FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

Federated learning in medical image analysis

Fabiana Rodrigues da Silva



MASTER IN DATA SCIENCE AND ENGINEERING

Supervisor: João Manuel R.S. Tavares

July 21, 2023

© Fabiana Rodrigues da Silva, 2023

Federated learning in medical image analysis

Fabiana Rodrigues da Silva

MASTER IN DATA SCIENCE AND ENGINEERING

July 21, 2023

Resumo

A análise de imagens médicas é crucial para o diagnóstico eficiente de muitas doenças. Os hospitais geralmente mantêm vastos repositórios de imagens, que podem ser aproveitados para diversos fins, incluindo pesquisa. No entanto, o acesso a essas coleções de imagens é amplamente restrito para proteger a privacidade das pessoas cujas imagens estão sendo armazenadas, pois surgem preocupações com a proteção de dados. Recentemente, o desenvolvimento da Análise Automatizada de Imagens Médicas tem ganhado significativa atenção, com o *Deep learning* sendo uma solução que tem alcançado resultados notáveis na análise de imagens médicas.

Uma abordagem promissora para análise de imagens médicas é o *Federated learning (FL)*, que permite o uso de um conjunto de repositórios de dados fisicamente distribuídos (os nós) para análise, satisfazendo a restrição de que os dados não saem do repositório. Sob essas condições, o FL pode construir modelos precisos de alta qualidade usando muitos dados disponíveis onde quer que estejam. Essa abordagem pode ajudar pesquisadores e clínicos a diagnosticar doenças e tomar decisões médicas de forma mais eficiente e robusta.

A detecção de pneumonia em imagens de radiografia de tórax (raio-X) é proposta em um ambiente de FL usando *Flower* como *framework* e *Federated Averaging (FedAvg)* como estratégia. Essa abordagem de aprendizado supervisionado faz uso de modelos pré-treinados em redes neurais convolucionais para alavancar a transferência de aprendizado: VGG-16, Resnet-18 e Resnet-50, e também técnicas de *data augmentation* foram aplicadas para ajustar os modelos. Simulando um ambiente de FL com 8 hospitais compartilhando suas próprias imagens, o Resnet-18 apresenta o melhor resultado com acurácia de 98,46%, seguido por Resnet-50 com 78,46% e VGG-16 também com 78,46%, todos avaliados no lado do servidor após 5 rodadas de treinamento.

Os experimentos sugeridos no trabalho de pesquisa apresentaram um alto custo computacional devido a um conjunto de dados não balanceado, uma restrição prevalente encontrada no estudo (27), que utiliza o mesmo conjunto de dados desta investigação aplicado em uma solução de FL.

Portanto, a pesquisa discute aplicações, contribuições, limitações e desafios, sendo adequada para aqueles que desejam entender como o FL pode contribuir para o domínio das imagens médicas.

Além disso, essa solução é aplicável como uma base para resolver outros problemas de classificação binária usando imagens médicas, como imagens de Ressonância Magnética (IRM), Tomografia Computadorizada (TC), Radiografia de Tórax (raio-X) e imagens histológicas.

Keywords: Pneumonia, Imagens de Radiografia de Tórax, *Federated Learning (FL)*, Análise de Imagens Médicas, Inteligência artificial (IA)

Abstract

Medical image analysis is crucial for the efficient diagnosis of many diseases. Hospitals typically maintain vast repositories of images, which can be leveraged for various purposes, including research. However, access to such image collections is largely restricted to safeguard the privacy of the individuals whose images are being stored, as data protection concerns come into play. Recently, the development of Automated Medical Image Analysis has gained significant attention, with Deep Learning being one solution that has achieved remarkable results in medical image analysis.

One promising approach for medical image analysis is Federated Learning (FL), which enables using a set of physically distributed data repositories (the nodes) for analysis, satisfying the restriction that data does not leave the repository. Under these conditions, FL can build high-quality accurate models using a lot of available data wherever it is. This approach can help researchers and clinicians to diagnose diseases and support medical decisions more efficiently and robustly.

Detection of pneumonia on chest X-radiography (X-ray) images is proposed in a FL environment using Flower as framework, and FedAvg as strategy. This supervised learning approach uses pre-trained Convolutional Neural Network (CNN) models to leverage transfer learning: VGG-16, Resnet-18 and Resnet-50, and also data augmentation techniques are applied to fine-tune the models. Simulated a FL environment having 8 hospitals sharing their own images, Resnet-18 shows the best result with 98.46% of accuracy, followed by Resnet-50 with 78.46% of accuracy, then VGG-16 with 78.46% of accuracy as well, all evaluated on the server-side after 5 rounds of training.

The experiments suggested in the research work exhibited significant computational expense owing to an uneven dataset, a prevalent constraint encountered in the study (27), presented in the current state-of-the-art, which utilizes the identical dataset as this inquiry applied in a FL solution as well.

Hence, it discusses applications, contributions, limitations, and challenges, and is suitable for those who want to understand how FL can contribute to the medical imaging domain.

Furthermore, this solution is applicable as a baseline to solve other binary classification problems using medical images, such as Magnetic resonance imaging (MRI), Computed tomography (CT), X-radiography (X-ray), and histology images.

Keywords: Pneumonia, Chest X-ray Images, Federated Learning (FL), Medical Image Analysis, Artificial Intelligence (AI)

Acknowledgements

I would like to thank my best friend, Jesus, for giving me the breath of life, health, wisdom, and the opportunity to go through this journey and for the accomplishment of this project.

I would like to express my sincere gratitude to my supervisor, João Manuel R.S. Tavares, and to my professor Rui Camacho, for providing guidance and feedback throughout this project.

I am also thankful to my mother, Márcia Gomes Rodrigues da Silva, and father, Nilson da Silva, for inspiring and encouraging me to move to another country to go through a master's degree.

To conclude, I cannot forget to thank my husband, Jônatas Ruiz Ferraz, for all the unconditional support during these very intense academic years.

Fabiana Rodrigues da Silva

"Each life is made up of mistakes and learning,
waiting and growing,
practicing patience and being persistent."

Billy Graham

Contents

1 Introduction **1**
- 1.1 Overview and motivation . . . 1
- 1.2 Objectives . . . 2
- 1.3 Work Plan . . . 3
- 1.4 Document structure . . . 4

2 Literature review **5**
- 2.1 Federated learning . . . 5
- 2.2 Searching Method . . . 7
  - 2.2.1 Magnetic resonance imaging . . . 23
  - 2.2.2 Computed tomography . . . 23
  - 2.2.3 X-ray . . . 23
  - 2.2.4 Histology . . . 24
- 2.3 Frameworks . . . 24
  - 2.3.1 OBiBa . . . 25
  - 2.3.2 DataSHIELD . . . 25
  - 2.3.3 Virtual Pooling and Analysis of Research data . . . 25
  - 2.3.4 Data Safe Havens . . . 26
- 2.4 Summary . . . 26

3 Pneumonia **28**
- 3.1 Overview . . . 28
- 3.2 Diagnosis exams . . . 31
- 3.3 Summary . . . 32

4 Dataset and Framework **33**
- 4.1 Introduction . . . 33
- 4.2 Dataset . . . 33
  - 4.2.1 Data Augmentation . . . 34
- 4.3 CNNs . . . 35
- 4.4 Frameworks . . . 37
  - 4.4.1 Friendly Federated Learning Framework . . . 39
- 4.5 Tools . . . 41
- 4.6 Summary . . . 42

5 Implementation and Results **44**
- 5.1 Introduction . . . 44
- 5.2 Methodology . . . 44

5.2.1 Cross-silo . . . 44
5.2.2 Batch size . . . 45
5.2.3 Server-side evaluation . . . 45
5.3 Metrics . . . 46
5.4 Federated Learning with Flower . . . 46
5.4.1 Strategy . . . 47
5.5 Results . . . 48
5.5.1 Training the model before FL . . . 48
5.5.2 Training the model with FL . . . 51
5.6 Discussion and Conclusion . . . 53

6 Conclusion and Future work **55**
6.1 Final Conclusion . . . 55
6.2 Future work . . . 56

References **58**

List of Figures

1.1 Macro Timetable for planned tasks. 3
1.2 Micro Timetable for planned and performed tasks. 4

2.1 Example of a Federated learning approach. 6
2.2 Steps of a common Federated learning approach. 6
2.3 Articles published in the last 10 years retrieved from the SCOPUS database, using the keywords "federated learning","federated machine learning", "medical image", "medical imaging". 8
2.4 PRISMA study flow diagram of the performed systematic literature review. . . . 9
2.5 Types of Medical images addressed by the selected articles per year. 9

3.2 Example of chest X-ray images of (a) healthy lungs, (b) pneumonia bacterial, and (c) pneumonia viral. 28
3.1 Healthy air sacs versus air sacs inflamed in the lungs. 29
3.3 Leading causes of death globally (adapted from WHO's Global Health Estimates, 2020), (38). 31

4.1 Samples of the selected dataset. 34
4.2 Resulting images from various data augmentation techniques. 35
4.3 Example of a CNN architecture. 37
4.4 Flower core framework architecture with both Edge Client Engine and Virtual Client Engine (adapted from Beutel et al., 2020), (53). 40

5.1 Classification matrix for (a) VGG-16, (b) Resnet-18 and (c) Resnet-50. 49
5.2 Training and validation curves for different architectures over one randomly organization for 10 epochs, (a) Accuracy, (b) F1-score and (c) Loss. 50

List of Tables

2.1 Total number of articles retrieved from the used electronic repository. 7
2.2 Summary of the studied articles related to MR images (1 of 4). 11
2.3 Summary of the studied articles related to MR images (2 of 4). 12
2.4 Summary of the studied articles related to MR images (3 of 4). 13
2.5 Summary of the studied articles related to MR images (4 of 4). 14
2.6 Summary of the studied articles related to CT images (1 of 2) 15
2.7 Summary of the studied articles related to CT images (2 of 2). 16
2.8 Summary of studied article related to CT and MR images. 17
2.9 Summary of studied article related to CT and X-ray images. 18
2.10 Summary of studied articles related to X-ray images (1 of 3). 19
2.11 Summary of studied articles related to X-ray images (2 of 3). 20
2.12 Summary of studied articles related to X-ray images (3 of 3). 21
2.13 Summary of studied articles related to histology images. 22

4.1 Description of the selected dataset. 34
4.2 Performed data augmentation techniques. 35
4.3 Hyper-parameters used for fine-tuning. 37

5.1 Testing Accuracy, Loss, and F1-score of different architectures. 48
5.2 Server-side evaluation using FedAvg with VGG-16 architecture. 52
5.3 Server-side evaluation using FedAvg with Resnet-18 architecture. 52
5.4 Server-side evaluation using FedAvg with Resnet-50 architecture. 53

Abbreviations and Symbols

AI	Artificial intelligence
ABIDE	Autism Brain Imaging Data Exchange
ASD	Autism spectrum disorders
CNN	Convolutional Neural Network
CNNs	Convolutional Neural Networks
CT	Computed tomography
DAGs	Dual Attention Gates
DICOM	Digital Imaging and Communications in Medicine
DSHs	Data Safe Havens
FedAdagrad	Federated Adagrad
FedAvg	Federated Averaging
FCN	Fully convolution network
FL	Federated learning
FLVBSF	Federated Learning-based Vertebral Body Segment Framework
GAN	Generative Adversarial Network
GCNs	Graph Convolutional Neural Networks
GDPR	General Data Protection Regulation
HC	Healthy control
IoT	Internet of Things
KD	Knowledge distillation
ML	Machine learning
MLP	Multi-layer perceptron
MMoE	Multi-gate Mixture-of-Experts
MoE	Mixture of Experts
MR	Magnetic resonance
MRI	Magnetic resonance imaging
MSD	Medical Segmentation Decathlon
Non-IID	Not independent and identically distributed
PA	Pixel-level Accuracy
PriMIA	Privacy-preserving Medical Image Analysis
Restnet	Residual Network
RGB	Red, green and blue
ROI	Region of interest
RPC	Remote Procedure Call
SBM	Shared-bottom Model
SGD	Stochastic Gradient Descent
Std	Standard deviation
STM	Single Task Model
ViPAR	Virtual Pooling and Analysis of Research data
X-ray	X-radiography

Chapter 1

Introduction

In this chapter, firstly, Section 1.1 gives an overview about main subjects of this study, followed by the motivation. Secondly, Section 1.2 describes the main objectives of this research work. Thirdly, in Section 1.3 is presented the work plan to accomplish all contributions in this research work. Finally, Section 1.4 presents the document structure so the readers know what to expect next.

1.1 Overview and motivation

Nowadays produced data, both in industry and public research centers, is huge and growing everyday. Some data may be of sensitive or personal nature and must be therefore protected from public access. This scenario is no different in the medical environment, where Health Information Systems and Electronic Health Records can collect and store patient data on a large scale.

Analyzing medical images is a daunting task for radiologists, due to the heavy workload, the need to analyze a large number of images, the intricate nature of cases, and the resemblance to other diseases, which often requires additional studies within a constrained timeframe. Despite the extensive training that healthcare experts undergo, they are prone to commit errors given the absence of information or insufficient time spent analyzing images, which may impact identifying the right treatment to prevent the growth and spread of the disease under analysis.

Different types of image modalities such as MRI, CT, X-ray, or histology images are demanded to establish a correct diagnosis. Therefore, a comprehensive study of the use of those modalities across several image repositories is crucial to develop accurate image analysis software.

Artificial Intelligence (AI) is nowadays in a development stage that can use medical images to detect and diagnose pathological conditions successfully (1). In this way, hospitals make use of technology by building models to discover new patterns of diseases and treatments from a given image dataset, mainly through data mining. However, institutions should be careful about the data privacy of their patients, since health data is sensitive and legally protected, and therefore, personal data cannot be exposed outside them (2).

Building very high-quality models for image analysis requires diversity and a lot of images, i.e., examples used for training them, especially for complex problems. One way to access such a large amount of medical images is to directly access or obtain the images in a set of hospital image repositories. Nevertheless, several problems may then occur. For example, acquiring the necessary permissions to access the images may entail lengthy bureaucratic procedures. Moreover, the images must be duly anonymized, and restrictions may prohibit their removal from the hospital repository.

Privacy-preserving is a subject that brings concern and awareness to various institutions once personal data cannot be shared without the person's consent. This subject is so important that the General Data Protection Regulation (GDPR) ensures that the European Union data protection rules are applied to guarantee the protection of your personal data.

In order to prioritize the privacy of patients' data and legal issues related to ethics, rises the approach called Federated learning (FL), which enables building better machine learning models focusing on privacy-preserving (3), (4), (5).

FL has been helping healthcare professionals to extract meaningful information from images by applying machine learning models, such as for detecting a disease based on the analysis of medical imaging, resulting in a quicker time for disease identification and maintaining a knowledge base to be applied in other new cases.

Furthermore, healthcare delivery can be improved by automating radiology procedures in hospitals where access to radiologists is minimal.

Therefore, this study aims to develop an FL solution to discuss how it contributes to medical systems in image analysis.

1.2 Objectives

The project aimed to develop a solution based on federated learning to detect pneumonia using chest X-ray images. It seeks to achieve the most accurate result for the proposed problem without sharing patient data.

The use of FL aims to support medical decisions by bringing more robustness and efficiency to disease diagnosis, and the following objectives are intended to be achieved during this project:

1. Data preparation: Select and extract collections of data into an FL environment, which represents various hospitals providing different images;

2. Data understanding: Identify metadata, visualize attributes, explore heterogeneity, and identify the need for data transformations, such as the applicability of data augmentation techniques;

3. Feature selection and extraction: Extract features that enable the identification of pneumonia and perform feature selection to identify the most relevant patterns of pneumonia behavior;

4. Evaluate ML architecture: Choose ML architectures that utilize pre-trained models and CNNs for the classification task;

5. Build, train, test, and tune hyperparameters of the ML models: Split the data into train and test sets, select an ML classifier model suitable for FL, and optimize hyperparameters;

6. Evaluate FL architecture: Choose a FL architecture, framework, and strategy for training the models and sharing the learning with the central server;

7. Evaluate privacy: Assess the solution based on GDPR to ensure no personal data is shared among the distributed system;

8. Evaluation of pneumonia detection: Measure the performance of the classification algorithms using metrics such as accuracy, f1-score, and loss. Additionally, re-examine previous steps if the results are unsatisfactory.

1.3 Work Plan

To accomplish the objectives stated by the research objectives, a timeline was created, which is composed of five macro tasks: 1. Literature review, 2. Data Sourcing, 3. Implementation, 4. Discussion & Validation, and 5. Dissertation writing.

The tasks are organized and prioritized according to their importance to facilitate project management. Figure 1.1 shows a macro timetable for the planned tasks, and Figure 1.2 details subtasks for each macro task, categorized as planned or done until the current calendar week. Considering the unpredicted delays or advanced progress faced in specific tasks, it was deemed that if a task was finished before the proposed time, the next one was picked up immediately. Although, if a task exceeded the proposed timeline, the next one had a shorter completion period. Because the tasks are all interconnected, the timetable was mostly pursued in order to conclude this work successfully in time.

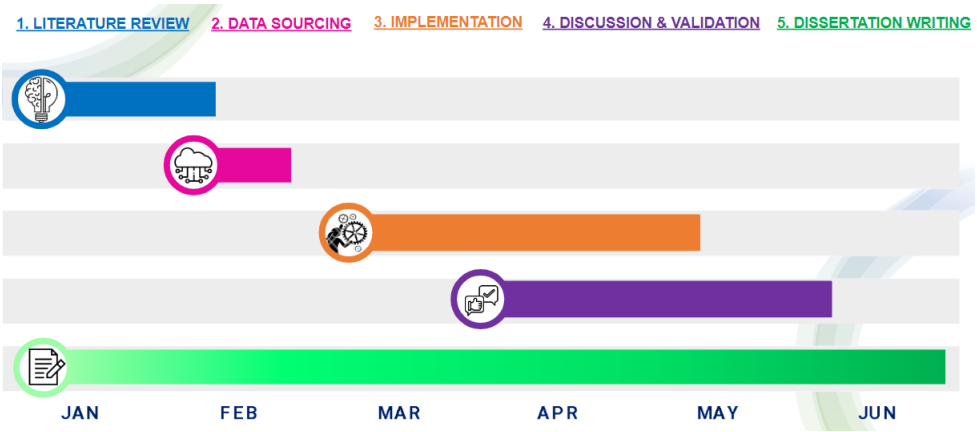


Figure 1.1: Macro Timetable for planned tasks.

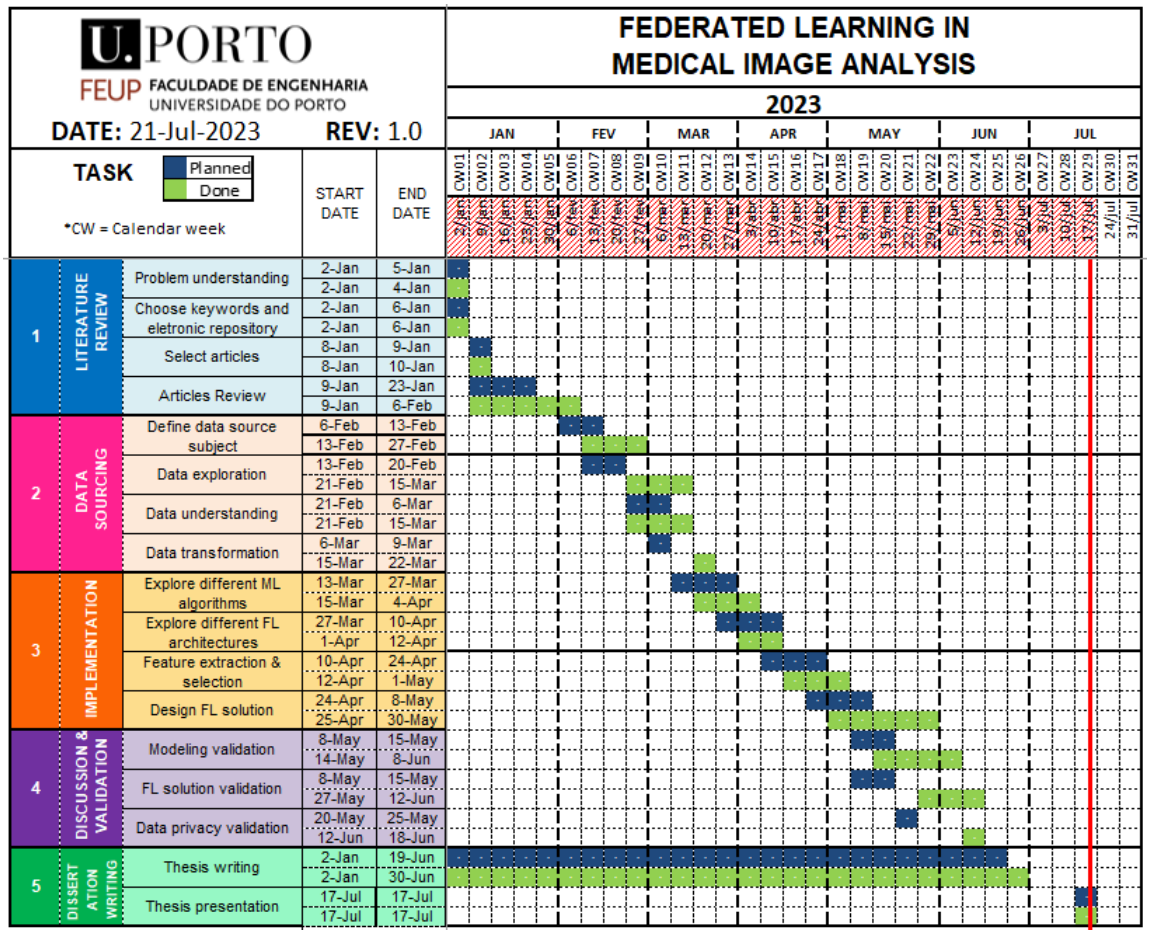


Figure 1.2: Micro Timetable for planned and performed tasks.

1.4 Document structure

In addition to the introduction, this report contains five more chapters. Chapter 2 describes the literature review. It includes several articles using different approaches to solve various problems using FL. Chapter 3 presents the problem with an overview of pneumonia, symptoms, survival rate, and diagnosis exam performed on patients for pneumonia identification. Chapter 4 presents the dataset and framework used in this research work. Chapter 5 describes the process of execution, methodology, and results of pneumonia detection through chest X-ray in an FL environment. In Chapter 6, conclusions are summarized and suggested the future work.

Chapter 2

Literature review

In this chapter, the literature review is presented. Section 2.1 presents a brief overview of FL. Searching Method is described in Section 2.2 containing selected and reviewed articles as a baseline for the research work. Frameworks that use FL in collaborative research are discussed in Section 2.3; and to finalize, Section 2.4 presents a summary of this chapter.

2.1 Federated learning

Federated learning is a term used for collaborative machine learning without centralized training data, which was introduced in a paper first published by Google in 2016 (6).

Data is distributed in different sites, locations and devices. In federated building, the process of training data happens locally, where the data is located, which is called decentralized learning. A device, i.e., an edge device, trains its model and stores its data locally. The server aggregates the result of updated models from each device and updates the centralized, i.e., global, model.

The global model is enriched by multiple trained models, once each local device provides feedback to the server based on its individually trained model, and the central server maintains a global shared model and disseminates it to all institutions. This is a typical FL workflow, called FL aggregation server.

The data exchanged is encrypted to ensure no other devices accesses private information. With that, personal information is never sent to a central server, but the weights, biases, and other parameters are learned by the local model trained in each device.

When new hospitals participate in this distributed learning environment, they bring more data and computational resources to the consortium. Nonetheless, some hospitals may generate a greater number of images than others, with minimal image heterogeneity. Consequently, the consortium may possess a substantial dataset in terms of capacity. Still, the models will not acquire fresh features because of deficient image variety, which causes models to produce inaccurate data. On the other hand, hospitals that possess a substantial quantity and diversity of images will contribute positively to developing better learning models and consequently producing more dependable results.

Figures 2.1 and 2.2 depict how federated learning works.

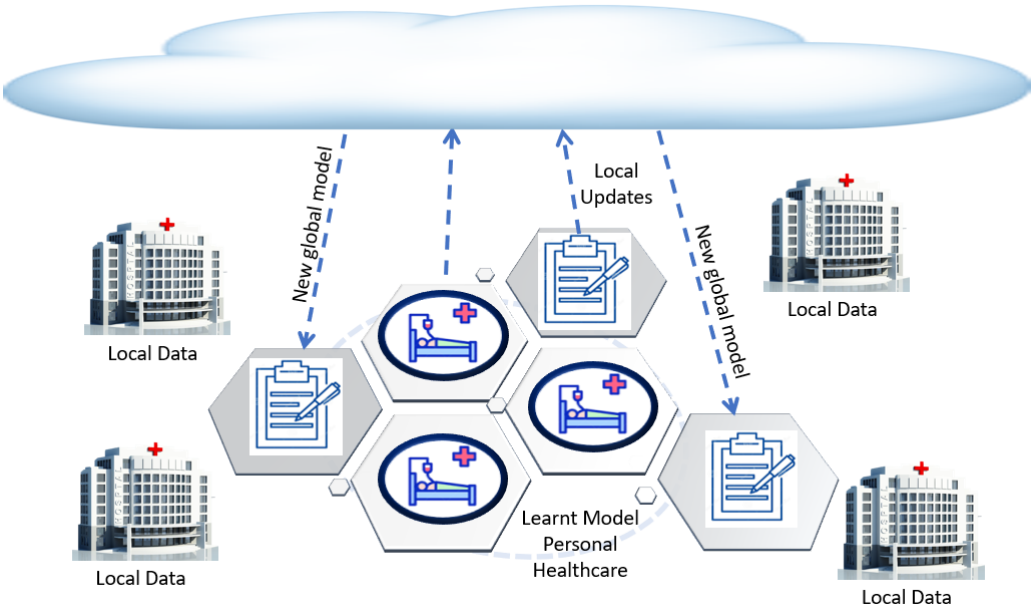


Figure 2.1: Example of a Federated learning approach.

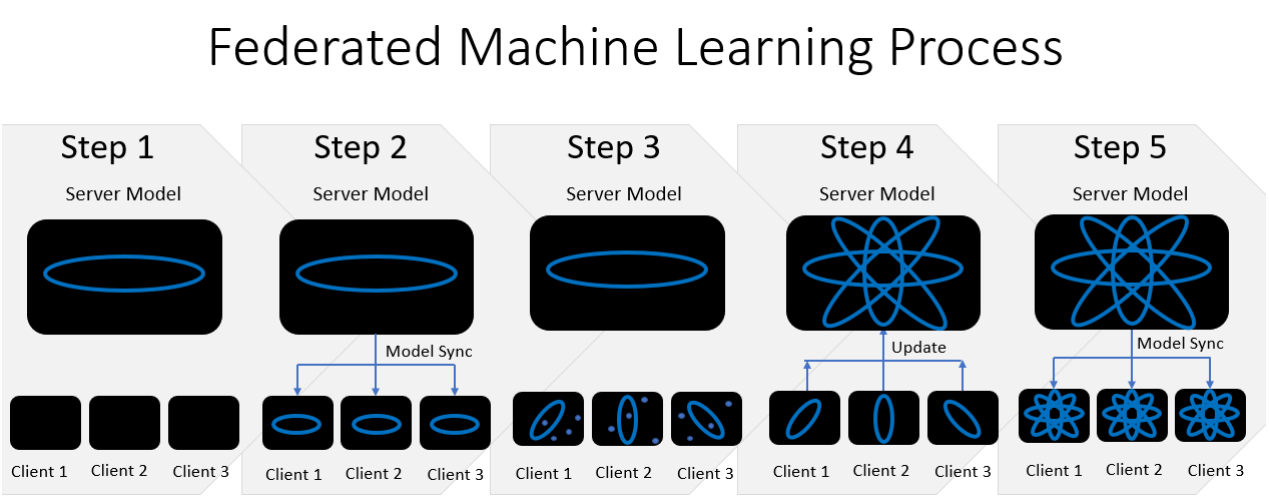


Figure 2.2: Steps of a common Federated learning approach.

Figure 2.2 shows that, in the first step, the server model selects the data sources and the machine learning model that will be used. The diverse data is stored locally on each client and are collected from those different sources. The second step involves sending an initial version of the machine learning model to each data source, which are also known as edge device or client, so the server and edge device models are synchronized. In the third step of the process, the edge devices use their respective datasets to train the model locally. The locally trained models have similar parameters (represented by ellipses), but differ in their weights (represented by dots) due to their locally collected dataset. In step four, each edge device sends the updated weights of its model

back to the central server. The central server aggregates the received information by averaging the updated model weights from each data source to create a new version of its model. In the fifth step, the updated server model is synchronized with the edge devices without accessing any data, then an individual model will have its model updated and ready to be evaluated on new and unseen data.

Federated learning is valuable when the data is sensitive or difficult to share due to regulatory or privacy concerns. Therefore, FL is being applied in diverse industries, especially in the health field, where there is a high demand for collaborative analysis of sensitive data.

2.2 Searching Method

This systematic literature review was carried out in the SCOPUS database. In order to find articles, the following queries were combined using the AND logical operator: (a) "federated learning" OR "federated machine learning"; (b) "medical image" OR "medical imaging". Table 2.1 presents the query and the total number of gathered articles.

Table 2.1: Total number of articles retrieved from the used electronic repository.

Repository	Query performed	No. of gathered articles
Scopus	TITLE-ABS-KEY ((federated learning) OR (federated machine learning) AND (medical image) OR (medical imaging)) AND (LIMIT-TO (DOCTYPE, "ar"))	84

The current research topic is relatively new and, therefore, it could be noticeable the growing number of researches being done on the topic every year, Figure 2.3. Actually, even though the conducted search was relatively the last 10 years, no articles published before 2019 were found.

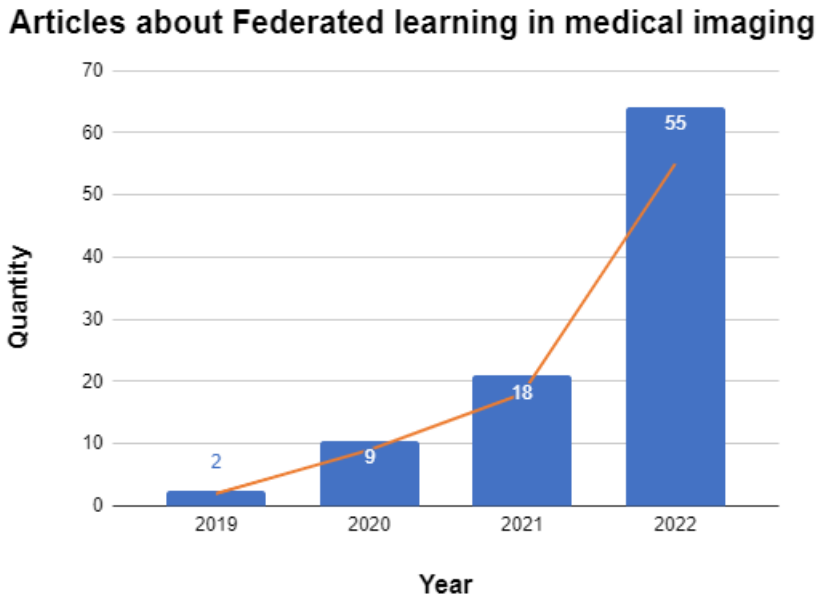


Figure 2.3: Articles published in the last 10 years retrieved from the SCOPUS database, using the keywords "federated learning","federated machine learning", "medical image", "medical imaging".

Based on an analysis of the gathered articles' title and abstract, a total of 84 articles were initially selected. Then, 42 articles were removed for not meeting the criteria: medical images acquired by Magnetic Resonance Imaging (MRI), Computed Tomography (CT) imaging, X-ray imaging or histology imaging, bringing the total to 42 articles. Then, 20 articles were excluded once they were focused on topics not related to this study. In the end, 22 original articles were selected for the review.

The Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) diagram (7) shown in Figure 2.4, represents the performed systematic search.

Figure 2.5 indicates the types of the medical images addressed by the selected articles per year, and provides a better understanding of what types of images have been recently studied particularly related to FL. In 2022, for example, MRI leads the number of research in this field, with 9 articles, followed by X-ray, with 4 articles, 1 (one) article involving histology images, and lastly CT, with 1 (one) article.

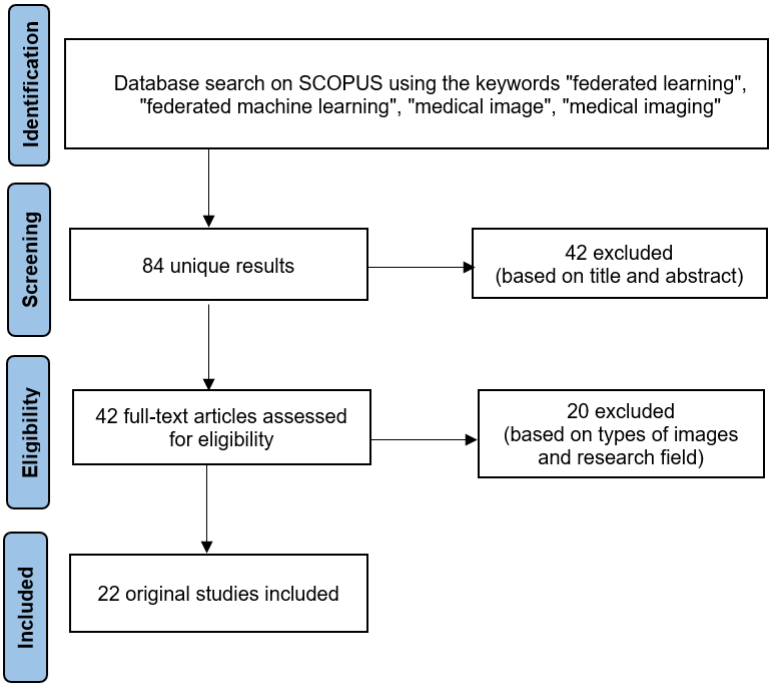


Figure 2.4: PRISMA study flow diagram of the performed systematic literature review.

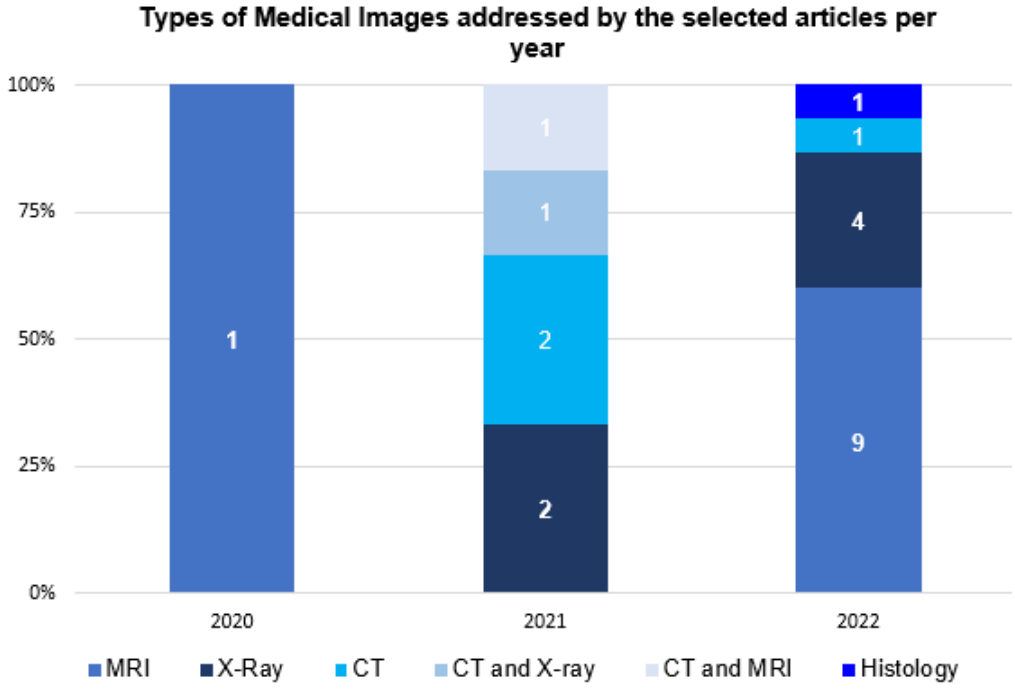


Figure 2.5: Types of Medical images addressed by the selected articles per year.

A summary of the selected and studied articles, as to the used models, datasets, architectures, input dimension and imaging modality, research goal, main purpose propose, contributions and benefits, limitations, and reported accuracy, is presented along Tables 2.2 to 2.13 according to

the used imaging type(s): Tables 2.2, 2.3, 2.4 and 2.5 are related to Magnetic Resonance (MR) images and include 10 articles, Tables 2.6 and 2.7 detail 3 articles related to CT images, Table 2.8 is related to CT and MR images with 1 (one) article, Tables 2.9 includes 1 (one) article addressing CT and X-ray images. Tables 2.10, 2.11 and 2.12 present 6 articles as X-ray images, and lastly, Table 2.13 concerns 1 (one) article related to histology images.

Ref	Datasets	Arch.	IDM	RG	Propose	Contributions	Limitations	Accuracy
8	BT-small-2c and BT-large-3c	CNN End-to-end	2D, MRI	Binary classification to diagnose a brain tumor disease in a federated environment.	Designing the CNN model for each client and exchanging the training parameters between clients and server.	FL framework to detect brain cancers using 3, 2 publicly datasets.	Proposed model limited to 4 different strategies.	Higher accuracy on the larger dataset (BT-large-3c), 96% than to the smallest one (BT-small-2c), 82%.
9	M&M and ACDC 2018	3D-CNN	3D, MRI	Diagnosis of hypertrophic cardiomyopathy, which is a cardiovascular disease.	Test CNN models on partitions of the centers seen during the training and also on unseen centers.	First simulated federated learning study on the modality of cardiovascular MRI.	Small dataset (M&M), several data augmentation techniques were needed to artificially increase the size of the training set.	-
10	SpineSagT2-Wdataset3, which comes from the Chinese Society of Biomedical Engineering	DAGs-U-Net	MRI	Segmentation of Vertebral Body images.	Create a Federated Learning-based Vertebral Body Segment Framework (FLVBSF) with a novel local Dual Attention Gates (DAGs)-based attention mechanism.	FLVBSF has a strong capability to classify each vertebral body pixel from the background.	Several approaches applied on the U-Net. When the number of iterations or institutions is set to smaller or larger, the performance of the federated model varies.	Pixel-level Accuracy (PA) of 98.29% is achieved by the U-Net with DAGs.

Table 2.2: Summary of the studied articles related to MR images (1 of 4).
Ref means Reference, IDM means Input Dimension and Modalities, Arch means Architecture, and RG means Research Goals

Table 2.3: Summary of the studied articles related to MR images (2 of 4).
Ref means Reference, IDM means Input Dimension and Modalities, Arch means Architecture, and RG means Research Goals

Ref	Datasets	Arch.	IDM	RG	Propose	Contributions	Limitations	Accuracy
11	ACDC MICCAI 2017	Momentum Contrast (MoCo)	3D, MRI	Segmentation and generalization performance on a cardiac MRI dataset.	Two federated self-supervised learning frameworks for volumetric medical image segmentation with limited annotations.	The first framework has high accuracy and fits high-performance servers with high-speed connections. The second framework addresses lower communication costs, applicable to mobile devices.	The proposed optimized method FCLOpt does not rely on negative samples, which reduces the communication cost of contrastive learning.	First framework, FCL, has high accuracy (82.4% $\pm$ 2.5%) with feature sharing. The second framework, FCLOpt, has an accuracy of 82.1% $\pm$ 2%.
12	BraTS 2017, Retina and BoneAge	ResNet34 as pre-trained on ImageNet as the base network for all methods on Retina and BoneAge datasets. Used pre-trained U-Net for the BraTS dataset.	MRI	Implement SplitAVG method to drop the performance from 3 heterogeneous datasets.	Method SplitAVG applied on a CNN to overcome the performance drop from data heterogeneity in federated learning.	Used heterogeneous data in real-world federated learning settings. The SpliAVG method does not need any complex hyper-parameter tuning, training heuristics, or additional training/fine-tuning.	Data privacy concerns of reconstructing raw images from shared feature maps of the cut layer.	Achieved 96.2% of accuracy.
13	MSLUB, MISBI, MSKRI, GBKRI, BraTS	Convolutional auto-encoders	MRI	Segmentation of an unsupervised brain pathology.	Train an unsupervised CNN using Federated Disentangled representation learning, called FedDis.	Apply healthy reconstructions by increasing the anomaly scores, leveraging the global anatomical structure, and detaching the parameters while mitigating domain shifts.	Trains the model locally, it is aggregated only the updated models of the local institutes.	Anomaly segmentation has the results improved by 99.74% for multiple sclerosis and 40.45% for tumors over locally trained models.

Ref	Datasets	Arch.	IDM	RG	Propose	Contributions	Limitations	Accuracy
14	ABIDE, ADHD-200 and COBRE	STM, SBM, and MMoE	MRI	Identify multiple related mental disorders.	Compare different architectures to help medical professionals set up early detection and personalized treatment in FL.	The FL framework has effective learning for those participating institutions with relatively small datasets.	Regardless of increasing available datasets and the majority of client models fail to converge during the training process, it turned out a significant performance degradation.	Accuracy of 69.48+-1.6% in autism spectrum disorder, 71.44+-3.2% in attention deficit/hyperactivity disorder, and 83.29+-3.2% in schizophrenia.
15	UK Data Service dataset	CNN Ensemble	Axial T2 and Coronal slices of MR images.	Brain tumor identification.	Used 6 pre-trained models, combined 3 best models based on the accuracy on an average CNN (InceptionV3, VGG16, DenseNet) for Voting Ensemble, in order to compare the performance to select the global model in FL.	FL approach achieved privacy-protected tumor classification with high accuracy.	Small dataset which degrades the global performance if many clients have really small datasets and may heavily overfit. Within the clients, no measure was taken to address class distribution imbalance.	FL with an accuracy of 91.05% compared to the 96.68% of accuracy for the base ensemble model.
16	ABIDE: Autism brain imaging data exchange ADNI: Alzheimer's disease neuroimaging initiative	GCNs	MRI	Predict neural diseases such as Autism and Alzheimer's.	Train a global Graph Convolutional Neural Networks (GCNs) node classifier in many institutions using a federated graph learning platform, called FedNI.	Designed a federated network where a missing node generator grants each institution the ability to generate missing nodes and edges, and federated network inpainting. FedAvg was used as an aggregation strategy.	Training an effective GCN model for node classification requires a bigger dataset.	Accuracy of 66.7% ($\pm$0.6%) on the ABIDE dataset, and 75.8% ($\pm$0.7%) on the ADNI dataset.

Table 2.4: Summary of the studied articles related to MR images (3 of 4).
Ref means Reference, IDM means Input Dimension and Modalities, Arch means Architecture and RG means Research Goals

Ref	Datasets	Arch.	IDM	RG	Propose	Contributions	Limitations	Accuracy
17	ABIDE	Multi-layer perceptron (MLP) 6105-16-2: 6105 nodes for the input (first) layer, 16 nodes for the hidden layer, and 2 nodes for the output layer)	MRI	Identify autism spectrum disorders (ASD) or healthy control (HC).	Applied 2 methods: Mixture of Experts (MoE), adaptation near the output layer, and Adversarial domain alignment, adaptation on the data knowledge representation level.	Overcome the domain shift issue; the federated learning model has revealed possible brain biomarkers for identifying ASD. Fed-MoE outperformed Fed on NYU, UM and UCLA site and Fed-Align outperformed Fed on NYU, UM and UCLA site by comparing mean accuracies. All Fed and Fed+Domain Adaptation strategies showed significant improvement compared to Cross, Single and Ensemble strategies.	Implications into other disease areas, particularly rare diseases with few patients.	Results of using different training strategies, but the best accuracy was 78.9% ($\pm$15.3%) when with Fed framework on ASD.

Table 2.5: Summary of the studied articles related to MR images (4 of 4).
Ref means Reference, IDM means Input Dimension and Modalities, Arch means Architecture and RG means Research Goals

Ref	Datasets	Arch.	IDM	RG	Propose	Contributions	Limitations	Accuracy
18	Private large dataset with unsegmented 2230 axial chest CT images from various institutions, divided into (a) COVID-19 (1016 images), (b) lung cancer, and non-COVID-19 lung infections (610 images), and (c) normal lung aspect (604 images)	CNN	CT lung images.	Multiclass classification of COVID-19, cancer, and non-COVID-19 lung infections or normal lung.	Used VGG-16 pre-trained on ImageNet dataset with transfer learning, and FedAvg method. 3 individual clients were deployed on their own instead of training the model on the entire dataset in a centralized way.	The proposed AI-assisted software is capable to identify a health lung in the absence of COVID-19 to disregard lung cancer or a non-COVID-9 lung infection.	High computation cost, being possible to train and validate only 2 clients from 3 rounds of the total of ten chosen for training.	The proposed method—FL VGG-16 had an accuracy of 83.82% on images during training, and 79.32% during validation.
19	3 private datasets from local hospitals (from Germany, China, and 1 public available dataset)	2D CNN based on an improved version of RetinaNet18	CT	Detecting lung abnormalities in COVID-19 patients.	Used DeepLesion CT lesion dataset as a pre-trained model, applying 3 independent networks: 1- Learning with limited annotations, 2 - Learning to segment COVID-19 CT scans from non-COVID-19 CT scans, 3 - Learning with both COVID-19 and non-COVID-19 CT scans. Then, a model ensemble of these 3 individual models, and lastly, a baseline of training a single joint model with all data centralized.	In order to support clinical disease management, the usage of diverse and large datasets led to a good impact of AI providing low-cost and scalable tools for lesion burden estimation.	Limitation of concept shift, where the manual annotations in the cohort were not directly compatible with training data.	-

Table 2.6: Summary of the studied articles related to CT images (1 of 2).
Ref means Reference, IDM means Input Dimension and Modalities, Arch means Architecture, and RG means Research Goals

Ref	Datasets	Arch.	IDM	RG	Propose	Contributions	Limitations	Accuracy
20	Private datasets from hospitals in China, Italy and Japan	CNN	CT	Detect COVID-19 infections.	Semi-supervised FL learning for segmentation of abnormal regions related to COVID-19 using 3D U-shape FCN as a baseline model.	Semi-supervision potentially reduces the annotation burden under a distributed setting. The proposed approach had a better performance in real-world datasets, compared to the default setting of federated learning.	Conventional data sharing instead of model weight sharing. The model is not trained to discriminate against other types of abnormalities, e.g., other pneumonia or cancer. Semi-supervised framework complexity is higher compared to regular FL.	-

Table 2.7: Summary of the studied articles related to CT images (2 of 2).
Ref means Reference, IDM means Input Dimension and Modalities, Arch means Architecture, and RG means Research Goals

Ref	Datasets	Arch.	IDM	RG	Propose	Contributions	Limitations	Accuracy
21	Two private datasets from Medical Segmentation Decathlon (MSD): pancreas and brain tumor segmentation.	CNN (MoNet, a shallow of U-Netlike architecture)	CT and MRI	Pancreatic segmentation in CT and brain tumor segmentation in MRI.	Proposed the MoNet architecture for federated learning applications in order to segment pancreatic in CT and brain tumor in MRI.	MoNet has a small number of parameters and less computational cost compared to U-Net-16, and extracts more robust features that generalize better to out-of-sample data.	Not considered the use of larger, multi-institutional training and validation sets, neither considered to use other larger models as well.	-

Table 2.8: Summary of studied article related to CT and MR images.
Ref means Reference, IDM means Input Dimension and Modalities, Arch means Architecture, and RG means Research Goals

Ref	Datasets	Arch.	IDM	RG	Propose	Contributions	Limitations	Accuracy
22	LIDC dataset and 2 other datasets from the National Institutes of Health	CNNs (GhostNet, ResNet50, and ResNet101)	CT and X-ray	COVID region segmentation in chest CT images.	Proposed an architecture for a dynamic fusion-based learning approach for semi-supervised medical image analysis to detect COVID-19 infections, in order to improve communication efficiency and model performance.	Conducted 18 groups of experiments using 3 different models: GhostNet, ResNet50, and ResNet101. Used aggregated global model. The fusion-based federated perform better in real-world data sets than the default setting of federated learning, which learning can ensure fault tolerance and robustness.	The study does not consider the communication efficiency and model accuracy issues of federated learning.	-

Table 2.9: Summary of studied article related to CT and X-ray images.
Ref means Reference, IDM means Input Dimension and Modalities, Arch means Architecture, and RG means Research Goals

Ref	Datasets	Arch.	IDM	RG	Propose	Contributions	Limitations	Accuracy
23	COVIDX-8A , COVIDX-8B	CNNs (ResNet18 and ResNet34)	Chest X-ray	Classification of COVID on chest X-ray images to ensure additional security and privacy features.	Propose MediSecFed—a secure framework for federated learning in a difficult environment applied on chest X-ray datasets. Developed Knowledge Distillation (KD), a new technique for compacting and accelerating neural networks, to maintain comparable performance while keeping the constraints in mind.	To generate the synthetic data used for KD, a publicly available frozen DenseNet-121 model trained on CheXpert for chest X-ray image generation was used. The method does not allow the server to perform model inversion attacks, since the model parameters are never shared.	The method outperforms FedAvg by 15% on both datasets in a difficult environment.	-
24	CheXpert and Mendeley	CNNs (DenseNet121 and ResNet50)	Chest X-ray	Federated learning for chest X-ray classification of pneumonia aiming the prevention of data privacy attacks.	Propose a FL as a solution for privacy-preserving distributed learning, by integrating Rényi differential privacy with a Gaussian noise mechanism into the federated learning process.	Used 2 architectures pre-trained on ImageNet data, showing that the classification produced better results on Mendeley than on CheXpert data. Images reconstruction from shared model updates within the FL setting from networks using the DLG attack. Used FedAvg aggregation strategy to send local models back to the server.	The effectiveness of model aggregation is limited by data heterogeneity and imbalance.	-

Table 2.10: Summary of studied articles related to X-ray images (1 of 3).
Ref means Reference, IDM means Input Dimension and Modalities, Arch means Architecture, and RG means Research Goals

Ref	Datasets	Arch.	IDM	RG	Propose	Contributions	Limitations	Accuracy
25	DarkCOVID , ChestCOVID	GAN , CNNs	COVID-19 X-ray	FL scheme (FedGAN) to generate realistic COVID-19 images, in order to facilitate privacy-enhanced COVID-19 detection with GANs in edge cloud computing.	COVID-19 multi-class classification with a blockchain-based FedGAN framework for security.	Enhances COVID-19 data privacy and detection performance with a blockchain-based FedGAN framework for secure COVID-19 data analytics. Compared to the state-of-the-art schemes, the proposal has high detection accuracy rate and low running time. Achieves a significant accuracy by combining FL and GAN, composing the FedGAN.	GAN models have a high computational cost. Limitations on Edge nodes on its resources to train image datasets, which are needed for a FL-GAN process to have a better result.	Accuracy of 99.2% on DarkCOVID, and 98.5% on Chest-COVID datasets. -
26	COVIDX8, composed of sub-datasets from open source chest radiography datasets	CNN (ResNet18)	Chest X-ray	COVID-19 detection.	Dynamic fusion FL framework to detect COVID-19, called FedFocus.	Applied a blockchain-based solution to decentralize the aggregation process. Highly restores the imbalance of the real dataset by simulating the division of the training set based on the population and the infected cases of 3 real cities. Compared to FedAvg, FedFocus had a significantly better stability.	Imbalance of samples in each training set of not independent and identically distributed (Non-IDD) data, resulting in different training losses. The aggregation method and the optimal setting have the same effect on the optimal dynamic factor and resilience factor.	FedFocus has achieved similar accuracy to FedAvg, both higher than 92%.

Table 2.11: Summary of studied articles related to X-ray images (2 of 3).
Ref means Reference, IDM means Input Dimension and Modalities, Arch means Architecture, and RG means Research Goals

Ref	Datasets	Arch.	IDM	RG	Propose	Contributions	Limitations	Accuracy
27	A public paediatric pneumonia dataset called MedNIST	CNN (ResNet18)	Chest X-ray	Classifies paediatric chest X-ray images into 1 of 3 categories: normal (no signs of infection), viral pneumonia or bacterial pneumonia, using large and multinational datasets.	Proposed an FL, which is created with the goal to be an open-source software framework, called Privacy-preserving Medical Image Analysis (PriMIA).	Compatible with a wide range of medical imaging data formats, user-configurable which introduces functional improvements to FL training.	High computational cost on an imbalanced dataset. Needed high data quality on nodes to have the success of FL models.	The best accuracy was 91% on training, and 92% on validation which occurred on the centrally trained experiment.
28	A public chest X-ray dataset acquired from the Department of Health and Human Services, Montgomery County, Maryland, USA and Shenzhen No. 3 People's Hospital in China.	CNNs (VGG-16 and ResNet50)	Chest X-ray	Chest X-ray images classification to identify COVID-19 from non-COVID-19 cases.	Proposed 2 FL architectures based on not independent and identically distributed (Non-IID) data, using unbalanced data.	The first study that addresses the problem of federated learning on X-ray images for COVID-19 detection. CNNs applied pre-trained weights on ImageNet. The FL framework led to comparable performance with a centralized learning process and remained robust.	Needed to apply data augmentation on the experiments once a small dataset was used.	Accuracy of 94.40% on FL-VGG16 with data augmentation, and 97.0% on FL-ResNet50 with data augmentation.

Table 2.12: Summary of studied articles related to X-ray images (3 of 3).
Ref means Reference, IDM means Input Dimension and Modalities, Arch means Architecture, and RG means Research Goals

Ref	Datasets	Arch.	IDM	RG	Propose	Contributions	Limitations	Accuracy
29	Private datasets from multiple institutions	ResNet50	Histo-logy whole slide imaging (WSI) with only slide-level labels.	FL for digitized gigapixel whole-slide images for binary and muti-class classification on breast cancer and renal cell cancer.	Proposed a FL solution focusing on weakly-supervised deep learning models to demonstrate the feasibility and effectiveness of privacy-preserving using only slide-level labels for supervision on survival prediction.	Software package available for usage on github website, enabling multiple institutions to integrate their WSI datasets and train their models. The developed FL framework has the clear potential to be utilized in numerous crucial computational pathology assignments beyond those demonstrated in this study.	Limited to weakly-supervised federated multiple instance learning.	Best balanced accuracy of 0.900±0.020 on Renal Cell Carcinoma (RCC) on subtyping test reported as a five-fold mean. Best balanced accuracy of 0.756±0.026 on Breast Invasive Carcinoma (BRCA) on subtyping test reported as a five-fold mean; both in a federated environment where $\alpha = 0.1$.

Table 2.13: Summary of studied articles related to histology images.
Ref means Reference, IDM means Input Dimension and Modalities, Arch means Architecture, and RG means Research Goals

2.2.1 Magnetic resonance imaging

Among the 10 selected articles related to MRI, half (5) proposed Convolutional Neural Networks (CNNs) containing several deep layers to extract features from the images. Of those 5 articles, 2 used pre-trained models: one used ResNet34 and U-Net (12), and the other article used 6 pre-trained models (15), and applied a CNN ensemble combining the 3 best models based on the average accuracy (InceptionV3, VGG16, DenseNet) for the Voting Ensemble model. From the group of CNNs, only 1 (one) article used convolutional auto-encoders (13), and 2 articles developed end-to-end CNNs (8), (13). The remaining articles did not make use of CNNs, and applied different architectures, including: one article used Single Task Model(STM), Shared-bottom Model (SBM), and Multi- gate Mixture-of-Experts (MMoE) (14), another Graph Convolutional Neural Networks (GCNs) (16), other Multi-layer perceptron(MLP) (17), another DAG's (10) and a last one MoCo (11).

In terms of used datasets, in (14), (16) and (17), the experiments were conducted on the same dataset, called Autism Brain Imaging Data Exchange (ABIDE). With small datasets, common techniques were applied in the pre-processing step, mainly of data augmentation (8), (9), (14).

2.2.2 Computed tomography

Among the 5 reviewed articles related CT imaging, in all of them, private large datasets were used and the research goal was related to COVID-19. They also made use of pre-trained models to train their CNN based models, such as VGG-16 (18), an improved version of RetinaNet18 (19). GhostNet, ResNet50, and ResNet101 were used as pre-trained models using 3D U-shape fully convolution networks (FCN) as baseline model (20). Those researches deal with a common topic, and the applied FL framework suffers from high computational cost and communication efficiency issues.

The authors of (22) proposed an architecture for a Dynamic-Fusion-Based Federated by conducting experiments in 2 types of images: CT and X-ray images. They used 3 public datasets and applied CNN based models: pre-trained GhostNet, ResNet50, and ResNet101 models.

Knolle et al. (21) used CT images for pancreatic segmentation and MR images for brain tumor segmentation, both on private datasets, and also applied a CNN in a pre-trained model called MoNet, a shallow U-Netlike architecture, giving the possibility to extract more robust features, which generalize better to out-of-sample data.

2.2.3 X-ray

Among the 7 selected articles about X-ray images, all of them use CNNs, 6 used pre-trained models, and almost half of them applied more than 1 (one) pre-trained model to compare results (23), (24), (28). They used ResNet18 and ResNet34, DenseNet121 and ResNet50, VGG-16 and ResNet50, respectively, and share some common models.

Nguyen et al. (25) used a generative adversarial network (GAN) that generates realistic COVID-19 images, in order to facilitate privacy-enhanced COVID-19 detection with GANs in edge cloud computing, and applied them on a FL scheme called FedGAN.

The authors in (26) and (27) used chest X-ray images for COVID-19 and Pneumonia detection, respectively, and both of them applied ResNet-18 as a pre-trained model. In article (26), it is proposed an architecture for a Dynamic Fusion FL Framework, and compared two strategies: FedAvg and FedFocus, where both had achieved similar results. In article (27), the authors proposed an open-source software framework called Privacy-preserving Medical Image Analysis (PriMIA), but the proposed solution had a high computational cost due to the used imbalanced dataset.

As shown in Tables 2.10, 2.11 and 2.12, all the selected articles had indicated high accuracy (if reported) on the proposed FL approach. Feki et al. (28), for example, had achieved a high accuracy after applying data augmentation to the training data.

2.2.4 Histology

The single article found in the performed literature search that is related to histology images used different private datasets with the goal to predict and solve two different diagnostic problems: classification of breast and renal cell cancers. It used thousands of histology whole-slide images with only slide-level labels, and a pre-trained ResNet50 CNN encoder.

There is a lack of annotations in the majority of real-world whole slide histopathology datasets, and as explained in (29), the proposed FL solution can be linked to weakly-supervised multiple instance learning in order to solve binary and multi-class classification problems. In addition, it presented accurate results without direct data exchange and its related intricacies, while likewise maintaining differential privacy through randomized noise generation.

2.3 Frameworks

FL strategies help AI solutions on enriching data on the training dataset, generating more accurate results by allowing multiple collaborators to build a robust machine learning model using a large dataset. That is possible because there is no direct data sharing, as federated learning prioritizes the privacy of patients' data.

In the health field, for example, when new hospitals participate in this collaborative environment, they bring more data and more computational resources. Although, individual hospitals which do not have a large dataset can benefit from the rich datasets without providing much data, but at the same time, it can be a concern to big hospitals. There can be another challenge of equitable allocation, where a hospital may produce considerably more images than another, but the diversity of its images may be low. Differences in images in terms of acquisition protocols and labeling methodologies, coming from different devices, may contribute poorly or even negatively to the central model as well.

Some open-source software solutions are available for secure bioscience collaboration based on FL, as the ones presented in the following sections.

2.3.1 OBiBa

According to the OBiBa website, OBiBa is an international project that offers an open source software for epidemiological studies (30), and suites applications for studies built on a federated database infrastructure. The software includes five main tools, which are:

- Collect with Onyx: web application to collect data in clinics or assessment centers and manage interviews;
- Store and Document with Opal: central data repository for epidemiological studies;
- Analyze with R and DataSHIELD: uses Rock server application (Rest API) to provide statistical analysis services without accessing individual-level data;
- Publish with Mica: Mica is a web application which can be used to create web data portals for epidemiological studies (30). This tool provides accessibility to query datasets stored in Opal databases, search variable dictionaries and create study catalogs;
- Manage Users with Agate: central authentication server, which can be used for email notification services. This might help notify researchers of new data available for studies.

OBiBa infrastructure enables researchers to have access to patients' medical information related to epidemiology, but does not allow access to personal data, such as patients' address or phone number. This infrastructure helps researchers to create studies or collect more data by respecting data privacy, and the process of enriching data improves the quality of the study, once there is data diversity for models to learn from.

2.3.2 DataSHIELD

DataSHIELD is a free and open-source software developed for biomedicine, social science and public health, which uses the OBiBa infrastructure. This analytical tool has been used to develop projects in collaboration with health by impacting social effects, lifestyle and healthcare (31).

A large dataset contains images suitable for analyzing "remotely", but transferring them has a high computational cost and also a risk for leakage once shared throughout the network. In addition, as already mentioned, privacy data is crucial in the health sector because institutions deal with ethical issues when dealing with patients' data, where security and data confidentiality are important factors to consider.

DataSHIELD is helpful for co-analysis of individual-level data from multiple studies or sources, but not physically sharing them.

2.3.3 Virtual Pooling and Analysis of Research data

Virtual Pooling and Analysis of Research data (ViPAR), is a software platform that implements database federation techniques. Researchers can analyze datasets located in different locations across the globe remotely with a web-based environment. ViPAR is an open source technology,

simple, powerful framework for centralized data management, and facilitates the sharing of research data respecting the ethical and privacy issue (32). The tool facilitates the standard analyses without accessing individual-level data, which is pre-configured for R, SAS and Stata programming languages, and can be shared with other researchers.

Regarding comparison, Vipar pools site-specific raw data, while DataShield pools site-specific statistics. In this way, DataShield can be considered a better tool in terms of collaborations since no data leaves a particular site on any occasion, while ViPAR analysis data leave a data-contributing site momentarily (32).

2.3.4 Data Safe Havens

According to Burton et al. (33), Data Safe Havens (DSHs) is a secure database for sensitive data related to biomedicine, health and healthcare systems, accessible only by authorized users with appropriate informatics systems and governance, who are working on projects and investigations that enhance these fields.

In a secure and reliable environment for performing medical studies, DSHs provides access to medical records, financial data, and other sensitive information. It follows compliance regulations, respects data protection policies, and helps reduce the risk of data breaches and cyberattacks, which facilitates collaboration among researchers by accessing sensitive data in a secure environment.

However, maintaining DSHs can be expensive once it often requires significant financial and technical resources. DSHs can also restrict or limit access to specific types of data, resulting in difficulties for researchers to have access to it. Researchers may also face difficulties regarding inconsistent policies, where each environment may be under data protection policies that are different from others.

The concept of DSHs in AI is becoming increasingly important as AI systems become more complex and integrated into everyday life.

2.4 Summary

There are several studies proposing different federated learning approaches for disease identification and classification, going from small to higher complexity.

Most of the reviewed articles were published last year, which proves that researchers are interested in this field, and FL in medical imaging AI based solutions has the potential to inspire and attract them to advance in this field. In addition, most authors use CNNs to extract features from the input images and use pre-trained models which enables neural networks models to learn faster.

The authors have used different approaches, datasets, and architectures in building their FL frameworks, but overall, MR images have led to better accuracy than CT and X-ray images, and might be the reason for MR images have had more attention in the studied articles, once MR

images are a better option than X-ray or CT images when specialists need to visualize soft tissues. Likewise, there are several publicly available datasets of MR and X-ray images, but just a few with CT images, and this difficulty of research should be addressed.

Histology images are another important type of medical imaging modality that affects clinical decision-making, but there is a lack of FL solutions using this type of image, as demonstrated in the presented literature search. In this way, there are opportunities and demands for those images to be explored in a FL environment for the purpose of solving different problems.

The direct quantitative comparisons between models applied on different datasets are not suitable because each research has a different context and goal, e.g., detecting lung abnormalities in COVID-19 patients, or binary classification to diagnose a brain tumor disease.

This chapter also presented a summary of free open sources for bioscience collaboration, which aims to improve human health through science. OBiBa, DataSHIELD and Vipar encourage collaborative learning by using federation principles and facilitating the research relationships. In addition, Data Safe Havens can support researchers to store, analyze and work on confidential individuals' data. These environments protect sensitive personal data, establish strict data protection policies, and are accessible only by authorized personnel.

Data Safe Havens provide a secure and controlled environment for collecting and storing sensitive data such as medical records, financial details, and personal information. It adheres to strict ethical and legal guidelines, ensuring that AI systems do not violate user privacy, consent, or discrimination; and facilitates such collaboration while maintaining data confidentiality and integrity.

From the current review, it can be realized that in the near future researchers, clinicians, and hospitals might use more frequent applications that utilize FL to predict mortality, propose treatments to start earlier improvement of patients' care, and maintain the security of patients' data.

Through its comprehensive review, this survey can be used as a reference for future explorations in FL medical domain.

Chapter 3

Pneumonia

This chapter describes an overview of pneumonia, symptoms, and the global survival rate are described in Section 3.1. The diagnosis exams applied to identify this disease are detailed in Section 3.2. A summary of this chapter is presented in Section 3.3.

3.1 Overview

Pneumonia is a respiratory infection commonly caused by viruses, bacteria, fungi, or parasites affecting millions of people worldwide. Considered a significant global contributor to mortality, this aggressive disease, if not adequately treated, can result in illness and severe health complications.

The air sacs in one or both lungs get infected and inflamed. They are filled with fluid or pus (purulent material), as shown in Figure 3.1, resulting in difficulty breathing by limiting oxygen intake, chills, cough with phlegm or pus, and fever. The infections are generally spread by direct contact with infected people, being more vulnerable in children and older people.

Figure 3.2 exemplifies chest X-rays from people diagnosed with bacterial and viral pneumonia, in which it is possible to see the infected parts of the lungs and compare them to healthy lungs.

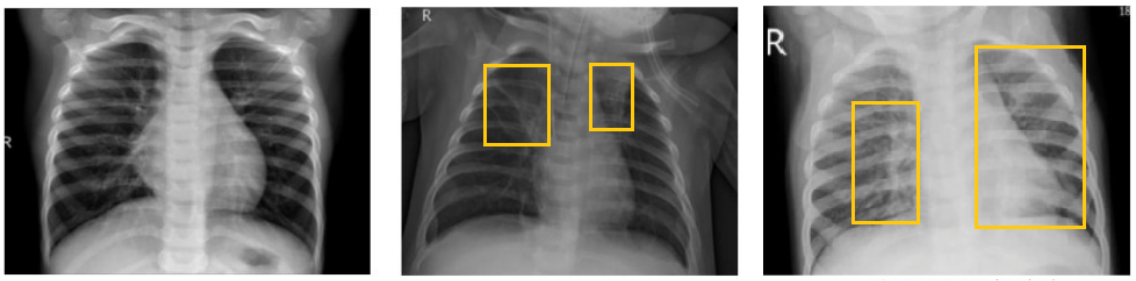


Figure 3.2: Example of chest X-ray images of (a) healthy lungs, (b) pneumonia bacterial, and (c) pneumonia viral.

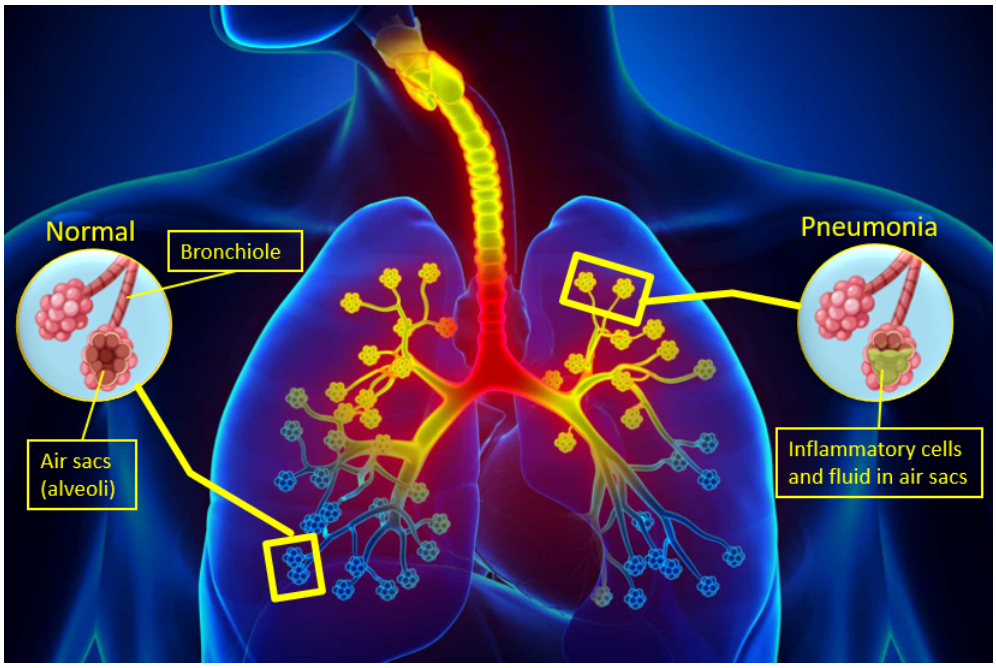


Figure 3.1: Healthy air sacs versus air sacs inflamed in the lungs.

Bronchial pneumonia (known as bronchopneumonia or focal pneumonia) involves patches throughout both lungs. It typically affects smaller areas of the lungs, such as one or more bronchioles (the small airways that lead to the alveoli or air sacs in the lungs). Lobar pneumonia involves larger sections of the lungs, usually one or more areas (lobes) of the lungs, and four different phases pursue them:

- Congestion: occurs around 24 hours after developing lung infection and lasts about 24 hours. At this first stage, the common symptoms are fever, tiredness, chest pain, and congestion (34).
- Red hepatization: happens around two to three days after the infection develops and usually lasts between two to four days. At this second stage, it is common to have difficulties breathing, the lungs are firm, and they may appear red because more red blood cells are fighting against the disease (34).
- Grey Hepatization: occurs on days four to six of infection and endures for four to eight days. The lungs appear grayish-brown or yellow because of the decomposition of red cells, leading to a more fluid-like exudate (34).
- Resolution: this is the final recovery stage and occurs for 8 to 10 days when the immune cells and enzymes fight the infection permitting air sacs to return to normal lung function. More coughing and sputum production are expected as the lungs operate to release the accumulated fluid and debris (34).

Additionally, pneumonia can be classified into various types, which are:

- Atypical or Walking Pneumonia: considered as a mild form of pneumonia, Gereige and Laufer (35) state that this type of pneumonia affects school-aged children and young adults, caused by a bacteria called *Mycoplasma pneumoniae*, delivers minor symptoms that do not disrupt regular activity.

- Community-Acquired Pneumonia (CAP): Brown (36) explains that it is an acute infection that frequently requires hospitalization, and can also be referred to as a community-acquired lung infection. The principal pathogens that cause CAP include influenza A, *Streptococcus pneumoniae*, *Chlamydophila pneumoniae*, and *Mycoplasma pneumoniae*. The primary risk factors associated with CAP are age, smoking and comorbidities. The incidence of CAP and its typical complications, including the need for intensive care and complex parapneumonic effusions, is on the rise.

- Nosocomial pneumonia or hospital-acquired pneumonia (HAP): Gereige and Laufer (35) explain that it occurs 48 hours or more after hospital admission. People who live in chronic care facilities, for example, are also vulnerable to acquiring this type of pneumonia due to pre-existing or long-term illness.

Cases of pneumonia are present throughout the year, but they are more common during the colder months in moderate climates without a specific cause. It is believed that the higher incidence is due to the transmission of respiratory droplets from person to person, increased indoor crowding, weakened mucociliary clearance, and the peak of viral infections, which can lead to viral pneumonia with secondary bacterial pneumonia (35).

According to Unicef (37), pneumonia globally kills more newborns and children than any other infectious disease. In 2019, pneumonia accounted for 14% of all deaths of children under 5 years old worldwide.

Unicef also states that lack of proper nutrition, indoor and outdoor air pollution, polluted drinking water, and lack of sanitation and healthcare are the main reasons that contribute to mortality due to childhood pneumonia. These factors are strongly associated with poverty (37).

Additionally, the WHO's Global Health Estimates (38) report that lower respiratory infections ranked 4th as a communicable disease causing death globally in 2019, and pneumonia falls within this category, as shown in Figure 3.3.

The survival rate of pneumonia varies depending on several factors such as age, health condition, type of pneumonia, and promptness of treatment. According to the World Health Organization (39), the average mortality rate for CAP is between 3-6%, while the mortality rate for HAP is higher at 20-50%. In general, younger individuals with no underlying health conditions have a better chance of survival compared to elderly individuals with pre-existing health issues.

Proper treatment and supportive care, which depend on the underlying cause, severity, and individual factors, can significantly improve the chances of recovery and survival. Antibiotics may be prescribed for bacterial pneumonia, antiviral medicine for viral pneumonia, antifungal medication for fungal pneumonia, and specific medications to kill parasites for parasitic pneumonia.

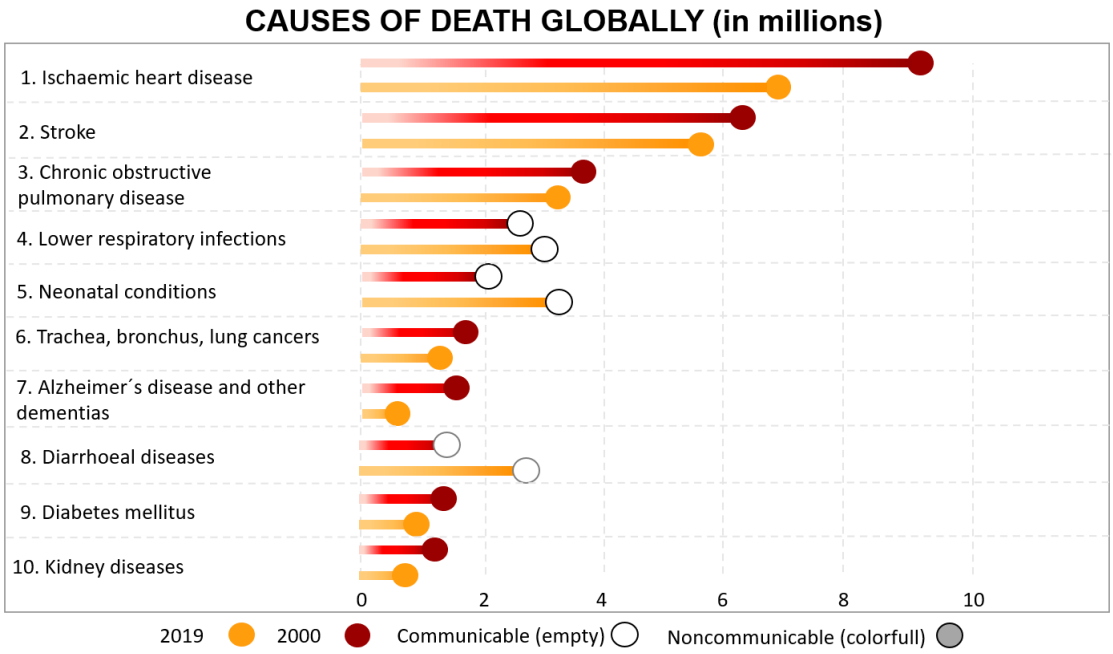


Figure 3.3: Leading causes of death globally (adapted from WHO's Global Health Estimates, 2020), (38).

Additionally, recovery requires plenty of rest, avoiding smoking, staying hydrated by drinking fluids, and controlling fever with prescribed medication.

3.2 Diagnosis exams

Several tests are usually performed to diagnose pneumonia, including blood tests, chest X-rays, pulse oximetry, and sputum tests. Further detailed testing, such as CT scans and pleural fluid culture, may be necessary for patients older than 65 years or those with serious symptoms or health conditions.

Considered one of the best methods for pneumonia detection, chest X-ray images are the primary test used for diagnosing this disease (40) because they can help doctors and radiologists to understand the extension and location of the infection by promoting the view of the lungs, heart, and blood vessels.

X-rays use electromagnetic waves to create images of the inside of the human body. In a chest X-ray of a healthy person, the radiation cannot pass through dense tissue and bone, resulting in white areas on the image. Softer tissues, like the lungs, can be seen as darker areas in the image.

To diagnose pneumonia, the infected tissue shows denser areas appearing as white spots in the darker background of the lungs. Depending on how big these spots are, radiologists and doctors can determine how severe the infection is.

Although, analyzing chest X-rays images is a challenge for radiologists due to the workload, pressure on analyzing higher volumes of images, the complexity of cases, and similarity to other pulmonary diseases, which requires them to do more studies in a short period. Even though

professionals in the health area are well trained, they are prone to making mistakes, and the lack of information or spending much time examining images is crucial to define a proper treatment to avoid the growth and spread of pneumonia.

In complex cases when chest X-rays are not sufficient to ensure the pneumonia diagnosis, a CT scan, for example, is a powerful test used to obtain a detailed image of the lungs and airway (trachea and bronchi), once it takes pictures from different angles and puts them together in a three dimensional (3D) image.

To facilitate and improve the disease diagnosis process, AI solutions have been continuously more popular lately (41), (42).

3.3 Summary

Pneumonia is a dangerous lung infection, and determining whether a patient has this condition is critical in making treatment decisions. It can range in severity from mild to life-threatening, but early diagnosis can prevent complications and help for a faster recovery.

Several diagnostic exams can be used to diagnose pneumonia, including physical examination, chest X-rays, blood tests, sputum tests, and pulse oximetry. Although, chest X-rays are the most frequently used method in diagnosing pneumonia.

The treatment may vary depending on the severity and underlying causes of the disease, but it often involves antibiotics, antivirals, and supportive care such as oxygen therapy and fluids.

There are different types of pneumonia, including community-acquired pneumonia (CAP), hospital-acquired pneumonia (HAP), and walking pneumonia. CAP is the most common type acquired outside of the hospital setting. In contrast, HAP is received during a hospital stay or in a healthcare facility, and walking pneumonia (or atypical pneumonia) is a milder form of the disease caused by certain bacteria, such as *Mycoplasma pneumoniae*.

Preventive measures, such as vaccines and good hygiene practices, can help reduce the risk of acquiring and spreading pneumonia. It is essential to seek prompt medical attention if experiencing symptoms such as fever, cough, chest pain, difficulty breathing, or fatigue, as untreated pneumonia can lead to severe complications or death.

In the context of infectious diseases, pneumonia is the top causing death in newborns and children. Exposure to certain environmental factors, such as pollution, toxic fumes, and chemicals, can increase the risk of developing pneumonia. Additionally, weakened immune systems (aged adults and young children are more vulnerable), chronic lung diseases, and smoking are also the main reasons why pneumonia can be developed.

Early detection and accurate diagnosis of pneumonia are crucial for effective treatment and management of the disease. To address it, machine learning-based approaches have emerged as powerful tools for improving pneumonia diagnosis with high accuracy and speed.

Chapter 4

Dataset and Framework

In this chapter, Section 4.1 provides an introduction to the problem that needs to be addressed and the specific type of medical imaging that will be utilized in a FL framework to facilitate problem-solving. Section 4.2 details the selected dataset. Section 4.3 presents CNN architectures used for chest X-ray imaging classification. Finally, Section 4.4 describes the framework applied in this research work.

4.1 Introduction

A computerized AI-assisted program was developed to identify pneumonia using chest X-ray images, which is expected to improve patient care through early disease detection and consequently limit the spread of the disease. In order to enhance the diagnosis, the solution aims to support radiologists and doctors in their decision-making process by extracting faster and more accurate results.

In this context, a chest X-ray imaging dataset was used in the proposed solution. The FL environment utilizes Flower as a framework for the development of a comprehensive classification model to identify patients as having pneumonia or not.

4.2 Dataset

As introduced, this research aims to build a federated learning approach over a supervised classification task to predict pneumonia based on chest X-ray images. Therefore, it was used a dataset (43) composed of chest X-ray images from pediatric patients from one to five years of age at the Guangzhou Women and Children's Medical Center.

The chest X-ray images were taken from routine clinical care, and according to Kermany et al. (44), all images went through a quality control process. Firstly, low-quality or unreadable scans were discarded. Secondly, two expert physicians analyzed and classified them for training in an AI system. Lastly, a third expert evaluated the images to look for any classifying errors.

As shown in Table 4.1, the dataset is divided into sets of training and testing, and each one is organized by two subsets of chest X-ray images: NORMAL and PNEUMONIA, where pneumonia images consist of both bacterial and viral pneumonia. In addition, all images are .jpeg format and there is no standardization in their sizes.

Table 4.1: Description of the selected dataset.

	Training	Testing
PNEUMONIA	3.875	390
NORMAL	1.341	234
Total	5.216	624

Having two categories in the dataset, normal and pneumonia, is the basic requirement for a binary classification problem. The machine learning model is trained on such labeled data to understand and distinguish the common features that those categories have. Then, the testing images are used to test the model, meaning that the model has never seen those new images before but should be able to comprehend if they contain features that classify them as having pneumonia or not based on what was learned.

Figure 4.1 shows samples of normal and pneumonia chest X-ray images from the selected dataset.

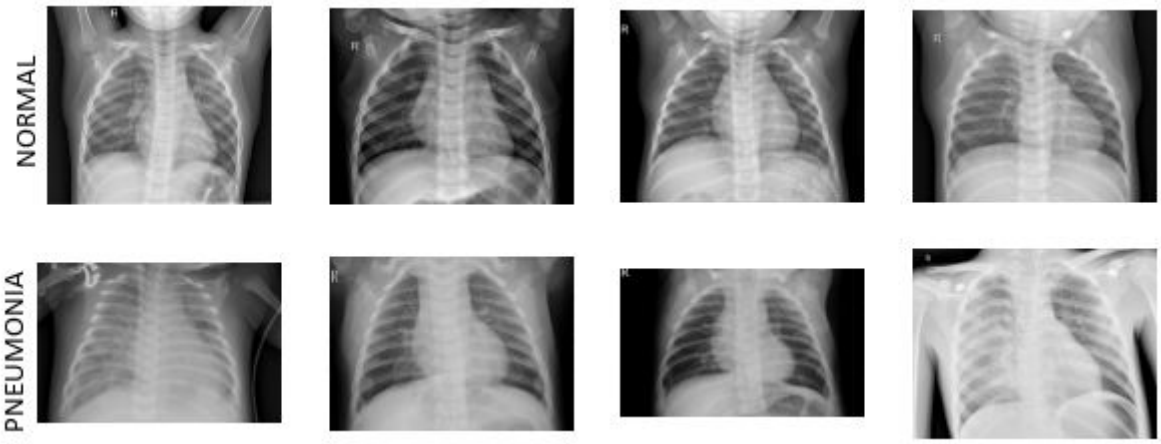


Figure 4.1: Samples of the selected dataset.

4.2.1 Data Augmentation

Data Augmentation is a technique used to create synthetic images based on original ones by applying modifications to them, for instance, geometric or color modifications.

This technique is used to increase the number of images in the dataset, especially when the dataset is small, which is the case of the selected dataset. In this way, the dataset gets enriched randomly by new images similar to the original ones, and the model is able to learn and extract new features, improving the model's performance and preventing overfitting.

The chest X-ray imaging dataset is considered small and not balanced (training represents 89.3% and testing represents 10.7% of the entire dataset). As a consequence, the applied model would not learn different and complex features with a few images, especially in a FL environment where the learned model is disseminated to the edge models. In order to have a larger dataset, it was applied data augmentation techniques in the entire dataset, as described in Table 4.2. The output images keep three channels of colors since they need to match the number of channels of the images used in the pre-trained models.

Table 4.2: Performed data augmentation techniques.

Technique	**Setting**
Resize	224x224x3
Normalization	mean and std of 0.5 for each RGB channel
Horizontal shift	0.15
Rotation	20

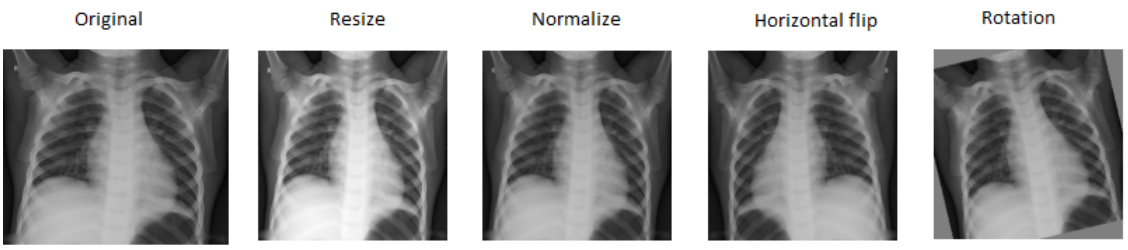


Figure 4.2: Resulting images from various data augmentation techniques.

It was considered the whole image as the region of interest (ROI), once the images do not have a standard picture frame and applying cropping might lose the most significant information from them.

After the raw chest X-ray images are processed in data augmentation procedures, the dataset is prepared and ready to be used in CNNs.

4.3 CNNs

A Convolutional Neural Network (CNN) is a type of deep neural network that is commonly used for image recognition and computer vision tasks. It is designed to analyze and process visual data in the form of images or videos. The network consists of a series of convolutional layers and pooling layers that extract features from the input images. These extracted features are passed through one or more fully connected layers to classify the input image. The trainable parameters of the network are tuned using a backpropagation algorithm to minimize the error between predicted and actual output. CNNs have been successful in a wide range of tasks including image classification, object detection, facial recognition, and natural language processing (45).

Pre-trained algorithms in CNNs, or Convolutional Neural Networks, encompass pre-trained models that have already learned from vast datasets, primarily for image classification tasks. These models significantly facilitate the recognition and classification of images. As a starting point, they possess meaningful features acquired from popular image datasets such as ImageNet, CIFAR-10, COCO, or Pascal VOC. In terms of architecture, pre-trained CNNs consist of a series of convolutional and pooling layers that extract relevant features and subsequent fully connected layers that contribute to efficient image recognition and classification. Apart from image classification, pre-trained CNNs find applications in transfer learning, object detection, and even in natural language processing tasks like sentiment analysis.

Pre-trained algorithms enable developers to use transfer learning, a technique where the pre-trained model is used as the initialization for the new task. Transfer learning allows for much faster training, as the network has already learned many features for image classification, reducing the need to start from scratch. Pre-trained models can be fine-tuned on the target dataset to further improve classification accuracy and performance.

Common pre-trained CNN models include VGG, ResNet, DenseNet, Inception, and MobileNet. These models can be imported and used in popular deep-learning frameworks like TensorFlow and PyTorch. Overall, pre-trained CNN models are a crucial component of the deep learning toolkit, allowing developers to accelerate their machine learning projects and produce accurate results.

Three different pre-trained algorithms were applied in this work for transfer learning: VGG-16, ResNet-18 and Resnet-50, which load weights and biases from a dataset of over 14 million images belonging to 1000 classes, called ImageNet dataset (46). Based on the state-of-art of this research work, as seen in Tables 2.10, 2.11 and 2.12, most studies on chest X-ray images apply those three architectures, which is a motivation for using them in the proposed methodology.

These pre-trained models are considered simple and applicable for solving deep learning problems, such as the classification of pneumonia or not through chest X-ray images.

The concepts of Visual Geometry Group-16 (VGG-16) (47) from the CNN are comprised of a network that features 16 layers. It uses only 3×3 convolutional layers stacked on top of each other in increasing depth, reduces volume size with max pooling, and uses two fully-connected layers.

Two ResNet models using both 18 and 50 layers were applied in this work since it skips connections to avoid vanishing and exploding gradients. It also holds 3 channels as input and makes use of global average pooling to help optimize the model (48).

Figure 4.3 details how a CNN architecture can be organized.

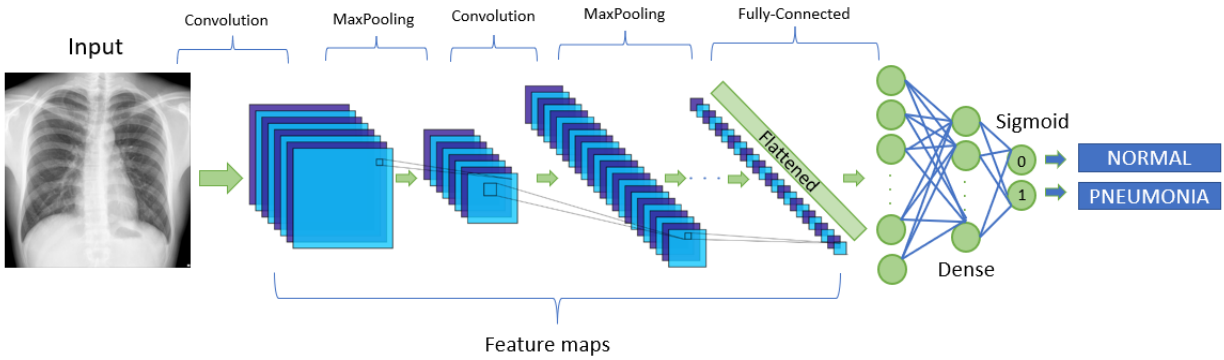


Figure 4.3: Example of a CNN architecture.

Certain parameter adjustments must be established to train the network in order to achieve the best or desired performance for pneumonia diagnosis. Table 4.3 details the hyper-parameters used for fine-tuning, where each architecture used the same fine-tuning settings for performance comparison purposes.

The neural network was compiled on top of the optimizer Stochastic Gradient Descent (SGD). As explained by Wilson et al. (49), SGD has a higher abstraction compared to adaptive optimizers, which are for example Adagrad, Adadelta, and Adam. Furthermore, BCEWithLogitsLoss was used as a loss function, combining a Sigmoid layer and the BCELoss in one single prediction class.

Table 4.3: Hyper-parameters used for fine-tuning.

Architecture	Image Size	Epochs	Optimizer	Learning Rate
VGG-16 ResNet-18 ResNet-50	224 x 224	10	Stochastic Gradient Descent	0.001

4.4 Frameworks

Implementing federated learning using Python language is possible through frameworks. According to Braungardt (50), a recent survey from January of 2023, stated that the most used open-source software for federated learning are the following:

- NVIDIA Federated Learning Application Runtime Environment (NVFlare): it is a FL framework developed by NVIDIA that supports efficient communication and coordination among participating devices. It enables privacy-preserving machine learning by allowing data to remain on local devices while training models on a central server. This framework is fast and scalable for large-scale federated learning, and supports various deep learning frameworks including TensorFlow and PyTorch. The cons of NVFlare framework are the limited documentation and lack of community support and requires expertise in GPU programming (51).

- Federated AI Technology Encapsulation (FATE): it is an open-source federated learning framework developed by WeBank that provides a secure computing environment for cross-institutional collaborative machine learning tasks. It includes several privacy-preserving technologies, such as homomorphic encryption and differential privacy. The pros consist of providing a complete end-to-end solution for federated learning, including privacy-preserving techniques, and supporting multiple machine learning algorithms and frameworks. The cons include a steep learning curve due to its complexity, limited documentation, and lack of community support (52).

- Flower: it is a lightweight Python library for building federated learning models. It provides a simple API for writing federated algorithms and integrating them with popular machine-learning libraries such as TensorFlow, PyTorch, and Keras. It is a great framework to get started with if developers are new to federated learning, as detailed in Section 4.4.1. The pros include simple and flexible API for federated learning, provides privacy-preserving techniques such as differential privacy. The Flower community is growing quickly and contains good documentation for beginners. It supports PyTorch, TensorFlow, Pandas, and Scikitlearn, among other libraries in Python (53). Although, FL is challenging to implement realistically, both in terms of magnitude and diversity of systems.

- PySyft: it is a Python library that allows you to write federated learning algorithms for distributed machine learning. It is an open-source project developed by OpenMined and is gaining popularity in the federated learning community. The pros consist of providing a high-level API for federated learning, and it supports different privacy-preserving techniques such as differential privacy and homomorphic encryption. The cons are composed of limited scalability for large-scale federated learning, plus, it requires expertise in PyTorch or TensorFlow (54).

- IBM FL Community Edition: it is a FL platform developed by IBM that enables privacy-preserving training of AI models across multiple devices and data centers. It includes a toolkit for developing and deploying distributed machine learning models using TensorFlow and PyTorch. As an advantage, it provides a good balance between ease of use and sophisticated features and supports multiple machine learning frameworks. As a disadvantage, it has limited scalability for large-scale federated learning, and limited documentation and lack of community support (55).

- OpenFL: it is an open-source federated learning framework developed by LF AI & Data that aims to simplify the deployment and management of federated learning systems. It includes tools for data partitioning, model aggregation, and client selection. OpenFL supports various machine learning frameworks, and provides privacy-preserving techniques including differential privacy. On the other side, it presents limited scalability for large-scale federated learning, limited documentation and lack of community support (56).

- TensorFlow Federated (TFF): it is an open-source framework developed by Google that allows running machine learning models using federated learning. It is mainly used for scalable and privacy-preserving collaborative machine learning. Tensorflow Federated can be used to build customized federated learning models in Python. Because it was developed by Google, there is strong community support and supports both TensorFlow and Keras packages in Python. Although, it presents limited scalability for very large-scale federated learning, and limited support for privacy-preserving techniques (57).

- Federated Learning Library (FedML): it is an open-source federated learning toolkit developed by Tencent that provides a set of reusable modules for building and evaluating federated learning systems. It includes algorithms for federated learning, as well as evaluation metrics and visualization tools. Also, it supports various machine learning frameworks and algorithms, and provides privacy-preserving techniques including differential privacy and secure aggregation. Although, it presents limited scalability for very large-scale federated learning, limited documentation and lack of community support as well (58).

Considering the main functionalities of a FL framework, such as client-side training, server-side aggregation and communication, the FL framework chosen for this work is called Flower, a Friendly Federated Learning Framework (53) that facilities the execution of large-scale FL experiments and contemplates heterogeneous edge devices.

Another motivation for using Flower is that it provides good documentation for beginners (59), which contains beneficial working examples as a starting point for creating a federated learning system.

4.4.1 Friendly Federated Learning Framework

Flower has the ability to run experiments between simulation and real-world devices by shifting or mixing those scenarios, which is handy for exploring each individual area, facilitating the simulation of a realistic federated learning environment using those elements.

Another strength of Flower is its capacity to use different ML frameworks. Multi-framework workloads accommodate updates from multiple clients with various ML frameworks, having an updated and integrated environment.

Beutel et al. (53) developed the Flower framework architecture organized in three main goals, which are:

- Core Framework Architecture: the global computations happen on the global server, then the learning is distributed over the available clients. The local computations happen on the client's side once they have access to the actual data, which is used to train or evaluate model parameters.

  The core framework architecture is composed of Edge Client Engine and Virtual Client Engine. Edge clients carry out the real edge devices and communicate with the server over Remote Procedure Call (RPC).

Virtual clients load models and data into memory only if the client is selected for training or evaluation; thus, inactive clients make the consumption of resources reduce dramatically close to zero. On one hand, Edge clients operate on tangible edge devices and utilize RPC to communicate with the server. On the other hand, Virtual clients are extremely resource-efficient while inactive, only loading data and models into memory when it is needed for training or evaluation.

RPC is a widely used technology in distributed computing and is commonly used in client-server systems, where clients request a server to execute a particular task or process. In simple terms, it is a process by which a computer program can execute code on a remote computer or server as if it were running on the local machine without having to know the details of its underlying network or communication technology.

Figure 4.4 exemplifies the Flower core framework architecture with both Edge Client Engine and Virtual Client Engine.

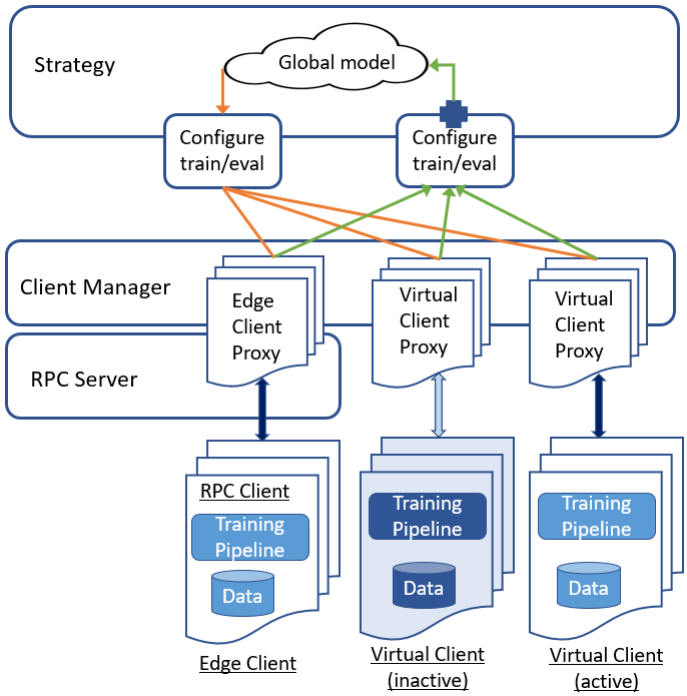


Figure 4.4: Flower core framework architecture with both Edge Client Engine and Virtual Client Engine (adapted from Beutel et al., 2020), (53).

Once the local models are trained, the global model takes the average of the local model parameters by applying FL algorithms such as FedAvg (6), so the global model gets updated. Besides the FedAvg, Flower also provides implementation of both SecAgg (60) and SecAgg+ (61) protocols.

Three components build the Server side: The server ClientManager, the Strategy that can be user-customizable, and the FL loop.

The ClientManager provides sample clients to the server components and manages a set of ClientProxy objects. Each object represents a single client connected to the server. The actual client communicates with the server by sending and receiving Flower Protocol messages through those components.

The FL loop coordinates the entire learning process, but the Strategy is the component that decides how to proceed. In addition, the connections, errors, and client life cycle are all managed by the framework.

- Virtual Client Engine

  The Virtual Client Enginee (VCE) is a tool for virtualizing Flower Clients, which maximizes the utilization of the available hardware.

  This component is responsible for scheduling jobs, initializing and running services of Flower Clients and Flower Server, avoiding unnecessary resource hardware, and re-configuring setups, such as GPU (single or multi-node) or desktop machine.

  VCE enables the Flower framework to use computational resources more robustly, improving the usage of computational resources to run FL workloads on a large scale.

- Edge Client Engine

  Flower is an open-source framework, and it was designed to be device agnostic, which means that it can have new connected computing components without requiring any specific adaptations. With that, this framework provides a low-level integration, offering the Flower Protocol to exchange messages on the client side.

Flower provides a customizable, understandable, extendable, and framework-agnostic for building federated learning systems. Besides the federated learning approach, Flower also presents federated evaluation and federated analytics unified in this framework, assisting the data analysis and evaluation of the model on the data.

Section 5.4 details how the Flower framework was used in the research.

4.5 Tools

The implementation of a federated learning system for this research work used Flower and PyTorch as Frameworks, coded in Python as programming languages.

The execution was carried out on Google Colaboratory (Google Colab) (62), which is a web-based environment for creating, storing, and accessing scripts coded in Python in a cloud-based storage service called Google Drive, which was also utilized for storing the acquired chest X-ray images.

Google Colab is commonly used for data science, machine learning, and deep learning tasks because it offers free access to computing resources such as CPUs, GPUs, and TPUs and facilitates easy collaboration and sharing with others.

To simulate a FL environment, the free version of Google Colab did not have the required computational resources to run the code. To solve this limitation, it was used a paid upgrade version of Google Colab, called Google Colab Pro, that offers faster GPUs and additional RAM to save computing time while computing the code.

Python is an interpreted high-level programming language used widely in data science, machine learning, and other scientific computing tasks due to its simplicity, readability, and many libraries and tools. It is commonly used for tasks such as data cleaning, data manipulation, model training and evaluation; plus, it has a large community of developers and users contributing to its open-source libraries and tools.

Pytorch is a framework based on Python and provides a simple and intuitive interface for building and training neural networks. It includes a series of powerful libraries for handling complex data structures and creating efficient data pipelines. PyTorch also supports distributed computing, which allows developers to train their models on multiple GPUs or across multiple nodes in a cluster.

The experimental results have been implemented using Python 3.10.0, Flower 1.4, PyTorch 2.0, and CUDA used to build PyTorch: 11.8. The specifications of the computer are Core i5 10th generation, Intel(R) UHD Graphics and 8GB Ram.

4.6 Summary

The supervised learning approach built in a FL environment used the Guangzhou Women and Children's Medical Center pneumonia dataset. It had its images transformed by the data augmentation technique, which is used in machine learning to increase the amount of data by generating new, slightly different versions of existing data.

In addition, the imaging dataset had its features extracted through pre-trained CNN models: VGG-16, Resnet-18, and Resnet-50, where CNNs are deep learning algorithms commonly used for image recognition tasks due to their ability to extract features from images and build complex models.

Currently, several free and open-source frameworks can be embedded in Python language to develop FL solutions, but the proposed research selected Flower as a federated learning framework, which is an open-source platform for building and running distributed machine learning systems. It enables individuals and organizations to collaborate on training machine learning models without sharing the underlying data, thereby preserving user privacy.

Tools such as Google Colab and Python are commonly used to run codes for implementations of data science and machine learning solutions. Although, it is significant to highlight that training models and sharing them among clients in a FL environment using Google Colab might be computationally expensive.

To facilitate the system development using Python, the Pytorch framework was used together for supporting distributed computing. Once this framework was designed to be flexible, it is suitable for researchers and scientists who need to rapidly prototype new neural network models.

Defining the dataset, data augmentation techniques (if required), frameworks, programming language, and additional tools are essential components to support the development of computer-aided diagnosis solutions. In this way, they can provide the necessary elements and infrastructure for the developer or data science engineer to build a FL solution.

Chapter 5

Implementation and Results

In this chapter, firstly, Section 5.1 introduces the content presented in this chapter. Secondly, Section 5.2 explains the methodology used to create a federated learning environment to identify pneumonia in chest X-ray images. Thirdly, Section 5.3 details the metrics used to evaluate the models' performance. Then, Section 5.4 explains the framework used in this research work. Section 5.5 describes the results obtained from the experiments in this work. Lastly, Section 5.6 discusses the results and a chapter summary.

5.1 Introduction

Developing and implementing a solution based on federated learning to support medical diagnosis requires defining the techniques, restrictions, and procedures for conducting the defined experiments. Based on that, it is possible to explore how the solution behaves, understand its outcomes and perform adjustments.

Having that in mind, this chapter provides meticulous details of each step involved during the project development. Moreover, the analysis of the results obtained from the developed solution is comprehensively examined to achieve effective and reliable diagnosis resolution to enhance clinical decision-making for the proposed use case.

5.2 Methodology

5.2.1 Cross-silo

To build a federated learning system, the experiments performed in this work include simulating different data coming from various institutions working together to train the classification models.

As only one dataset was utilized in this experiment, a specific setting in federated learning called "cross-silo" (63) was employed. This setting entailed dividing the primary dataset into multiple partitions to mimic diverse datasets from various organizations. Therefore, each partition contained information from a distinct entity. The federated learning environment would be connected to various organizations with their own data in the real world.

The FL environment was simulated as having eight clients, where each client is represented by an institution connected to the federated learning server. Thus the central model is trained over these eight organizations.

Each client has its own training (90%) and validation (10%) data, meaning that the original 5.216 training images were split into eight clients containing 586 images for training and 66 images for validation.

A single test set (not split) containing the original 624 images was used to evaluate each organization's network with its own data.

5.2.2 Batch size

DataLoader in PyTorch is a data loading utility used to load data from a dataset and create minibatches for training. It is used to simplify the data loading process and allows for efficient and effective training of machine learning models.

This data loading utility was used to create the training, validation, and testing DataLoaders. For that, it was required to specify the number of samples in each minibatch for the neural network.

The batch size is a significant hyperparameter that affects how the model is trained and the computation resources required to train a neural network. It determines how many samples are processed at one time in a single forward and backward pass through the network.

The larger the batch size, the faster the training process, but it also requires more memory. On the other hand, by using a smaller batch size, the network is updated more frequently with smaller gradient steps, which can lead to a more accurate model. However, smaller batch sizes can also result in slower convergence and noisy updates, especially if the data has a high variance.

In this particular case, a batch size of 16 was chosen because it strikes a balance between efficiency and model performance by allowing a reasonable number of iterations in each epoch, which helps reduce the variance in estimating the gradients.

Ultimately, the data shuffling feature was applied to the training DataLoader to reduce bias during training, in which the data is randomly shuffled after each epoch.

5.2.3 Server-side evaluation

The method used to evaluate federated learning for a chest X-ray classification scenario is server-side evaluation. In server-side evaluation, the model is trained on data from multiple clients in a distributed manner, and the aggregated model is evaluated on a held-out test set. This method ensures that the model is evaluated on a representative sample of data from all clients, rather than just the data from a single client in a client-side evaluation.

This approach is often used when there are privacy concerns and the clients want to keep their data locally. Although, server-side evaluation requires significant computational resources on the central server, which can cause performance issues.

In terms of data privacy and security, evaluation on the server-side means that sensitive data is not exposed or transmitted to the client side, reducing the risk of data breaches and theft. With

server-side evaluation, only the necessary results or information are sent to the client side, reducing the amount of data that needs to be transmitted and eliminating the need to store sensitive information on the client side. Additionally, security features such as encryption and authentication can be implemented on the server side to ensure that data is protected from unauthorized access.

Overall, server-side evaluation provides a more secure and privacy-conscious way of handling data, making it a preferred solution for applications dealing with sensitive information, such as chest X-ray images.

5.3 Metrics

All the experiments were assessed using the test dataset once the model was trained. Their efficacy was evaluated using metrics such as accuracy and F1-score.

In the following definitions and equations, when distinguishing between healthy and pneumonia patients, TP represents the number of normal images properly recognized as normal (healthy), TN means the number of pneumonia images correctly identified as pneumonia, FP stands for the number of pneumonia images inaccurately recognized as normal, and FN refers to the number of normal images wrongly recognized as pneumonia.

- Accuracy: It represents how the ML model is classified correctly.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{5.1}$$

- F1: It is defined as the harmonic mean of precision and recall. Precision represents how accurate the model is respecting to which were predicted positive. The recall represents the number of actual positives the model was capable of capturing by labeling it as positive (true positive).

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall} = \frac{2 * TP}{2 * TP + FP + FN} \tag{5.2}$$

In pneumonia detection, it is essential to reduce the occurrence of False Negatives, as it can lead to various medical implications when the forecast suggests a healthy lung, but in reality, it is pneumonia. Those medical implications may hinder the initiation of proper treatment and protocols in a timely manner.

5.4 Federated Learning with Flower

In federated learning, the central server transmits the global model parameters to individual clients, which then modify their local models accordingly. These local models are trained using the client's own data, resulting in local adjustments to the model parameters. The client then transmits the **updated model parameters** back to the server, or, in an alternate approach, only the gradients are sent back rather than the complete set of parameters (64).

The experiment replicated a federated learning system with eight clients on a single machine, applying the **virtual client engine**. It implies that the resources such as CPU, GPU, and memory need to be shared among the server and all eight clients. However, running eight instances of FlowerClient in memory on a single machine can quickly exhaust the available memory resources, even if only a subset of these clients participate in a single round of federated learning. Therefore, Flower provides unique simulation features that create instances of FlowerClient only when the need arises for either training or evaluations, allowing Flower to create clients when required. That facilitates the creation of a FlowerClient instance that performs fit or evaluation roles for a given client ID, where each client ID loads distinct local data partitions for diverse clients.

5.4.1 Strategy

The strategy applied in this research work is a built-in strategy of Flower called Federated Averaging (FedAvg).

FedAvg is a basic algorithm for federated learning that uses simple averaging of the local client models to update the global model. In FedAvg, each client trains a local model on its own data, then sends the local model weights to the server, where they are averaged to obtain a global model.

In addition, FedAvg works well when the datasets of the clients are similar in distribution because it assumes that the clients' datasets are independent and identically distributed. On the other hand, if the datasets are not similar in distribution, the model may become overfit to certain clients' data, resulting in poor generalization performance.

The strategy had the following parameters customized on the initialization:

- Sample 30% of available clients for training;
- Sample 30% of available clients for evaluation;
- Minimum of 3 clients available for training;
- Minimum of 3 clients available for evaluation;
- Minimum of all 8 clients available in the system.

The simulation starts by retrieving the pre-trained model parameters, either VGG-16, Restnet-18, or Restnet-50, and using them to create an instance of the model, which are the initial parameters for the global model. Therefore, Flower ensures that customized parameters for global model initialization are valid by randomly selecting two clients (30% of available clients) to train the model.

The server obtains the updated model parameters from the clients and delivers them to the FedAvg strategy for consolidation. The strategy combines the updates to produce the new global model, which is then utilized in the next round of federated training.

5.5 Results

5.5.1 Training the model before FL

Firstly, in order to evaluate the network performance before implementing FL, the model was trained on the dataset of one of our organizations, selected randomly and not in a distributed learning.

Conducting a 10-epoch training of the CNN model using our pneumonia split has led to a test set accuracy of roughly between 73% and 77% in all architectures, but the VGG-16 has shown the best result, as presented in Table 5.1. Even though those accuracies are high, they may be examined carefully, once the dataset is not class-balanced, meaning that each class of the dataset does not have the same number of samples. Nevertheless, real-world datasets are heavily class-imbalanced, and the purpose of this research work is to preserve the unbalanced dataset to mimic a dataset used in the real world.

The main objective of training those models before applying FL is establishing a rudimentary centralized training process laying the foundation for the upcoming federated learning process.

Table 5.1: Testing Accuracy, Loss, and F1-score of different architectures.

Architecture	**Accuracy**	**F1-Score**	**Loss**
VGG-16	77.24%	0.842	0.593
ResNet-18	75.96%	0.835	0.504
ResNet-50	73.23%	0.821	0.486

To evaluate the scores, confusion matrices were generated for all architectures represented in Figure 5.1, allowing the assessment of model efficacy through the calculation of true positives, true negatives, false positives, and false negatives. Precision and recall have an inverse relationship, where increasing recall results in decreasing precision, and vice versa.

The confusion matrices show false negatives greater than false positives, which means that the model is more prone to incorrectly detecting positive cases as negative (false negatives) than incorrectly detecting negative cases as positive (false positives).

By analyzing Figure 5.1, it is possible to see that all three models have classified mostly correctly the pneumonia images, but mostly wrongly the normal images, resulting from an imbalanced dataset with more pneumonia images than normal images.

In medical contexts, it is critical to identify all patients with a given condition requiring maximized recall; however, a low recall may be acceptable if the cost of follow-up medical examination is not high. Hence, the F1-score provides an optimal balance between precision and recall.

In Figure 5.2, it is possible to see the training and validation curves obtained while training the models along 10 ten epochs. The training accuracy for all the models exceeded 90%, the training loss for all the models is below 0.58, and the training f1-score for all the models is below 0.96.

Having in mind that the tested epochs are from 0 (zero) to 10, the accuracy graphs show that for all models the training plot is increasing mostly until epoch number 8, then it starts having

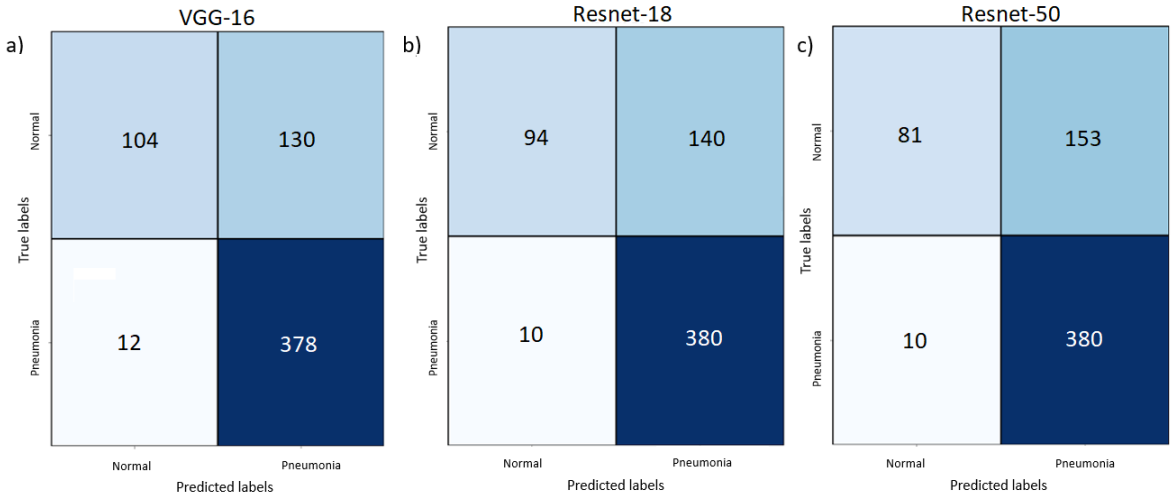


Figure 5.1: Classification matrix for (a) VGG-16, (b) Resnet-18 and (c) Resnet-50.

a constant trend until epoch number 10. Although, in Resnet-50 the training accuracy starts decreasing in epoch 8. For the validation accuracy, for all models, the metric reaches above 95% increasing from the start until epoch 7, then starts decreasing. In this way, training and validation accuracy follows a similar trend. Once the training and validation accuracy increase along the epochs, it indicates that the model is learning from the data and improving its performance.

In the F1-score graphs, the training plot increases until epoch 6 for VGG-16 and Resnet-18, but for Resnet-50 it increases until epoch 8. For the validation, all of them start decreasing when they reach epoch 7. The training and validation F1-score increase along the epochs, meaning that the model finds relevant patterns in the data and generalizes well to unseen data.

In the loss graphs, all models reach a validation loss close to 0.1 at the end of the training, starting to decrease since epoch zero. The train losses have a similar trend and stay below 0.2 for Resnet-18 and Resnet-50 after epoch 8. Although, for VGG-16, the training loss stays between 0.1 and 0.2 starting on epoch 6.

A sign that the model is acquiring knowledge is evident in the loss graphs, where the training and validation losses are diminishing.

Once the train loss and validation loss decrease along the epochs, it means that the model is improving and becoming more accurate. The training loss represents the error between the predicted output and the actual output during the training of the model. The validation loss represents the error between the predicted output and the actual output on validation data that the model has not seen before. If both losses decrease, it means that the model is generalizing well and is not overfitting to the training data. This is a desirable outcome for any machine learning model as it means that it can accurately predict new data that the model has not seen before.

However, it is important to note that increasing the accuracy and F1-score on the training and validation sets does not guarantee good performance on the test set or in the real world. It is necessary to evaluate the model on new, unseen data to ensure that it can generalize well and

perform accurately in real-world scenarios.

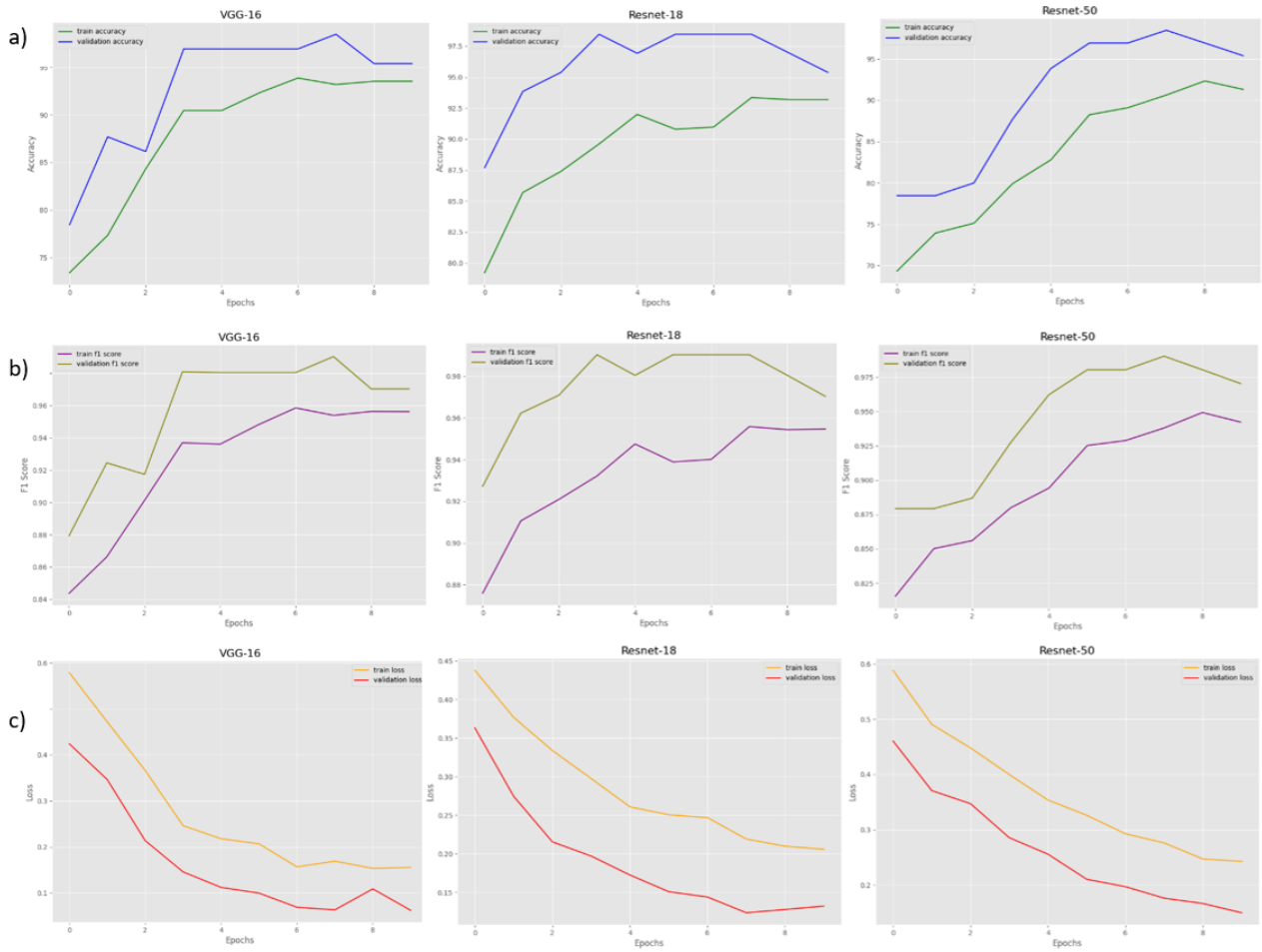


Figure 5.2: Training and validation curves for different architectures over one randomly organization for 10 epochs, (a) Accuracy, (b) F1-score and (c) Loss.

The experiments proposed in the research presented high computational cost due to an imbalanced dataset, a common limitation experimented in study (27) as well, detailed in Table 2.12, which makes use of the same dataset as this study.

The run experiments took around 3-4 hours to train each model on 5 epochs in the simulated FL system, which is considered an elevated time if this solution would be considered to be used in the real world. With that, various factors can cause the high computational costs of running a federated learning environment, including:

- Communication: Federated learning requires frequent communication between the client devices and the central server. This communication involves sending updates and receiving the updated model parameters. The transmission of these parameters and data requires significant computational power and bandwidth.

- Data Heterogeneity: In a federated learning environment, the data used for training comes from different sources. This diversity makes it difficult to harmonize the data and construct an efficient model.
- Security: Federated learning requires that client devices upload their updates to a central server, which may raise privacy and security concerns. Federated learning algorithms must therefore take measures to secure the data and protect the privacy of the clients. In terms of impartiality, certain circumstances in federated learning might promote prejudice towards a few parties, for example, the devices carrying the most regular model patterns.

  Another security challenge is that becoming distributed in federated learning enhances the probability of facing various attacks (e.g. model poisoning). Regular defense systems are hardly adequate in federated learning and additionally require specialized approaches to be made and tried.
- Resource Constraints: Federated learning models are often trained on mobile or IoT devices with limited resources. These limited computational resources can make it challenging to train complex models or process large amounts of data.
- Federated Learning Architecture: The architecture of federated learning systems can also contribute to high computational costs. The central server must process and transmit data from multiple client devices, which requires significant computational power and bandwidth.
- Model Complexity: Federated learning must manage the trade-off between model complexity and communication efficiency. Complex models require more computation power to train and more data to process, which increases the computational costs of the entire system

Overall, high computational costs are a tradeoff for the benefits of federated learning, which include improved privacy and scalability.

5.5.2 Training the model with FL

In the run experiments, the FedAvg strategy is applied, meaning that in each training round, the average of the clients' weights is transferred to the global model, as detailed in Section 5.4.1.

For that, it is simulated 5 rounds of training in the FL system using the FedAvg strategy. Once the evaluation of the aggregated model parameters happens on the server side, it is possible to see how the global model behaves in each CNN architecture, as shown in Tables 5.2, 5.3 and 5.4.

Applying the VGG-16 architecture, Table 5.2 shows that the first accuracy was 55.38%, then it increased and remained at 78.46% for the next 4 rounds. Even though the accuracy remains the same, the metrics F1-score has an oscillation from rounds 2 to 5.

The loss metrics illustrate how poorly the model performs. In the run experiments, this metric remains at 0.87 from rounds 2 to 5 as well, meaning that the model had a bad classification and had not gotten worse or better.

Table 5.2: Server-side evaluation using FedAvg with VGG-16 architecture.

VGG-16			
Round	Accuracy (%)	F1-score	Loss
1	55.38	0.70	0.68
2	78.46	0.66	0.87
3	78.46	0.53	0.87
4	78.46	0.50	0.87
5	78.46	0.49	0.87

The harmonic mean, F1-score, starts at 0.70 and keeps decreasing until the last round achieving 0.49. The F1-score is an evaluation metric being used in conjunction with accuracy once it is based on the performance of all classes rather than overall performance, as it is accomplished by accuracy. The F1-score metric is low because it implies that the precision (exactness of correct predictions) is the main factor in the overall accuracy measure - almost all of the instances in the dataset are identified as positive, thus the accuracy among predicted positives is nearly the same as the accuracy among all instances, which is well known in datasets with imbalanced classes.

Table 5.3: Server-side evaluation using FedAvg with Resnet-18 architecture.

Resnet-18			
Round	Accuracy (%)	F1-score	Loss
1	21.53	0.0	1.01
2	78.46	0.87	0.54
3	78.46	0.87	0.41
4	96.92	0.98	0.06
5	98.46	0.99	0.04

The Resnet-18 architecture, shown in Table 5.3, has a great improvement on the accuracy metric starting on round 2. The first round has only 21.53% accuracy, 0 (zero) as F1-score, and 1.01 as a loss. Overall, the first round gathered an average of all client's models and they performed actually harmful. Although, starting in the second round, the global model metrics started improving, showing 78.46% of accuracy for rounds 2 and 3, and F1-score and loss really similar, with 0.87 for F1-score and loss between 0.41 and 0.54. This model has presented an interesting result because, at rounds 4 and 5, it is possible to see that the accuracy is really high above 95%, F1-score above 0.97, and the loss is lower than 0.07. It is assumed that the model classified mostly chest X-ray images correctly.

Analyzing the results of Resnet-50 presented in Table 5.4, the server-side evaluation has the same accuracy of 78.46% and a loss of 0.87 in all five rounds. The F1-score starts at round 1 (one) with 0.57 and decreases throughout other rounds, reaching the minimum of 0.33 in the fifth round. Even though this experiment had the same accuracy and loss for the first four rounds, the F1-score shows that the model has been classifying better in each round, once the new updated global model is returned to get used in the next round of federated learning.

Table 5.4: Server-side evaluation using FedAvg with Resnet-50 architecture.

Resnet-50			
Round	Accuracy (%)	Loss	F1-score
1	78.46	0.87	0.57
2	78.46	0.87	0.48
3	78.46	0.87	0.44
4	78.46	0.87	0.49
5	78.46	0.87	0.33

5.6 Discussion and Conclusion

Comparing the proposed architectures after implementing the FL solution, it is possible to observe that the Resnet-18 architecture yields the best results, followed by Resnet-50 and VGG-16. Resnet-18 and Resnet-50 utilize skip connections, which enable the network to avoid the "vanishing gradient" problem and enhance the information flow between layers. In contrast, VGG-16 does not employ skip connections.

Resnet-18 features only one type of shortcut connection, whereas Resnet-50 incorporates multiple types of shortcut connections. These connections enable the network to learn more intricate patterns by combining information from various layers.

Currently, VGG-16 is primarily used for image classification purposes, while Resnet-18 and Resnet-50 are employed for a wider range of tasks, such as object detection, segmentation, and image captioning. Resnet-50, particularly, is well-suited for tasks requiring high accuracy, such as medical image analysis.

The number of training epochs applied in this study is relatively small compared to existing machine learning models in the literature. Despite this, it is evident that the global model achieved superior performance after finishing training compared to the initial training. Additionally, comparing the results of the training models before and with FL, we can notice that the training rounds have concluded with better results in the FL. This demonstrates that applying only five training rounds in the FL environment allows the models to learn swiftly without significant improvements over subsequent epochs. This results in a qualified model for utilization within the healthcare community.

Overall, applying the FedAvg strategy on the server side effectively improves the robustness and performance of the global model for pneumonia detection with each round of training. However, the model may exhibit overfitting as the server achieves remarkable results in just five rounds.

Hence, all three models demonstrated improved results after the final round of training the FL model compared to their initial performances before FL implementation.

By distributing the model training across multiple devices, FL preserves the data privacy of each user while enhancing model accuracy. These promising results reinforce the potential benefits of federated learning, particularly in scenarios involving distributed or sensitive data. By continuously improving and advancing federated learning, we can contribute to developing smarter

and more reliable machine learning models while ensuring privacy and security remain of utmost importance.

In conclusion, with its strong performance, the FL-trained model has proven to be a practical and innovative approach to overcoming some challenges associated with traditional machine learning techniques. However, the limited computational resources, specifically the GPUs provided by Google, posed significant performance challenges in this research work.

Chapter 6

Conclusion and Future work

This chapter presents the findings of the proposed research work, including the obtained knowledge and insights. Section 6.1 contains the final conclusion of the presented research work, and Section 6.2 proposes future work.

6.1 Final Conclusion

Technological advancements have greatly enhanced the medical field's ability to detect and predict illnesses by combining artificial intelligence with human expertise. Federated learning is an emerging area in machine learning that enables multi-institutional collaborations in healthcare, improving the accuracy of disease diagnosis. FL also addresses essential considerations such as privacy, large-scale machine learning, and distributed optimization.

This study proposes a FL system that utilizes the Flower framework and effectively solves binary classification problems. Specifically, the system can accurately classify new images as either pneumonia or normal using a supervised learning algorithm. This system can also be applied to collaborative learning for other binary classification tasks, such as cancer detection or organ anomalies.

While existing literature has focused on FL applications in diagnosing brain tumors and lung abnormalities, there is a clear need to explore FL's potential in pneumonia detection, as it is a prevalent disease. To support the diagnosis of pneumonia, this study develops, analyzes, and designs a FL solution using a chest X-ray dataset. The methodology involves selecting appropriate algorithms and architectures for FL, evaluating the solution's ability to prevent data leakage, and assessing performance using metrics such as accuracy, loss, and F1-score. To promote project management, macro and micro timetables were also created to supply a visualization of task realization and time based on its complexity.

Developing a computer-aided diagnostics system using FL requires significant computational complexity, especially when simulating multiple organizations connected to a single machine. However, this approach optimally utilizes computational resources to simulate multiple institutions.

Health data, including chest X-rays, contain sensitive personal information and are subject to regulatory requirements such as the General Data Protection Regulation (GDPR). Organizations processing such data must implement appropriate measures to ensure confidentiality, integrity, and availability and respect individuals' rights. Protecting sensitive personal data is crucial in maintaining trust and confidence in healthcare providers. Individuals must consent to process chest X-rays, and the data must be processed only for specific and legitimate purposes, ensuring accuracy and up-to-dateness. GDPR also grants individuals the right to access, rectify, and erasure of personal health information.

FL has gained significant interest as a privacy-preserving collaborative learning system that complies with GDPR. FL can support disease diagnosis in the medical domain, particularly in resource-constrained environments or cases of high misdiagnosis rates. Also, FL is a powerful tool for diagnosing rare diseases, where a single establishment may lack the diversity of instances necessary to train a competent model.

The proposed FL solution provides three model results, representing different doctors' perspectives, to generate a shared diagnosis and treatment decision. This FL approach is a solid strategy that should be further refined in the future. The best-performing model, using the Resnet-18 architecture, achieved an accuracy of 98.46% and an F1-score of 0.99. However, it is possible that the datasets used for training, validation, and testing contained similar images, indicating model overfitting despite the application of data augmentation techniques. This possibility is supported by the successful results obtained from the server side, which only required five rounds to obtain the aggregated average model from the clients.

6.2 Future work

In the literature search presented in Section 2.2, only 1 (one) article studied histology images in FL. This type of medical image heavily impacts clinical decision-making. Consequently, it is proposed to test the FL system on histology images to increase the number of researchers in this field and determine if this system performs better using this type of image or other ones.

Additionally, it is suggested to conduct new experiments using the presented pneumonia dataset, e.g., using other pre-trained CNNs and simulating different numbers of edges in the FL system. Another suggestion is to run experiments with different batch sizes in the data loaders and select the one that yields the best performance on the validation data, such as 32 or 64. This will help understand the effect of adjusting these parameters in the collaborative learning system.

Zech et al. (65) have stated that models trained with a single institutional dataset are more prone to data bias. Having that in mind, it is proposed to use the FL system with other datasets to avoid bias and obtain more accurate results.

Moreover, depending on the structure of the new dataset, data augmentation, and pre-processing techniques may be necessary. Therefore, deeper analysis and adjustment of procedures are required. Other strategies on the server side can be implemented using Flower in order to understand

which strategy solves the problem better, such as Adaptive Federated Optimization using Adagrad (FedAdagrad),

Furthermore, evaluating the models on the client side can prevent "weak" models from being incorporated into the system, thus avoiding negative contributions to the FL.

Simulating an FL environment with multiple physical machines can address computational limitations. In this way, each organization can share its computational resources, facilitating collaborative learning and simulating real-case FL scenarios.

For better metric results, it is concluded that improvements should be applied in future experiments. This includes selecting the best client nodes to participate in the next training round and ensuring that the best models are sent to the central model.

References

[1] Ng, D., Lan, X., Yao, M.M.-S., Chan, W.P., Feng, M.: Federated learn- ing: a collaborative effort to achieve better medical imaging models for individual sites that have small labelled datasets. Quantitative Imaging in Medicine and Surgery 11(2), 852 (2021). DOI: https://doi.org/10.21037/QIMS-20-595

[2] Mouhni, N., Elkalay, A., Chakraoui, M., Abdali, A., Ammoumou, A., Amalou, I.: Federated learning for medical imaging: An updated state of the art. Journal homepage: http://iieta.org/journals/isi 27(1), 143–150 (2022). DOI: https://doi.org/10.18280/isi.270117

[3] Gomathisankaran, M., Yuan, X., Kamongi, P.: Ensure privacy and secu- rity in the process of medical image analysis. In: 2013 IEEE International Conference on Granular Computing (GrC), pp. 120–125 (2013). IEEE. DOI: https://doi.org/10.1109/GrC.2013.6740392

[4] Li, W., Milletarì, F., Xu, D., Rieke, N., Hancox, J., Zhu, W., Baust, M., Cheng, Y., Ourselin, S., Cardoso, M.J., et al.: Privacy-preserving federated brain tumour segmentation. In: Machine Learning in Medical Imaging: 10th International Workshop, MLMI 2019, Held in Conjunction with MICCAI 2019, Shenzhen, China, October 13, 2019, Proceedings 10, pp. 133–141 (2019). Springer. DOI: https://doi.org/10.1007/978-3-030-32692-0_16

[5] Gkoulalas-Divanis, A., Loukides, G.: Introduction to medical data privacy. Medical data privacy handbook, 1–14 (2015). DOI: https://doi.org/10.1007/978-3-319-23633-9_1

[6] McMahan, B., Moore, E., Ramage, D., Hampson, S., y Arcas, B.A.: Communication-efficient learning of deep networks from decentralized data. In: Artificial Intelligence and Statistics, pp. 1273–1282 (2017). PMLR. DOI: https://doi.org/10.48550/arXiv.1602.05629

[7] Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., ... & Moher, D. (2021). The PRISMA 2020 state- ment: an updated guideline for reporting systematic reviews. International journal of surgery, 88, 105906. DOI: https://doi.org/10.1136/bmj.n71

[8] Mahlool, D.H., Abed, M.H.: Distributed brain tumor diagnosis using a federated learning environment. Bulletin of Electrical Engineering and Informatics 11(6), 3313–3321 (2022). DOI: https://doi.org/10.11591/eei.v11i6.4131

[9] Linardos, A., Kushibar, K., Walsh, S., Gkontra, P., Lekadir, K.: Federated learning for multi-center imaging diagnostics: a simulation study in cardiovascular disease. Scientific Reports 12(1), 3551 (2022). DOI: https://doi.org/10.1038/s41598-022-07186-4

[10] Liu, J., Liang, X., Yang, R., Luo, Y., Lu, H., Li, L., Zhang, S., Yang, S.: Federated learning-based vertebral body segmentation. Engineering Applications of Artificial Intelligence 116, 105451 (2022). DOI: https://doi.org/10.1016/j.engappai.2022.105451

[11] Wu, Y., Zeng, D., Wang, Z., Shi, Y., Hu, J.: Distributed contrastive learning for medical image segmentation. Medical Image Analysis 81, 102564 (2022). DOI: https://doi.org/10.1016/j.media.2022.102564

[12] Zhang, M., Qu, L., Singh, P., Kalpathy-Cramer, J., Rubin, D.L.: SplitAVG: A heterogeneity-aware federated deep learning method for medical imaging. IEEE Journal of Biomedical and Health Informatics 26(9), 4635–4644 (2022). DOI: https://doi.org/10.48550/arXiv.2107.02375

[13] Bercea, C.I., Wiestler, B., Rueckert, D., Albarqouni, S.: Federated disen- tangled representation learning for unsupervised brain anomaly detection. Nature Machine Intelligence 4(8), 685–695 (2022). DOI: https://doi.org/10.1038/s42256-022-00515-2

[14] Huang, Z.-A., Hu, Y., Liu, R., Xue, X., Zhu, Z., Song, L., Tan, K.C.: Fed- erated multi-task learning for joint diagnosis of multiple mental disorders on MRI scans. IEEE Transactions on Biomedical Engineering (2022). DOI: 10.1109/TBME.2022.3210940

[15] Islam, M., Reza, M.T., Kaosar, M., Parvez, M.Z.: Effectiveness of federated learning and CNN ensemble architectures for identifying brain tumors using MRI images. Neural Processing Letters, 1–31 (2022). DOI: https://doi.org/10.1007/s11063-022-11014-1

[16] Peng, L., Wang, N., Dvornek, N., Zhu, X., Li, X.: Fedni: Federated graph learning with network inpainting for population-based disease prediction. IEEE Transactions on Medical Imaging (2022). DOI: https://doi.org/10.1109/TMI.2022.3188728

[17] Li, X., Gu, Y., Dvornek, N., Staib, L.H., Ventola, P., Duncan, J.S.: Multi- site fMRI analysis using privacy-preserving federated learning and domain adaptation: ABIDE Results. Medical Image Analysis 65, 101765 (2020). DOI: https://doi.org/10.48550/arXiv.2001.05647

[18] Florescu LM, Streba CT, Şerbănescu MS, Mămuleanu M, Florescu DN, Teică RV, Nica RE, Gheonea IA.: Federated learning approach with pre-trained deep learning models for COVID-19 detection from unsegmented CT images. Life 12(7), 958 (2022). DOI: https://doi.org/10.3390/life12070958

[19] Dou, Q., So, T.Y., Jiang, M., Liu, Q., Vardhanabhuti, V., Kaissis, G., Li, Z., Si, W., Lee, H.H., Yu, K., et al.: Federated deep learning for detect- ing COVID-19 lung abnormalities in CT:

a privacy-preserving multinational validation study. NPJ digital medicine 4(1), 60 (2021). DOI: https://doi.org/10.1038/s41746-021-00431-6

[20] Zhang, W., Zhou, T., Lu, Q., Wang, X., Zhu, C., Sun, H., Wang, Z., Lo, S.K., Wang, F.-Y.: Dynamic-fusion-based federated learning for COVID-19 detection. IEEE Internet of Things Journal 8(21), 15884–15891 (2021). DOI: https://doi.org/10.1109/JIOT.2021.3056185

[21] Knolle, M., Kaissis, G., Jungmann, F., Ziegelmayer, S., Sasse, D., Makowski, M., Rueckert, D., Braren, R.: Efficient, high-performance semantic segmentation using multi-scale feature extraction. Plos one 16(8), 0255397 (2021). DOI: https://doi.org/10.1371/journal.pone.0255397

[22] Yang, D., Xu, Z., Li, W., Myronenko, A., Roth, H.R., Harmon, S., Xu, S., Turkbey, B., Turkbey, E., Wang, X., et al.: Federated semi-supervised learning for covid region segmentation in chest CT using multi-national data from China, Italy, Japan. Medical image analysis 70, 101992 (2021). DOI: https://doi.org/10.48550/arXiv.2011.11750

[23] Kumar, A., Purohit, V., Bharti, V., Singh, R., Singh, S.K.: MediSecFed: private and secure medical image classification in the presence of malicious clients. IEEE Transactions on Industrial Informatics 18(8), 5648–5657 (2021). DOI: https://doi.org/10.1109/TII.2021.3138919

[24] Ziegler, J., Pfitzner, B., Schulz, H., Saalbach, A., Arnrich, B.: Defend- ing against reconstruction attacks through differentially private federated learning for classification of heterogeneous Chest X-ray data. Sensors 22(14), 5195 (2022). DOI: https://doi.org/10.3390/s22145195

[25] Nguyen, D.C., Ding, M., Pathirana, P.N., Seneviratne, A., Zomaya, A.Y.: Federated learning for COVID-19 detection with generative adversarial net- works in edge cloud computing. IEEE Internet of Things Journal 9(12), 10257–10271 (2021). DOI: https://doi.org/10.1109/JIOT.2021.3120998

[26] Li, Z., Xu, X., Cao, X., Liu, W., Zhang, Y., Chen, D., Dai, H.: Integrated CNN and Federated Learning for COVID-19 Detection on Chest X-ray images. IEEE/ACM Transactions on Computational Biology and Bioinformatics (2022). DOI: https://doi.org/10.1109/TCBB.2022.3184319

[27] Kaissis, G., Ziller, A., Passerat-Palmbach, J., Ryffel, T., Usynin, D., Trask, A., Lima Jr, I., Mancuso, J., Jungmann, F., Steinborn, M.-M., et al.: End-to-end privacy preserving deep learning on multi-institutional medical imaging. Nature Machine Intelligence 3(6), 473–484 (2021). DOI: https://doi.org/10.1038/s42256-021-00337-8

[28] Feki, I., Ammar, S., Kessentini, Y., Muhammad, K.: Federated learning for COVID-19 screening from Chest X-ray images. Applied Soft Computing 106, 107330 (2021). DOI: https://doi.org/10.1016/j.asoc.2021.107330

[29] Lu, M. Y., Chen, R. J., Kong, D., Lipkova, J., Singh, R., Williamson, D. F., ... & Mahmood, F. (2022). Federated learning for computational pathology on gigapixel whole slide images. Medical image analysis, 76, 102298. DOI: https://doi.org/10.1016/j.media.2021.102298

[30] OBiBa: Open source software for epidemiology. (2020). Access at http://www.obiba.org/. Last access in 31-Jan-2023

[31] DataSHIELD: A software solution for secure bioscience collaboration. (n.d). Access at https://www.datashield.org/. Last access in 31-Jan-2023

[32] Carter, K.W., Francis, R.W., Carter, K., Francis, R., Bresnahan, M., Gissler, M., Grønborg, T., Gross, R., Gunnes, N., Hammond, G., et al.:Vipar: a software platform for the virtual pooling and analysis of research data. International journal of epidemiology 45(2), 408–416 (2016). DOI: https://doi.org/10.1093/ije/dyv193

[33] Burton, P. R., Murtagh, M. J., Boyd, A., Williams, J. B., Dove, E. S., Wallace, S. E., ... & Knoppers, B. M. (2015). Data Safe Havens in health research and healthcare. Bioinformatics, 31(20), 3241-3248. DOI: https://doi.org/10.1093/bioinformatics/btv279

[34] Dock, E. (2021). What are the Stages of Lobar Pneumonia? Healthline. https://www.healthline.com/health/pneumonia/pneumonia-stages. Last access in 31-May-2023

[35] Gereige, R. S., & Laufer, P. M. (2013). Pneumonia. Pediatrics in review, 34(10), 438-456. DOI: https://doi.org/10.1542/pir.34-10-438

[36] Brown, J. S. (2012). Community-acquired pneumonia. Clinical Medicine, 12(6), 538. DOI: https://doi.org/10.7861/clinmedicine.12-6-538

[37] UNICEF, D. (2019). Pneumonia via https://data.unicef.org/topic/child-health/pneumonia/. Last access in 29-Apr-2023

[38] WHO's Global Health Estimates. (2020). The top 10 causes of death. Retrieved from https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death. Last access in 01-May-2023

[39] World Health Organization. (2022). Pneumonia in children. Retrieved from https://www.who.int/news-room/fact-sheets/detail/pneumonia. Last access in 01-May-2023

[40] World Health Organization. (2001). Standardization of interpretation of chest radiographs for the diagnosis of pneumonia in children (No. WHO/V&B/01.35). World Health Organization.

[41] Kallianos, K., Mongan, J., Antani, S., Henry, T., Taylor, A., Abuya, J.,& Kohli, M. (2019). How far have we come? Artificial intelligence for chest radiograph interpretation. Clinical radiology, 74(5), 338-345. DOI: https://doi.org/10.1016/j.crad.2018.12.015

[42] Liu, N., Wan, L., Zhang, Y., Zhou, T., Huo, H., & Fang, T. (2018). Exploiting convolutional neural networks with deeply local description for remote sensing image classification. IEEE access, 6, 11215-11228. DOI: https://doi.org/10.1109/ACCESS.2018.2798799

[43] Kermany, Daniel; Zhang, Kang; Goldbaum, Michael (2018), "Labeled Optical Coherence Tomography (OCT) and Chest X-Ray Images for Classification", Mendeley Data, V2. DOI: https://doi.org/10.17632/rscbjbr9sj.2

[44] Kermany, D. S., Goldbaum, M., Cai, W., Valentim, C. C., Liang, H., Baxter, S. L., ... & Zhang, K. (2018). Identifying medical diagnoses and treatable diseases by image-based deep learning. cell, 172(5), 1122-1131. DOI: https://doi.org/10.1016/j.cell.2018.02.010

[45] OŚhea, K., & Nash, R. (2015). An introduction to convolutional neural networks. arXiv preprint arXiv:1511.08458. DOI: https://doi.org/10.48550/arXiv.1511.08458

[46] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In 2009 IEEE conference on computer vision and pattern recognition (pp. 248–255). https://doi.org/10.1109/CVPR.2009.5206848

[47] Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. arXiv 2014, arXiv:1409.1556. DOI: https://doi.org/10.48550/arXiv.1409.1556

[48] Philipp, G., Song, D., Carbonell, J. G. (2018). Gradients explode-deep networks are shallow-resnet explained.

[49] Wilson, A.C.; Roelofs, R.; Stern, M.; Srebro, N.; Recht, B. The marginal value of adaptive gradient methods in machine learning. In Advances in Neural Information Processing Systems; Neural Information Processing Systems: CA, USA; 2017; pp. 4148–4158.

[50] Braungardt, Alex. FATE, Flower, PySyft & Co. — Federated Learning Frameworks in Python. (2023). Access at https://medium.com/elca-it/flower-pysyft-co-federated-learning-frameworks-in-python-b1a8eda68b0d. Last access in 18-May-2023

[51] NVIDIA FLARE — NVIDIA FLARE 2.2.1 documentation. (2022). Access at https://nvflare.readthedocs.io/en/2.2.1/index.html. Last access in 31-May-2023

[52] What is FATE. (n.d.). Access at https://fate.fedai.org/overview/. Last access in 31-May-2023

[53] Beutel, D. J., Topal, T., Mathur, A., Qiu, X., Parcollet, T., de Gusmão, P. P., & Lane, N. D. (2020). Flower: A friendly federated learning research framework. arXiv preprint arXiv:2007.14390. DOI: https://doi.org/10.48550/arXiv.2007.14390

[54] PySyft's documentation. (2021). Access at https://openmined.github.io/PySyft/index.html. Last access in 31-May-2023

[55] Wynter, L. (2020). IBM Federated Learning Community Edition: An Interactive Demonstration. IBM Research Publications. Access at https://research.ibm.com/publications/ibm-federated-learning-community-edition-an-interactive-demonstration. Last access in 31-May-2023

[56] OpenFL - Creative expression for desktop, mobile, web and console platforms. (2023.). Access at https://www.openfl.org/. Last access in 31-May-2023

[57] TensorFlow Federated. (n.d.). TensorFlow. Access at https://www.tensorflow.org/federated. Last access in 31-May-2023

[58] He, C., Li, S., So, J., Zeng, X., Zhang, M., Wang, H., ... & Avestimehr, S. (2020). Fedml: A research library and benchmark for federated machine learning. arXiv preprint arXiv:2007.13518. DOI: https://doi.org/10.48550/arXiv.2007.13518

[59] Flower Documentation. (2022). Access at https://flower.dev/docs/index.html. Last access in 18-May-2023

[60] Bonawitz, K., Ivanov, V., Kreuter, B., Marcedone, A., McMahan, H. B., Patel, S., Ramage, D., Segal, A., and Seth, K (2017, October). Practical secure aggregation for privacy-preserving machine learning. In proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security (pp. 1175-1191). DOI: https://doi.org/10.1145/3133956.3133982

[61] Bell, J. H., Bonawitz, K. A., Gascón, A., Lepoint, T., & Raykova, M. (2020, October). Secure single-server aggregation with (poly) logarithmic overhead. In Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security (pp. 1253-1269). DOI: https://doi.org/10.1145/3372297.3417885

[62] Bisong, E., & Bisong, E. (2019). Google colaboratory. Building machine learning and deep learning models on google cloud platform: a comprehensive guide for beginners, 59-64. DOI: https://doi.org/10.1007/978-1-4842-4470-8_7

[63] Huang, C., Huang, J., & Liu, X. (2022). Cross-Silo Federated Learning: Challenges and Opportunities. arXiv preprint arXiv:2206.12949. DOI: https://doi.org/10.48550/arXiv.2206.12949

[64] An Introduction to Federated Learning - Flower 1.5.0. (2022). Access at https://flower.dev/docs/tutorial/Flower-1-Intro-to-FL-PyTorch.html. Last access in 03-May-2023

[65] Zech, J. R., Badgeley, M. A., Liu, M., Costa, A. B., Titano, J. J., & Oermann, E. K. (2018). Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. PLoS medicine, 15(11), e1002683. DOI: https://doi.org/10.1371/journal.pmed.1002683