/>

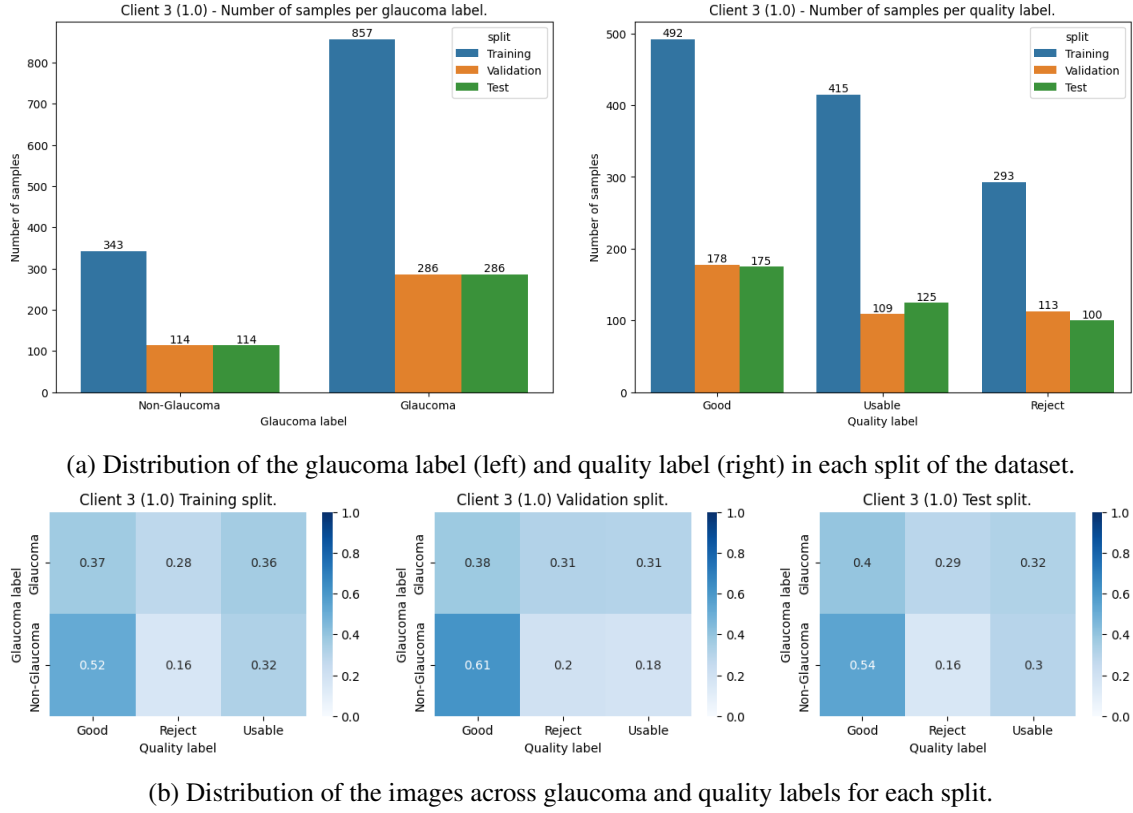
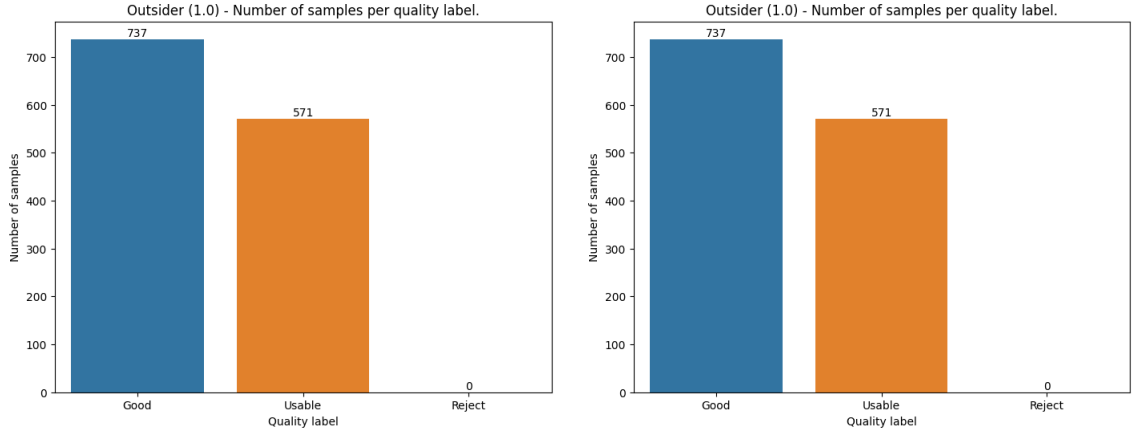


Figure 4.10: Client 3's dataset 1.0.

if necessary. Another major feature that influenced our choice was the ease of implementing new federated strategies and the expansion of already implemented strategies. The local training process of the clients was implemented with TensorFlow⁴ with the glaucoma classification model mentioned above as the base model for initialization. For the model, we used class-weighted binary cross-entropy loss and Adam optimizer with a learning rate $2.5e - 5$, decay of 0.0 and first and second momentum estimates of 0.9 and 0.999, respectively. Furthermore, to conduct experiments on the impact of image quality on the training of a federated model, each client data manager had a togglable filter for images labelled "Reject" by the RetQualEvaluator. With this, each strategy implemented will be tested twice on each dataset distribution generated above, with the first being with no filter in place and the second with the filter enabled. As such, the numbers to identify each experiment are detailed in Table 4.8.

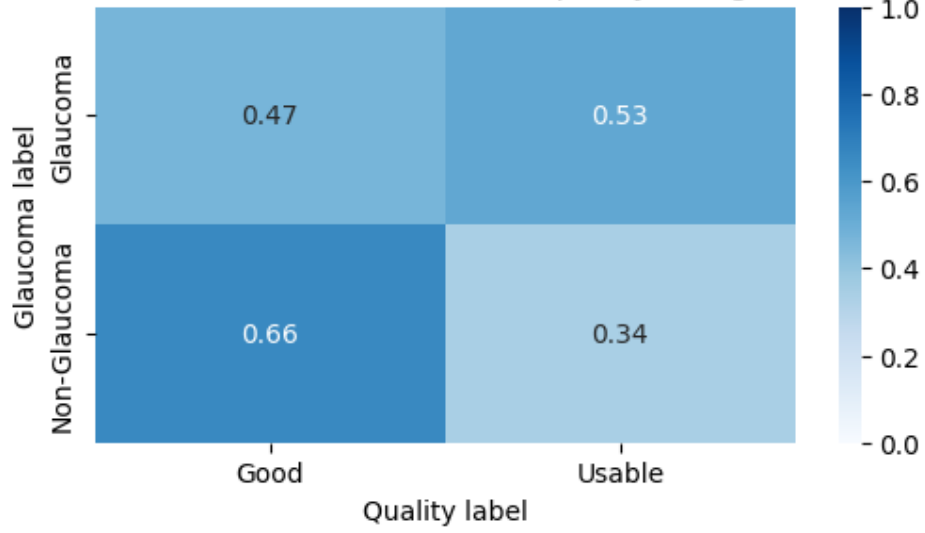
To measure the performance of the models, a collection of metrics was collected by both clients and server at the end of the local training and federated round, respectively. We measured the models and federated strategies regarding training time and performance on the classification task. The following sections will discuss each federated strategy implementation details and its experimental setup.

⁴<https://www.tensorflow.org/>



(a) Distribution of the glaucoma label (left) and quality label (right) of the dataset.

Outsider (1.0) - Dataset distribution over quality and glaucoma labels.



(b) Distribution of the images across glaucoma and quality labels.

Figure 4.11: Outsider's dataset 1.0.

Table 4.8: Experimental setups details.

Experiment	Dataset	Filter
1.x	1.0	✗
2.x	1.0	✓
3.x	2.0	✗
4.x	2.0	✓

FedAvg and FedSGD

With FedAvg [40], each client was trained locally for 25 epochs for five server rounds. At the end of the round, all clients send the updated model weights in their totality to the server, which

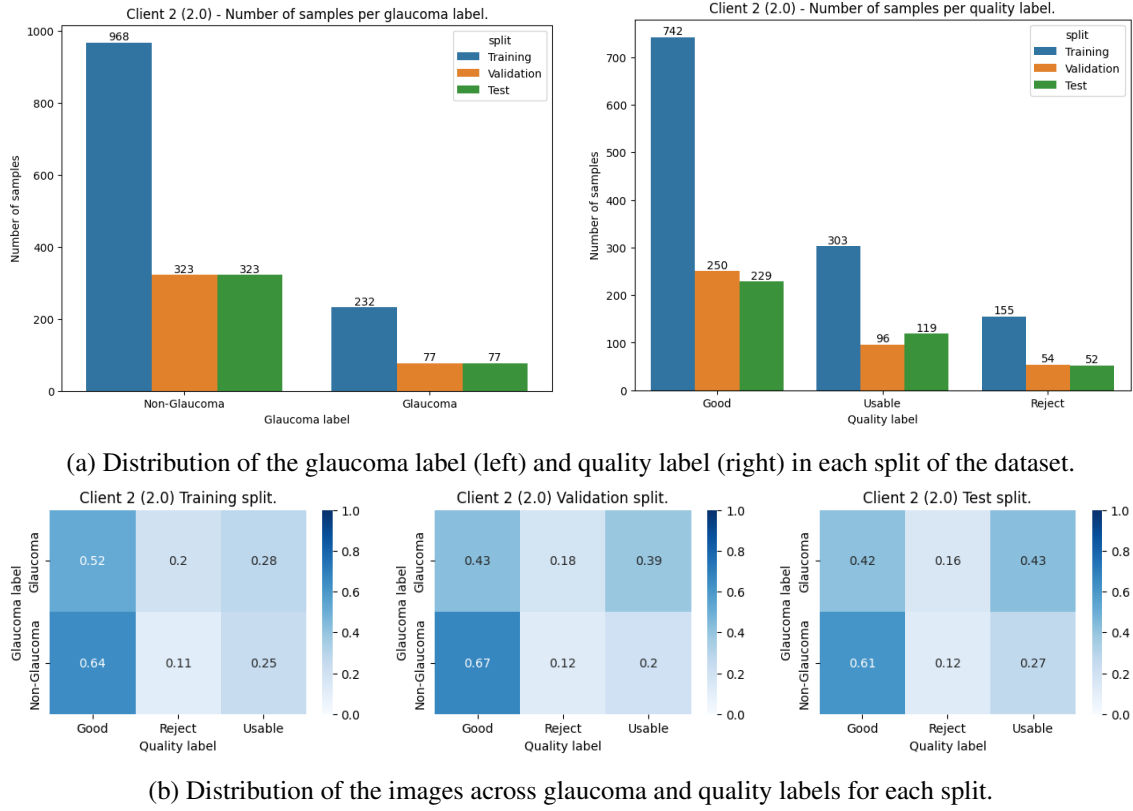


Figure 4.12: Client 2's dataset 2.0.

aggregates them with a weighted average and proceeds to send the aggregated weights to each client. This strategy serves as the baseline for federated solutions.

The FedSGD [40] algorithm shares the implementation of FedAvg with one modification: each client now performs only one local step (epoch) per federated round. To account for this adjustment, the number of federated rounds was extended to 125.

FedProx

The FedProx [28] algorithm has no alterations compared to FedAvg server-side. This is, on each federated round, the clients' weights are aggregated using a weighted average. However, one key modification happens client-side: a proximal term (μ) is added to the local updates to limit the divergence of the local model from the global model in each federated round. We implemented an optimizer following the author's implementation and description to achieve this. This optimizer replaces the Adam optimizer described above while keeping the learning rate at $2.5e - 5$. Larger values of μ lead to more restrictive updates by forcing it to remain closer to the starting point, which can help the aggregation of statistically heterogeneous updates from the clients. Following the author's study, we test three different values for μ . The Table 4.9 presents the values used for μ on each experiment.

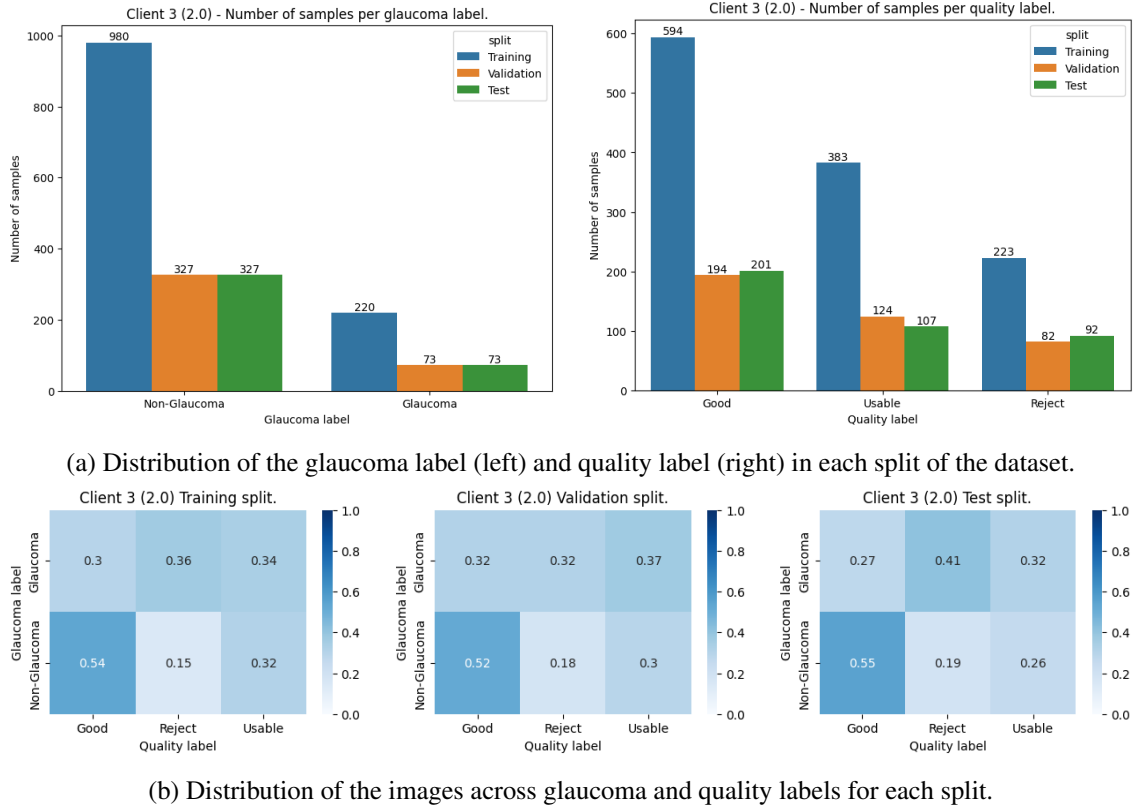


Figure 4.13: Client 3's dataset 2.0.

Table 4.9: Values for FedProx's proximal term (μ) on each experiment.

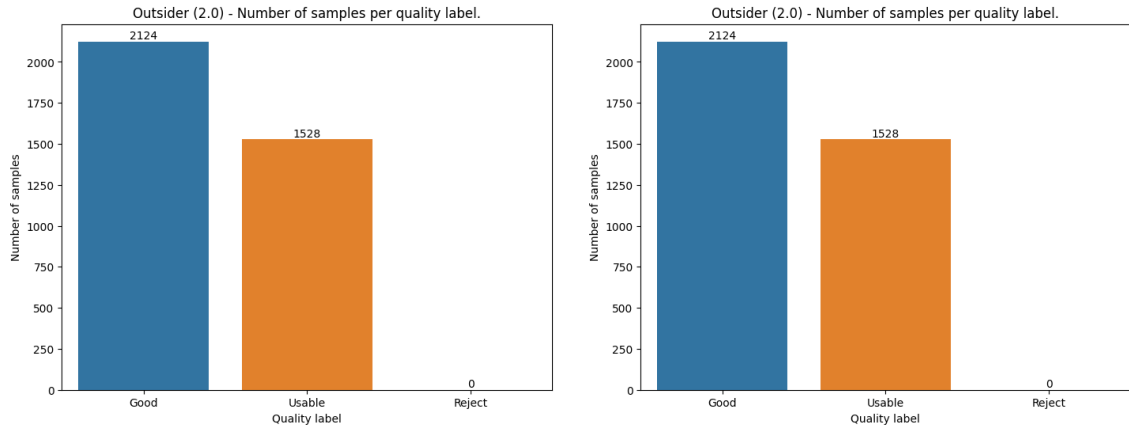
Experiment	Proximal Term (μ)
x.0	1.0
x.1	0.01
x.2	0.1

FedBN

The FedBN [30] algorithm, like FedProx, has no alterations compared to FedAvg server-side. Client-side, there's one key modification to the weights sent to the server: the batch normalization layers are excluded from being transmitted to the server. Everything else is kept the same, and the federation is run for five federated rounds, with each client training for 25 epochs on each round.

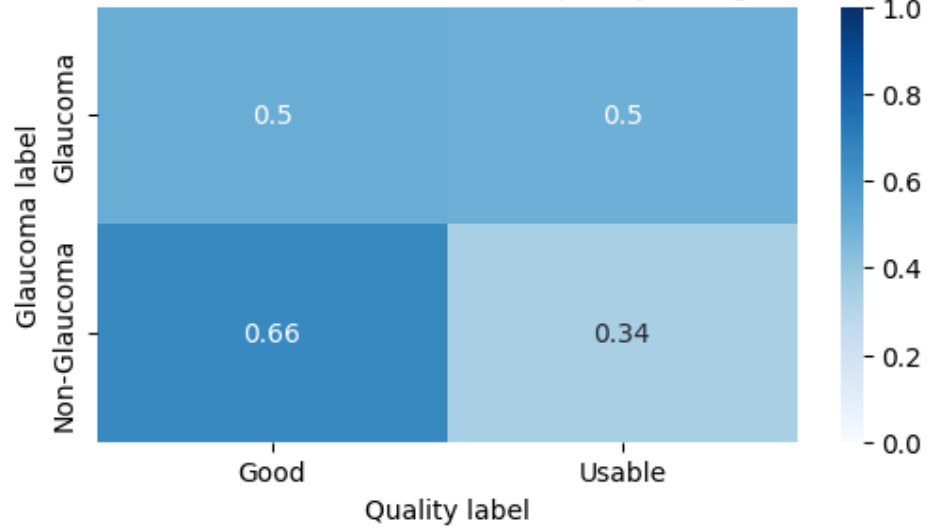
FedYogi

The FedOpt [50] algorithm is a family of federated strategies based on adaptive optimizers such as Adam, Adagrad and Yogi. In this study, we implemented the strategy that obtained the best results on the authors' study, the FedYogi variant. This strategy requires no modifications client-side.



(a) Distribution of the glaucoma label (left) and quality label (right) of the dataset.

Outsider (2.0) - Dataset distribution over quality and glaucoma labels.



(b) Distribution of the images across glaucoma and quality labels.

Figure 4.14: Outsider's dataset 2.0.

However, in the server aggregation of the clients' models, this strategy transforms the formulation for FedAvg weighted average into a pseudo-gradient update. This allows the introduction of optimization on the server side from a global perspective. In this study, we test with two different learning rates (η) for the strategy based on the best values obtained by the authors. The remaining hyperparameters of the strategy remained fixed: first (β_1) and second (β_2) estimates with values 0.9 and 0.999, respectively, and the parameter τ that controls the degree of adaptability with the value of 0.001. Table 4.10 presents the values used for the hyperparameters in each experiment.

Table 4.10: Values for FedYogi’s hyperparameters on each experiment.

Experiment	η	β_1	β_2	τ
x.0	0.01	0.9	0.999	0.001
x.1	0.001	0.9	0.999	0.001

Local-Only and Centralized

To compose the non-federated approaches, we train a local model on each client solely using the client’s local dataset. We also train a centralized model on all the client’s data combined into a single data centre. This serves as a comparison point of non-federated approaches to the federated approach. Both models were trained for 125 epochs, similar to the approach taken on FedSGD, with an early-stopping mechanism to avoid overfitting the training data. The early stopping had a tolerance of 15 epochs without improvement in the validation loss. We also measured the performance of the models without early stopping to ensure the mechanism wasn’t negatively impacting the results.

Chapter 5

Results and Analysis

In this chapter, we will analyse and discuss the results obtained from the experiments described. To evaluate and compare the different federated strategies, we tracked a collection of metrics. However, not all metrics will be displayed on every table and chart discussed in this chapter.

5.1 Convergence Analysis

In this section, we analyse the evolution of the training and validation curves for the loss and accuracy of the various strategies. Fig. 5.1 shows the evolution of the loss and accuracy curves in the training phase. FedYogi shows the most variability of all the strategies on both variations of η , i.e., the server learning rate. While FedYogi with $\eta = 0.001$ presents higher spikes of loss and lower training accuracy than FedYogi with $\eta = 0.01$, the latter diverged on the validation dataset, labelling all the cases with negative labels (i.e. non-glaucomatous). The same behaviour is shown on the test dataset, both locally and on the outsider dataset, as presented in Table 5.4. Meanwhile, FedYogi with $\eta = 0.001$ converges at a slower rate on the training dataset but, in turn, achieves a better performance on the validation dataset. Despite that, its performance achieved is lower than the rest of the strategies. The slower convergence rate of this strategy results in the requirement of longer training times with a larger amount of rounds compared to the rest.

The remaining strategies all exhibit a similar convergence rate, with FedSGD presenting the slowest among them. This is to be expected, as FedSGD only performs a single local epoch in each federated round, leading to a higher degree of variability as the aggregation of weights from heterogeneous sources occurs more frequently. Nevertheless, all federated strategies, except FedYogi, had practically no difference in the validation dataset, with the centralized and local-only approaches showing the fastest convergences on the local dataset, achieving equivalent or superior results in half the number of epochs.

Although the faster convergence compared to federated strategies, the local-only model does not achieve the best performance on the test datasets, as will be discussed in the next sections. To verify that the early stopping mechanism wasn't the cause of the reduced performance, we also analysed the performance of the local-only model. Table 5.1 shows the difference in performance

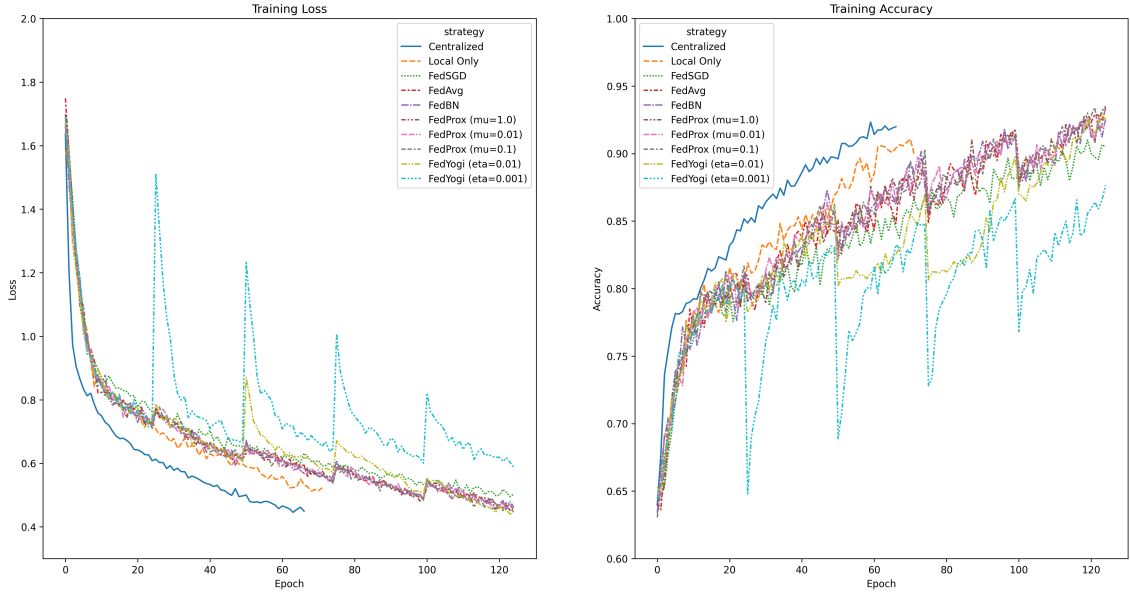


Figure 5.1: Evolution of the training loss curve and accuracy of each strategy on experiment 3.0 in client 1.

For federated strategies, every 25 epochs correspond to the end of a federated round. Thus the trend of variation spikes for all federated strategies at those breakpoints as the models from the clients are aggregated in that instant. FedSGD is the exception due to only having one local epoch on each federated round.

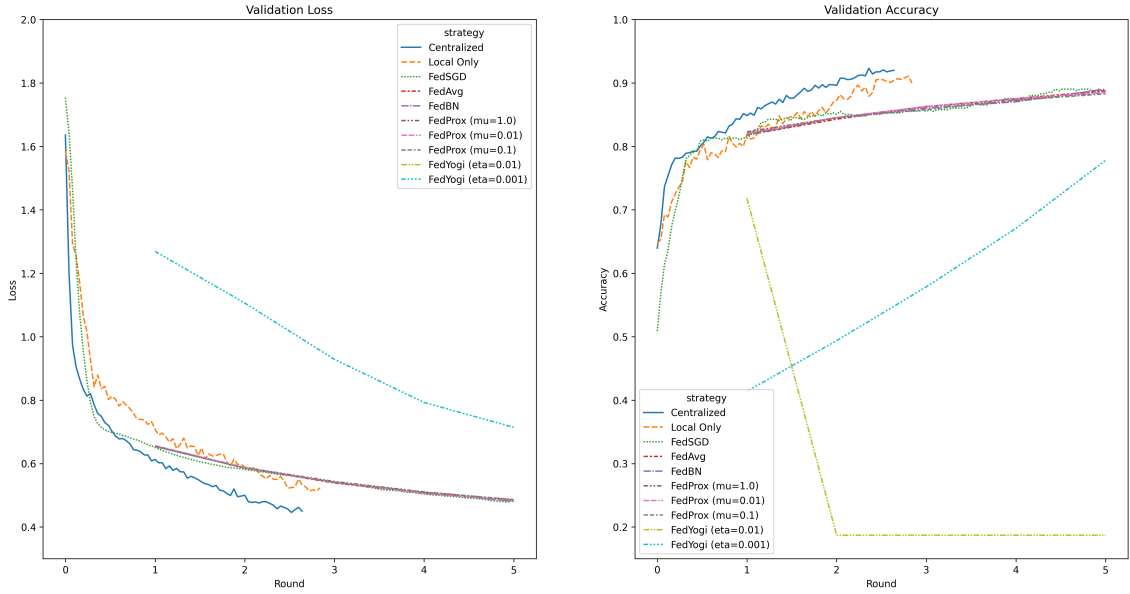


Figure 5.2: Evolution of the validation loss curve and accuracy of each strategy on experiment 3.0 in client 1.

The federated strategies validation only begins after the end of the first round, thus not showing any results between round 0 and round 1, except for FedSGD. Centralized and local-only trained do not have any rounds. Still, for comparison purposes, we scaled the evolution down by 25, as each federated strategy trained for 25 local epochs on each federated round. FedSGD is an exception to the federated strategies, as it only trains for one epoch on each federated round but, in turn, trains for a larger number of rounds. For this reason, we also scaled FedSGD down by a factor of 25 to be comparable to the other strategies.

between the local-only models trained with and without early stopping. While the early stopping resulted in lower performance, the difference is marginal and does not alter the conclusions obtained.

Table 5.1: Effect of early-stopping on the performance of the local-only model.

Strategy	Fit Time (h)	Local Performance					Outsider Performance				
		Accuracy	Precision	Recall	F1 Score	Specificity	Accuracy	Precision	Recall	F1 Score	Specificity
Local-only (no early-stopping)	2.7	0.855	0.901	0.924	0.912	0.548	0.698	0.632	0.951	0.759	0.446
Local-only (early-stopping)	1.8	0.850	0.899	0.920	0.909	0.534	0.696	0.630	0.946	0.756	0.445

Performance was measured in the third client using dataset 2.0 and no filtering during training (experiment 3.0).

5.2 Performance Analysis

In this section, we analyse the performance of the different strategies in the test environments. For this, we compare the performance of the models on its local test dataset that will match the distribution and quality of the data used for training the model. These results allow us to observe the performance of the models when optimising the model for the client. We also conduct an additional comparison on a test dataset to represent an outsider institution of the federation, in which we evaluate the model’s generalizability. We also compare the performance of the federated strategies compared to the local-only and centralized approaches. For this, we trained a local-only model on each client’s data and a centralized model on all the aggregated data. For the locally trained models, only the best of the models will be displayed on the results.

5.2.1 Analysis on the local-only approach

Table 5.2 and Table 5.3 contain the performances of the local-only models of each client on its local test dataset and the outsider test dataset. In the first experiment, denoted by Table 5.2, client 2 achieved the best performance overall on the outsider dataset, followed closely by client 1. This is to be expected, as the distribution of client 2 presents similar characteristics to the outsider dataset, in which both contain a close to even distribution of glaucomatous and non-glaucomatous cases. And both client 1 and client 2 are composed of retinal images with high quality, which matches more closely to the outsider test dataset, which contains no images rejected by the RetQualEvaluator. On the other hand, client 3 contained a higher percentage of images of "Reject" quality, as well as presenting a distribution of mostly positive cases, which diverged the most from the outsider test dataset, and, consequently, shows the lowest performance of the three.

By analysing the precision, recall and specificity of the three models, we can observe a noteworthy difference in the performance of clients 1 and 3. Client 1 was trained mostly on negative cases, while client 3 was trained mostly on positive cases. Despite that, client 1’s model achieves higher recall and lower specificity, while client 3’s model achieves the opposite. While counterintuitive, this can be attributed to the difference in the dataset’s characteristics in which the models were trained compared to the outsider test dataset. Client 3’s model learnt to be more specific on

the positive cases, which led to the opposite effect of the expected when faced with the outsider dataset whose characteristics diverged from the local dataset, resulting in more false negatives and, consequently, lower recall. The same behaviour can be seen on client 1's model but in the opposite direction.

Table 5.2: Performance of the clients' local-only model on its local test dataset and the outsider test dataset.

Client	Local Performance					Outsider Performance				
	Accuracy	Precision	Recall	F1 Score	Specificity	Accuracy	Precision	Recall	F1 Score	Specificity
Client 1	0.825	0.908	0.874	0.891	0.608	0.751	0.704	0.864	0.776	0.638
Client 2	0.752	0.751	0.797	0.733	0.702	0.783	0.794	0.765	0.779	0.801
Client 3	0.738	0.532	0.649	0.585	0.733	0.739	0.759	0.699	0.728	0.778

Performance was measured using dataset 1.0 and no filtering during training (experiment 1.0).

In the second experiment, denoted by Table 5.3, all the clients were trained on datasets with similar distributions of mostly negative cases, with the key difference between each of the clients being the overall quality of the images, with client 1 presenting the highest quality and client 3 the lowest. Client 1 performs best on the outsider test dataset, being significantly ahead of the other clients except for exhibiting lower recall. Under the same distribution across clients, this is expected as the characteristics of client 1's dataset match closest with the outsider test dataset, in which both are composed of high-quality images. Clients 2 and 3 exhibited higher recall but lower precision and specificity, which means a higher percentage of false positives. This can be attributed to the quality of images on which the model was trained, where a higher percentage of the lower quality images are of glaucomatous cases (i.e. "Reject" and "Usable" quality) and the higher-quality of non-glaucomatous cases (i.e. "Good" quality). This led to a higher percentage of false positives when tested on a dataset with a higher percentage of "Usable" cases on non-glaucomatous cases.

Table 5.3: Performance of the clients' local-only model on its local test dataset and the outsider test dataset.

Client	Local Performance					Outsider Performance				
	Accuracy	Precision	Recall	F1 Score	Specificity	Accuracy	Precision	Recall	F1 Score	Specificity
Client 1	0.827	0.916	0.868	0.891	0.649	0.754	0.712	0.854	0.777	0.655
Client 2	0.817	0.903	0.867	0.885	0.610	0.683	0.620	0.944	0.749	0.422
Client 3	0.850	0.899	0.920	0.909	0.534	0.696	0.630	0.946	0.756	0.445

Performance was measured using dataset 2.0 and no filtering during training (experiment 3.0).

5.2.2 Analysis of the federated approach

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

Table 5.4 shows the performance of all the strategies on client 3's local test dataset and outsider test dataset. We highlighted with orange the best performance between the centralized and

the local-only model, representing the non-federated scenario, and the best-performing strategies overall were marked with a green highlight. As we had verified earlier during the convergence analysis, all federated strategies, except FedYogi, performed similarly with only marginal differences between them. The variations of the hyperparameter of FedProx (μ) didn't result in significant differences in performance, and the best-performing value for the proximal term would vary depending on what client and what dataset were to be used. Nevertheless, this hyperparameter optimization for FedProx can, based on our experiments, improve the performance by 1%, and in other scenarios, it would require another search for the optimal value.

Despite the faster convergence on the training and validation datasets, as shown earlier, the local-only model performance is lower compared to all the federated strategies, with the performance gain from the federation ranging from a marginal difference on 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 5.5, 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 and those that fit better into a more controlled environment. In Table 5.4 and Table 5.5 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 variables of heterogeneity, where both the distribution and quality of the data varied across every client.
2. FedSGD and FedProx achieve the most generalizable model 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 models from diverging too much from the global model. However, the FedProx approach does not result in increased communication costs.

Despite not achieving the best performance, FedAvg remained close to every strategy in all scenarios, making it a viable choice in scenarios where there's 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, as depending on the amount of data available on a

client after filtering, the filter can negatively affect the performance, as observed in client 3 on experiment 1.0, which had a reduction of 24% on 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.

Table 5.4: Performance of the different strategies on the local test dataset and the outsider test dataset.

Strategy	Fit Time (h)	Local Performance					Outsider Performance				
		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 in the third client using dataset 2.0 and no filtering during training (experiment 3.0).

Additionally, we can also analyse the impact of the heterogeneity of the clients on the performance in 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 of the target label were similar across the clients, as shown in Table 5.5. Clients in both experiments had similar quality distributions; this is, 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 label, 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.

Finally, we can also compare the federated approach compared to the centralized approach. 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 thesis. 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.

Table 5.5: Accuracy achieved on the test datasets for each federated strategy.

Strategy	Filter	Dataset 1.0				Dataset 2.0			
		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

¹For federated strategies with multiple experiments, only the best is represented in the table.

²The 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.

5.3 Final Remarks

In this section, we will summarise the results obtained and answer the research questions proposed. Firstly, we analysed the convergence of the various strategies and observed a similar pattern on all strategies except for FedYogi, which diverged in one experiment and had a significantly slower convergence in the other. We also observed the centralized and local-only models converge at a faster rate compared to all the federated models and that early stopping doesn't negatively affect the performance of the models.

Despite that, we later observed the performance of the local-only model on the test datasets was lower than the performance of the majority of the federated strategies, especially on whose local data was comprised of lower-quality images. With this, we can answer one of the research questions proposed: "Does the federated approach have performance gains over non-federated approaches?"; from our results, we conclude 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 also can 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 the centralized while maintaining the data secure and isolated in their own client's network.

Furthermore, we also analysed the performance differences of each strategy for each of the experiments. From that, we observed the model's performance would be lower if trained in a more heterogeneous environment, as can be seen in client 1's tests, whose data remained identical across experiments. However, eliminating lower-quality images from the training process to achieve a more consistent quality distribution across clients doesn't always benefit the model's performance, as we observed in our experiments. With this, we answer our second research question: "What is the impact of the diverse properties of the clients and the nature of the data on the performance of the FL algorithms?"; from our results, 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, as we'll discuss in the next paragraph.

The last observation from our experiments is the adaptability of certain strategies to each context. For example, we observed FedBN would generally perform the best given a scenario where the federated model goal was to optimise the performance of the client's local data, given those clients followed similar distributions on their data. On the other hand, FedSGD and FedProx exhibited higher performance under scenarios where the clients' data is more diverse. These strategies also showed higher generalizability, as we can observe by the performances on the outsider's dataset. With this, we can answer our final research question: "Given the nature of the data, such as data distribution or data quality, and the model's objective, such as optimising local performance or generalized performance, what is the most suitable FL algorithm?"; from our results, we concluded that FedSGD and FedProx would be the most suitable strategies on scenarios where 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.

Chapter 6

Conclusions

FL approaches show great potential in digital healthcare due to their nature of ensuring a more secure environment when working with sensitive data. Research on this topic has been rapidly growing in the past years, and we expect to grow even further with the frameworks such as NVIDIA Flare [53], which aims to provide a scalable framework for federated learning, along with new algorithms to generate more optimal global models.

In this thesis, 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 utilising public datasets of retinal fundus images from different sources to create clients with different characteristics. The model developed by Leonardo et al. [27] was used as the base model for the federated setup. We evaluated different FL algorithms in these settings and analysed the results obtained. And from the experiments, we concluded that the use of federated learning in this area of application resulted in an improvement in the model performance over the centralized approach. We also concluded from our experiments that federated learning is 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. For example, 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.

Finally, we also delineate some limitations of our experiments and evaluation: (i) the experiments only consist of a limited set of federated strategies, and further evaluation on more recent approaches, such as incorporating reinforcement learning to learn the aggregation of the clients' models [17] and incorporation of domain generalization and personalization approaches into the federation [32, 22]; (ii) the incorporation of cases from 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; and (iii) 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 in different tasks.

References

- [1] Optical Coherence Tomography: What It Is & How It's Used. URL <https://my.clevelandclinic.org/health/diagnostics/17293-optical-coherence-tomography>.
- [2] What data can we process and under which conditions? URL https://commission.europa.eu/law/law-topic/data-protection/reform/rules-business-and-organisations/principles-gdpr/what-data-can-we-process-and-under-which-conditions_en.
- [3] Federated Learning: Collaborative Machine Learning without Centralized Training Data. URL <https://ai.googleblog.com/2017/04/federated-learning-collaborative.html>.
- [4] Tonometry: Purpose, Procedure, and Results. URL <https://www.healthline.com/health/tonometry#procedure>.
- [5] Optical Coherence Tomography | Department of Ophthalmology. URL <https://ophthalmology.med.ubc.ca/patient-care/ophthalmic-photography/optical-coherence-tomography/>.
- [6] Learn Why the Pachymetry Test Is Important for Diagnosing Glaucoma. URL <https://www.verywellhealth.com/understanding-pachymetry-3421820>. Section: Verywell.
- [7] Adekanmi Adegun and Serestina Viriri. Deep learning techniques for skin lesion analysis and melanoma cancer detection: a survey of state-of-the-art. *Artificial Intelligence Review*, 54(2):811–841, February 2021. ISSN 1573-7462. doi: 10.1007/s10462-020-09865-y. URL <https://doi.org/10.1007/s10462-020-09865-y>.
- [8] Adnan Ahmad, Wei Luo, and Antonio Robles-Kelly. Robust federated learning under statistical heterogeneity via hessian-weighted aggregation. *Machine Learning*, 112(2):633–654, February 2023. ISSN 1573-0565. doi: 10.1007/s10994-022-06292-8. URL <https://doi.org/10.1007/s10994-022-06292-8>.
- [9] Karen Allison, Deepkumar Patel, and Omobolanle Alabi. Epidemiology of Glaucoma: The Past, Present, and Predictions for the Future. *Cureus*, 12(11):e11686, November 2020. ISSN 2168-8184. doi: 10.7759/cureus.11686. URL <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7769798/>.
- [10] José Camara, Alexandre Neto, Ivan Miguel Pires, María Vanessa Villasana, Eftim Zdravevski, and António Cunha. Literature Review on Artificial Intelligence Methods for Glaucoma Screening, Segmentation, and Classification. *Journal of Imaging*, 8(2):19, January 2022. ISSN 2313-433X. doi: 10.3390/jimaging8020019.

- [11] Jianmin Chen, Xinghao Pan, Rajat Monga, Samy Bengio, and Rafal Jozefowicz. Revisiting Distributed Synchronous SGD, March 2017. URL <http://arxiv.org/abs/1604.00981>. arXiv:1604.00981 [cs].
- [12] Mark Christopher, Akram Belghith, Christopher Bowd, James A. Proudfoot, Michael H. Goldbaum, Robert N. Weinreb, Christopher A. Girkin, Jeffrey M. Liebmann, and Linda M. Zangwill. Performance of Deep Learning Architectures and Transfer Learning for Detecting Glaucomatous Optic Neuropathy in Fundus Photographs. *Scientific Reports*, 8(1):16685, November 2018. ISSN 2045-2322. doi: 10.1038/s41598-018-35044-9. URL <https://www.nature.com/articles/s41598-018-35044-9>. Number: 1 Publisher: Nature Publishing Group.
- [13] Kumale Tolesa Daba, Girum W Gessesse, and Sisay Bekele Sori. Proportion of Glaucoma among Voluntary People Coming for Glaucoma Screening Program at Jimma University Department of Ophthalmology, Jimma, Ethiopia. *Ethiopian Journal of Health Sciences*, 30(1):13–22, January 2020. ISSN 1029-1857. doi: 10.4314/ejhs.v30i1.3. URL <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7036452/>.
- [14] Coen de Vente, Koenraad A. Vermeer, Nicolas Jaccard, He Wang, Hongyi Sun, Firas Khader, Daniel Truhn, Temirgali Aimyshev, Yerkebulan Zhanibekuly, Tien-Dung Le, Adrian Galdran, Miguel Ángel González Ballester, Gustavo Carneiro, Devika R. G, Hrishikesh P. S, Densen Puthussery, Hong Liu, Zekang Yang, Satoshi Kondo, Satoshi Kasai, Edward Wang, Ashritha Durvasula, Jónathan Heras, Miguel Ángel Zapata, Teresa Araújo, Guilherme Aresta, Hrvoje Bogunović, Mustafa Arikian, Yeong Chan Lee, Hyun Bin Cho, Yoon Ho Choi, Abdul Qayyum, Imran Razzak, Bram van Ginneken, Hans G. Lemij, and Clara I. Sánchez. AIROGS: Artificial Intelligence for RObust Glaucoma Screening Challenge, February 2023. URL <http://arxiv.org/abs/2302.01738>. arXiv:2302.01738 [cs, eess].
- [15] Malay Kishore Dutta, Amit Kumar Mourya, Anushikha Singh, M. Parthasarathi, Radim Burget, and Kamil Riha. Glaucoma detection by segmenting the super pixels from fundus colour retinal images. In *2014 International Conference on Medical Imaging, m-Health and Emerging Communication Systems (MedCom)*, pages 86–90, November 2014. doi: 10.1109/MedCom.2014.7005981.
- [16] Gozde N. Gunesli, Mohsin Bilal, Shan E. Ahmed Raza, and Nasir M. Rajpoot. Fed-DropoutAvg: Generalizable federated learning for histopathology image classification, November 2021. URL <http://arxiv.org/abs/2111.13230>. arXiv:2111.13230 [cs].
- [17] Pengfei Guo, Dong Yang, Ali Hatamizadeh, An Xu, Ziyue Xu, Wenqi Li, Can Zhao, Daguang Xu, Stephanie Harmon, Evrim Turkbey, Baris Turkbey, Bradford Wood, Francesca Patella, Elvira Stellato, Gianpaolo Carrafiello, Vishal M. Patel, and Holger R. Roth. Auto-FedRL: Federated Hyperparameter Optimization for Multi-institutional Medical Image Segmentation, August 2022. URL <http://arxiv.org/abs/2203.06338>. arXiv:2203.06338 [cs, eess].
- [18] Yuki Hagiwara, Joel En Wei Koh, Jen Hong Tan, Sulatha V. Bhandary, Augustinus Laude, Edward J. Ciaccio, Louis Tong, and U. Rajendra Acharya. Computer-aided diagnosis of glaucoma using fundus images: A review. *Computer Methods and Programs in Biomedicine*, 165:1–12, October 2018. ISSN 0169-2607. doi: 10.1016/j.cmpb.2018.07.012. URL <https://www.sciencedirect.com/science/article/pii/S0169260718305492>.

- [19] Adam Hanif, Charles Lu, Ken Chang, Praveer Singh, Aaron S. Coyner, James M. Brown, Susan Ostmo, Robison V. Paul Chan, Daniel Rubin, Michael F. Chiang, Jayashree Kalpathy-Cramer, John Peter Campbell, Michael F. Chiang, Susan Ostmo, Sang Jin Kim, Kemal Sonmez, John Peter Campbell, Robert Schelonka, Aaron Coyner, R. V. Paul Chan, Karyn Jonas, Bhavana Kolli, Jason Horowitz, Osode Coki, Cheryl-Ann Eccles, Leora Sarna, Anton Orlin, Audina Berrocal, Catherin Negron, Kimberly Denser, Kristi Cumming, Tammy Osentoski, Tammy Check, Mary Zajechowski, Thomas Lee, Aaron Nagiel, Evan Kruger, Kathryn McGovern, Dilshad Contractor, Margaret Havunjian, Charles Simmons, Raghu Murthy, Sharon Galvis, Jerome Rotter, Ida Chen, Xiaohui Li, Kent Taylor, Kaye Roll, Mary Elizabeth Hartnett, Leah Owen, Darius Moshfeghi, Mariana Nunez, Zac Wennber-Smith, Jayashree Kalpathy-Cramer, Deniz Erdogmus, Stratis Ioannidis, Maria Ana Martinez-Castellanos, Samantha Salinas-Longoria, Rafael Romero, Andrea Arriola, Francisco Olguin-Manriquez, Miroslava Meraz-Gutierrez, Carlos M. Dulanto-Reinoso, and Cristina Montero-Mendoza. Federated Learning for Multicenter Collaboration in Ophthalmology: Implications for Clinical Diagnosis and Disease Epidemiology. *Ophthalmology Retina*, 6(8):650–656, August 2022. ISSN 2468-6530. doi: 10.1016/j.oret.2022.03.005. URL <https://www.sciencedirect.com/science/article/pii/S246865302200104X>.
- [20] Trang-Thi Ho, Khoa-Dang Tran, and Yennun Huang. FedSGDCOVID: Federated SGD COVID-19 Detection under Local Differential Privacy Using Chest X-ray Images and Symptom Information. *Sensors*, 22(10):3728, January 2022. ISSN 1424-8220. doi: 10.3390/s22103728. URL <https://www.mdpi.com/1424-8220/22/10/3728>. Number: 10 Publisher: Multidisciplinary Digital Publishing Institute.
- [21] Chao Huang, Jianwei Huang, and Xin Liu. Cross-Silo Federated Learning: Challenges and Opportunities, June 2022. URL <http://arxiv.org/abs/2206.12949>. arXiv:2206.12949 [cs].
- [22] Meirui Jiang, Hongzheng Yang, Chen Cheng, and Qi Dou. IOP-FL: Inside-Outside Personalization for Federated Medical Image Segmentation, April 2022. URL <http://arxiv.org/abs/2204.08467>. arXiv:2204.08467 [cs, eess].
- [23] Yong Woo Kim, Young In Yun, and Hyuk Jin Choi. Screening fundus photography predicts and reveals risk factors for glaucoma conversion in eyes with large optic disc cupping. *Scientific Reports*, 13(1):81, January 2023. ISSN 2045-2322. doi: 10.1038/s41598-022-26798-4. URL <https://www.nature.com/articles/s41598-022-26798-4>. Number: 1 Publisher: Nature Publishing Group.
- [24] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, March 2017. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1611835114. URL <http://arxiv.org/abs/1612.00796>. arXiv:1612.00796 [cs, stat].
- [25] David B. Larson, David C. Magnus, Matthew P. Lungren, Nigam H. Shah, and Curtis P. Langlotz. Ethics of Using and Sharing Clinical Imaging Data for Artificial Intelligence: A Proposed Framework. *Radiology*, 295(3):675–682, June 2020. ISSN 0033-8419. doi: 10.1148/radiol.2020192536. URL <https://pubs.rsna.org/doi/10.1148/radiol.2020192536>. Publisher: Radiological Society of North America.

- [26] Constance D. Lehman, Adam Yala, Tal Schuster, Brian Dontchos, Manisha Bahl, Kyle Swanson, and Regina Barzilay. Mammographic Breast Density Assessment Using Deep Learning: Clinical Implementation. *Radiology*, 290(1):52–58, January 2019. ISSN 1527-1315. doi: 10.1148/radiol.2018180694.
- [27] Ricardo Leonardo, João Gonçalves, André Carreiro, Beatriz Simões, Tiago Oliveira, and Filipe Soares. Impact of Generative Modeling for Fundus Image Augmentation With Improved and Degraded Quality in the Classification of Glaucoma. *IEEE Access*, 10:111636–111649, 2022. ISSN 2169-3536. doi: 10.1109/ACCESS.2022.3215126. Conference Name: IEEE Access.
- [28] Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated Optimization in Heterogeneous Networks, April 2020. URL <http://arxiv.org/abs/1812.06127>. arXiv:1812.06127 [cs, stat].
- [29] Wenqi Li, Fausto Milletari, Daguang Xu, Nicola Rieke, Jonny Hancox, Wentao Zhu, Maximilian Baust, Yan Cheng, Sébastien Ourselin, M. Jorge Cardoso, and Andrew Feng. Privacy-preserving Federated Brain Tumour Segmentation, October 2019. URL <http://arxiv.org/abs/1910.00962>. arXiv:1910.00962 [cs].
- [30] Xiaoxiao Li, Meirui Jiang, Xiaofei Zhang, Michael Kamp, and Qi Dou. FedBN: Federated Learning on Non-IID Features via Local Batch Normalization, May 2021. URL <http://arxiv.org/abs/2102.07623>. arXiv:2102.07623 [cs].
- [31] Jiang Liu, Zhuo Zhang, Damon Wing Kee Wong, Yanwu Xu, Fengshou Yin, Jun Cheng, Ngan Meng Tan, Chee Keong Kwoh, Dong Xu, Yih Chung Tham, Tin Aung, and Tien Yin Wong. Automatic glaucoma diagnosis through medical imaging informatics. *Journal of the American Medical Informatics Association: JAMIA*, 20(6):1021–1027, 2013. ISSN 1527-974X. doi: 10.1136/amiajnl-2012-001336.
- [32] Quande Liu, Cheng Chen, Jing Qin, Qi Dou, and Pheng-Ann Heng. FedDG: Federated Domain Generalization on Medical Image Segmentation via Episodic Learning in Continuous Frequency Space, March 2021. URL <http://arxiv.org/abs/2103.06030>. arXiv:2103.06030 [cs].
- [33] Shengfeng Liu, Yi Wang, Xin Yang, Baiying Lei, Li Liu, Shawn Xiang Li, Dong Ni, and Tianfu Wang. Deep Learning in Medical Ultrasound Analysis: A Review. *Engineering*, 5(2):261–275, April 2019. ISSN 2095-8099. doi: 10.1016/j.eng.2018.11.020. URL <https://www.sciencedirect.com/science/article/pii/S2095809918301887>.
- [34] Yang Liu, Yan Kang, Chaoping Xing, Tianjian Chen, and Qiang Yang. Secure Federated Transfer Learning. *IEEE Intelligent Systems*, 35(4):70–82, July 2020. ISSN 1541-1672, 1941-1294. doi: 10.1109/MIS.2020.2988525. URL <http://arxiv.org/abs/1812.03337>. arXiv:1812.03337 [cs, stat].
- [35] Yang Liu, Yan Kang, Tianyuan Zou, Yanhong Pu, Yuanqin He, Xiaozhou Ye, Ye Ouyang, Ya-Qin Zhang, and Qiang Yang. Vertical Federated Learning, November 2022. URL <http://arxiv.org/abs/2211.12814>. arXiv:2211.12814 [cs].
- [36] Julian Lo, Timothy T. Yu, Da Ma, Pengxiao Zang, Julia P. Owen, Qinqin Zhang, Ruikang K. Wang, Mirza Faisal Beg, Aaron Y. Lee, Yali Jia, and Marinko V. Sarunic. Federated Learning for Microvasculature Segmentation and Diabetic Retinopathy Classification of OCT Data.

- Ophthalmology Science*, 1(4):100069, December 2021. ISSN 2666-9145. doi: 10.1016/j.xops.2021.100069. URL <https://www.sciencedirect.com/science/article/pii/S2666914521000671>.
- [37] Charles Lu, Adam Hanif, Praveer Singh, Ken Chang, Aaron S. Coyner, James M. Brown, Susan Ostmo, Robison V. Paul Chan, Daniel Rubin, Michael F. Chiang, John Peter Campbell, Jayashree Kalpathy-Cramer, Michael F. Chiang, Susan Ostmo, Sang Jin Kim, Kemal Sonmez, John Peter Campbell, Robert Schelonka, Aaron Coyner, R. V. Paul Chan, Karyn Jonas, Bhavana Kolli, Jason Horowitz, Osode Coki, Cheryl-Ann Eccles, Leora Sarna, Anton Orlin, Audina Berrocal, Catherin Negron, Kimberly Denser, Kristi Cumming, Tammy Osentoski, Tammy Check, Mary Zajechowski, Thomas Lee, Aaron Nagiel, Evan Kruger, Kathryn McGovern, Dilshad Contractor, Margaret Havunjian, Charles Simmons, Raghu Murthy, Sharon Galvis, Jerome Rotter, Ida Chen, Xiaohui Li, Kent Taylor, Kaye Roll, Mary Elizabeth Hartnett, Leah Owen, Darius Moshfeghi, Mariana Nunez, Zac Wennber-Smith, Jayashree Kalpathy-Cramer, Deniz Erdogmus, Stratis Ioannidis, Maria Ana Martinez-Castellanos, Samantha Salinas-Longoria, Rafael Romero, Andrea Arriola, Francisco Olguin-Manriquez, Miroslava Meraz-Gutierrez, Carlos M. Dulanto-Reinoso, and Cristina Montero-Mendoza. Federated Learning for Multicenter Collaboration in Ophthalmology: Improving Classification Performance in Retinopathy of Prematurity. *Ophthalmology Retina*, 6(8): 657–663, August 2022. ISSN 2468-6530. doi: 10.1016/j.oret.2022.02.015. URL <https://www.sciencedirect.com/science/article/pii/S2468653022001038>.
- [38] Tirivangani Magadza and Serestina Viriri. Deep Learning for Brain Tumor Segmentation: A Survey of State-of-the-Art. *Journal of Imaging*, 7(2):19, February 2021. ISSN 2313-433X. doi: 10.3390/jimaging7020019. URL <https://www.mdpi.com/2313-433X/7/2/19>. Number: 2 Publisher: Multidisciplinary Digital Publishing Institute.
- [39] José Martins, Jaime S. Cardoso, and Filipe Soares. Offline computer-aided diagnosis for Glaucoma detection using fundus images targeted at mobile devices. *Computer Methods and Programs in Biomedicine*, 192:105341, August 2020. ISSN 0169-2607. doi: 10.1016/j.cmpb.2020.105341. URL <https://www.sciencedirect.com/science/article/pii/S0169260719312015>.
- [40] H. Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-Efficient Learning of Deep Networks from Decentralized Data, February 2017. URL <http://arxiv.org/abs/1602.05629>. arXiv:1602.05629 [cs].
- [41] H. Brendan McMahan, Daniel Ramage, Kunal Talwar, and Li Zhang. Learning Differentially Private Recurrent Language Models, February 2018. URL <http://arxiv.org/abs/1710.06963>. arXiv:1710.06963 [cs].
- [42] Michael V. Drake MD. The Importance of Corneal Thickness | glaucoma.org, March 2022. URL <https://glaucoma.org/the-importance-of-corneal-thickness/>. Section: Eye Exams.
- [43] Marriam Nawaz, Tahira Nazir, Ali Javed, Usman Tariq, Hwan-Seung Yong, Muhammad Attique Khan, and Jaehyuk Cha. An Efficient Deep Learning Approach to Automatic Glaucoma Detection Using Optic Disc and Optic Cup Localization. *Sensors (Basel, Switzerland)*, 22(2):434, January 2022. ISSN 1424-8220. doi: 10.3390/s22020434.
- [44] John Neter. *Applied Linear Statistical Models*. Irwin, 1996. ISBN 978-0-256-11736-3. Google-Books-ID: q2sPAQAAMAAJ.

- [45] Christopher Nielsen, Anup Tuladhar, and Nils D. Forkert. Investigating the Vulnerability of Federated Learning-Based Diabetic Retinopathy Grade Classification to Gradient Inversion Attacks. In Bhavna Antony, Huazhu Fu, Cecilia S. Lee, Tom MacGillivray, Yanwu Xu, and Yalin Zheng, editors, *Ophthalmic Medical Image Analysis*, Lecture Notes in Computer Science, pages 183–192, Cham, 2022. Springer International Publishing. ISBN 978-3-031-16525-2. doi: 10.1007/978-3-031-16525-2_19.
- [46] Luke Oakden-Rayner. The Rebirth of CAD: How Is Modern AI Different from the CAD We Know? *Radiology. Artificial intelligence*, 1(3):e180089, May 2019. ISSN 2638-6100. doi: 10.1148/ryai.2019180089. URL <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8017402/>.
- [47] José Ignacio Orlando, Huazhu Fu, João Barbosa Breda, Karel van Keer, Deepti R. Bathula, Andrés Diaz-Pinto, Ruogu Fang, Pheng-Ann Heng, Jeyoung Kim, JoonHo Lee, Joonseok Lee, Xiaoxiao Li, Peng Liu, Shuai Lu, Balamurali Murugesan, Valery Naranjo, Sai Samarth R. Phaye, Sharath M. Shankaranarayana, Apoorva Sikka, Jaemin Son, Anton van den Hengel, Shujun Wang, Junyan Wu, Zifeng Wu, Guanghui Xu, Yongli Xu, Pengshuai Yin, Fei Li, Xiulan Zhang, Yanwu Xu, and Hrvoje Bogunović. REFUGE Challenge: A unified framework for evaluating automated methods for glaucoma assessment from fundus photographs. *Medical Image Analysis*, 59:101570, January 2020. ISSN 1361-8415. doi: 10.1016/j.media.2019.101570. URL <https://www.sciencedirect.com/science/article/pii/S1361841519301100>.
- [48] Vijayaraghavan Prathiba and Mohan Rema. Teleophthalmology: A Model for Eye Care Delivery in Rural and Underserved Areas of India. *International Journal of Family Medicine*, 2011:683267, 2011. ISSN 2090-2042. doi: 10.1155/2011/683267. URL <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3263845/>.
- [49] An Ran Ran, Clement C. Tham, Poemen P. Chan, Ching-Yu Cheng, Yih-Chung Tham, Tyler Hyungtaek Rim, and Carol Y. Cheung. Deep learning in glaucoma with optical coherence tomography: a review. *Eye*, 35(1):188–201, January 2021. ISSN 1476-5454. doi: 10.1038/s41433-020-01191-5. URL <https://www.nature.com/articles/s41433-020-01191-5>. Number: 1 Publisher: Nature Publishing Group.
- [50] Sashank Reddi, Zachary Charles, Manzil Zaheer, Zachary Garrett, Keith Rush, Jakub Konečný, Sanjiv Kumar, and H. Brendan McMahan. Adaptive Federated Optimization, September 2021. URL <http://arxiv.org/abs/2003.00295>. arXiv:2003.00295 [cs, math, stat].
- [51] Luc Rocher, Julien M. Hendrickx, and Yves-Alexandre de Montjoye. Estimating the success of re-identifications in incomplete datasets using generative models. *Nature Communications*, 10(1):3069, July 2019. ISSN 2041-1723. doi: 10.1038/s41467-019-10933-3. URL <https://www.nature.com/articles/s41467-019-10933-3>. Number: 1 Publisher: Nature Publishing Group.
- [52] Holger R. Roth, Ken Chang, Praveer Singh, Nir Neumark, Wenqi Li, Vikash Gupta, Sharut Gupta, Liangqiong Qu, Alvin Ihsani, Bernardo C. Bizzo, Yuhong Wen, Varun Buch, Meesam Shah, Felipe Kitamura, Matheus Mendonça, Vitor Lavor, Ahmed Harouni, Colin Compas, Jesse Tetreault, Prerna Dogra, Yan Cheng, Selnur Erdal, Richard White, Behrooz Hashemian, Thomas Schultz, Miao Zhang, Adam McCarthy, B. Min Yun, Elshaimaa Sharaf, Katharina V. Hoebel, Jay B. Patel, Bryan Chen, Sean Ko, Evan Leibovitz, Etta D. Pisano,

- Laura Coombs, Daguang Xu, Keith J. Dreyer, Ittai Dayan, Ram C. Naidu, Mona Flores, Daniel Rubin, and Jayashree Kalpathy-Cramer. Federated Learning for Breast Density Classification: A Real-World Implementation. volume 12444, pages 181–191. 2020. doi: 10.1007/978-3-030-60548-3_18. URL <http://arxiv.org/abs/2009.01871>. arXiv:2009.01871 [cs, eess].
- [53] Holger R. Roth, Yan Cheng, Yuhong Wen, Isaac Yang, Ziyue Xu, Yuan-Ting Hsieh, Kristopher Kersten, Ahmed Harouni, Can Zhao, Kevin Lu, Zhihong Zhang, Wenqi Li, Andriy Myronenko, Dong Yang, Sean Yang, Nicola Rieke, Abood Quraini, Chester Chen, Daguang Xu, Nic Ma, Prerna Dogra, Mona Flores, and Andrew Feng. NVIDIA FLARE: Federated Learning from Simulation to Real-World, December 2022. URL <http://arxiv.org/abs/2210.13291>. arXiv:2210.13291 [cs].
- [54] Justin Schweitzer, Matt Geller, and Gold Eleanor. The 2022 Glaucoma Report, June 2022. URL <https://eyesoneyecare.com/resources/2022-glaucoma-report/>.
- [55] Sertan Serte and Hasan Demirel. Deep learning for diagnosis of COVID-19 using 3D CT scans. *Computers in Biology and Medicine*, 132:104306, May 2021. ISSN 0010-4825. doi: 10.1016/j.compbimed.2021.104306. URL <https://www.sciencedirect.com/science/article/pii/S0010482521001001>.
- [56] Micah J. Sheller, G. Anthony Reina, Brandon Edwards, Jason Martin, and Spyridon Bakas. Multi-Institutional Deep Learning Modeling Without Sharing Patient Data: A Feasibility Study on Brain Tumor Segmentation. *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries. BrainLes (Workshop)*, 11383:92–104, 2019. doi: 10.1007/978-3-030-11723-8_9.
- [57] Neta Shoham, Tomer Avidor, Aviv Keren, Nadav Israel, Daniel Benditkis, Liron Mor-Yosef, and Itai Zeitak. Overcoming Forgetting in Federated Learning on Non-IID Data, October 2019. URL <http://arxiv.org/abs/1910.07796>. arXiv:1910.07796 [cs, stat].
- [58] Reza Shokri and Vitaly Shmatikov. Privacy-Preserving Deep Learning. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security, CCS '15*, pages 1310–1321, New York, NY, USA, October 2015. Association for Computing Machinery. ISBN 978-1-4503-3832-5. doi: 10.1145/2810103.2813687. URL <https://doi.org/10.1145/2810103.2813687>.
- [59] Jayanthi Sivaswamy, S. R. Krishnadas, Gopal Datt Joshi, Madhulika Jain, and A. Ujjwaft Syed Tabish. Drishti-GS: Retinal image dataset for optic nerve head (ONH) segmentation. In *2014 IEEE 11th International Symposium on Biomedical Imaging (ISBI)*, pages 53–56, April 2014. doi: 10.1109/ISBI.2014.6867807. ISSN: 1945-8452.
- [60] Mukesh Soni, Nikhil Kumar Singh, Pranjit Das, Mohammad Shabaz, Piyush Kumar Shukla, Partha Sarkar, Shweta Singh, Ismail Keshta, and Ali Rizwan. IoT-Based Federated Learning Model for Hypertensive Retinopathy Lesions Classification. *IEEE Transactions on Computational Social Systems*, pages 1–10, 2022. ISSN 2329-924X. doi: 10.1109/TCSS.2022.3213507. Conference Name: IEEE Transactions on Computational Social Systems.
- [61] Janey L. Wiggs and Louis R. Pasquale. Genetics of glaucoma. *Human Molecular Genetics*, 26(R1):R21–R27, August 2017. ISSN 0964-6906. doi: 10.1093/hmg/ddx184. URL <https://doi.org/10.1093/hmg/ddx184>.

- [62] D.W.K. Wong, J. Liu, J. H. Lim, N.M. Tan, Z. Zhang, S. Lu, H. Li, M.H. Teo, K.L. Chan, and T.Y. Wong. Intelligent fusion of cup-to-disc ratio determination methods for glaucoma detection in ARGALI. In *2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 5777–5780, September 2009. doi: 10.1109/IEMBS.2009.5332534. ISSN: 1558-4615.
- [63] Junde Wu, Huihui Fang, Fei Li, Huazhu Fu, Fengbin Lin, Jiongcheng Li, Lexing Huang, Qinji Yu, Sifan Song, Xinxing Xu, Yanyu Xu, Wensai Wang, Lingxiao Wang, Shuai Lu, Huiqi Li, Shihua Huang, Zhichao Lu, Chubin Ou, Xifei Wei, Bingyuan Liu, Riadh Kobbi, Xiaoying Tang, Li Lin, Qiang Zhou, Qiang Hu, Hrvoje Bogunovic, José Ignacio Orlando, Xiulan Zhang, and Yanwu Xu. GAMMA Challenge: Glaucoma grading from Multi-Modality images, December 2022. URL <http://arxiv.org/abs/2202.06511>. arXiv:2202.06511 [cs] version: 4.
- [64] Yingda Xia, Dong Yang, Wenqi Li, Andriy Myronenko, Daguang Xu, Hirofumi Obinata, Hitoshi Mori, Peng An, Stephanie Harmon, Evrim Turkbey, Baris Turkbey, Bradford Wood, Francesca Patella, Elvira Stellato, Gianpaolo Carrafiello, Anna Ierardi, Alan Yuille, and Holger Roth. Auto-FedAvg: Learnable Federated Averaging for Multi-Institutional Medical Image Segmentation, April 2021. URL <http://arxiv.org/abs/2104.10195>. arXiv:2104.10195 [cs, eess].
- [65] Soheil Yousefi, Tooba Abbassi-Daloui, Thirsa Kraaijenbrink, Martijn Vermaat, Hailiang Mei, Peter van 't Hof, Maarten van Iterson, Daria V. Zhernakova, Annique Claringbould, Lude Franke, Leen M. 't Hart, Roderick C. Sliker, Amber van der Heijden, Peter de Knijff, R. Jansen, J. B. van Meurs, B. T. Heijmans, D. I. Boomsma, J. van Dongen, J. J. Hot-tenga, P. E. Slagboom, H. E. D. Suchiman, M. van Iterson, E. W. van Zwet, P. A. 't Hoen, R. Pool, M. M. van Greevenbroek, C. D. Stehouwer, C. J. van der Kallen, C. G. Schalkwijk, C. Wijmenga, A. Zhernakova, E. F. Tigchelaar, M. Beekman, J. Deelen, D. van Heemst, J. H. Veldink, L. H. van den Berg, C. M. van Duijn, B. A. Hofman, A. G. Uitterlinden, P. M. Jhamai, M. Verbiest, M. Verkerk, R. van der Breggen, J. van Rooij, N. Lakenberg, H. Mei, J. Bot, D. V. Zhernakova, P. Van't Hof, P. Deelen, I. Nooren, M. Moed, M. Vermaat, M. J. Bonder, F. van Dijk, M. van Galen, W. Arindrarto, S. M. Kielbasa, M. A. Swertz, A. Isaacs, L. Franke, Peter A. C. 't Hoen, and BIOS consortium. A SNP panel for identification of DNA and RNA specimens. *BMC Genomics*, 19(1):90, January 2018. ISSN 1471-2164. doi: 10.1186/s12864-018-4482-7. URL <https://doi.org/10.1186/s12864-018-4482-7>.
- [66] Bocong Yuan and Jiannan Li. The Policy Effect of the General Data Protection Regulation (GDPR) on the Digital Public Health Sector in the European Union: An Empirical Investigation. *International Journal of Environmental Research and Public Health*, 16(6):1070, March 2019. ISSN 1661-7827. doi: 10.3390/ijerph16061070. URL <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6466053/>.
- [67] Zhuo Zhang, Feng Shou Yin, Jiang Liu, Wing Kee Wong, Ngan Meng Tan, Beng Hai Lee, Jun Cheng, and Tien Yin Wong. ORIGA-light: An online retinal fundus image database for glaucoma analysis and research. *2010 Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBC'10*, pages 3065–3068, 2010. ISSN 1557-170X. doi: 10.1109/IEMBS.2010.5626137.
- [68] Zhuo Zhang, Ruchir Srivastava, Huiying Liu, Xiangyu Chen, Lixin Duan, Damon Wing Kee Wong, Chee Keong Kwoh, Tien Yin Wong, and Jiang Liu. A survey on computer aided

- diagnosis for ocular diseases. *BMC medical informatics and decision making*, 14:80, August 2014. ISSN 1472-6947. doi: 10.1186/1472-6947-14-80.
- [69] Yue Zhao, Meng Li, Liangzhen Lai, Naveen Suda, Damon Civin, and Vikas Chandra. Federated Learning with Non-IID Data. 2018. doi: 10.48550/arXiv.1806.00582. URL <http://arxiv.org/abs/1806.00582>. arXiv:1806.00582 [cs, stat].