The Data Heterogeneity Issue Regarding COVID-19 Lung Imaging in Federated Learning

by

Fatimah Saeed Alhafiz

A thesis submitted for the requirements of the degree of Doctor of Philosophy in Computer Science

Faculty of Computing and Information Technology
King Abdulaziz University
Jeddah, Saudi Arabia
Dhu al-Qi'dah 1446 H - May 2025 G

The Data Heterogeneity Issue Regarding COVID-19 Lung Imaging in Federated Learning

by

Fatimah Saeed Alhafiz

A thesis submitted for the requirements of the degree of Doctor of Philosophy in Computer Science

Advisor

Prof. Kamal M. Jambi

Faculty of Computing and Information Technology
King Abdulaziz University
Jeddah, Saudi Arabia
Dhu al-Qi'dah 1446 H - May 2025 G

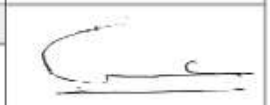

بسم الله الرحمن الرحيم

The Data Heterogeneity Issue Regarding COVID-19 Lung Imaging in Federated Learning

by

Fatimah Saeed Alhafiz

A thesis submitted for the requirements of the degree of Doctor of Philosophy in Computer Science

Name	Rank	Field	Signature
Advisor			
Kamal M. Jambi	Professor	Computer Sciences	
Internal Examiner			
Fathy Eassa	Professor	Computer Sciences	
External Examiner			
Jamel FEKI	Professor	Computing and Information Technology	

Faculty of Computing and Information Technology
King Abdulaziz University
Jeddah, Saudi Arabia
Dhu al-Qi'dah 1446 H - May 2025 G

Copyright

Dedication

This thesis is lovingly dedicated to my parents, whose unwavering support, encouragement, and sacrifices have been the foundation of my journey. To my husband for his patience, understanding and steadfast belief in me throughout every step of this process.

To my beloved children, Abdulhadi, Ali, Saeed, and the flower of our family, Reema, whose love and laughter have been a source of endless joy and motivation.

And to Professor Abdullah Basuhail, whose boundless wisdom, patience, and encouragement have guided me through my doctoral journey. From the earliest conception of research ideas to the final stages of manuscript preparation, his mentorship has been a constant source of inspiration and strength. Though he retired before my defense presentation, his influence endures on every page of this work. I am forever grateful for his dedication to my growth as a scholar and for his profound impact on my professional and academic life.

Each of you has played a unique and invaluable role in this achievement and for that I am forever grateful.

Acknowledgments

First and foremost, I am profoundly grateful to Allah for granting me the strength, perseverance, and wisdom to complete this thesis.

I would like to express my deepest gratitude to my primary supervisor, Professor Abdullah Basuhail, for his invaluable guidance, encouragement, and unwavering support throughout my doctoral journey. His expertise, patience, and dedication to my academic and personal growth have been instrumental in shaping this work.

I am also deeply thankful to Professor Kamal Jambi for graciously accepting the role of supervisor in the final stages of this research. His thoughtful feedback, constructive criticism, and confidence in my abilities have been essential in bringing this thesis to its successful conclusion.

My heartfelt thanks go to my parents, whose love, prayers, and steadfast belief in me have been my constant source of strength. Their sacrifices and encouragement have made every milestone possible.

To my husband, thank you for your support—your understanding, patience, and encouragement have carried me through every challenge and triumph.

To my beloved children—Abdulhadi, Ali, Saeed, and Reema—your laughter, love,

and innocence have reminded me of life's true joys and kept me balanced throughout this journey.

I am also grateful to my brothers, sisters, colleagues, and friends for their insights, camaraderie, and moral support, which have enriched my experience in countless ways.

Finally, I extend my appreciation to King Abdulaziz University, my department, and all members of the faculty and staff who provided the resources and encouragement necessary for this research.

To you all, thank you for being an integral part of this accomplishment.

Abstract

Federated Learning (FL) has emerged as a transformative paradigm for privacy-preserving collaborative machine learning, particularly well-suited for domains such as healthcare, where data sensitivity and institutional silos pose significant barriers to centralized model training. This thesis investigates the application of FL for COVID-19 diagnosis using lung imaging, focusing on how non-independent and non-identically distributed (non-IID) data across institutions affects model performance, convergence, and reliability. A central contribution of this work is the systematic characterization of medical data heterogeneity that arises in real-world federated settings, especially under the constraints and variability introduced during the COVID-19 pandemic. Six distinct types of data heterogeneity are defined, mathematically modeled, and evaluated within a simulated FL framework. These include variations in label distribution, sample quantity, imaging modality, and demographic attributes. Experimental results demonstrate that extreme label skew, where a single institution holds data with a unique label, leads to the most significant degradation in global model performance. In contrast, other forms of skew, such as quantity skew and label distribution skew, allow the federated model to retain a higher degree of generalizability and maintain acceptable diagnostic performance. This thesis offers a detailed exploration of the comparison between global generalization and local

personalization in FL under non-IID conditions. It identifies critical challenges and provides insights into designing more valuable federated frameworks for medical imaging. By evaluating FL in diverse heterogeneity settings, this work contributes to researchers developing more robust, generalizable, and clinically applicable diagnostic models in healthcare AI.

Key Word: *Federated learning, Data heterogeneity, Non-IID, Generalization metric, Personalization metric*

Contents

List of Tables

List of Figures

List of Abbreviations

AI	Artificial Intelligence
API	Application Programming Interface
CAD	Computer-Aided Diagnosis
CNN	Convolutional Neural Network
CWT	Cycling Weight Transfer
DL	Deep Learning
DP	Differential Privacy
ELS	Extreme Label Skew
FedAvg	Federated Average
FL	Federated Learning
fMRI	functional Magnetic Resonance Imaging
GAN	Generative Adversarial Network

GDPR General Data Protection Regulation

GPU Graphics Processing Unit

HE Homomorphic Encryption

HFL Horizontal Federated Learning

HIPAA Health Insurance Portability and Accountability Act

IID Independent and non-Identically Distributed

ML Machine Learning

MRI Magnetic Resonance Imaging

NLP Natural Language Processing

Non-IID Non-Independent and non-Identically Distributed

P2P Peer-to-Peer

SMOTE Synthetic Minority Oversampling Technique

SMPC Secure Multi-Party Computation

SWT Single Weight Transfer

VFL Vertical Federated Learning

WHO World Health Organization

Chapter 1

Introduction

1.1 Introduction

COVID-19 is a worldwide pandemic that was first reported in December 2019 and has continuing effects on people all over the world, with the virus transforming into multiple strains [1]. The number of infected cases has increased exponentially, and the number of reported deaths reached more than 7 million in 2025, according to the World Health Organization (WHO) [2]. The pulmonary system serves as the principal target of infection for severe acute respiratory syndrome COVID-19, with profound hypoxemia identified as the leading cause of mortality in the most severe instances.

The manifestation of COVID-19 exhibits extremely heterogeneous regarding its severity, clinical presentation, and, significantly, its worldwide prevalence[3]. While a majority of individuals diagnosed with the acute infection ultimately achieve

recovery[4], a substantial number persist in experiencing long-term complications that impact multiple organ systems, including the lung[5]. The precise pathobiological mechanisms underlying the pulmonary vascular complications associated with COVID-19, particularly in both the acute and chronic phases of the disease, remain inadequately elucidated and are not fully comprehended within the current scientific literature [3]. The lack of reliable information about the behavior of COVID-19 (e.g., how it is spread, variants of individual symptoms, and unstable responses to treatment strategies) has created a need for collaboration to collect more data about the virus.

Imaging equipment is one way to understand the virus behaviors and diagnosing the lung damage in the initial phase of the infection provided justifiable relations over different populations. Also, screening for uncertain, risky cases or to reduce the time and complexity of manual reverse transcription-polymerase chain reaction RT-PCR tests reduces the misdiagnosis rate of manual RT-PCR tests by 30% [6]. However, the continuous increase in the number of COVID-19 infections, many medical images have been produced. The increase in these images leaves hospitals with two challenges. First, interpreting medical images requires radiology experts for manual labeling, segmenting, and annotating. The limited number of radiologists in hospitals creates challenges for accumulating these data. There is a need to assist radiologists working alone to develop individual deep learning (DL) models locally as an embedded software in computer-aided diagnosis CAD systems to interpret these images promptly for the diagnosis, treatment, and prognosis of COVID-19 from lung testing. Deep learning is a subset of machine learning that uses multi-layered artificial neural networks to model complex patterns in data. It

helps machines recognize patterns of different diseases from images, as shown in Figure 1.1 (a) [5].

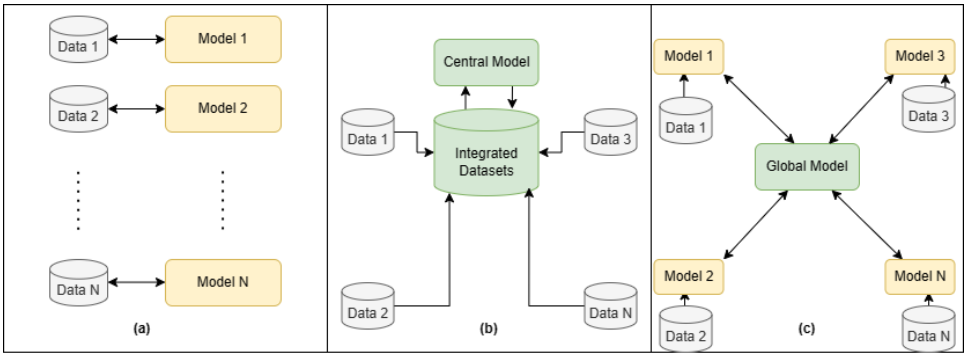


Figure 1.1: Training techniques for distributed data: (a) individual training technique, (b) centralizing technique, and (c) federated learning technique.

The second challenge is the reported results of an individual training model of data from single sources, which lack diversity of infected cases and produce bias due to the locality of the patient population. Hence, the generalization measure of that model will be inefficient for testing new data from out-of-sample [6]. The alternative solution is collecting and standardizing medical images at a single point may be effective for improving diversity and generalization problems as shown in Figure 1.1 (b). However, this requires large amounts of storage capacity, computational resources, communication bandwidth and security management for accessing and retrieving data at the central data point [7]. Furthermore, data governance policies preserve patient privacy and prevent the sharing of medical data that may reveal sensitive information about patients, even with anonymization technology, which limits the efforts of medical and health institutions to benefit from DL systems [8]. FL is a paradigm that allows for distributed points to train their data locally without the need to share actual data as shown in Figure 1.1 (c). It collects only the weights

(i.e., learning parameters/gradients) from different parties and computes the global model from distributed training. FL techniques are successful for medical image applications, such as disease diagnosis [9], segmentation [6] and treatments [10], and they result in higher accuracy than centralized learning [11]. Distributed processing of sensitive information has motivated researchers to utilize FL to overcome the lack of information about COVID-19 [12].

FL's capacity to generate learning models from large, diverse datasets makes it particularly effective in mitigating the widespread distribution challenges pertaining to COVID-19 imaging data. For instance, FL systems have successfully identified potential COVID-19 cases before the first reported patient date [13]. However, despite its effectiveness in analyzing distributed data, FL faces significant challenges related to data heterogeneity, which can negatively affect overall model performance. Diverse imaging equipment and a lack of uniformity in medical imaging acquisition have led to disorganized and scattered data collections worldwide [8]. This variability further complicates the problem of virus heterogeneity manifestation, making it challenging to establish consistent diagnostic and analytical models.

In FL, data heterogeneity is referred to as the non-independent and non-identically distributed (non-IID) issue, where variations in data distributions across institutions hinder model convergence and reduce generalization performance. Several studies have evaluated non-IID data by partitioning the datasets in different non-IID settings. While some studies report stable global model performance [14], Nguyen et al. [15] observed that performance can degrade by up to 50% due to non-IID data. These discrepancies often result from variations in data skewness across distributed datasets. Medical imaging datasets, in particular, exhibit unique characteristics, including

differences in patient demographics, imaging modalities, acquisition techniques, and storage formats. Proper consideration of these factors during FL partitioning is essential to accurately interpret results under non-IID conditions.

This research focuses on the challenges associated with data heterogeneity in FL and explores potential opportunities for leveraging FL on available COVID-19 lung imaging datasets for diagnosis and classification applications. Additionally, it defines real-world data distributions for COVID-19 lung imaging and analyzes their impact under different metrics, highlighting the obstacles posed by data heterogeneity in FL.

The study adopts a horizontal FL architecture, which enhances control over data privacy, security, and participant evaluation. Within this framework, two key metrics assess model performance: the generalization metric evaluates the global model's ability to handle unseen external data [16], while the personalization metric measures how well the updated model aligns with the specific characteristics of local datasets. While some studies prioritize improving generalization at the central node [17], others focus on enhancing personalization [18]. We describe the open issues related to understanding the impact of non-IID data in FL, as achieving an optimal balance between generalization and personalization is crucial for robust model performance.

1.2 Research Problem

The effectiveness of Federated Learning (FL) in medical imaging, particularly for COVID-19 lung diagnosis, is significantly impacted by data heterogeneity, com-

monly referred to as the non-independent and non-identically distributed (non-IID) issue. Variability in imaging equipment, acquisition protocols, patient demographics, and data storage formats leads to inconsistencies across distributed datasets, affecting model convergence and generalization. While FL enables collaborative model training without sharing sensitive patient data, the presence of non-IID data introduces biases that degrade performance, potentially reducing diagnostic accuracy and limiting the applicability of FL in real-world medical settings.

Despite existing research efforts, there is no comprehensive understanding of how different types of non-IID data impact FL performance, nor is there a standardized approach to mitigating these effects. Studies have reported conflicting results, with some demonstrating stable global model performance, while others indicate performance degradation due to data heterogeneity. These discrepancies arise from variations in data distribution skewness, making it challenging to generalize FL models across diverse medical institutions.

Therefore, this study aims to systematically analyze the impact of non-IID data on FL performance in COVID-19 lung imaging by defining real-world data distributions and evaluating their effects under various types of skewness. Additionally, it explores potential strategies for balancing global generalization and local personalization to enhance FL model robustness. Addressing these challenges is crucial for improving the reliability of FL in medical imaging applications and ensuring its effectiveness in collaborative healthcare environments. Understanding the implications of these findings could lead to more effective deployment of federated learning frameworks, ultimately advancing diagnostic accuracy and patient outcomes in a variety of clinical settings.

1.3 Research Motivation

Accurate diagnosis of COVID-19 remains a challenge due to the virus's dynamic nature, which manifests heterogeneously in lung imaging across diverse patient populations. Variations in imaging equipment, acquisition protocols, and demographic factors further complicate the development of reliable diagnostic models. FL has emerged as a promising approach to leverage DL to extract global insights from distributed and privacy-sensitive medical datasets. Unlike traditional centralized learning, FL allows institutions to collaboratively train models without sharing raw patient data, addressing critical privacy concerns in medical AI applications. However, the heterogeneity of medical imaging data, poses a significant challenge to FL model convergence and generalization.

This research is motivated by the need to:

- Evaluate the impact of realistic data heterogeneity on FL performance by simulating different non-IID scenarios in COVID-19 lung imaging.
- Assess FL's ability to balance global generalization and local adaptation to ensure both diagnostic accuracy and site-specific relevance.
- Visualize and analyze key performance metrics to provide a detailed assessment of how various non-IID distributions affect model convergence and predictive accuracy.

By addressing these critical aspects, this study aims to advance the understanding of FL's limitations and optimization strategies in medical imaging, paving the way for more reliable and privacy-preserving AI-driven healthcare solutions.

1.4 Research Questions

This research aim to answer the following questions:

- What are the types of non-IID data distributions in HFL have the effect on FL model convergence and generalization?
- How does data heterogeneity (non-IID distribution) impact the performance of FL models in COVID-19 lung imaging?
- Which skew type have the most negative effects of non-IID data in medical imaging applications?
- Which evaluation metrics best capture the performance of FL models under different non-IID conditions in real-world medical imaging datasets?
- What are the key technical and computational challenges in implementing FL for heterogeneous COVID-19 imaging data?

These questions will guide the study in analyzing the impact of non-IID data in FL while exploring potential solutions to improve model robustness and clinical applicability.

1.5 Research Objectives

The primary objective of this study is to investigate the impact of data heterogeneity on FL performance in COVID-19 lung imaging and explore strategies to enhance

model robustness. The specific objectives are as follows:

1. **Assess the Applicability of FL in COVID-19 Lung Imaging**
    - Evaluate the feasibility and benefits of implementing FL for processing and training distributed COVID-19 lung imaging data.
    - Identify the imaging modalities, equipment variations, and institutional differences that influence FL model performance.
2. **Analyze FL System Design and Implementation Constraints**
    - Provide a comprehensive overview of FL frameworks, outlining key variables affecting implementation in medical imaging.
    - Examine the practical constraints and technical challenges of deploying FL in real-world healthcare settings.
3. **Investigate Data Heterogeneity and Its Impact on FL**
    - Define and characterize different types of data heterogeneity (non-IID) in FL, including their effects on model convergence, generalization, and personalization.
    - Formulate mathematical representations of various types of data skewness to quantify their impact on FL performance.
4. **Evaluate the Effects of Non-IID Data on FL Metrics**
    - Analyze how different types of data heterogeneity influence global model generalization and local model personalization.

- Conduct systematic experiments under multiple non-IID conditions to assess FL's adaptability and performance stability.

5. **Interpret FL Model Performance Across Distributed Institutions**
   - Investigate the implications of real-world data heterogeneity on FL outcomes in collaborative medical imaging.
   - Compare results across different institutions and evaluation metrics to ensure reliable and unbiased model assessments.
6. **Identify Research Gaps and Optimization Strategies for FL**
   - Highlight limitations in existing FL research related to medical imaging, particularly in handling non-IID data.
   - Propose potential solutions and optimization techniques to improve FL model robustness in heterogeneous data environments.
7. **Address Broader Challenges in FL for Medical Imaging**
   - Summarize key challenges in FL, such as data privacy, communication efficiency, and system heterogeneity.
   - Provide a holistic perspective on improving FL frameworks for broader medical applications beyond COVID-19.
8. **Explore Future Directions for FL in COVID-19 and Beyond**
   - Investigate the potential of FL in analyzing COVID-19 effects on lung and internal organ imaging.

- Suggest pathways for extending FL applications in medical diagnostics and advancing collaborative AI-driven healthcare solutions.

By achieving these objectives, this study aims to enhance the understanding of FL's potential in medical imaging, address critical challenges posed by data heterogeneity, and contribute to the development of more robust, generalizable, and personalized FL frameworks for healthcare applications.

1.6 Research Contributions

The significant contributions of this research are the following:

1. Find out the applicability and benefits of implementing FL to process and train the distributed data of COVID-19 lung using various imaging modalities and equipment, identifying the imaging types and modes available in distributed hospitals and medical institutions.

2. Provide an overview of the FL system and describe the variables of implementation and the practical constraints in the medical field. Also, investigates the progress made in developing FL frameworks to train medical images and identifies areas that require further effort to overcome the pitfalls of distributed learning performance.

3. Provide detailed descriptions of the data heterogeneity issue, identify the metrics that could be affected by that issue, and offer a mathematical description of

the problem for each type of skewness, along with valuable research directions to mitigate the impact of data heterogeneity.

4. Illustrate the effects of various types of skewness on both generalization and personalization metrics.

5. Interpret the results and highlight the implications of real-world data heterogeneity for FL model performance across all participating institutions and evaluation metrics.

6. Emphasize the other prevalent FL issues concisely to offer a comprehensive perspective for research on the FL environment.

7. Describe potential avenues for future research to explore how COVID-19 affects the lung and internal organs through imaging data, referencing ongoing studies that take into account relevant factors from a medical and radiology point of view.

By addressing these critical aspects, this study comprehensively analyzes data heterogeneity's impact on FL performance, offering valuable insights into model robustness under varying conditions. These findings will advance the development of FL systems in COVID-19 lung imaging and beyond.

1.7 Thesis Organization

The organizational structure of this thesis is as follows:

Chapter 2 presents literature review which provides a comprehensive overview of the federated learning (FL) paradigm and its applications in medical imaging, with a particular focus on COVID-19 datasets. It examines the challenges posed by data heterogeneity, categorizes different types of skewness, and identifies key sources of bias. Furthermore, it reviews existing approaches aimed at mitigating these issues and highlights current research trends and potential opportunities for leveraging FL to enhance the robustness of COVID-19 image analysis.

Chapter 3 presents a design analysis of the proposed FL algorithm, emphasizing the impact of non-independent and identically distributed (non-IID) data. Various types of data heterogeneity and their implications for model convergence and performance are explored.

Chapter 4 details the implementation framework, outlining the system architecture, computational setup, and key design choices made to facilitate federated learning in the medical imaging domain, and we describes the strategies used for partitioning data in a non-IID manner, as well as the evaluation metrics employed to assess the performance of the proposed approach.

Chapter 5 presents the experimental results, providing quantitative and qualitative analyses of the algorithm's effectiveness under different data distributions. We discusses the findings derived from the experimental results, offering insights into the observed trends, challenges, and the impact of non-IID data on FL performance.

Chapter 6 concludes the thesis by summarizing the key contributions and findings, highlighting their significance in the broader context of FL for medical imaging, and proposing directions for future research. We outlines limitations and provides

recommendations for improving the efficacy of FL in medical image analysis. It discusses potential avenues for future work, including algorithmic enhancements, data augmentation strategies, and domain-specific optimizations.

Chapter 2

Literature Review

Creating data silos in medical imaging has demonstrated significant potential for achieving high generalizability and generating valuable insights, as exemplified by the (ENIGMA) project [19]. By integrating medical imaging data from 70 distributed sites, ENIGMA identified factors associated with brain diseases that would have been unattainable by individual sites working in isolation. However, this process required substantial computational resources and a considerable investment in security to ensure the safe handling of data and the development of generalizable models. Furthermore, it posed challenges to privacy regulations, which are primarily designed to prevent the disclosure of sensitive patient information.

FL offers an innovative solution for researchers in the medical field to leverage large-scale datasets while avoiding the risks and costs associated with centralized data storage. This paradigm enables hospitals and clinical institutions to collaborate effectively without directly sharing patient data, addressing privacy concerns

and compliance with governance regulations. As a result, many institutions are increasingly adopting or developing FL software to facilitate such collaborations.

In this chapter, we provide an overview of FL and delve into the specific context of COVID-19 medical imaging datasets. This sets the stage for a discussion of the challenges posed by data heterogeneity, which are explored in greater detail in the subsequent subsections. Finally, we review related works at the end of this chapter, highlighting key studies that have addressed similar challenges.

2.1 FL Overview

FL is a technology for implementing DL over distributed data in multiple distinct sites without the need to share data from hospitals or medical institutions. To compute a global model, nodes upload only the local weights, also known as learning parameters or theta values, to a central point and then download them again [20]. FL takes one of two common design architectures: central aggregation or peer-to-peer architecture (P2P). In centralized aggregation, a central server coordinates the participant nodes, computes the global model, and communicates with the nodes, resulting in higher performance and flexibility, as illustrated in Figure 2.1. P2P proposes more a reliable and secure architecture to prevent a central point of failure in aggregation. It is a fully distributed architecture in which each P2P node coordinates itself and communicates with other nodes to compute the global model locally [21], as shown in Figure 2.2.

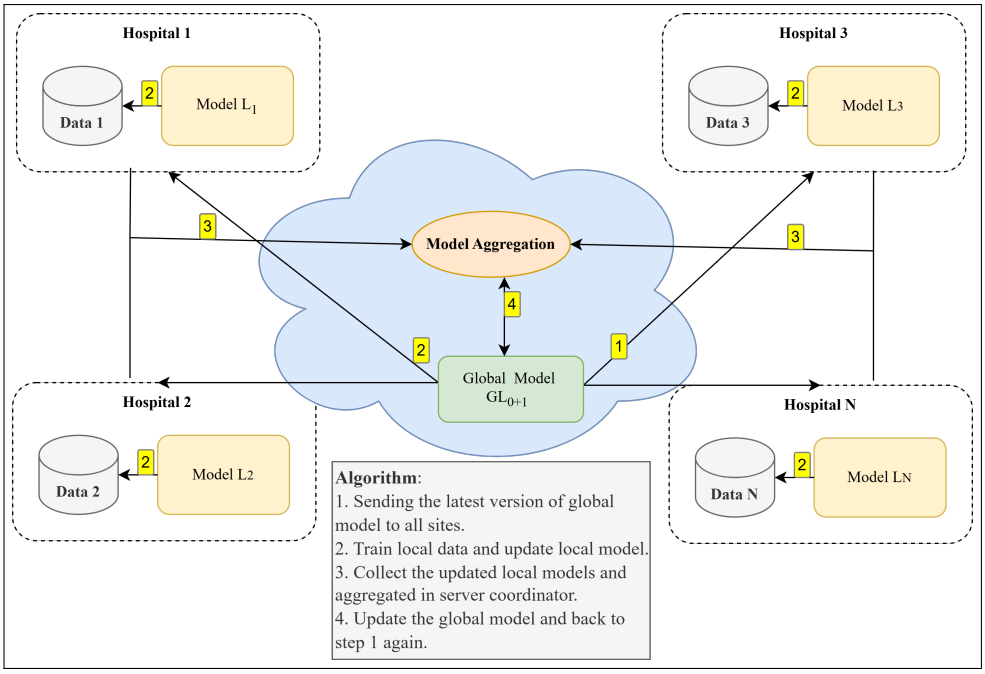


Figure 2.1: The algorithm of central FL architecture.

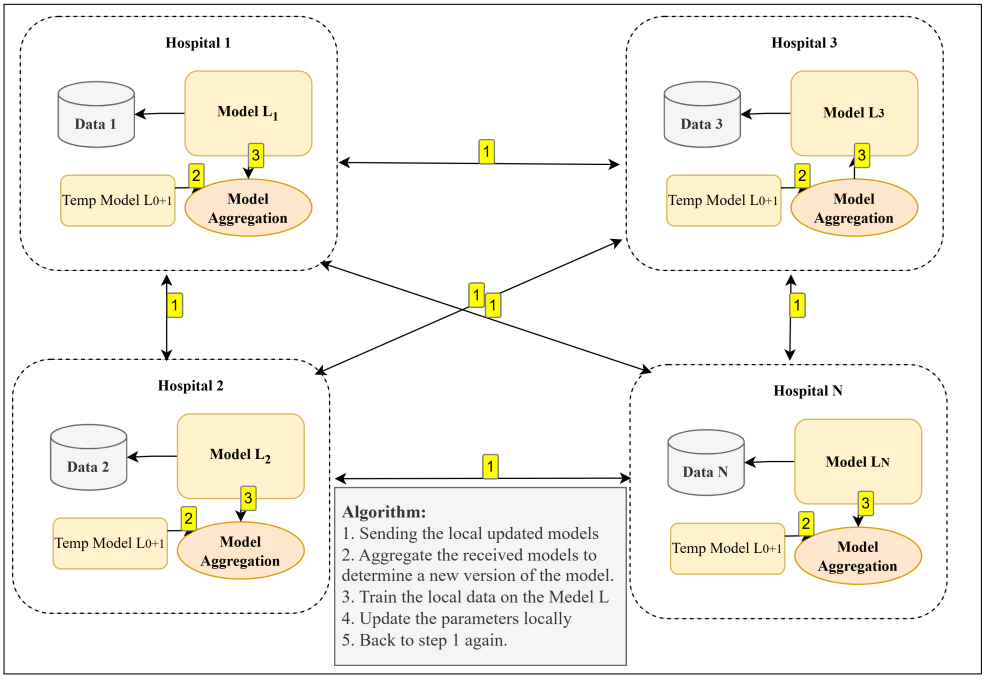


Figure 2.2: The algorithm of peer-to-peer architecture.

Depending on the federation's goals, FL categorized into horizontal FL (HFL) and vertical FL (VFL) methods in distributed data contexts. If the objective of the federation is to expand the data sample to provide a greater volume and variety of training datasets, this necessitates the use of HFL. HFL is satisfied when the distributed data sample contains similar features, particularly in medical settings where different patient populations share the same disease, such as COVID-19 and cancer. On the other hand, if the goal is to increase the feature dimensions for identical records in distinct datasets, VFL is appropriate [22]. That is applicable when the distributed data have different sets of features for the same sample or records have the same ID, like when researchers need to work together between obstetrics and gynecology clinics and pulmonary clinics to investigate how COVID-19 affects pregnancy. Practically speaking, VFL is rarely used in the medical field because of the need for multiple datasets that must be identical, arranged, and standard. This is unrealistic due to the complex nature of medical image datasets and the lack of standardization management [23]. This study concentrates on the HFL, a commonly utilized format for medical images [24, 25]

For the FL scenario in a medical image application, suppose the dataset D_h for a hospital or medical institution, where h has data for S disease $D(S_e)$. $h = 0, 1, 2, \ldots, N$, and $N \in \mathbb{R}$, where N is the number of participant nodes for a e medical application (e.g., diagnosing, segmenting, or other applications) for a disease S by training a model $L_k(S)$ for k epochs. Initially, the central server system offers a predefined learning model $L_0(S_e)$ and sends it to all or a subset of N nodes in parallel. Each node updates the model weights or gradients value θ based on the local training of its own image data if the number of training images in the local

dataset is d and $d = \text{size}(D(S_e))$ update the learning weight given by Eq.2.1:

$$\theta_j \coloneqq \theta_j + \alpha \frac{1}{d} \sum_{i=1}^{d} (L_0(S_e(x_i)) - \hat{y})\, x_j \tag{2.1}$$

once the θ_j values are updated with the learning rate α on image j for number of k epochs by the node h, these values send-back to the central server. After central server collected updated weights from all hospital nodes, the central computes the global model by aggregating received models using Eq.2.2:

$$L_g(S_e(x))_T = \frac{1}{N} \sum_{i=0}^{N} L_i(S_e) \tag{2.2}$$

Aggregates the global model by averaging each weight over the number of sites, a process known as the FedAvg method [20]. For T rounds FL follows the same process to compute a last version of the global model.

Notably, two types of proposed aggregation strategies exist: sequential aggregation and parallel aggregation. Sequential aggregation trains the distributed data site by site, updating the global model after each site completes its local training and moves on to the next. It repeats the training iteration until the model converges, like in cycling weight transfer (CWT), or train it at a site for a predetermined duration or number of epochs, like in single weight transfer (SWT). It is efficient to perform these steps in a fully distributed FL framework. The second aggregation strategy, known as a parallel aggregation, shares the same initial model with the distributed sites, enabling them to train data simultaneously. It then collects the local updated models from all sites before aggregating a global model. In HFL, averaging the weights in the global model is a vanilla method because it supports the convergence

of updated weights. There are variant settings in the computations of local sides to preserve the personality of each hospital [26].

Considerations of FL in Medical Imaging

FL has demonstrated its applicability in various medical imaging tasks. For instance, FL has been successfully applied to brain segmentation in MRI images [27] and biomarker detection across multiple fMRI sites [28]. Previous research has underscored the importance of secure collaboration in medical AI, highlighting FL as a valuable approach, particularly in the context of combating COVID-19. Several studies have shown that FL can enhance diagnostic accuracy in multiple aspects, including: Automated COVID-19 Diagnosis [25, 29, 30], Lung Damage Assessment [31], and Prediction of Oxygen Requirements [32] .

However, the implementation of FL in medical imaging requires distinct considerations that set it apart from other domains, such as IoT-based FL. We identified and analysis their impacts in FL framework as the following:

- **Limited Number of Participating Institutions**: Managing medical imaging data is expensive, and many healthcare institutions lack the infrastructure for standard data management [33]. This limits data sources for FL in medical research.

  Positive impact: The limited number of participating institutions in FL for medical imaging offers certain advantages. Unlike FL in IoT applications, where thousands of edge devices may frequently disconnect or fail, hospitals and research centers are more stable participants. This ensures consistent

availability of training nodes, reducing disruptions in model updates and improving overall reliability. Moreover, a smaller number of institutions makes it easier to identify and characterize the source and type of data skewness [34].

Negative impact:A smaller number of participating sites leads to reduced data diversity, which may hinder the generalization ability of the FL model, particularly for handling rare conditions or unseen patient demographics and it may not perform well across a broader population.

- **High-Quality and Reliable Datasets**: Only medium-to-large hospitals or medical research institutions own the repositories of standardized medical images. These institutions contribute reliable and high-quality data for training models [22].

  Positive Impact: Standardized and curated datasets improve model training by reducing the influence of noisy or low-quality data. This enhances feature extraction, leading to better model performance, faster convergence in lower number of rounds, and more clinically relevant predictions.

  Negative Impact: While high-quality data improves reliability, it may also result in overfitting to specific imaging protocols or patient demographics. If the dataset lacks sufficient variability, the FL model might not generalize well to data from institutions with different imaging techniques or patient populations.

- **Computational and Communication Efficiency**: Medical imaging models are inherently large, it is exceeding 10 million parameters [35], making

their training computationally intensive. Integrated models with privacy-preserving techniques such as differential privacy (DP), homomorphic encryption (HE), or secure aggregation, the computational further increases. These methods are essential for protecting sensitive patient data but introduce significant overhead in terms of processing power and communication latency.

Positive Impact: These methods enhance security and guarantee patient privacy, ensuring compliance with strict medical data regulations. The use of complex architecture models captured intricate medical patterns more effectively. In medical applications, where precision is critical, higher model complexity translates into better clinical decision support.

Negative Impact: This leads to higher latency and increased hardware demands, which may delay model aggregation and slow down convergence.

In summary, the high data quality and trusted participants in FL contribute to improved reliability and faster convergence. However, challenges such as limited data diversity and computational constraints can hinder generalization and increase training complexity.

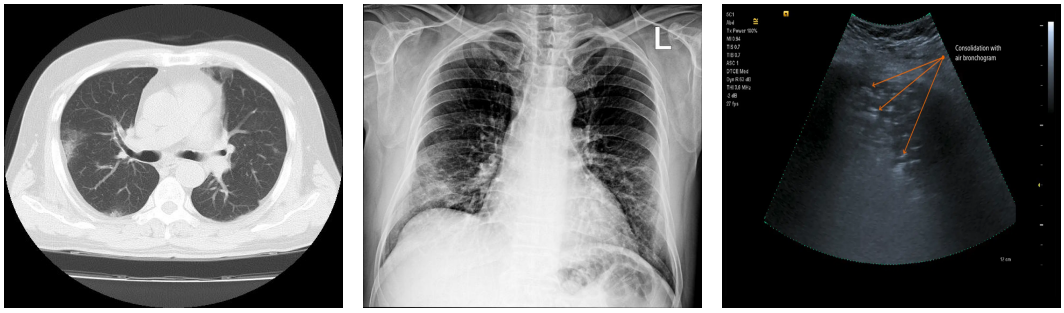
In this work, we focus on the first two key conditions for FL implementation in medical imaging, particularly addressing data heterogeneity in distributed datasets. Furthermore, to advance research on COVID-19 lung imaging, we identify and analyze the available data from a radiological perspective in the next subsection.

2.2 COVID-19 Medical Data Challenges

Since the pandemic began in late 2019, ongoing collective scientific efforts have aimed to build reliable and consistent information about COVID-19, creating many medical applications from textual [36], tabular [37], audio [38], and medical image datasets [39]. This section provides a brief description of the types of common imaging modalities, how AI systems identify the biomarkers in the lungs, and the selection criteria of available databases that DL uses for diagnosing, treating, extracting, and analyzing information about the infected lung. The research focuses on the lung medical image datasets for COVID-19 patients.

From a technical perspective, Junaid Shuja et al. [39] classified the imaging modalities used by learning models for COVID-19 applications into CT images and X-ray images, mentioning the low availability and quality of ultrasound datasets as example shown in Figure 2.3. Both X-ray images and CT images can capture COVID-19 symptoms in the lungs, but CT images use a multifocal view that more clearly distinguishes COVID-19 symptoms from other types of viral pneumonia [40]. This provides DL models with high applicability for diagnosing COVID-19 in the early stages, around days 2–4 of onset [6].

Different hospitals and clinical institutions store and label X-ray and CT images in various formats, leading to heterogeneity in shared medical image data. Hospitals and clinical institutions identified heterogeneity issues and provided early steps to overcome this by developing a standard format for CT images, X-rays, and other medical image modalities. DICOM is a medical technical tool that combines medical images with structured patient information reports. A study reported the



(a) Chest CT scan showing ground-glass opacities. [41]

(b) Chest X-ray showing lung consolidation. [41]

(c) Lung ultrasound showing B-lines. [42]

Figure 2.3: COVID-19 imaging modalities: CT, X-ray, and ultrasound.

successful collection, visualization, and diagnosis of COVID-19 from four hospitals remotely over the DICOM network [43]. The subsequent section explores the challenges of COVID-19 lung imaging data that produce high heterogeneity issues. These challenges could not be solved by individual DL approaches,and they might be considered opportunities in the implementation of FL in medical imaging.

2.2.1 FL Opportunities for COVID-19 Lung Datasets

In the medical field, deep learning (DL) has reduced the time and cost of clinical decisions for diagnosis, prognosis, and treatment [44], providing high accuracy and confidence. However, DL models require large and diverse datasets, which often necessitate combining data from various hospitals globally.

The COVID-19 pandemic has revealed challenges in structuring medical imaging databases, with one study showing that a large-scale DL model for COVID-19 medical imaging, while promising, exposed biases [45]. In this section, we highlight the factors that make FL more suitable than DL, particularly for addressing data heterogeneity and privacy concerns.

Data Availability

The rapid spread of COVID-19 across many countries, coupled with its unclear cause of infection and the wide range of symptoms in different populations [3, 46], led to the collection of scattered COVID-19 medical data. This fragmentation contributed to misclassifications in annotated data, particularly in the early stages of the pandemic [47]. Moreover, due to the large volume of patients and limited storage capacity in some medical facilities, testing protocols focused on archiving and labeling images solely for confirmed COVID-19 cases [48]. This hindered the ability to differentiate COVID-19 from other similar infections, such as pneumonia or SARS, at a localized level.

Training DL models with small, high-quality datasets from local hospitals is insufficient for generalization, as these datasets often carry biases due to incomplete

feature representation [10]. Local models tend to overfit, yielding high accuracy but low reliability. In contrast, several studies have integrated diverse medical imaging datasets to train DL models for COVID-19, aiming to improve generalization [47, 49]. However, the poor quality of these datasets resulted in underperforming models with significant biases [39]. This is because DL models struggle to recognize COVID-19-specific features, relying instead on common patterns within a single dataset [50].

FL, also known as distributed learning, addresses these issues by training models using diverse image acquisitions, rare cases, and varying attributes from multiple sources, without the need to exchange actual data. This approach reduces bias in local models by leveraging a broader, high-quality dataset while maintaining data privacy.

Hot/Cold-Start Problem

Historically, epidemiological diseases in specific geographical areas have evolved gradually, with symptoms and transmission patterns becoming more uniform over time. However, the COVID-19 pandemic differed when the World Health Organization (WHO) declared it a global pandemic on March 11, 2020, approximately two and a half months after the first cases were reported in Wuhan, China [36]. The rates of confirmed cases and deaths varied significantly across countries due to different social, political, and healthcare circumstances. For example, while the virus had devastating effects in China, other countries, such as North Korea, experienced less severe outcomes, similar to seasonal flu [51].

Due to varying levels of access to epidemiological data, some countries possess more comprehensive datasets than others, creating an urgent need for collaboration to analyze these disparate data sources without sharing patient records. Identifying patterns across a larger dataset would accelerate diagnosis, treatment, and healthcare services. FL allows global collaboration on heterogeneous data sources, enabling knowledge sharing and insights without breaching privacy. This collaborative model can expedite response efforts, improving public health outcomes and addressing infectious diseases more effectively [52].

Time and Cost of Processing

In the medical field, hospitals and institutions value patient data highly, leading to local processing and analysis to prevent unauthorized access. Data governance rules, such as those governing health data privacy, make local processing costly in terms of human resources and hardware [46].

Labeling also adds to these costs, especially in regions with unequal population density. Larger towns typically have busy hospitals with higher screening rates, leading to workload peaks during certain hours. As a result, radiologists spend more time analyzing and labeling images from these overcrowded hospitals. The rapid spread of COVID-19 further exacerbated this issue, leaving many screening tests pending labeling. By utilizing these unlabeled data as a testing set, FL can train a global model on distributed datasets to classify infections [53].

Security and Privacy

Health data is highly protected in most countries, with strict governance rules to prevent leaks. Privacy concerns are a major obstacle to sharing patient information, as regulations such as HIPAA in the United States [54] and the GDPR in Europe [55] impose strict conditions on data sharing and use, even for research purposes. Privacy-preserving methods have proven inadequate, as anonymized patient data is vulnerable to re-identification attacks, such as linkage attacks. Recent studies have even developed machine learning methods capable of predicting personal details like age, gender, and name from chest X-rays [56].

FL offers a solution by enabling local processing under the control of the medical institution, ensuring data privacy [10]. Collaborative research via FL can provide consistent, reliable insights into COVID-19 while maintaining data privacy. It has already shown promising results in managing medical imaging data efficiently [57], and this approach could be extended to other diseases in the future.

2.3 Data Heterogeneity Issue

FL allows each site to train its local model on its own data and optimize it according to the features that arise internally. The model weights are adjusted to align with the local results. This implies that updates in each local model weights align with its unique features. In the aggregation phase, the global model tries to balance the received weights by combining these local updates and averages. This makes the optimization very unstable and leads to a high degree of divergence over rounds. Moreover, it produces an optimization drift between the local and global models, which leads the global model to bias [58]. Additionally, the distributed sites assign the global model to train local data for the next round, leading to a deficiency in personalization metrics and degradation of the local performance model.

Bias, a commonly reported issue in FL, negatively impacts both the generalizability of the global model and the personality of local models. The generalization measure refers to the global model's ability to make more comprehensive decisions about out-sampling or external training data, which may contain new features and come from a variety of sources. Each round's aggregation step progresses to collect additional weights from local updates.

Conversely, FL personalization metric refers to the global model's ability to effectively validate local data after each round across distributed sites. The global model performance for the site that has the lowest size of data achieved the lowest accuracy due to the averaging process of weights biased to the sites having a larger training sample [59]. Further computations are suggested to adapt each local model depending on the specifications corresponding to the participant sites (e.g., PPML [18]

and FedAMP [60]). However, this measure required more investigation to specify effective results and findings about variant non-IID situations.

Specifically, controlling a larger divergence between these two measures becomes more challenging in FL when dealing with more heterogeneous data. The large variation between distributed datasets across FL participants is called the non-identical independent distribution (non-IID) data issue [31].Therefore, we aim to investigate the types of non-IID data and examine the impact of each type on various FL metrics.

The COVID-19 pandemic facilitated the creation of extensive medical imaging datasets using modalities such as CT scans, X-ray and ultrasound images. These resources have motivated researchers to apply FL frameworks to COVID-19 imaging datasets for various medical purposes, including lung disease classification [14, 61], COVID-19 diagnosis [62, 60], severity identification [17, 63], lung damage segmentation [18, 64], and oxygen-need prediction [32].

By investigating related works that have considered the non-IID data, we initially mentioned the variant process of partitioning experimented data before viewing their proposed solutions to mitigate this issue in the first subsection of related works. In the second subsection of related works, we reviewed the studies based on the evaluation metrics that have been employed to assess FL performance in non-IID scenarios, focusing on both generalization and personalization.

2.3.1 Related work based-Skewness Types

Data skewness refers to the extent to which data distributions shift or deviate across different distributed environments, which collaborate to enhance the overall learning process. The type of skewness or shift in the data directly influences how the data is partitioned across FL participants. In some proof-of-concept experiments, the partitioning may reflect real-world data heterogeneity, but in many cases, the manual partitioning process across nodes is not explicitly detailed [29]. Additionally, these studies do not always specify the exact types of skewness used in designing their experiments.

This section aims to identify the scenarios in which distribution skew or data shifts have been investigated in COVID-19 lung imaging data. We adopt the categorization framework outlined in the systematic review by Shyu et al. [22], which examines the impact of data skewness in medical imaging and classifies it into three primary types: quantity skew, label distribution skew, and data acquisition skew. These classifications highlight fundamental challenges of non-IID of medical imaging in FL. However, COVID-19 imaging studies provides that data heterogeneity in FL extends beyond these three categories.

Building upon this foundation, our review of COVID-19 medical imaging datasets reveals additional forms of data skewness that have not been explicitly addressed in prior work. Specifically, we extend the existing classification by introducing three additional types: extreme label skew, modality skew, and feature skew. These new categories capture unique challenges specific to medical imaging datasets in FL, where inconsistencies in labels, imaging modalities, and feature representations

can impact model training and performance. The six types of data skewness are described below, along with related studies that have investigated their impact.

Quantity Skew

Quantity skew occurs when larger hospitals have larger datasets based on the number of patients, and the other participants have small datasets with a large difference between the size of each. The different number of medical images affects the model, with an extreme increase or decrease in data quantity and they suggested limiting the accepted ratio of quantity skew between two sites to 1:100 [22]. Sites with a small data quantity may lead to data ignorance due to the averaging of the model weights over sampling data per training site. As a result, the existing emphasis on COVID-19 features would fade with smaller datasets and bias the global model toward a higher quantity of data.

The researchers proposed the following methods to alleviate quantity skew across distributed data:

- Using the augmentation method to expand the size of the image dataset is a simple and common solution for highly training the model on the same data features, which can be achieved by changing various scales such as transformation, zooming, and rotation. However, the transformation methods used for generating data are not always effective in training, which may degrade the model's performance [62].
- Alternatively, a generative adversarial network (GAN) [29] can be utilized [65]. It offers small improvements with a high computational time.

- In such aggregation strategies, quantity skew issues are improved by assigning a learning rate or batch number to each client variant based on the quantity of data [66].
- The FedAMP model exhibits resistance to quantity skew because its aggregate weights are adaptively learned throughout the training process [60].

Label Distribution Skew

Label distribution skew could be a combination of two types of skewness: quantity and label. Label distribution skews occur when datasets have the same labels but different quantities for each site. The larger hospitals have access to larger datasets simply because they have more patients. For example, a large urban hospital with a significant patient population may have a much larger dataset than a smaller rural clinic. On the other hand, label distribution skew refers to a situation where the total quantities of data across all FL sites are similar, but the distribution of labels varies. That means, at one site, the majority of a dataset might be labeled as COVID-19-positive, while the majority might be COVID-19-negative cases at another site. This skew could lead to imbalanced training, wherein a model becomes biased toward the more frequent labels, hindering its ability to generalize across different sites. This skew has a negative impact on FL generalizability, personality, and the computational time [65]. FedAvg [20] is proposed to address this type of skewness, but it fails when the distribution skewness ratio is high [14]. To mitigate this issue, researchers have followed one of three directions:

- A first-direction solution is implemented before training data by preprocessing

data to ensure a uniform distribution across sites using the local augmentation method [21], GANs [22], and the synthetic minority oversampling technique (SMOTE)[67]. However, these solutions require more communication between parties to fine-tune the number of labels and the distribution of images in each. This may also lead to information leakage from participant data with slightly improvements in accuracy, approximately around 1.6% [14].

- The second direction is to improve convergence between updated models locally, which focuses on the monitoring of local updates per batch in FedBN [68], per round in FedProx [26], or by normalizing both local and global updates in HarmoFL [58]. These methods report efficient bias mitigation and improve the global model's generalizability. However, they may have a negative impact on the model's personality. In other words, the model may yield poor results due to its incompatibility with local population data.

- The third direction is examining the local updates on the server side before accepting the updated models. This may rely on various calculations of acceptance priority [46], the use of voting methods [7], and the implementation of smart contracts in blockchain-based systems [20, 17]. However, these methods have additional computational overheads.

Data Acquisition Skew

Under this skew type, the same modality can be considered, but with image datasets with different attributes. All sites have X-ray modality but with different scanning devices, resolutions, light conditions, protocol for capturing (e.g., patient with up-

per/lower hands posed), and protocol for storing (e.g., with/without compression and with/without radiologist annotations). This type is the most commonly considered skew. Many studies have applied the same modality of acquisition in several hospitals [64]. A previous study investigated X-Ray datasets with a different image volume thickness in each site of FL. The authors reported results with satisfactory performance, which could be collaborated to identify COVID-19 features, even with a variety of image attributes [18]. The researchers followed one of the following directions to fix acquisition skew:

- A self-adaptive hyperparameter is proposed to adopt the variability of data between distributed sites [66].
- In one study, the authors used a self-adaptive aggregation function with meta-transfer modules to manage the settings of training data locally [64].
- In two further studies, the authors made distributed datasets uniform by incorporating regularization methods and leveraging ensemble techniques such as normalize image intensity, resizing, and geometric transforming methods [29, 32].

Extreme Label Skew

This type of extreme skewness occurs when one or more labels completely disappear from one or more sites; this phenomenon is known as extreme label distribution skew. This situation is common because, during the COVID-19 pandemic, some hospitals stored only confirmed infection cases due to storage capacity limitations

[48]. Additionally, certain medical image datasets categorize their data into two [64], three [25], or more [63] labels.

This leads to poor convergence and degrades global accuracy in the central testing node of the FL framework for COVID-19 X-ray images [69].The following are valuable solutions proposed by different studies for this type of skew:

- The authors used a semi-supervised method to label unlabeled data and reported satisfactory accuracy [31]. As a recommendation to the uniform label name in the FL framework, their method could be useful in this situation.
- To address the word variants issue, radiologists could also analyze meta-data using natural language processing (NLP) [53].

Modality Skew

Medical image equipment varies between single sites and different sites. Different devices produce different imaging modalities (e.g., X-ray, CT, and ultrasound), with variations in volume, resolution, color intensity, type, and amount of noise. Different image modalities require different models and analysis procedures, which are difficult to unify [31]. It is a rare situation in an FL framework (as Figure 2.4 shows), where most studies considered single-modality X-ray data [24, 14, 64, 70] or CT data [63, 46, 31, 71, 72], some studies combined both and then redistributed randomly [73] or assumed variant modalities in clustered edge nodes [74], and one study employed emerging X-ray imaging with simple vital signs [32].

- Practically, Qayyum et al. [74] conducted an experiment using the same model

for X-ray, CT, and ultrasound images for COVID-19 diagnosis and fixing the impact of modality skew by clustering sites of FL that have a similar modality. They described the model's ability to recognize COVID-19 features without prior knowledge of the image modality. However, this type of skewness is challenging the performance of the model that applied for training, which comparing features per pixel. Researchers thus need to further investigate this skew type and define the impact of varying image qualities on FL accuracy [62].

- One feasible approach worth further exploration involves implementing an additional layer inside a neural network. This layer would be capable of determining the most suitable FL for deployment based on medical imaging and clinical similarities between variant dataset features [60].

Feature Skew

The feature skew of datasets relates each participant's training sample to distinct features like age range, patient gender, smoking patients, and severity of patients. The hospital or medical institution may acquire data for an intended feature, or it may depend on the virus's behavior in their population.

This type of feature skew is applicable to horizonal FL, where the non-overlapping sample at each site differs from the one defined in vertical FL. However, it is a real situation and rare in vertical experiments studies. A unique study investigated the FL performance by distributing the dataset into four sites, each with a range of ages [75]

By expanding the classification of data skewness in FL for medical imaging, this study provides a more comprehensive framework for understanding the unique challenges posed by non-IID data distributions. Addressing these issues is crucial for improving the reliability and fairness of FL models in medical applications, particularly in the context of COVID-19 imaging.

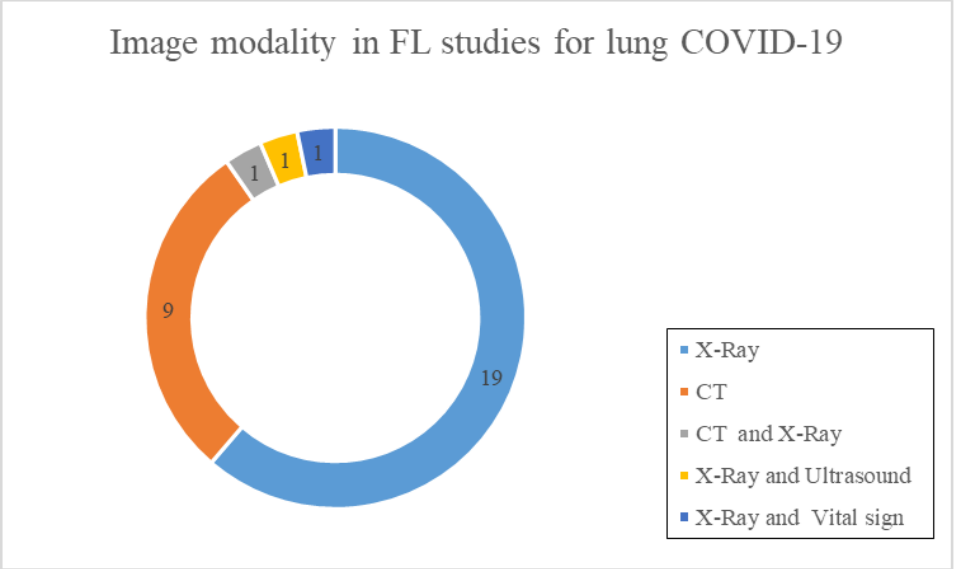


Figure 2.4: Type of lung imaging dataset modalities used in FL studies for COVID-19

As mentioned above, a real medical data situation may consist of one or many skewness types in the training or testing data in the FL framework. It is crucial to examine the data specifications for each participant, as this information can be quantified and shared as meta-data prior to the start of training. These meta-data were used to identify the degree of non-IID per skewness, which provides the manager with the ability to assess the source of bias and help them to take advantage of FL, depending on their goal.

2.3.2 Related Works Based-Evaluation Metrics

Researchers have employed evaluation metrics to assess FL performance in non-IID scenarios, focusing on both generalization and personalization as Table 2.1 shown the comparison between studies that evaluate different types of skew and their results. To evaluate generalizability, two partitioning methods are commonly employed. In the first, datasets are redistributed randomly across sites, while subsets of local data are retained for central evaluation. Florescu et al. [13] demonstrated effective generalizability under such settings. In the second method, realistic scenarios are simulated by assigning unique training resources to each site, as Bhattacharya et al. [61] achieved through P2P communication without central evaluation. However, Peng et al. [60] reported significant accuracy degradation when using external testing data in the central node, highlighting the challenges posed by high heterogeneity across sites. Similarly, personalization metrics evaluate local model performance by tracing the compatibility of updated models with local data. Some studies use subsets of local data for both validation and testing [76], while others rely on external testing data to assess personalization [62, 60].

Zhou et al. [18] proposed strategies for mitigating the degradation of personalization metrics, while Dou et al. [29] evaluated both generalization and personalization metrics, focusing solely on acquisition skew.

Therefore, technical research shows confusing diagnosing of the COVID-19 infection due to dynamic behavior exhibited which causes heterogeneity impact on the lung imaging from variant populations [3]. FL offers a framework that enables deep learning (DL) to extract global knowledge from huge, distributed, and sen-

Table 2.1: A comparison of related studies that evaluated non-IID types along with their measured metrics.

Ref	Heterogeneity Type	Evaluated Factor	Contribution
[18]	Acquisition skew	Personalization metric	Focused on a personalized metric fixed by an adaptive local epoch is an effective method.
[29]	Acquisition skew	Generalization and personalization metrics	Distributed data used for training were made visible by pretrained CNN models.
[65, 77, 66, 67]	Quantity and label distribution skew	Generalization metric	Maximizing the size of data is an effective way of improving the generalization metric.
[69]	Extreme label skew	Generalization metric	Skew can be managed via hyperparameter settings.
[75]	Feature skew	Generalization metric	Non-IID data had a significant negative impact.
[74]	Data acquisition and modality skew	Generalization and localization metrics	The global model could be successful regardless of the image modality.
Our work	IID vs. 6 different skewness types	Generalization, personalization, and localization metrics	FL exhibited effective performance when there was a maximum of one skew type. However, a mixture of different skew types caused high divergence of the training model for all metrics.

sitive data in an efficient and private manner. However, the heterogeneity of data takes different sittings because of variant variabilities in medical imaging situation which may further complicate that diagnose confusing. Furthermore, Assessing the influence of heterogeneous data absent in certain investigations [14] as their findings suggested to dismiss the effect of non-IID on the model's efficacy. This advocates for the execution of experiments to examine its influence in a deeper manner across each site with varying non-IID scenarios. In this experimental study, the most real distributions of imaging medical situation are intended to recognize their impact on the FL results. Additionally, understanding and visualizing all the related metrics in the FL framework provides a comprehensive evaluation performance for each non-IID case. Thus enable the researchers to propose suitable solutions to address the degradation of FL performance based on the heterogeneity of hospitals' data.

Chapter 3

System Design

FL has emerged as a promising approach for training machine learning models in a decentralized manner while preserving data privacy. However, the effectiveness of FL in medical imaging, particularly for COVID-19 lung imaging, is significantly influenced by system architecture, data heterogeneity, and the presence of non-IID distributions across clients. This chapter presents the design of our FL system, detailing its key components, operational phases, core algorithm, and a mathematical analysis of the impact of non-IID data types on model performance.

The system is composed of multiple components, including a central server responsible for global model aggregation, local clients that perform training on private datasets, and a secure communication framework that ensures efficient model updates. The workflow of FL is divided into several phases: the initialization phase, where the global model is shared with clients; the local training phase, where each client updates model parameters using local data; and the aggregation phase, where

client updates are merged to refine the global model. The FL algorithm implemented in this study follows an iterative training process that accounts for data imbalance and distribution skew among participants, employing optimization techniques such as adaptive aggregation and personalization.

To further understand the challenges introduced by non-IID data, a mathematical analysis is conducted to quantify the effects of different types of data skew, including quantity skew, label distribution skew, and acquisition protocol skew. These variations lead to discrepancies in local model updates, impacting convergence rates and generalization. The following sections elaborate on each aspect of the system design, providing a comprehensive framework for FL in COVID-19 medical imaging.

3.1 System Phases

The proposed FL framework consists of five main phases, each responsible for a crucial step in the overall system workflow. These phases ensure efficient model training across distributed datasets and measure the impact of challenges related to non-IID data as shown in Figure 3.1. The phases are as follows:

1. **System Configuration Settings Phase:** This phase initializes the FL environment by configuring the central server and distributed clients. It defines hyperparameters such as learning rate, batch size, and communication rounds. Additionally, it establishes secure communication protocols and assigns computational resources to ensure efficient training.

2. **Data Preprocessing Phase:** Before training begins, local datasets undergo a series of preprocessing steps to ensure consistency and enhance model performance. First, resizing is performed to standardize image dimensions across different sources, ensuring that all input data conform to a fixed resolution while preserving aspect ratios where necessary. Next, normalization is applied to mitigate variations in pixel intensity caused by differences in acquisition devices and protocols. This process scales pixel values to a common range for improving stability and convergence during training. Additionally, data augmentation techniques are utilized to improve model generalization and address potential data imbalance issues.

3. **Distributed Training Phase:** Once the data preprocessing phase is completed, the FL process enters the distributed training phase. In this phase, each participating client (e.g., hospitals or medical institutions) independently trains a local model using its preprocessed dataset without sharing raw data, thereby preserving data privacy. The training process follows a predefined global model architecture initialized by the central server. Each client performs multiple local training iterations. To enhance efficiency, the system employs parallel training, where multiple clients simultaneously train their local models on separate computing resources. This approach accelerates the learning process and reduces training time.

   After local training, each client validates its model using a separate local validation set in a parallel validation process. This locality measure aim to ensure that each local training of model is evaluated before being sent to the central server, allowing for early detection of potential overfitting or poor

generalization.

4. **Model Aggregation Phase:** After each round of distributed training, the local models from all participating sites are sent to the central server for aggregation. The central server computes the global model by applying an aggregation strategy. This aggregated global model represents a more generalized version that incorporates knowledge from all distributed datasets. Once the global model is updated, it is shared back with the participating nodes, allowing them to replace their previous local models with the newly aggregated version. Each site then initializes the next training round using the updated global model, continuing the iterative process of FL. This cyclic process ensures progressive model improvement over multiple communication rounds, gradually enhancing performance while maintaining data privacy.

5. **Performance Evaluation Phase:** Once the global model is updated through aggregation, it undergoes thorough evaluation to assess its generalizability and effectiveness across distributed datasets. The evaluation is performed for two metrics generalizability using an independent test dataset maintained at the central server and personalizability by redistributing the global model back to participating sites for local validation. The primary goal is to ensure that the global model performs consistently across diverse medical imaging data and does not exhibit biases toward specific institutions.

   To enhance reliability, parallel validation is employed, where each client evaluates the global model using its own dataset. This step ensures that the model maintains high performance across different distributions and enables

early detection of potential biases.

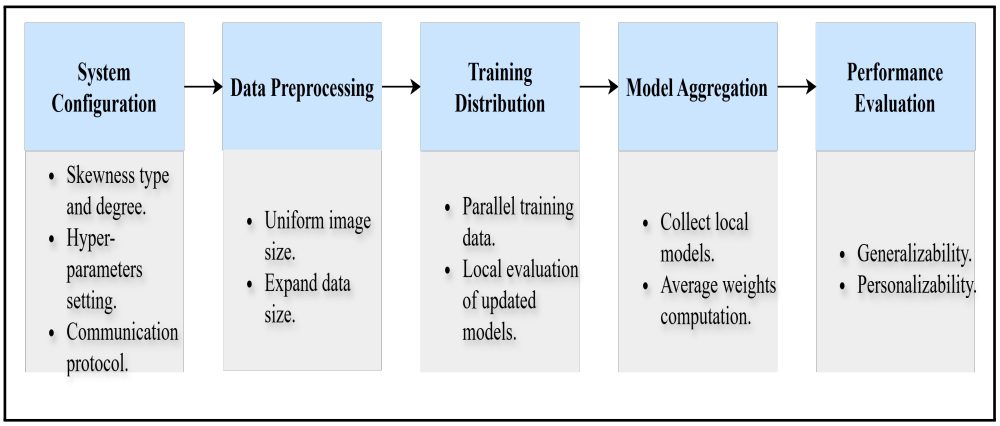


Figure 3.1: System phases and their corresponding steps in the FL

3.2 System Overview

FL is a technique that can be used to increase the generalizability of learning results for distributed data. It facilitates distributed deep learning that can be used to train models without requiring access to local data in distinct sites. In FL, the parameters of models are only exchanged between parties and aggregated to build a global model in a private and secure manner. It provides a global model that can overcome model bias generated by training local data with considering the heterogeneity of data from different resources.

3.2.1 System Components

The system is designed to study the non-IID issue pertaining to COVID-19 medical imaging in the FL framework. Figure 3.2 provides a high-level overview of the system's design, comprising four components, each with distinct roles and modules shown in low-level overview of system as Figure 3.4 , as outlined below:

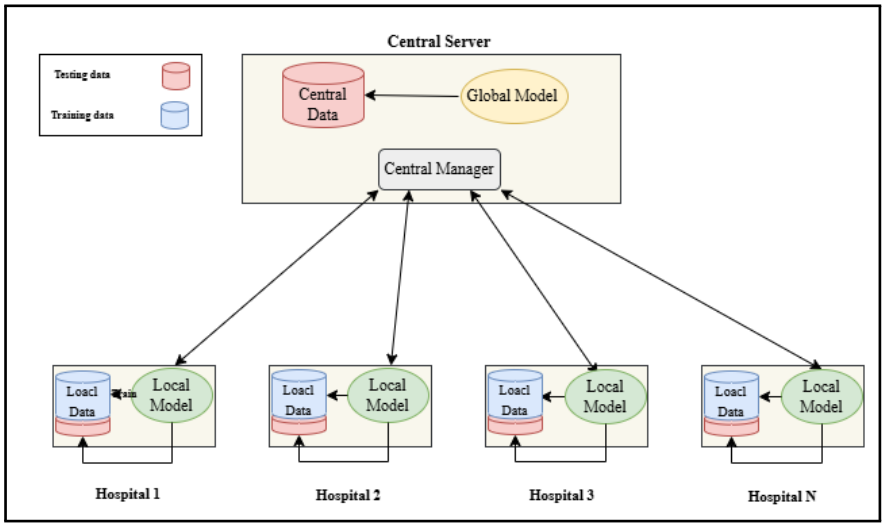


Figure 3.2: System components

1. **Central server**

   The central server functions as a trusted node, incorporating three key modules: the central manager, the aggregator, and the validator. The central manager is responsible for initializing communication between distributed sites, coordinating training rounds, and ensuring secure data exchanges. The aggregator performs global model computation by merging local updates received from clients, applying optimization techniques to mitigate the ef-

fects of data heterogeneity. Lastly, the validator evaluates the generalization capability of the global model using out-sampled data, assessing its performance across diverse distributions to ensure robustness and reliability. These modules collectively enable efficient communication, model aggregation, and performance validation within the FL framework.

2. **Hospital nodes**

The participating hospitals are responsible for three critical tasks within the FL framework: Local model training, Upon receiving the global model from the central server, each hospital performs local training using its proprietary dataset to update the model parameters. The Second is local model validation, after generating the updated local model, each hospital conducts validation to assess its performance prior to submitting it to the central server for aggregation. The third one is global model testing. Following the aggregation of all local models at the central server to produce the global model, each hospital evaluates the received global model to determine its performance and applicability to the local data distribution.

This systematic approach ensures the integrity of the FL process by measuring model generalization and personalization while accommodating data heterogeneity across participating institutions.

3. **Models**

All the sites in the proposed system share similar architectures from central server, with the only difference being that the weights update according to the local data of the hospital node. In this study, we employed a simple a convolutional neural network (CNN), as depicted in Figure 3.3. Model architecture consisting of seven primary layers, excluding the input layer. The model's first layer is a convolutional layer featuring six filters with dimensions of 5 × 5; it processes the input's three channels, typically representing RGB images. This is followed by a max-pooling layer with a 2 × 2 kernel and a stride of 2; it down-samples the feature maps. Subsequently, there is a second convolutional layer, which utilizes sixteen 5 × 5 filters and receives the six channels generated by the previous layers. After this layer, a second max-pooling layer is applied with the same configuration as the first one. The output from the convolutional block is then flattened and passed to the first fully connected layer, which is dynamically initialized based on the output of the previous layers and includes 120 neurons. This layer is followed by a second fully connected layer with 84 neurons. Finally, the output layer is another fully connected layer, with several neurons corresponding to the number of classes specified for the classification task. The activation function of the fully connected layers is ReLU.

The essential aim of this research to explore the behavior of both local and global models with respect to non-IID data acquired through FL. Notably, the choice of a CNN architecture was not our primary concern with respect to achieving higher accuracy; a simple model is enough to achieve our goal.

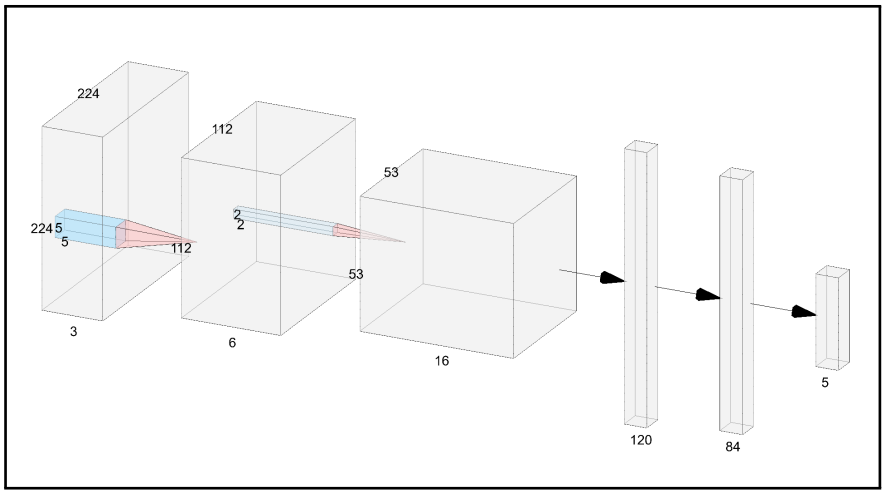


Figure 3.3: CNN model used for training distributed hospital datasets in FL framework.

4. **Distributed local datasets**

   We divided each local set of data into two subsets: training data and testing data. The testing data were used to validate the updated models twice per round. The first validation occurred after the local models had been trained, aiming to measure the accuracy of the localization model. The second validation was performed using the same test data used for the model that was generated after the global model was averaged over all the models from other participating training nodes.

The Figure 3.4 illustrates a low-level system view of an FL framework comprising two main components: the central server node and the participant node. The central server node coordinates the training process and aggregates the models to produce a global model. It includes several key components. The FL coordinator manages participant selection through a site selector and sets training configurations

via a hyperparameter configuration. The central round manager orchestrates the communication by using a model sender to distribute the global model, a model receiver to collect updated local models, and a results receiver to gather local evaluation metrics. The model aggregator then combines the received local models into a new global model. This model is assessed by the global evaluator using a dedicated central testing dataset. The results from each round are saved in the round results module for performance tracking. On the other hand, the participant node

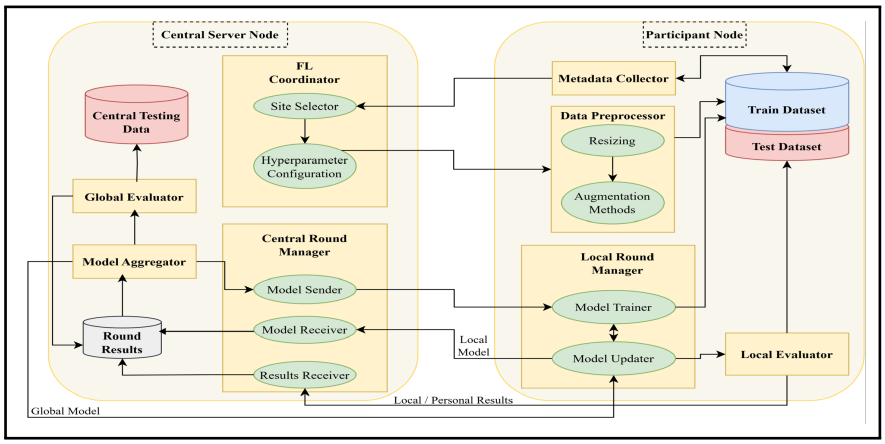


Figure 3.4: Illustration of the system's workflow with low-level architecture view of system nodes.

handles localized training and evaluation tasks. It starts by collecting environment-specific data characteristics via the metadata collector, which are then shared with the central server. The data preprocessor component standardizes the input data using resizing and applies augmentation methods to improve generalization. The core training activities are handled by the local round manager, which includes a model trainer to perform training on the participant's private train dataset, and a model updater to refine the model parameters. The updated model is subsequently evaluated by the local evaluator using the local test dataset, and both the updated

model and its evaluation results are sent back to the central server to complete the round.

3.2.2 System Algorithm

The proposed system achieves its functional requirements based on the following scenario, as illustrated in Figure 3.5.

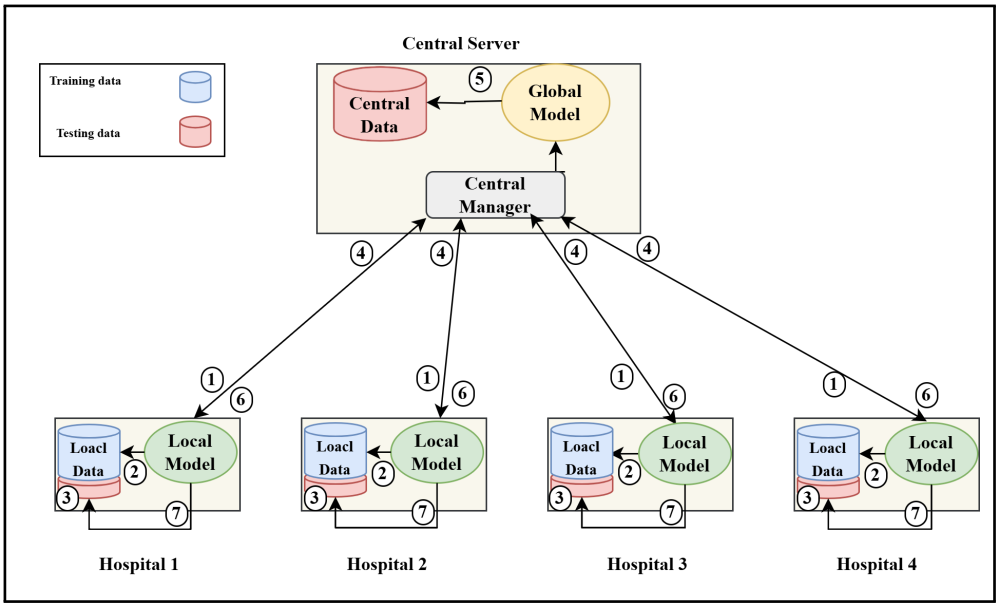


Figure 3.5: Illustration of the system's workflow and numbered steps.

1. The system's scenario begins with sending a simple CNN model architecture and configuration settings to hospitals or medical institutions in parallel.

2. Once the participant node receives the model, the CNN model is trained on local data, and the weights of the received model are updated using the number of concurrent local training epochs.

3. Each participating node evaluates the last version of the local model after updating it for the last epoch to evaluate locality performance.

4. All distributed nodes send the locality evaluation metrics, local training sizes, and the latest versions of their models back to the central server.

5. After the training session, the central manager provides a new update of the global model using the FedAvg algorithm and evaluates it against the central testing data to determine the generalization metric.

6. All participating nodes share the computed global model.

7. Finally, each local site evaluates the global model using local testing data to measure the personalization metric. The round number is then incremented by one, initiating a new training session from step 2.

3.2.3 System Deployment

The illustrated architecture, as in Figure 3.6, represents the implementation of FL, which is designed to allow collaborative model training between multiple healthcare institutions without requiring the exchange of sensitive patient data. Utilizing the Flower framework as the middleware, this system supports decentralized learning by orchestrating communication between a central coordinating server and a set of distributed client nodes, such as hospitals. Each client performs local training using its private dataset and transmits model updates—rather than raw data—to the central server for aggregation. The Flower framework facilitates this process

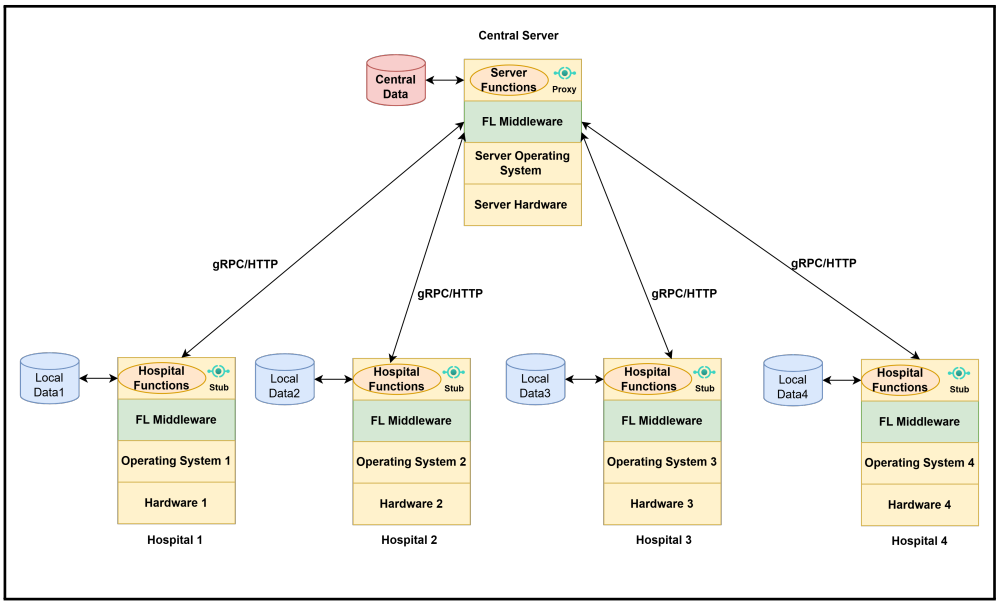


Figure 3.6: Illustration of the deployment system in an actual environment

by offering a lightweight, extensible, and device-agnostic infrastructure. A device-agnostic infrastructure ensures that all participants—regardless of their hardware specs—can still participate in the federated learning process. Also, that supports both synchronous federated learning scenarios [78].

Communication between the server and clients is managed using gRPC, a high-performance RPC protocol that ensures secure, reliable, and low-latency transmission of model parameters. gRPC uses HTTP/2 and protocol buffers for serialization, making it well-suited for scalable federated systems with dynamic client availability [79]. This architecture therefore, enables robust training workflows that preserve data locality, accommodate heterogeneous environments, and ensure modular interoperability across different platforms and institutions.

3.3 Mathematical Analysis of Data Heterogeneity

In this study, we investigate various types of data skew, focusing on how they impact both generalization and personalization metrics within the context of FL in relation to COVID-19 medical imaging. Below, we describe several common types of data skew, providing mathematical analysis and real-world case studies for each to illustrate their implications on evaluation metrics more clearly.

3.3.1 Quantity Skew

Quantity skew occurs when larger hospitals have access to larger datasets simply because they have more patients. By assuming N hospitals have the same image modality (i.e., X-ray, CT, and ultrasound) regardless of the captured equipment version, there is a dataset D_i of hospital i that does not have a size equal to that of the other dataset, where all the datasets have the same set of labels $L(D)$, and the sum of the lengths of each label k in all datasets is constant C for all labels m. An example is shown in Figure3.7 (a), where each cell represents the number of images for label L_m that are included in the corresponding dataset D_i. The mathematical expression of quantity skew is given via conditional Equation 3.1

$$\begin{aligned} \exists i, j \in \{1, 2, \ldots, N\} : \quad & |D_i| \neq |D_j| \\ & L(D_i) \equiv L(D_j) \\ & \sum_{i=1}^{N} |L_k(D_i)| = C \quad \text{for all } k \in \{1, 2, \ldots, m\} \end{aligned} \tag{3.1}$$

3.3.2 Label Distribution Skew

Label distribution skews occur when datasets have the same labels but different quantities for each site. By assuming N hospitals have the same image modality (i.e., X-ray, CT, or ultrasound) regardless of the captured equipment version, there is a dataset D_i of hospital i with a size equal to that of dataset D_j, where all the datasets have the same set of labels $L(D)$, and the summation of the lengths of each label k over all datasets is variable for other labels in the set. An example is shown in Figure 3.7 (b), where each cell represents the number of images for label L_m that are included in the corresponding dataset D_i.The mathematical expression of label distribution skew is given via conditional Equation 3.2:

$$\begin{aligned} \exists i, j \in \{1, 2, \ldots, N\} : |D_i| \equiv |D_j| \wedge L(D_i) \equiv L(D_j) \wedge \\ \sum_{i=1}^{N} |L_k(D_i)| \neq \sum_{i=1}^{N} |L_{k+1}(D_i)| \\ \text{for all } k \in \{1, 2, \ldots, m\}. \end{aligned} \tag{3.2}$$

3.3.3 Extreme Label Distribution Skew

By assuming N hospitals have the same image modality (i.e., X-ray, CT, or ultrasound) regardless of the captured equipment version, there exists a dataset D_i of hospital i with a different size for dataset D_j. $L(D)$ is the set of labels in all the datasets, and the summation of the lengths of each label k over all datasets is extremely variable for other labels in the set. An example is shown in Figure 3.7 (c), where each cell represents the number of images for label L_m that are included in

the corresponding dataset D_i.The mathematical expression of extreme label skew is given via conditional Equation 3.3

$$\exists i, j \in \{1, 2, \ldots, N\} : |D_i| \neq |D_j| \wedge L(D_i) \neq L(D_j) \wedge \sum_{i=1}^{N} |L_k(D_i)| \neq \sum_{i=1}^{N} |L_{k+1}(D_i)| \quad \text{for all } k \in \{1, 2, \ldots, m\}. \tag{3.3}$$

3.3.4 Data Acquisition Skew

Data acquisition skew occurs when medical imaging equipment differs across sites, leading to variations in the quality and characteristics of the corresponding images. For example, one hospital may use high-resolution CT scanners, while another may use older equipment that produces lower-quality X-ray images.

By assuming N hospitals have the same image modality (i.e., X-ray, CT, or ultrasound) regarding the captured equipment version and the protocol of collecting the local data, where $P = \{\text{volume, resolution, format, equipment, patient pose, etc.}\}$, there exists a dataset D_i of hospital i with a size almost equal to that of dataset D_j. $L(D)$ is an identical set of labels in each of the datasets, and the summation of the lengths of each label k over all datasets is approximately equal for other labels in the set.

An example is shown in Figure3.7 (d), where each cell represents the number of images for label L_m that are included in the corresponding dataset D_i. The mathematical expression of the data acquisition protocol skew is given via conditional Equation 3.4:

$$\exists i, j \in \{1, 2, \ldots, N\} : |D_i| \cong |D_j| \wedge P(D_i) \neq P(D_j) \wedge L(D_i) \equiv L(D_j) \wedge \sum_{i=1}^{N} |L_k(D_i)| \cong \sum_{i=1}^{N} |L_{k+1}(D_i)|, \text{ for all } k \in \{1, 2, \ldots, L_m\}. \tag{3.4}$$

3.3.5 Modality Skew

Modality skew arises when different imaging modalities are used across sites or even within the same site in the FL framework. For example, one medical center may primarily use CT scans to diagnose COVID-19, while another may use X-rays or ultrasound images. In some cases, a single site may use a combination of these modalities for diagnosis. These varying modalities lead to different features in the corresponding medical images, potentially confusing a model during training, especially when it has to integrate and generalize knowledge from multiple imaging sources.

By assuming N hospitals have different image modalities M where it is set of types ($M = X - ray, CT, ultrasound, etc.$), there exists a dataset D_i of hospital i with a size almost equal to that of dataset D_j. $L(D)$ is an identical set of labels in each of the datasets, and the summation of the lengths of each label k over all datasets is approximately equal for other labels in the set. An example is shown in Figure 3.7 (e), where each cell represents the number of images for label L_m that are included in the corresponding dataset D_i.The mathematical expression of modality skew is given via conditional Equation 3.5:

$$\exists i, j \in \{1, 2, \ldots, N\} : |D_i| \cong |D_j| \wedge M(D_i) \neq M(D_j) \wedge$$
$$L(D_i) \equiv L(D_j) \wedge \sum_{i=1}^{N} |L_k(D_i)| \cong \sum_{i=1}^{N} |L_{k+1}(D_i)|, \tag{3.5}$$

for all $k \in \{1, 2, \ldots, L_m\}$.

3.3.6 Feature Skew

Feature skew occurs when each participant's training sample is associated with distinct features, such as age range, patient gender, smoking status, or mortality rates. Hospitals or medical institutions may collect data based on specific intended features, or the distribution of these features may reflect the virus's behavior in different populations. For instance, COVID-19 may disproportionately affect elderly patients in certain regions compared to others.

Assuming that N hospitals utilize the same image modality (i.e., X-ray, CT, or ultrasound), regardless of the equipment version used, let D_i represent the dataset of hospital i, where all datasets have the same set of labels $L(D)$, and the sum of the number of instances for each label k across all datasets is identical for all labels m. An example is illustrated in Figure 3.7(f), where each cell represents the number of images for label L_m in the corresponding feature F for dataset D_i, with $F = \{\text{age, gender, country, etc.}\}$. The mathematical formulation of feature skew is defined as:

$$\exists\, i, j \in \{1, 2, \ldots, N\} : |D_i| \equiv |D_j| \wedge F(D_i) \neq F(D_j) \wedge$$
$$L(D_i) \equiv L(D_j) \wedge \sum_{i}^{N} |L_k(D_i)| \equiv \sum_{i}^{N} |L_{k+1}(D_i)|, \tag{3.6}$$
$$\forall k \in \{1, 2, \ldots, L_m\}$$

where:

- $|D_i|$ represents the size of dataset D_i.
- $F(D_i)$ denotes the set of features associated with dataset D_i.
- $L(D_i)$ represents the set of labels in dataset D_i.
- $|L_k(D_i)|$ is the number of instances corresponding to label L_k in dataset D_i.

This condition states that although two datasets D_i and D_j may have the same number of samples and label distributions, their underlying features differ, causing feature skew in FL.

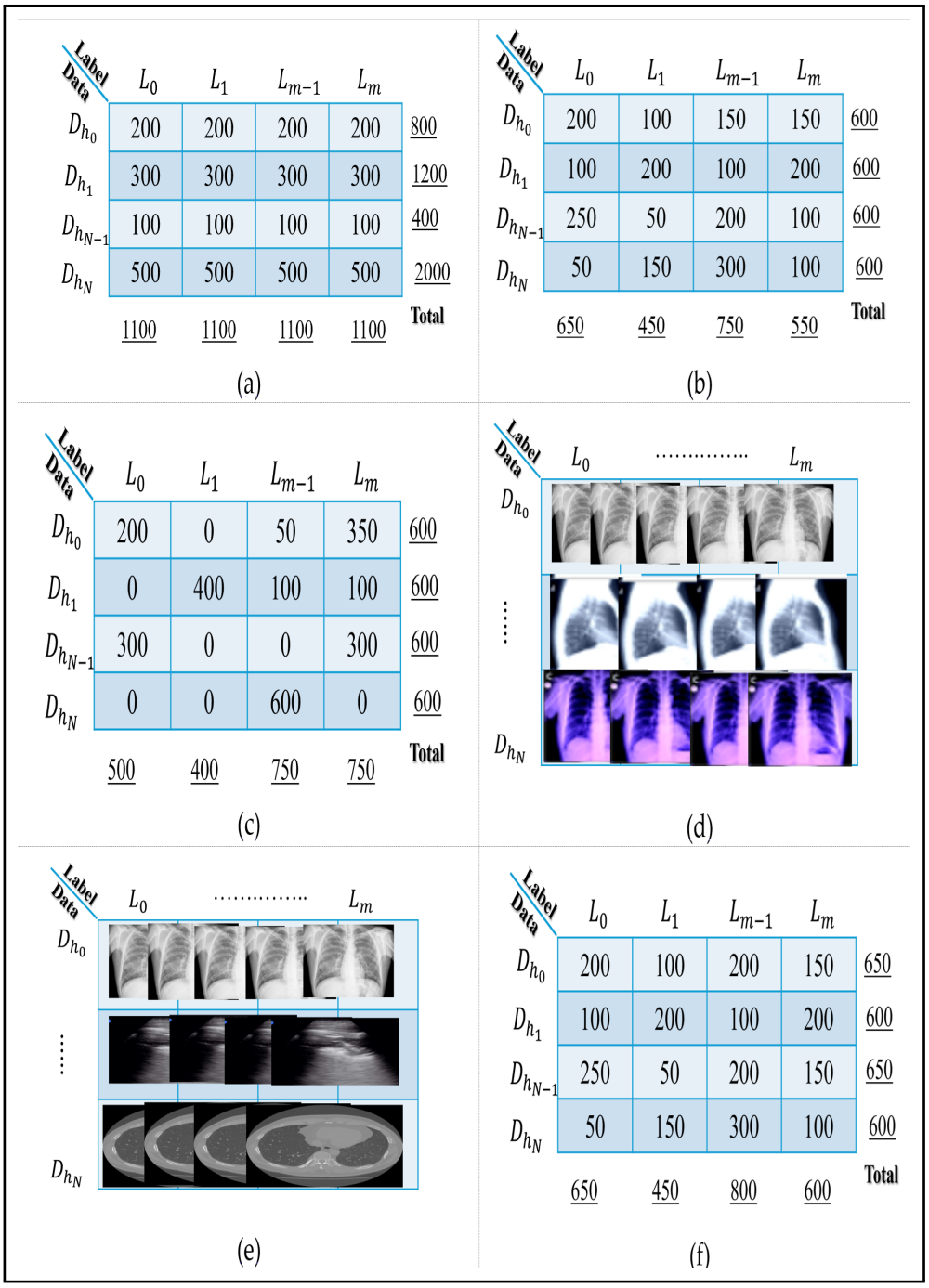


Figure 3.7: Skewness type examples including (a) quantity skew example, (b) label distribution skew example, (c) extreme label skew example, (d) acquisition protocol skew example, (e) modality skew, and (f) feature skew example.

In this chapter, we presented the detailed design of our FL system, outlining its key components, operational phases, and algorithmic workflow. The system is structured around a central server that manages communication, performs global model aggregation, and evaluates generalizability. The design incorporates multiple phases, including system configuration, data preprocessing, distributed training, model updates, and model evaluation, ensuring a robust and efficient learning process across distributed medical imaging datasets.

In the next chapter, we evaluate the effectiveness of this system through extensive experiments, analyzing its performance under various settings and comparing it against baseline methods.

Chapter 4

System Implementation

The aim of the proposed system is to allow the design of realistic experiments in medical situations, potentially assisting in delivering more commendatory clinical treatments. It introduces datasets with different acquisition devices, formatting, annotation protocols, and data skewness in seven different experiment settings. These settings are crucial for understanding how the models generalize and personalize in the distributed environments based on the real data heterogeneity settings. Furthermore, this section underscores the significance of experiment reproducibility in validating and enhancing findings in subsequent research.

This chapter details the implementation process of the proposed FL framework as following: First, the dataset characteristics and preprocessing steps are described, considering different types of non-IID data distributions. Next, the experimental setup is outlined, including hyperparameter configurations and local training strategies across participating nodes. Finally, the evaluation metrics used to assess model

performance, such as accuracy, loss, and convergence behavior, are presented. These elements collectively ensure a comprehensive analysis of the proposed approach under varying data heterogeneity conditions.

4.1 Material Data

This experimental study utilizes seven datasets, five of which have X-ray modality, and the remaining two have CT modality, as summarized in Table 4.1.

Table 4.1: Summary of COVID-19 imaging datasets and class distributions

Modality	Dataset Name	Normal	COVID-19	Viral Pneumonia	Lung Opacity	Bacterial Pneumonia
X-ray data	COVID-19 Radiography Dataset	10,192	3,616	1,345	6,012	-
	COVID-19 X-Ray with 3-classes	71	711	711	-	-
	COVID-19 X-Ray with 5-classes	71	711	711	711	711
	COVID-QU-Ex Dataset	1,456	2,913	1,457	-	-
	Pakistani Chest X-rays Dataset	60	390	-	-	-
CT data	Large COVID-19 CT Dataset	6,893	7,593	2,618	-	-
	MedRxiv Dataset	1,230	1,252	-	-	-

X-ray datasets as the following:

- The first dataset is the COVID-19 Radiography dataset [80, 81], which contains 3616, 10192, 1345, and 6012 images for COVID-19, normal, viral pneumonia, and lung opacity cases, respectively.

- The second dataset is the COVID-19 X-ray dataset , which comprises three classes: COVID-19, pneumonia, and normal, each containing 711 X-ray scans [82].
- The third dataset is the COVID-19 X-ray dataset with 5 classes [82], which are COVID-19, bacterial pneumonia, viral pneumonia, lung opacity, and normal; each class has 711 X-ray images.
- The fourth dataset is the COVID-QU-Ex dataset [83], which consists of 6,667 X-rays, including three classes: COVID-19 with 2,913 images, pneumonia with 2,618 images, and normal with 1,456 images.
- The fifth dataset, a COVID-19 Pakistani patients X-ray image dataset [84], gathers chest X-rays from 390 COVID-19-positive patients and 60 normal patients from a local hospital.

CT data sets are:

- The first dataset is the large COVID-19 CT dataset [85], which comprises two classes: 7,593 COVID-19 images from 466 patients, 6,893 normal images from 604 patients, and 2,618 CAP images from 60 patients.
- The second dataset, the MedRxiv dataset [86], comprises 1,252 CT scans positive for COVID-19 infection, 1,230 CT scans for normal lungs.

All the datasets are open-sources and accessed by the API of the Kaggle repository.

4.2 Experiment Settings

4.2.1 Data heterogeneity Settings

To evaluate the performance of the proposed FL framework, the experimental setup was categorized into three groups based on the degree of data distribution skewness. These groups simulate diverse and realistic scenarios to comprehensively analyze the framework's performance. The experiments were categorized into three groups based on data distribution: (A) IID settings for benchmark comparison, distributing data equally across four hospital nodes; (B) simple non-IID scenarios, including data quantity skew and extreme label skew to simulate real-world hospital variations; and (C) hyper-skewness scenarios with acquisition and modality skew to assess generalization in heterogeneous conditions. For A and B experiment groups, we used COVID-19 Radiography dataset with different distribution settings across training nodes. However,the central testing datasets remained consistent across all A and B experiment groups. The other datasets used in group C for hyper-skewness evaluations. In all the experimental configurations, no patient overlap was allowed between hospital subsets and central server to ensure data authenticity and privacy.

Group A: Benchmark Data Distribution (IID Settings)

In this experiment, we simulated an IID (independent and identically distributed) data distribution across four hospital nodes. The testing set was held centrally which 20% of X-ray Radiology Dataset, as depicted in Figure 4.1, to assess the global model's performance. The remaining 80% of data is for training were divided

almost equally among the four hospitals, with each hospital receiving same number of images for each class, as shown in Figure 4.2.

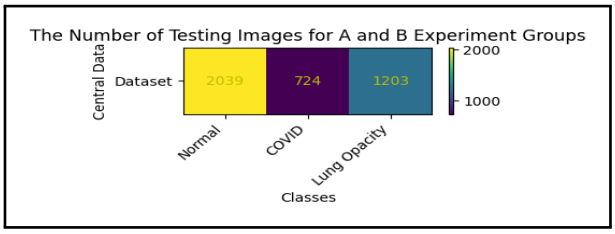


Figure 4.1: Testing set assigned in central node, experiment A and B.

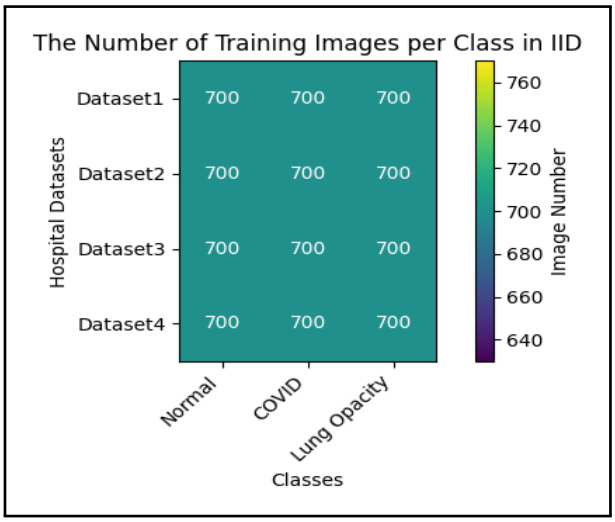


Figure 4.2: IID data setting, experiment A.

To evaluate the locality and personalization of the FL models, each hospital excluded 10% of its local training data, which were assigned as a validation set. This experiment serves as a baseline for understanding how the framework performs under ideal conditions with balanced data across all nodes.

Group B: Simple Non-IID Scenarios

The second group explores non-IID distributions by introducing two types of heterogeneity:

1. **Data Quantity Skew Experiment**

   In this experiment, we simulated a non-IID (non-independent and identically distributed) setting wherein the data distribution across the hospital sites varies in size. Specifically, Hospitals 1 and 2 were assigned larger datasets, while Hospital 3 had the smallest dataset, and Hospital 4 had a medium-sized dataset. This distribution reflects real-world scenarios, as shown in Figure 4.3, where larger urban hospitals (Hospitals 1 and 2) typically have access to more medical data, while smaller rural hospitals (hospital 3) have fewer patient records across all classes [22]. To evaluate the FL framework's performance under these conditions, 10% of the local data for each hospital were excluded from the training data for validation; these data were used to test the locality and personalization of the updated FL models. The central testing data used to assess global model performance remained the same as in the first experiment.

2. **Label Distribution Skew Experiment**

   This experiment was conducted to investigate the impact of label distribution skew across different hospital sites, a common scenario in real-world medical data settings. The same dataset was partitioned among four hospitals, with each site containing a varying number of images per label. Despite the skewed label distribution, the overall quantity of data per hospital remained relatively

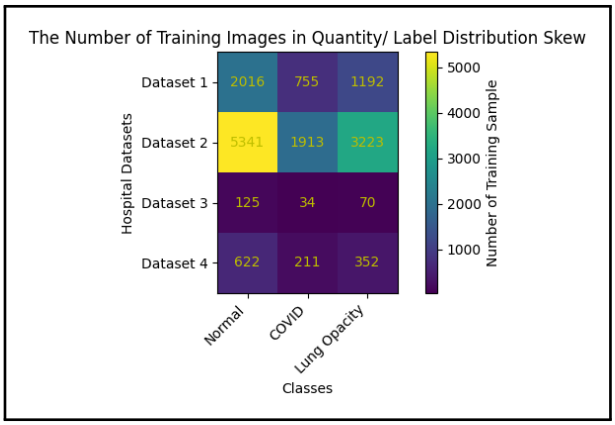


Figure 4.3: Quantity skew data setting, experiment B (1).

balanced, as illustrated in Figure4.4. The Central testing data is the same in IID experiment.

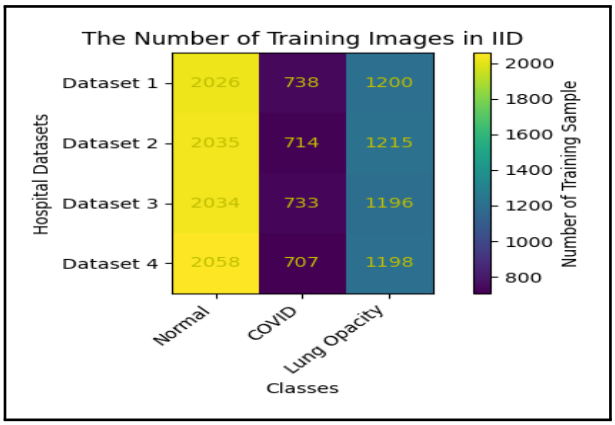


Figure 4.4: Label distribution skew data setting, experiment B (2).

3. **Extreme-Label-Skew Experiment**

   This experiment was conducted to explore label distribution skew across different hospital sites. Each hospital had a unique set of labels, with the following distribution: Hospital 1 only had data labeled as COVID-19; Hospital 2 included COVID-19 and normal labels; Hospital 3 had COVID-19, lung

opacity, and normal labels; and Hospital 4 had all three labels (COVID-19, normal, and lung opacity), as shown in Figure 4.5. This setup introduced extreme label distribution skew across the hospitals, which challenged the FL framework to adapt to such imbalanced label distributions. As in the previous experiment, we used the same central testing data to evaluate the global model's performance. This experiment helps assess how the framework handles label absent across distributed datasets in a real-world medical scenario.

Figure 4.5: Data settings for extreme label skew, experiment B (3).

Group C: Hyper-skewness Experiments

In the hyper-skewness scenarios, experiments were designed to reflect realistic conditions, wherein each hospital site trains its model on datasets exhibiting multiple dimensions of variability, such as differing capture protocols, types of imaging equipment, label distributions, and storage formats. These experiments were divided into two categories: one assumed that all training sites use similar imaging modalities, while the other introduced modality variations across at least one site. The primary aim was to assess the FL framework's ability to handle extreme data heterogeneity in distributed environments.

1. **Experiments Involving Acquisition Skew with and without Extreme Label Skew**

   In this experiment, we aim to evaluate the generalization and personalization metrics in relation to fully distributed datasets with and without extreme label skew. Each hospital site was assigned distinct dataset resources, comprising the COVID-19 Radiography dataset, the COVID-19 X-ray datasets (with three and five classes), and the COVID-QU-Ex dataset, with no data reserved for central testing. The central server instead evaluates the global model using external testing datasets as the following:

   a Extreme label skew:

   Each site contains varying numbers of labels. The data for each label are split into 90% for training and 10% for local validation, as shown in Figure 4.6. For central testing, an external dataset of COVID-19 Pakistani patients, comprising two labels, was utilized, as illustrated in

Figure 4.7.

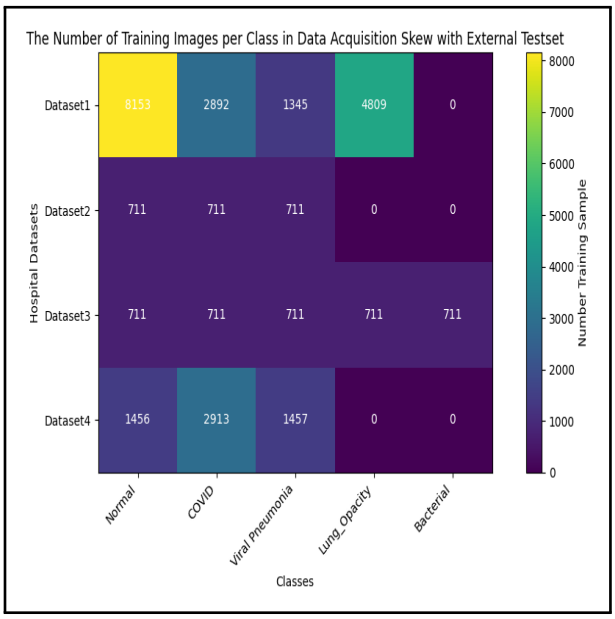


Figure 4.6: Training data settings used in data acquisition, experiment C (1.a).

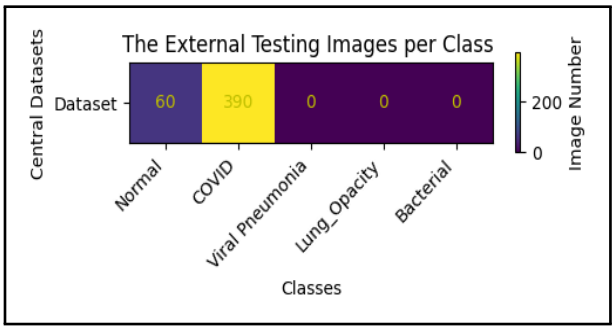


Figure 4.7: External data for testing the central model, experiment C(1.a).

b Fixed label distribution:

The same experiment was repeated with a standardized label set across all sites, retaining only “normal” and “COVID” cases. This fixed-label setup ensured a consistent evaluation of the model’s ability to handle reduced label variability, as shown in Figure 4.8 and Figure 4.9

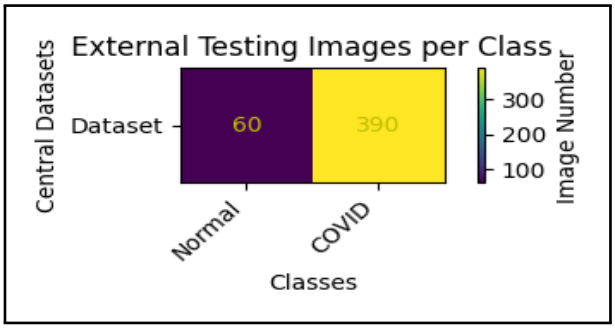


Figure 4.8: External testing data in the central node, experiment C (1.b).

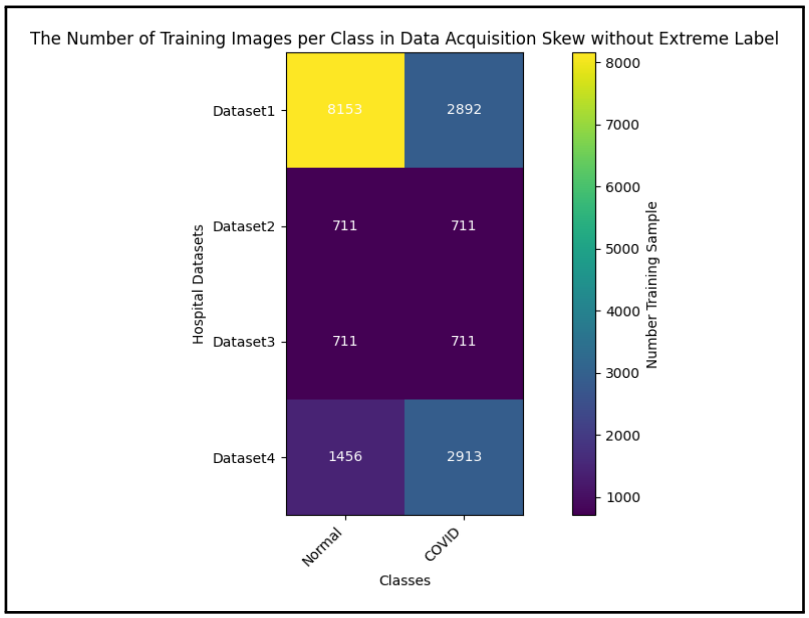


Figure 4.9: Training data settings for data acquisition, which have the same modality without extreme label skew, experiment C (1.b).

2. **Experiments regarding Modality Skew with Internal and External Data for Central Test Set**

   In this series of experiments, we investigated the impact of modality skew on the performance of FL by simulating scenarios where different hospital sites utilize datasets obtained using varying imaging modalities. Inspired by similar work conducted by Qayyum et al. [75], in which X-rays and ultrasounds were used across distributed nodes, we extended this approach by incorporating both X-ray and CT modalities into our experiment. Each hospital site was assigned a unique dataset resource: three sites used X-ray modalities sourced from the COVID-19 Radiography dataset and the COVID-19 X-ray datasets (with three and five classes), while one site used CT modality data, specifically from the large COVID-19 CT scan dataset.

   a Internal data testing:

   In the first step of this experiment, the four datasets mentioned were distributed across four hospital sites, as shown in Figure 4.10. Subsequently, 20% of each dataset was reserved for testing at the central server using internal data settings, allowing the evaluation of the global model's performance, as depicted in Figure 4.11.

   b External data testing:

   Here, the same experiment setup was repeated, but this time, the training data across the four sites consisted of the full datasets with common labels (normal and COVID-19), as shown in Figure 4.12. External datasets—comprising images obtained via both X-ray and CT modalities—were then used for testing at the central server which are COVID-

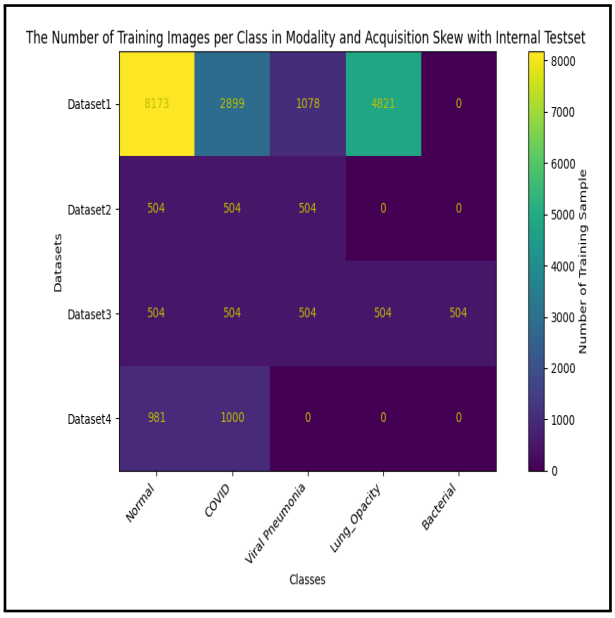


Figure 4.10: Training data for modality Skew experiment with internal testing set, experiment C (2.a).

Figure 4.11: Central testing data for experiment C (2.a)

19 Pakistani X-ray dataset and MedRxiv dataset, as illustrated in Figure 4.13. This approach evaluated the model's ability to generalize across modalities, with the global model being tested on external data sources.

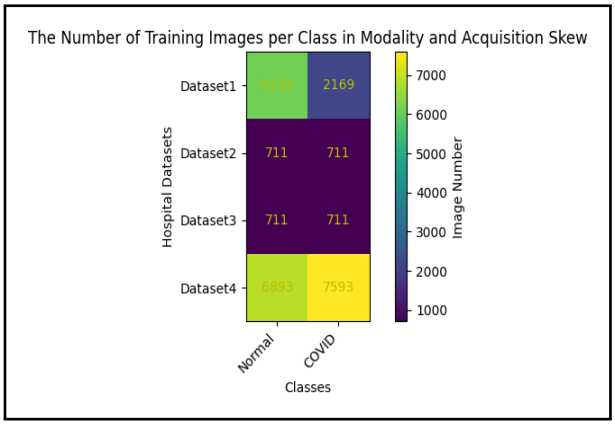


Figure 4.12: Training data for modality skew with external data test, experiment C (2.b).

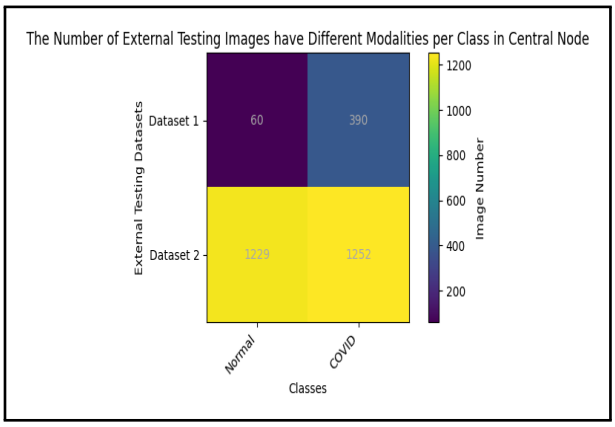


Figure 4.13: External central test data with modality skew x-ray and CT datasets, experiment C (2.b).

These experiment setups were designed to assess the robustness of the FL framework

in handling data heterogeneity, specifically when imaging modalities vary across distributed sites.

4.2.2 Hyperparameters Settings

All the image datasets were resized such that they had dimensions of 224 × 224, collectively representing the three-color channels commonly associated with RGB images and making up the input data. The images were then subjected to common advanced data augmentation techniques [87], such as random rotations, flips, and scaling, to enhance the diversity of the training data.

The model hyperparameters were the same for all experiments, as summarized in Table 4.2, FL was used to train the distributed data for 15 rounds to prevent any disruption to the Google collab time limit. The local epoch was set to 10 to prevent overtraining bias [88]. The batch size for all training sites was set to 32 to prevent RAM overflow and crushing, and the batch size for central testing was set to 64. The learning rate was tuned with a cost-consuming consideration of 0.001 and an SGD optimization function, resulting in a lower loss value [25]. In order to prevent randomness in sampling [32], we assumed there was a complete sampling of all training sites in each round. This allowed us to evaluate the aggregated model centrally and across all hospital sites, allowing for precise measurements of generalization and personalization.

Table 4.2: The settings of the model hyperparameters.

Parameters	**Values**
Round number (R)	15
Local epoch (l)	10
Train/Test batch size (BS)	32 for distributed training / 64 for central testing
Learning rate (λ)	0.001
Optimizer function	SGD
Number of participants	4, assumed to be working in parallel
Aggregation strategy	FedAvg
Image size	224×224
Augmentation methods	Training set: Images undergo random horizontal flips and are normalized with the specified mean and standard deviation for each channel (RGB). Testing set: The procedures applied are similar to those used in training but exclude random horizontal flips.

4.3 Evaluation Metrics

To evaluate the FL system of categorizing lung diseases and COVID-19 infection based on lung image datasets, we used statistical performance metrics like accuracy, which is commonly used in epidemiological and medical research, for all sites [46]. The confusion factors are the true-positive (TP), false-positive (FP), false-negative (FN), and true-negative (TN) values, which are defined below:

- TP—correctly predicted result;
- FP—incorrectly predicted result;
- TN—correctly predicted no-event value;
- FN—incorrectly predicted no-event value.

Based on the aforementioned factors, the accuracy metric regarding FL was calculated for each site, as shown in Equation 4.1:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{4.1}$$

Furthermore, the loss value is a numerical value generated and calculated via the loss function during the training process. It serves as a crucial indicator, revealing whether a model effectively converges towards an optimal solution. Loss values provide vital insights into the performance and learning path of a model over the training rounds in the FL framework. The optimizer then receives these values, identifying updated weights that enhance the model's accuracy and overall performance, thereby ensuring a robust and efficient learning process is maintained [13]. We used a multiple cross-entropy function, as shown in Equation 4.2, where y denotes the true labels for class i, p is the predicted probability for class i, and the summation is taken over all classes C.

$$H(y, p) = -\sum_{i=1}^{C} y_i \cdot \log(p_i) \tag{4.2}$$

In summary, we provided a detailed overview of the implementation process, covering material data, experimental settings for different types of non-IID data distributions, hyperparameter configurations, and evaluation metrics. These components form the foundation for assessing the effectiveness of the proposed FL approach. The next chapter presents the results and discusses their implications in relation to model performance, generalization, and robustness under varying non-IID data conditions.

Chapter 5

Results and Discussion

This chapter presents the experimental findings of the proposed FL framework. The results are analyzed based on key performance metrics, including accuracy, loss, and model convergence. Additionally, the impact of data heterogeneity on model performance is examined. The discussion further interprets these results, highlighting research directions, challenges, and potential implications in FL for medical imaging.

5.1 Results

We categorized the framework implementation results based on the experiments discussed in the implementation chapter . As mentioned earlier, our goal was not to improve the performance of the FL system. This chapter presents a comparative analysis of the global model's generalization performance and localization perfor-

mance and the personalization performance of the updated models across all training sites in each experiment, organizing them into the following result groups.

5.1.1 Group A Results (IID Setting)

Under IID settings, all evaluation types show high and balanced accuracy, reflecting stable and reliable generalization across participants. Accuracy values are consistently high and uniform across all sites.The difference between local evaluation and personalization is minimal, showing stable learning behavior. Generalization on central testing data achieves 0.87, slightly lower but very close to local/personalized performance as shown in Figure 5.1.

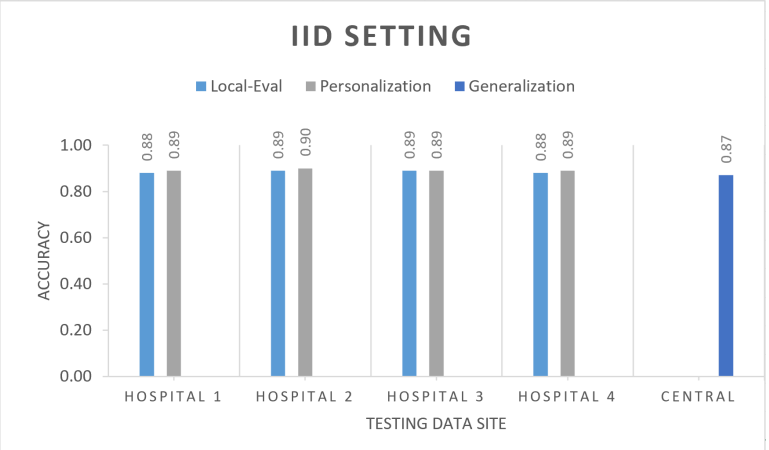


Figure 5.1: Results regarding IID distribution.

5.1.2 Group B Results (Simple Skew)

Results Regarding Quantity Skew

The local model evaluations were similar for all hospitals (85–87), as shown in Figure 5.2. At most hospital sites, the personalization model evaluations were slightly better than the local ones after aggregation, indicating an improvement in the global model's ability to highlight a useful lung disease feature collected from distributed sites. The IID setting led to a generalization metric that closely matched the average personalization accuracy across all sites. All the metrics yielded similar results, indicating the applicability of model updates for local data, both before and after aggregation, across all sites as well as for the central testing data. These findings reveal the stability of the improvements at all sites over the training rounds in the case of IID homogeneously distributed data.

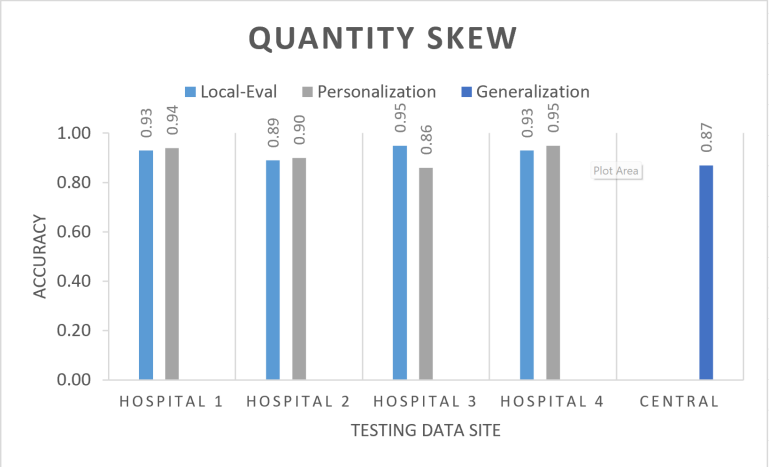


Figure 5.2: Results regarding quantity skew.

Results Regarding Label Distribution Skew

The evaluations of the local model vary depending on the specific set of hospital data. Figure 5.3 shows that most hospital sites (1, 2, and 4) achieved a higher personalization. In contrast, Hospital 3 observed a reduction in personalization due to local overfitting in training caused by limited training data. The accuracy of the generalization metric was lower than the average of personalization across all sites, that is, 91. This suggests that the global model is more robust for sites with numerous and varied data, such as Hospitals 1, 2, and 3.

This finding indicates that the reported generalization metric, with a quantity/label distribution skew, yields performance similar to that under IID settings, while a personalization metric yields better results for this skew type. The generalization metric achieved similar results in the IID case; references may have been made to the same datasets with different partitioning strategies.

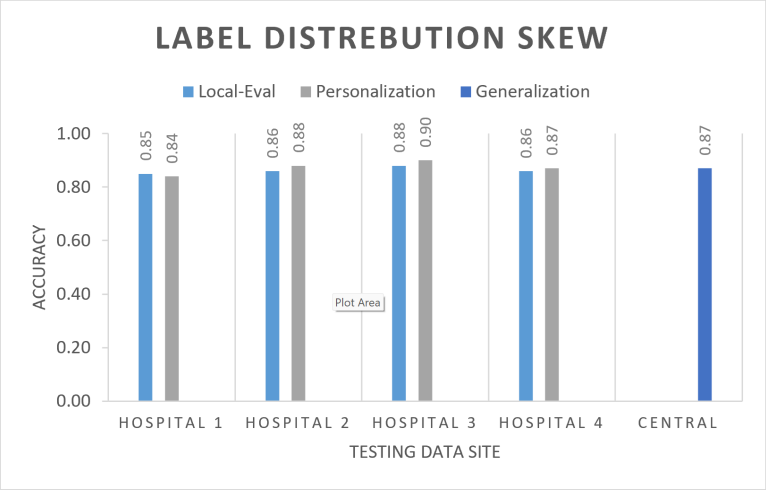


Figure 5.3: Results regarding label distribution skew.

Results Regarding Extreme Label Skew

As shown in Figure 5.4, most hospitals are missing one or more labels, and the locality metric exhibits overfitting in training, leading to a biased model. Personalization models address the issue of bias for sites with fewer data and labels. Hospital 1 has little data and only one data label, causing overfitting in the local model (reported accuracy = 100). However, the aggregative approach mitigated the impact of model bias and resulted in a more robust model, as demonstrated by the reported result of 93% for the personalization model. At Hospital 2, there are two labels, namely, normal and COVID-19, and the results of the local evaluation and personalization are convergent. Also, Hospital 3 has two labels but shows variations between the local and personal models. Hospital 4 has all data labels and shows convergent results for both local evaluation and personalization. Thus, Hospitals 1 and 3 have more data falling under the normal class in their datasets; therefore, the aggregative model was trained to a greater degree on the features of normal images at distributed sites. So, higher accuracy was achieved on the testing set that contains normal images. However, the smaller size of lung opacity images at the distributed sites resulted in a larger gap between the global and local models.

The experiments showing results with extreme non-IID reveal the lowest accuracy in terms of generalization. However, the global averaging model can enhance the local model's performance for participants trained on a specific label. Because the variety of their own data is very limited (i.e., their data are split from the same resource), overfitting during the training process leads to high bias in the DL model. This type of skewness necessitates further investigations into the behavior of updated models

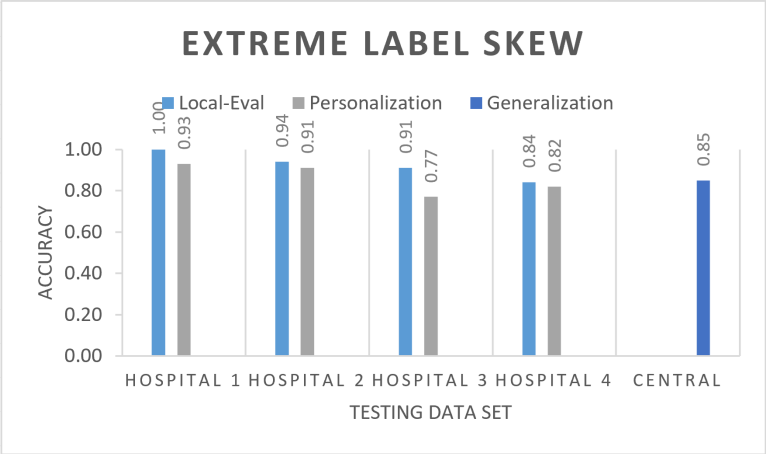


Figure 5.4: Results regarding non-IID distribution with extreme label skew.

at each site, both before and after aggregation.

5.1.3 Group C Results (Hyper-Skewness)

Results Regarding Acquisition Skew with and Without Extreme Labels

The results of the experiments with and without extreme label skew revealed the great stability of the training process over the rounds, as shown in Figure 5.5. With the extreme label skew, the prediction accuracy for external data decreased to less than 1% (as shown by the blue curve in Figure 18 for rounds 7, 11 and 15) due to distractions in the training process caused by variant labels in distributed data with insufficient numbers, which were often available in one place and absent in another.

Moreover, Figure 5.6 shows a comparison between the loss prediction values for the central model and the average of the personal model in all training sites with

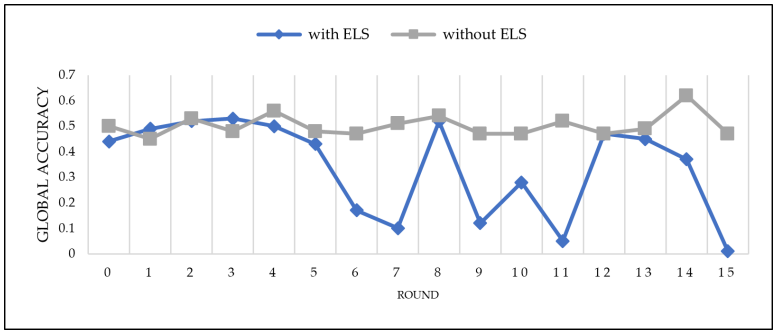


Figure 5.5: Generalization accuracy with and without extreme label skew.

and without extreme label skew, revealing lower loss prediction in the internal data locally and centrally in the external data without extreme label skew. Also, in this case, extreme label skew led to an oscillating loss curve, which reflects bad training data. Therefore, in the case of data acquisition skew, it could be helpful to generalize the global model on a new dataset that has the same training labels in all distributed sites and has undergone more training rounds. That means extreme label skew causes divergence and instability in improving accuracy over training rounds in FL.

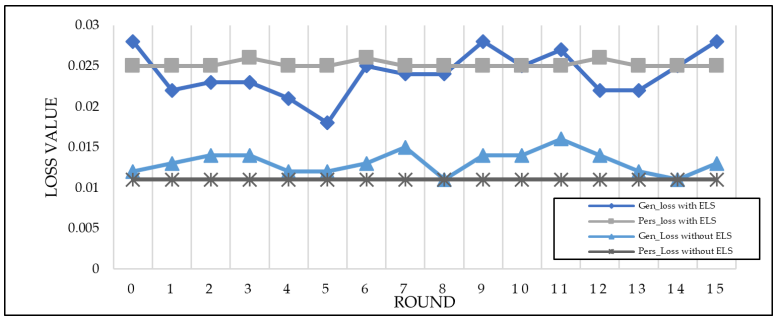


Figure 5.6: Generalization and personalization loss values with and without extreme label skew.

Results Regarding Modality Skew with Internal and External Data

In Figure 5.7, the local internal data seems similar to the local external data for all hospitals, except that Hospital 3's local internal accuracy was reduced to 0.19%, which might have occurred because there were more labels in this hospital's dataset. Moreover, the values of the personal external data for all the hospitals are significantly higher than those for the personal metric subjected to internal data testing. However, the personalized metric that underwent internal testing in Hospital 3 was used to train a model using one more unique label: bacterial pneumonia. This site combined two skew-type extreme labels with acquisition skews.

The generated personalization model could not accurately classify the local data. However, the ratio of the locality to personalization metrics in the external testing case better reflects the ability of all sites to contribute to building a global model and benefit from the resulting local updates. The generalization metric for the internal testing data is 6% higher than that for the external test data. This result was expected because the global model was tested on the same data that it was previously trained on. However, both of them exhibited insignificant improvements over 10 local epochs and 15 rounds.

Figure 5.8 shows a comparison between the loss prediction values for the central model and the average loss of the personal model at all training sites with internal and external testing data. It reveals lower loss prediction locally and centrally for the external. This result might be due to the limited prediction labels in the external data, which consist of only two labels in each modality, while there were five prediction labels for the internal data. Therefore, modality skew with extreme label

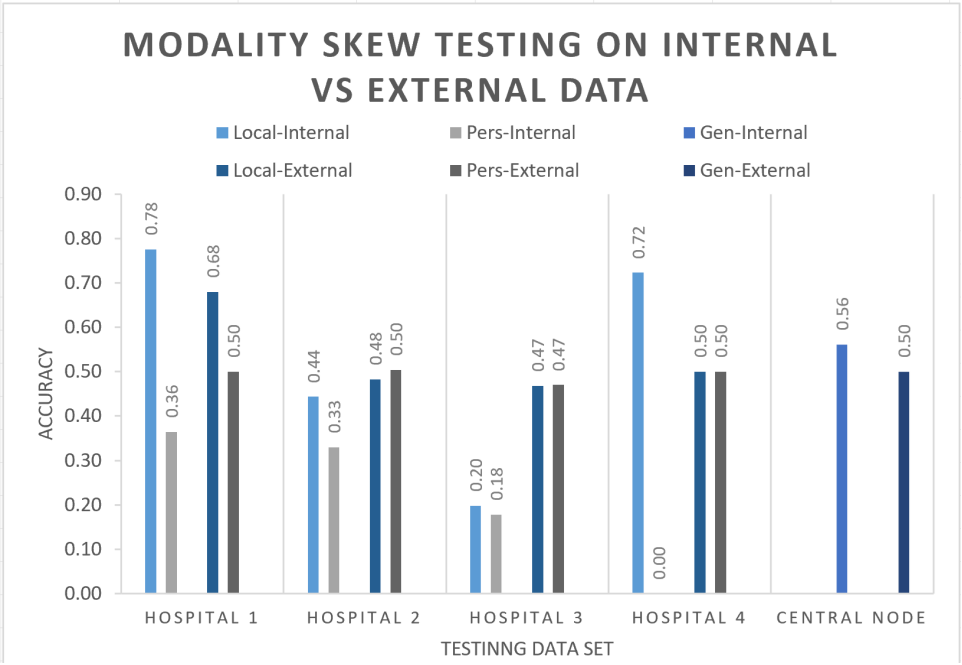


Figure 5.7: Accuracy results regarding modality skew testing for internal vs. external data.

skew led to higher degradation in the performance accuracy of FL, especially for the personalization metric.

Figure 5.9 shows the confusion matrix of the internal testing data; it visualizes the prediction of five labels, and the label with the truest prediction has the viral pneumonia label. The prediction number for the bacterial pneumonia label is zero, and most images were erroneously predicted to be viral pneumonia. That means that the aggregate model could not generalize the features that appear at the single hospital site. Therefore, the prediction of unique labels in FL is an unsolved problem that requires further investigation.

On other hand, the confusion matrix for the external test data revealed a higher true positive prediction rate for the CVOID-19 label, as Figure 5.10 shows. However, it

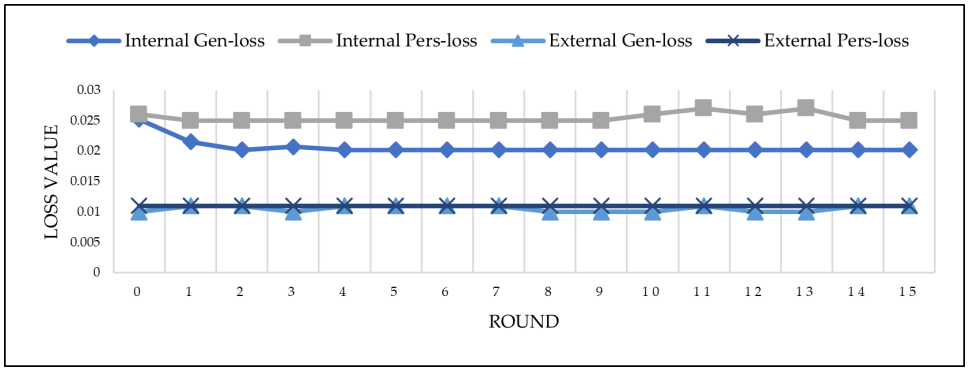


Figure 5.8: Results regarding generalization and personalization loss for internal and external datasets.

did not identify normal cases. The precision is 50% for both classes, indicating a severe imbalance in prediction behavior. Reflects a lack of robustness and fails to generalize to real-world situations.

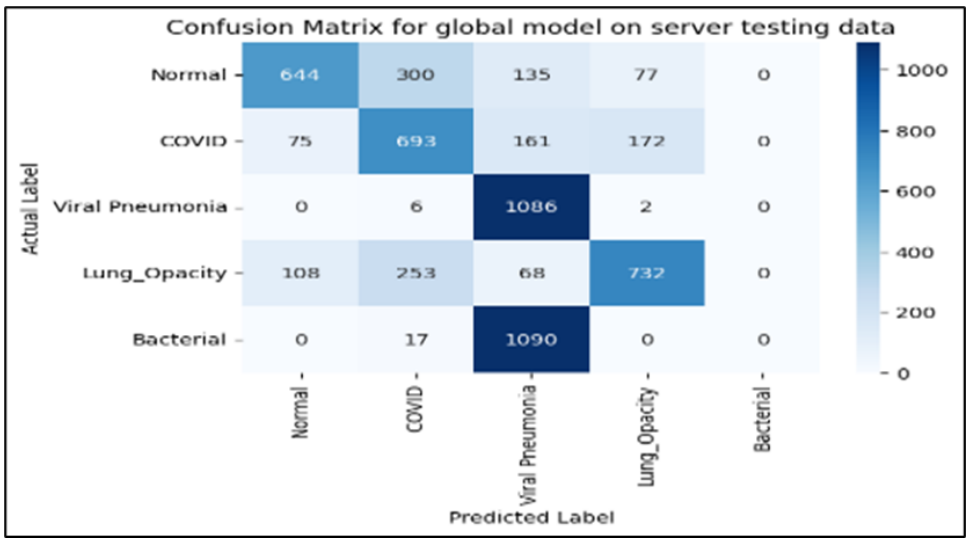


Figure 5.9: Confusion matrix of modality skew for testing using external data.

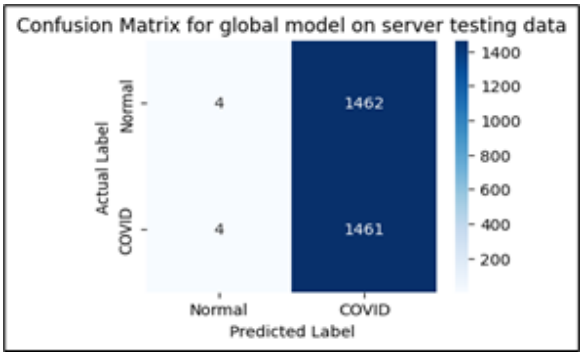


Figure 5.10: Confusion matrix for modality skew regarding testing on the internal dataset.

5.2 Discussion

When FL is used for COVID-19 lung imaging data, data heterogeneity can significantly impact classification performance. In our results, we observed that the generalization metric yielded similar results in the IID case, indicating that a skew in the quantity/label distribution for similar features of distributed data—such as in this experiment, where a single dataset was distributed across multiple sites—does not have a negative impact on the global model. However, an extreme label distribution was found to have a more negative impact on the generalization metric. The averaging equation interpreted this by aggregating data of the same size and model architecture in various ways throughout the rounds. Therefore, the global models appeared to be similar in experiments of group B, all of which were conducted using internal central data. This issue primarily affected the personalization metric, potentially leading to biased results for certain sites.

From the results, we gleaned that the larger gap between the results of locality and personalization metrics indicated an issue in the training process for the site in ques-

tion; if the locality accuracy was higher, that means there was local model overfitting and aggregation models from the other site would be able to overcome this locality issue. However, if the personalization metric was higher, then the aggregation strategy of the updated models succeeded in creating robust global models, and the opposite of the above is true. Thus, FL provides a solution to two common issues: trying to obtain a better balance of biased weights for sites with overfitting when training using a small dataset and when sites have a large dataset with scarce or vague features.

Therefore, hospital sites with limited datasets can acquire significant benefits from participating in FL partnerships. This is most likely a result of FL's ability to reduce bias in models trained on homogenous data and capture greater variation than local training. In one hospital or population, a demographic or age group that is underrepresented may be well-represented in another, such as children who may reflect features different from COVID-19, including other diseases detected via lung imaging [3].

An instance where data acquisition skew may occur is when one has two X-ray images: one from a hospital with modern, high-resolution equipment and another from a hospital with older, less sophisticated technology. The image from the first hospital might have clear, sharp features, making lung abnormalities like opacity more visible, whereas the image from the second one could appear grainy or blurred, obscuring those same features. This variation in image quality directly influences how well a model can classify the data. A model trained primarily on high-resolution images may perform well for similar high-quality data but struggles when faced with lower-quality images. This mismatch can reduce a model's ability

to generalize across different hospitals with varying imaging systems. Such challenges have been extensively observed in medical imaging studies, where equipment and protocol differences can lead to biased model performance, favoring data from higher-quality sources while neglecting others [21].

In regard to modality skew, shifting the application of FL training from X-ray to CT data challenges a model's adaptability. A model trained exclusively on one modality often struggles to interpret data from another, as features appear in different forms. This gap highlights the necessity of training across multiple modalities to ensure robust performance across varying imaging technologies [74].

In the hyper-skewness experiments, we dealt with medical imaging data that were highly varied because they were obtained from various sources and using different modalities, such as X-ray and CT imaging. Aggregation methods such as FedAvg are likely to fail when extreme label skew is accompanied by data acquisition or modality skew. The results show that broadcasting the model from the central node to the distributed sites with different modalities led to worse performance, leading to findings that indicate the infeasibility of training a model using distributed data obtained with varying data modalities. We could even conduct an infare experiment for generalizing result on out-sample even with a clustering architecture [74] or by integrating more than one modality at a training site [73]. This is because sharing the same global model with different modalities leads to significant differences in the updated models over various training rounds. Variations in image texture and lung view angle and the incorporation of different organs across modalities lead to the creation of bad training data and, thus, the generation of an unstable prediction model. Most image classification models update weights based on detecting edges

during the convolutional process of the training model. The local training model repeats this process over several epochs to optimize the weights corresponding to the features of edge pixels. Over rounds, the aggregation strategy averages these weights, resulting in a global model that suffers from underfitting. This explains the significant decline in the generalization and personalization metrics for group C's outcomes.

In summary, as illustrated in Figure 5.11, extreme label skew constitutes a critical red flag for distributional fairness. The figure depicts a spectrum of data skew types ordered by their impact on FL performance, ranging from "Worst Skew Performance" on the left to "Adequate Skew Performance" on the right. Modality skew ranks as the second most detrimental, since discrepancies in imaging modalities lead the model to learn features that do not generalize across domains. By contrast, acquisition skew and quantity/label distribution skew are comparatively more tolerable or amenable to correction.

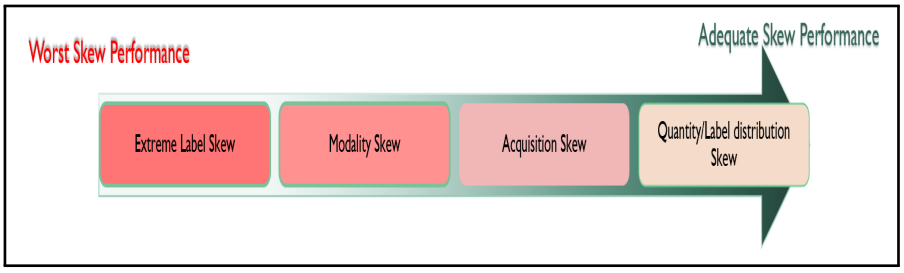


Figure 5.11: A spectrum of data skew types ranked by their impact on model performance

Furthermore, training with hyper-skew types with extreme label skew led to significant degradation in the personalization metric. The results of experiments C revealed there were no predictions of a unique label for a single site because of the aggregation of distributed models averaged over the size of the data. Therefore, the

unique label size that exists for one site across all the training samples might have a low ratio. This requires more investigation with variant data samples to make a fair judgment about the feasibility of training real distributed data with extreme label skew without preprocessing to unify the prediction labels. Further investigations required to shift hyper-skewed datasets from the worst-skew region toward the adequate-skew region of the heterogeneity spectrum.

The final finding supports the notion that the purpose of distributed training data is to enhance the model's generalization and lessen the influence of overfitting caused by acquiring variable data obtained with similar modalities. However, we should prioritize preserving the specialization of local features like labeling, annotation, acquisition protocols, and capture devices, all of which are associated with personalization models.

Chapter 6

Conclusion and Future Works

This chapter summarizes the key findings and contributions of this research, highlighting the effectiveness of understanding data heterogeneity challenges of FL in medical imaging. The study explored the impact of different non-IID data distributions on model performance and demonstrated the effectiveness of training COVID-19 lung imaging data under varying conditions. The implementation details, including dataset preparation, hyperparameter tuning, and evaluation metrics, were systematically designed to ensure a comprehensive analysis of FL in a realistic COVID-19 lung imaging scenario. Furthermore, this chapter discusses the limitations of the current work and proposes potential directions for future research. It summarizes the challenges posed by data heterogeneity, privacy concerns, and computational constraints. Most of the findings could be generalized to the field of medical images for practical deployment in real-world healthcare applications.

6.1 Conclusion

FL provides promising solutions to mitigate privacy risks that hinder the full utility of medical data. It significantly enhances the applicability of various hospitals and medical institutions worldwide, enabling the construction of a global model. This provides the medical field with more generalizable results, emphasizing the relationships between the various COVID-19 variables. This is crucial for ensuring the results' generalizability across a larger population. Furthermore, it is crucial for the FL facility to maintain updated models with newly generated data from various sources, ensuring they align with current trends and strains, particularly given the evolving nature of COVID-19 [89].

This study discussed one of the main obstacles challenging FL's applicability, which is the lack of standardization of medical imaging data, and this rise in heterogeneous data across distributed hospitals and medical institutions is called non-IID. The non-IID issue arises when data distributions across different sites are heterogeneous, leading to challenges in model training and generalization. In this research, the types of skewness in data heterogeneity are described as the real situation mentioned in the radiologist field [90]. As described, the types and sources of data skewness can hinder the aggregated model's performance. This variety can lead to unfair model updates, where some hospital site data have a greater impact on the overall model than they should, making the model less useful across different patient populations and medical settings. Non-IID data can also worsen overfitting problems because models may become too specific to the features of a few medical resources, and not enough attention is paid to the wider range of variations in the smaller resources.

6.2 System Limitation and Future Works

This work presents an experimental study on the impact of non-IID data in FL for medical imaging, analyzing how different types of data heterogeneity affect model performance. Through extensive experiments, we evaluate the challenges posed by skewed data distributions and assess their implications for FL convergence and accuracy. Our findings provide insights into the limitations of FL under non-IID conditions and highlight potential directions for future improvements.

One limitation of this study is the inability to implement feature skew due to the anonymization process of open-source datasets. Privacy regulations enforce strict measures to prevent patient record leakage, restricting access to certain features in medical images and limiting the ability to analyze feature distribution imbalances effectively.

Another major challenge is the quality of training data. According to radiology experts, many open-source DICOM datasets obtained from published studies or educational platforms do not adhere strictly to DICOM standards [4]. These datasets often exhibit low-quality acquisition [90] and weak standardization, which can impact the reliability of model training and evaluation.

Furthermore, computational constraints posed a significant challenge. The paid Google Colab server, which limits execution time to a maximum of 24 hours per session, restricted our ability to conduct prolonged training experiments. Despite utilizing the highest available GPU and RAM configurations, execution failures were encountered when the training batch size exceeded 23, underscoring the platform's computational limitations.

Future research should focus on addressing data heterogeneity challenges in FL by developing formal metrics to quantify the degree of data skew, enabling precise characterization of label, modality, and acquisition disparities. Adaptive strategies should be implemented to dynamically adjust key training parameters based on the specific skew type, ensuring more stable convergence. Additionally, introducing a Trainability Score could help predict the likelihood of successful training outcomes, guiding model selection and configuration. Finally, creating benchmark datasets with controlled skew types would allow for systematic evaluation of model resilience and generalizability under diverse heterogeneity scenarios.

6.3 Research Directions

Table A.1 in the appendix 1 shows the summarization of original papers that used FL framework to analysis and training COVID-19 lung medical imaging. Several research studies on the non-IID problem focused on comparing the performance of different models and the model's ability to achieve stable accuracy results over the training rounds [24, 89]. Other studies looked at how different hyperparameter settings affected the results of FL for different types of data skewness and tried to find a logical relationship between these factor settings in different situations of data heterogeneity [25]. In addition, those research studies focused on examining the various types of distributed data skewness and providing solutions to mitigate their impacts.

Many other common issues within the FL framework, such as privacy and security, communication, and system resources. To ensure FL's success in various medical

applications beyond the diagnosis and segmentation of COVID-19 infection from lung imaging data.The following outlines the directions of common technical issues.

6.3.1 Communication Issues

The communication links between servers and participating hospitals were used to share the weights of DL models for local medical image training. In the COVID-19 situation, reliable/synchronized communication is required to ensure the ability of participants to measure each contribution to the global model computation. Therefore, improvements in communication might adjust the links based on three factors: the size of the model, the number of rounds, and the available bandwidth. Medical image training typically involves a large model with millions of weights, necessitating the upload and download of numerous megabytes. For example, ResNet models are the most successful for COVID-19 images, and 3D-ResNet101 is 85.21 MB in size [62].

Similarly, size presents a challenge to FL performance, especially with the encryption overheads of privacy-preserving methods. The fusion model successfully minimizes the classification model, yielding satisfactory outcomes [73]. However, further research is required to understand its influence on various non-IID skewness. The integration of FL and knowledge distillation boosts the volume of computational data, leading to an improvement in communication efficiency and training throughput. When evaluated on two sets of medical image segmentation datasets, their results demonstrate data privacy, reduced communication costs, and improved accuracy when using TransUNet and ResUNet as teacher models [91].

The number of rounds is essential to consider; as mentioned, the COVID-19 medical situation requires a lower number of rounds than the FL implementation in another domain. However, many experimental studies [25, 92, 30, 32, 62] provided results after 15 rounds as a minimum for their evaluation, with some studies requiring up to 200 rounds. Observing the ratio of improvements in global model accuracy with increasing numbers of rounds revealed that it became insignificant after a few rounds in non-IID situations. However, in the IID situation, the global model convergence significantly improved after a few rounds, starting from the first 5–10 training rounds. Chowdhury et al. [70] observed stable convergence after the third round. To determine the stopping point of systems with low-performance improvements, further experiments are required.

6.3.2 Privacy and Security Issues

FL frameworks provide partial privacy and security. Guaranteed privacy is required to avoid reverse engineering attacks against local model updates, and numerous attacks have been reported in the FL system, as illustrated in Table 6.1. It is possible to implement privacy-preserving methods over three entities in FL systems, namely, images, local updates and global models, and communication channels, which can prevent any re-identification and reconstructor attacks. Reducing image quality by adding an amount of noise significantly degrades model performance [35]. Finding methods to preserve patient privacy requires more effort. This method works for both local and global models. It adds noise to the model's weights when it collects models from sites or uses on the weights that are shared (step 3 of Figure 2.1 or step

Table 6.1: Types of Attacks in FL

Attack Name	Description of Impact	Methods
Reconstructor Attacks	The image features are retrieved from the local updated weights.	DP/HE.
Poisoning Model	A local model trained on fake labels or irrelevant datasets aimed at harming a global model is uploaded.	Measure the quality of local updates, which is still an open issue in FL systems.

1 of Figure 2.2). HE has a lower impact on model accuracy because it depends on the encryption on the local side and the decryption on the server side. However, it has higher computational overheads than DP [8]. The Secure Multi-Party Channel or method encrypts communication channels between servers and participant nodes. Reports have indicated that it boasts a high number of communication channels [35], while blockchain-based systems are also used to maintain secure communication channels [17] and improve the accountability and accessibility of FL frameworks [30, 93]. For further details about FL attacks in the medical image domain, Kessie et al. [10] provided a comprehensive review of secure and private methods for FL in medical imaging.

6.3.3 System Resource Issues

For hospitals and medical institutions, computational resources are usually limited. High computational resources, GPUs, bandwidth, and secure storage are essential for an efficient FL framework. The complexity of tuning hyperparameters in an FL setting and detecting errors in deployment and configuration over various resources

may necessitate the involvement of numerous technical experts [60]. In FL, it is crucial to understand that every single training round functions completely independent of the participating sites. The selecting of optimal local epoch, batch size, learning rate, training model, and number of rounds are factors significantly influence on the robustness performance [48]. Abdul Salam et al. [25] identified which factors affect model accuracy and loss, such as activation function, model optimizer, learning rate, number of rounds, and data size in IID settings. However, in the context of non-IID medical imaging, further investigation is required. We suggest classifying the configuration factors into high impact and low impact, with the aim of reducing negative performance.

6.4 Recommendations

FL in medical imaging succeeded in creating a private environment to process these sensitive data. However, it suffers from inconsistent results because the training relies on insufficient medical imaging data for COVID-19. Many open-source datasets are published in technical communities such as Kaggle and GitHub [94]; unfortunately, these datasets have low quality and are not confirmed by validation or testing radiology methods to support measurement experiments. Data quality plays a vital role in the analysis of information about COVID-19 and patients, and to avoid the pitfalls of FL applications, the garbage in, garbage out theory must be considered. To be useful for technical applications, each collection of medical images must pass evaluation and validation checks performed by more than one radiology expert. In addition, a dataset must have annotated labels, segmentation, and reports based on

AI application needs, such as the RSNA dataset [95] and BIMCV dataset [96]. The findings advise medical institutions and hospitals to collaborate with ML researchers to access high-quality datasets. Researchers note that the DICOM standard provides a step forward for FL because it eliminates the challenges of data heterogeneity in distributed environments [10].

As well as heterogeneity of software and hardware, are variables that could be difficult to tune efficiently without in-depth investigation and practical experimentation. Identifying the FL framework with effective hyperparameter settings for medical images needs further work. It is important to study the trade-off settings in the case of non-IID data, such as the number of local epochs against GPU consumption, the number of rounds against accuracy, the number of participants, and the convergence rate against the batch size. Furthermore, more experiments are needed to define the proper aggregation strategy, activation function, and generalizability of the global model without compromising the personality of the local model.

Moreover, researchers are required to provide further investigations about the solutions to data heterogeneity by studying the mentioned skewness issues. In addition, more efforts are required to reduce the communication cost and find a recovery strategy in case of lost connections or dropping one or more communication sites during the training round.

Improving FL frameworks and overcoming these challenges can provide promising results for valuable studies in the medical field. This requires researchers to broaden the benefits of FL for COVID-19 data beyond just identifying the disease or extracting the affected part of the lung to encompass a wide range of applications, especially studies identifying the relationship between COVID-19 and patient im-

munity, genetic impact, and chronic diseases. FL enables studies on larger samples to gain reliable results.

Besides the technical perspective, clinical efforts are required to reflect the theoretical and simulation results in real-life situations. However, validation by independent medical researchers is crucial for predictive analytics [97]. Clinical recommendations should advocate publicly available and verified algorithms, and adequate and in-depth analysis of data complexity is necessary. For example, in survival cases, some prognostic results should be tracked for 30 days. Monitoring patients and analyzing their updates, whether they have recovered or died within that period, could be necessary as part of the risk management framework.

We highly recommend broadening the contact between researchers from both the medical and technical fields to facilitate studies about COVID-19 virus behaviors. The FL method could effectively process the variety of COVID-19 viral sequences, which aids medical studies in tracing the origins and transmission pathways of infections [3]. These data are crucial for improving global contact tracking, guiding epidemiological studies, and supporting broader public health campaigns to stop the virus's spread. The distributed samples that FL trains are also useful for studying and understanding how quickly COVID-19 sequences change. This has caused a lot of worry about the appearance of new variants that might be more virulent or more easily spread than the strains that are already out there.

Additionally, it could provide an answer to a crucial question in the context of viral evolution: could specific mutations within the Viral RNA Sequence potentially undermine the effectiveness of existing vaccines? This could potentially aid in the development of new immune strategies to combat the virus.

In conclusion, this work presents an experimental study on the impact of non-IID data in FL for medical imaging, analyzing how different types of data heterogeneity affect model performance. Through extensive experiments, we evaluate the challenges posed by skewed data distributions and assess their implications for FL convergence and accuracy. Our findings provide insights into the limitations of FL under non-IID conditions and highlight potential directions for future improvements.

Bibliography

[1] World Health Organization. Coronavirus disease (covid-19). https://www.who.int/emergencies/diseases/novel-coronavirus-2019. Accessed on 25 May 2024.

[2] Esteban Ortiz-Ospina Edouard Mathieu, Hannah Ritchie, Lucas Rodés-Guirao, Cameron Appel, Daniel Gavrilov, Charlie Giattino, Joe Hasell, Bobbie Macdonald, Saloni Dattani, Diana Beltekian and Max Roser. Coronavirus (COVID-19) Deaths - Our World in Data.

[3] S. Halawa, S. S. Pullamsetti, C. R. M. Bangham, K. R. Stenmark, P. Dorfmüller, M. G. Frid, G. Butrous, N. W. Morrell, V. A. de Jesus Perez, D. I. Stuart, et al. Potential long-term effects of sars-cov-2 infection on the pulmonary vasculature: A global perspective. *Nature Reviews Cardiology*, 19:314–331, 2022. [CrossRef] [PubMed].

[4] R. Li, S. Pei, B. Chen, Y. Song, T. Zhang, W. Yang, and J. Shaman. Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (sars-cov-2). *Science*, 368:489–493, 2020. [CrossRef].

[5] E.J. Williamson, A.J. Walker, K. Bhaskaran, S. Bacon, C. Bates, C.E. Morton, H.J. Curtis, A. Mehrkar, D. Evans, P. Inglesby, et al. Factors associated with covid-19-related death using opensafely. *Nature*, 584:430–436, 2020. [CrossRef].

[6] R. Aljondi and S. Alghamdi. Diagnostic value of imaging modalities for covid-19: Scoping review. *J. Med. Internet Res.*, 22:e19673, 2020. [CrossRef].

[7] T. Bahadur, K. Verma, B. Kumar, and D. Jain. Coronavirus disease (covid-19) detection in chest x-ray images using majority voting based classifier ensemble. *Expert Syst. Appl.*, 165:113909, 2021. [CrossRef].

[8] K.V. Sarma, S. Harmon, T. Sanford, H.R. Roth, Z. Xu, J. Tetreault, D. Xu, M.G. Flores, A.G. Raman, R. Kulkarni, et al. Federated learning improves site

performance in multicenter deep learning without data sharing. *J. Am. Med. Inform. Assoc.*, 28:1259–1264, 2021. [CrossRef].

[9] M. Shen, Y. Deng, L. Zhu, X. Du, and N. Guizani. Privacy-preserving image retrieval for medical iot systems: A blockchain-based approach. *IEEE Netw.*, 33:27–33, 2019. [CrossRef].

[10] G.A. Kaissis, M.R. Makowski, D. Rückert, and R.F. Braren. Secure, privacy-preserving and federated machine learning in medical imaging. *Nat. Mach. Intell.*, 2:305–311, 2020. [CrossRef].

[11] L. Li, L. Qin, Z. Xu, Y. Yin, X. Wang, B. Kong, J. Bai, Y. Lu, Z. Fang, Q. Song, et al. Using artificial intelligence to detect covid-19 and community-acquired pneumonia based on pulmonary ct: Evaluation of the diagnostic accuracy. *Radiology*, 296:E65–E71, 2020. [CrossRef].

[12] J.L. Raisaro, F. Marino, J. Troncoso-Pastoriza, R. Beau-Lejdstrom, R. Bellazzi, R. Murphy, E.V. Bernstam, H. Wang, M. Bucalo, Y. Chen, et al. Scor: A secure international informatics infrastructure to investigate covid-19. *J. Am. Med. Inform. Assoc.*, 27:1721–1726, 2020. [CrossRef].

[13] L.M. Florescu, C.T. Streba, M.S. ¸Serbănescu, M. Mămuleanu, D.N. Florescu, R.V. Teică, R.E. Nica, and I.A. Gheonea. Federated learning approach with pre-trained deep learning models for covid-19 detection from unsegmented ct images. *Life*, 12:958, 2022. [CrossRef] [PubMed].

[14] I. Feki, S. Ammar, Y. Kessentini, and K. Muhammad. Federated learning for covid-19 screening from chest x-ray images. *Appl. Soft Comput.*, 106:107330, 2020. [CrossRef] [PubMed].

[15] D.C. Nguyen, M. Ding, P.N. Pathirana, and Y. Jin. Federated learning for covid-19 detection with generative adversarial networks in edge cloud computing. *IEEE Internet Things J.*, 9:10257–10271, 2020. [CrossRef].

[16] G. Kaissis, A. Ziller, J. Passerat-Palmbach, T. Ryffel, D. Usynin, A. Trask, I. Lima, J. Mancuso, F. Jungmann, M.M. Steinborn, et al. End-to-end privacy-preserving deep learning on multi-institutional medical imaging. *Nat. Mach. Intell.*, 3:473–484, 2021. [CrossRef].

[17] R. Kumar, J. Kumar, A. Aman, H. Ali, C.M. Bernard, R. Ullah, and S. Zeng. Blockchain and homomorphic encryption based privacy-preserving model aggregation for medical images. *Computerized Medical Imaging and Graphics*, 102:102139, 2022.

[18] J. Zhou, L. Zhou, D. Wang, X. Xu, H. Li, Y. Chu, W. Han, and X. Gao. Personalized and privacy-preserving federated heterogeneous medical image analysis with pppml-hmi. *Comput. Biol. Med.*, 169:107861, 2024. [CrossRef].

[19] P. M. Thompson, J. L. Stein, S. E. Medland, D. P. Hibar, A. A. Vasquez, M. E. Renteria, R. Toro, N. Jahanshad, G. Schumann, B. Franke, and et al. The enigma consortium: Large-scale collaborative analyses of neuroimaging and genetic data. *Brain Imaging Behav.*, 8:153–182, 2014.

[20] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B.A. y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54, page 10, Ft. Lauderdale, FL, USA, 2017.

[21] E. Darzidehkalani. Federated learning in medical image analysis. *Pattern Recognition*, 151:110424, 2024.

[22] C. Shyu, K.T. Putra, H. Chen, Y. Tsai, K.S.M.T. Hossain, W. Jiang, and Z. Shae. A systematic review of federated learning in the healthcare area: From the perspective of data properties and applications. *Applied Sciences*, 11:11191, 2021.

[23] Ehsan Darzidehkalani, Mohammad Ghasemi-rad, and Pieter M.A. van Ooijen. Federated learning in medical imaging: Part ii: Methods, challenges, and considerations. *Journal of the American College of Radiology*, 19:975–982, 2022.

[24] B. Liu, B. Yan, Y. Zhou, Y. Yang, and Y. Zhang. Experiments of federated learning for covid-19 chest x-ray images. *arXiv*, 2020.

[25] M. Abdul, S. Id, S. Taha, and M. Ramadan. Covid-19 detection using federated machine learning. *PLoS ONE*, 16:e0252573, 2021.

[26] T. Li, A.K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith. Federated optimization in heterogeneous networks. In *Proceedings of Machine Learning Systems*, volume 2, pages 429–450, 2020.

[27] A. Guha Roy, S. Siddiqui, S. Pölsterl, N. Navab, and C. Wachinger. Braintorrent: A peer-to-peer environment for decentralized federated learning. *arXiv*, 2019.

[28] X. Li, Y. Gu, N. Dvornek, L.H. Staib, P. Ventola, and J.S. Duncan. Multi-site fmri analysis using privacy-preserving federated learning and domain. *Medical Image Analysis*, 65:101765, 2020.

[29] Q. Dou, T.Y. So, M. Jiang, Q. Liu, V. Vardhanabhuti, G. Kaissis, Z. Li, W. Si, H.H.C. Lee, K. Yu, et al. Federated deep learning for detecting covid-19 lung abnormalities in ct: A privacy-preserving multinational validation study. *NPJ Digital Medicine*, 4:60, 2021.

[30] Z. Wang, L. Cai, X. Zhang, C. Choi, and X. Su. A covid-19 auxiliary diagnosis based on federated learning and blockchain. *Computational and Mathematical Methods in Medicine*, 2022:7078764, 2022.

[31] D. Yang, Z. Xu, W. Li, A. Myronenko, H.R. Roth, S. Harmon, S. Xu, B. Turkbey, E. Turkbey, X. Wang, 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.

[32] I. Dayan, H.R. Roth, A. Zhong, A. Harouni, A. Gentili, A.Z. Abidin, A. Liu, A.B. Costa, B.J. Wood, C.S. Tsai, et al. Federated learning for predicting clinical outcomes in patients with covid-19. *Nat. Med.*, 27:1735–1743, 2021. [CrossRef] [PubMed].

[33] M. Badawy, N. Ramadan, and H.A. Hefny. Big data analytics in healthcare: data sources, tools, challenges, and opportunities. *Journal of Electrical Systems and Information Technology*, 11:63, 2024.

[34] L. Qu, N. Balachandar, and D. L. Rubin. An experimental study of data heterogeneity in federated learning methods for medical imaging. *arXiv preprint arXiv:2107.08371*, 2021.

[35] G. Kaissis, A. Ziller, J. Passerat-Palmbach, T. Ryffel, D. Usynin, A. Trask, I. Lima, J. Mancuso, F. Jungmann, M.M. Steinborn, et al. End-to-end privacy preserving deep learning on multi-institutional medical imaging. *Nature Machine Intelligence*, 3:473–484, 2021.

[36] World Health Organization. A timeline of who's covid-19 response in the who european region: A living document (version 3.0, from 31 december 2019 to 31 december 2021), 2022. Licence: CC BY-NC-SA 3.0 IGO.

[37] J.M. Banda, R. Tekumalla, G. Wang, J. Yu, T. Liu, Y. Ding, E. Artemova, E. Tutubalina, and G. Chowell. A large-scale covid-19 twitter chatter dataset for open scientific research—an international collaboration. *Epidemiologia*, 2:315–324, 2021.

[38] T. Xia, D. Spathis, C. Brown, J. Chauhan, A. Grammenos, J. Han, A. Hasthanasombat, E. Bondareva, T. Dang, A. Floto, et al. Covid-19 sounds: A large-scale audio dataset for digital respiratory screening. In *Proceedings of the Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, pages 1–13, 2021.

[39] J. Shuja, E. Alanazi, W. Alasmary, and A. Alashaikh. Covid-19 open source data sets: A comprehensive survey. *Applied Intelligence*, 51:1296–1325, 2021.

[40] D. Kvak, M. Bendik, and A. Chromcova. Towards clinical practice: Design and implementation of convolutional neural network-based assistive diagnosis system for covid-19 case detection from chest x-ray images. *arXiv*, 2022.

[41] B. Rasuli. Covid-19 pneumonia: Case study. https://doi.org/10.53347/rID-75330, 2025. Radiopaedia.org. Accessed on 10 Mar 2025.

[42] MD Dirk-André Clevert. White paper: Lung ultrasound in patients with coronavirus covid-19 disease. Siemens Healthineers, 2025. Professor of Radiology, Section Chief of the Interdisciplinary Ultrasound Center at the Department of Clinical Radiology.

[43] A. Golubev. Dicom network implementation and usage in the context of the covid-19 pandemic. *Archives of the Balkan Medical Union*, 56:80–87, 2021.

[44] N. Peiffer-Smadja, R. Maatoug, F.-X. Lescure, E. D'Ortenzio, J. Pineau, and J.-R. King. Machine learning for covid-19 needs global collaboration and data-sharing. *Nature Machine Intelligence*, 2:293–294, 2020.

[45] M. Roberts, D. Driggs, M. Thorpe, J. Gilbey, M. Yeung, S. Ursprung, A.I. Aviles-Rivero, C. Etmann, C. McCague, L. Beer, et al. Common pitfalls and recommendations for using machine learning to detect and prognosticate for covid-19 using chest radiographs and ct scans. *Nature Machine Intelligence*, 3:199–217, 2021.

[46] R. Kumar, A.A. Khan, J. Kumar, N.A. Golilarz, S. Zhang, Y. Ting, C. Zheng, and W. Wang. Blockchain-federated-learning and deep learning models for covid-19 detection using ct imaging. *IEEE Sensors Journal*, 21:16301–16314, 2021.

[47] A. Loddo, F. Pili, and C. di Ruberto. Deep learning for covid-19 diagnosis from ct images. *Applied Sciences*, 11(17):8227, 2021.

[48] S. Naz, K.T. Phan, and Y.P.P. Chen. A comprehensive review of federated learning for covid-19 detection. *International Journal of Intelligent Systems*, 37:2371–2392, 2022.

[49] M. Frid-Adar, R. Amer, O. Gozes, J. Nassar, and H. Greenspan. Covid-19 in cxr: From detection and severity scoring to patient disease monitoring. *IEEE Journal of Biomedical and Health Informatics*, 25:1892–1903, 2021.

[50] E. Tartaglione, C.A. Barbano, C. Berzovini, M. Calandri, and M. Grangetto. Unveiling covid-19 from chest x-ray with deep learning: A hurdles race with small data. *International Journal of Environmental Research and Public Health*, 17(18):6933, 2020.

[51] Johns Hopkins Coronavirus Resource Center. Mortality analyses—johns hopkins coronavirus resource center, 2024. Available online: https://coronavirus.jhu.edu/data/mortality (accessed on 8 May 2024).

[52] J. Pang, Y. Huang, Z. Xie, J. Li, and Z. Cai. Collaborative city digital twin for the covid-19 pandemic: A federated learning solution. *Tsinghua Science and Technology*, 26:759–771, 2021.

[53] D. Ng, X. Lan, M.M.S. Yao, W.P. Chan, and M. Feng. Federated learning: A collaborative effort to achieve better medical imaging models for individual sites that have small labelled datasets. *Quantitative Imaging in Medicine and Surgery*, 11:852–857, 2021.

[54] HHS.Gov. Privacy, 2022. Available online: https://www.hhs.gov/hipaa/for-professionals/privacy/index.html (accessed on 8 November 2022).

[55] General Data Protection Regulation (GDPR). Processing, 2022. Available online: https://gdpr-info.eu/issues/processing/ (accessed on 8 November 2022).

[56] P.H. Yi, J. Wei, T.K. Kim, J. Shin, H.I. Sair, F.K. Hui, G.D. Hager, and C.T. Lin. Radiology "forensics": Determination of age and sex from chest radiographs using deep learning. *Emerging Radiology*, 28:949–954, 2021.

[57] F. Qian and A. Zhang. The value of federated learning during and post-covid-19. *International Journal for Quality in Health Care*, 33:mzab010, 2021.

[58] M. Jiang, Z. Wang, and Q. Dou. Harmofl: Harmonizing local and global drifts in federated learning on heterogeneous medical images. In *Proceedings of*

the AAAI Conference on Artificial Intelligence, volume 36, pages 1087–1095, 2022.

[59] A. Bhattacharya, M. Gawali, J. Seth, and V. Kulkarni. Application of federated learning in building a robust covid-19 chest x-ray classification model. *arXiv*, 2022.

[60] L. Peng, G. Luo, A. Walker, Z. Zaiman, E.K. Jones, H. Gupta, K. Kersten, J.L. Burns, C.A. Harle, T. Magoc, et al. Evaluation of federated learning variations for covid-19 diagnosi using chest radiographs from 42 us and european hospitals. *Journal of the American Medical Informatics Association*, 30:54–63, 2023.

[61] A. Bhattacharya, M. Gawali, J. Seth, and V. Kulkarni. Application of federated learning in building a robust covid-19 chest x-ray classification model. *arXiv*, 2022.

[62] X. Bai, H. Wang, L. Ma, Y. Xu, J. Gan, Z. Fan, F. Yang, K. Ma, J. Yang, S. Bai, et al. Advancing covid-19 diagnosis with privacy-preserving collaboration in artificial intelligence. *Nat. Mach. Intell.*, 3:1081–1089, 2021. [CrossRef].

[63] Y. Xu, L. Ma, F. Yang, Y. Chen, K. Ma, J. Yang, X. Yang, Y. Chen, C. Shu, Z. Fan, et al. A collaborative online ai engine for ct-based covid-19 diagnosis. *medRxiv*, 2020. [CrossRef].

[64] N. Dong and I. Voiculescu. Federated contrastive learning for decentralized unlabeled medical images. In *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2021*, volume 12903 LNCS, pages 378–387. Springer, 2021.

[65] L. Zhang, B. Shen, A. Barnawi, S. Xi, N. Kumar, and Y. Wu. Feddpgan: Federated differentially private generative adversarial networks framework for the detection of covid-19 pneumonia. *Information Systems Frontiers*, 23:1403–1415, 2021.

[66] S.K. Lo, Y. Liu, Q. Lu, C. Wang, X. Xu, H.-Y. Paik, and L. Zhu. Blockchain-based trustworthy federated learning architecture. *arXiv*, 2021.

[67] H. Malik, A. Naeem, R.A. Naqvi, and W.K. Loh. Dmfl-net: A federated learning-based framework for the classification of covid-19 from multiple chest diseases using x-rays. *Sensors*, 23:743, 2023.

[68] X. Li, M. Jiang, X. Zhang, M. Kamp, and Q. Dou. Fedbn: Federated learning on non-iid features via local batch normalization. *arXiv*, 2021.

[69] T.T. Ho, K.D. Tran, and Y. Huang. Fedsgdcovid: Federated sgd covid-19 detection under local differential privacy using chest x-ray images and symptom information. *Sensors*, 22:3728, 2022.

[70] D. Chowdhury, S. Banerjee, M. Sannigrahi, A. Dey, A. Dhar, A. Chakraborty, and A. Das. Federated learning based covid-19 detection. *Expert Systems*, 40:e13173, 2023.

[71] R. Kumar, W. Wang, C. Yuan, J. Kumar, C. Zheng, and A. Aman. Blockchain based privacy-preserved federated learning for medical images: A case study of covid-19 ct scans. *arXiv*, 2021.

[72] L.M. Florescu, C.T. Streba, M.S. Şerbănescu, M. Mămuleanu, D.N. Florescu, R.V. Teică, R.E. Nica, and I.A. Gheonea. Federated learning approach with pre-trained deep learning models for covid-19 detection from unsegmented ct images. *Life*, 12:958, 2022.

[73] W. Zhang, T. Zhou, Q. Lu, X. Wang, C. Zhu, H. Sun, Z. Wang, S.K. Lo, and F.-Y. Wang. Dynamic fusion-based federated learning for covid-19 detection. *IEEE Internet of Things*, 8:15884–15891, 2021.

[74] A. Qayyum, K. Ahmad, M.A. Ahsan, A. Al-Fuqaha, and J. Qadir. Collaborative federated learning for healthcare: Multi-modal covid-19 diagnosis at the edge. *IEEE Open Journal of the Computer Society*, 3:1–10, 2022.

[75] R. Adhikari and C. Settles. Secure federated learning approaches to diagnosing covid-19. *arXiv*, 2024.

[76] N.K.; Kumar S.; Ali M.; Choi Y.R.; Joo M. II; Kim H.C. Aich, S.; Sinai. Protecting personal healthcare record using blockchain federated learning technologies. In *In Proceedings of the International Conference on Advanced Communication Technology (ICACT), PyeongChang, Republic of Korea, 13–16 February*, 2020.

[77] P. Jabłecki, F. Ślazyk, and M. Malawski. Federated learning in the cloud for analysis of medical images—experience with open source frameworks. In *Clinical Image-Based Procedures, Distributed and Collaborative Learning, Artificial Intelligence for Combating COVID-19 and Secure and Privacy-Preserving Machine Learning, Proceedings of the 10th Workshop, CLIP 2021*, volume 12969 LNCS, pages 111–119. Springer, 2021.

[78] D. Beutel, T. Topal, A. Mathur, X. Qiu, T. Parcollet, and N. D. Lane. Flower: A friendly federated learning research framework. *arXiv preprint arXiv:2207.06975*, 2022.

[79] Google. grpc: A high-performance, open-source universal rpc framework, 2023.

[80] M. E. H. Chowdhury, T. Rahman, A. Khandakar, R. Mazhar, M. A. Kadir, Z. B. Mahbub, K. R. Islam, M. S. Khan, A. Iqbal, N. A. Emadi, et al. Can ai help in screening viral and covid-19 pneumonia? *IEEE Access*, 8:132665–132676, 2020.

[81] T. Rahman, A. Khandakar, Y. Qiblawey, A. Tahir, S. Kiranyaz, S. B. A. Kashem, M. T. Islam, S. Al Maadeed, S. M. Zughaier, M. S. Khan, et al. Exploring the effect of image enhancement techniques on covid-19 detection using chest x-ray images. *Comput. Biol. Med.*, 132:104319, 2021.

[82] E. Vantaggiato, E. Paladini, F. Bougourzi, C. Distante, A. Hadid, and A. Taleb-Ahmed. Covid-19 recognition using ensemble-cnns in two new chest x-ray databases. *Sensors*, 21(5):1742, 2021.

[83] Anas M. Tahir, Muhammad E. H. Chowdhury, Yazan Qiblawey, Amith Khandakar, Tawsifur Rahman, Serkan Kiranyaz, Uzair Khurshid, Nabil Ibtehaz, Sakib Mahmud, and Maymouna Ezeddin. Covid-qu-ex. https://doi.org/10.34740/kaggle/dsv/3122958, 2021. Kaggle dataset.

[84] M. Umair, M. S. Khan, F. Ahmed, F. Baothman, F. Alqahtani, M. Alian, and J. Ahmad. Detection of covid-19 using transfer learning and grad-cam visualization on indigenously collected x-ray dataset. *Sensors*, 21(17):5813, 2021.

[85] M. Maftouni, A. C. C. Law, B. Shen, Z. J. K. Grado, Y. Zhou, and N. A. Yazdi. A robust ensemble-deep learning model for covid-19 diagnosis based on an integrated ct scan images database. In *Proceedings of the 2021 IISE Annual Conference*, Montreal, QC, Canada, 2021.

[86] E. Soares, P. Angelov, S. Biaso, M. H. Froes, and D. K. Abe. Sars-cov-2 ct-scan dataset: A large dataset of real patients ct scans for sars-cov-2 identification. *medRxiv*, 2020.

[87] N. Hernandez-Cruz, P. Saha, M.K. Sarker, and J.A. Noble. Review of federated learning and machine learning-based methods for medical image analysis. *Big Data Cogn. Comput.*, 8:99, 2024.

[88] Q. Li, Y. Diao, Q. Chen, and B. He. Federated learning on non-iid data silos: An experimental study. In *Proceedings of the 2022 IEEE 38th International Conference on Data Engineering (ICDE)*, pages 965–978, Kuala Lumpur, Malaysia, May 2022.

[89] M. Chetoui and M.A. Akhloufi. Federated learning approach for early detection of covid-19. *Computers*, 12:106, 2023.

[90] M. Aiello, G. Esposito, G. Pagliari, P. Borrelli, V. Brancato, and M. Salvatore. How does dicom support big data management? investigating its use in medical imaging community. *Insights into Imaging*, 12:164, 2021.

[91] N. Balachandar, K. Chang, J. Kalpathy-Cramer, and D.L. Rubin. Accounting for data variability in multi-institutional distributed deep learning for medical imaging. *Journal of the American Medical Informatics Association*, 27:700–708, 2020.

[92] D.C. Nguyen, M. Ding, and P.N. Pathirana. Federated learning for covid-19 detection with generative adversarial networks in edge cloud computing. *IEEE Internet of Things Journal*, 9:10257–10271, 2021.

[93] E. Jothimurugesan, K. Hsieh, J. Wang, G. Joshi, and P.B. Gibbons. Federated learning under distributed concept drift. In *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics*, volume 206, pages 5834–5853. PMLR, 2023.

[94] Joon-Hee Yoo, Hojoon Jeong, Jongho Lee, and Tae Myoung Chung. Federated learning: Issues in medical application. In *Future Data and Security Engineering, Proceedings of the 8th International Conference, FDSE 2021, Virtual Event, 24–26 November 2021*, Lecture Notes in Computer Science. Springer, 2021.

[95] E.B. Tsai, S. Simpson, M.P. Lungren, M. Hershman, L. Roshkovan, E. Colak, B.J. Erickson, G. Shih, A. Stein, J. Kalpathy-Cramer, et al. The rsna international covid-19 open radiology database (ricord). *Radiology*, 299:E204–E213, 2021.

[96] M.d.l.I. Vayá, J.M. Saborit, J.A. Montell, A. Pertusa, A. Bustos, M. Cazorla, J. Galant, X. Barber, D. Orozco-Beltrán, F. García-García, et al. Bimcv covid-19+: A large annotated dataset of rx and ct images from covid-19 patients. *arXiv*, 2020.

[97] S.O. Hwang and A. Majeed. Analysis of federated learning paradigm in medical domain: Taking covid-19 as an application use case. *Applied Sciences*, 14:4100, 2024.

[98] D.R. Kandati and T.R. Gadekallu. Federated learning approach for early detection of chest lesion caused by covid-19 infection using particle swarm optimization. *Electronics*, 12, 2023.

Appendix 1

A Additional Research Directions

Table A.1 presents a summary of notable studies that have explored the issue of data heterogeneity in federated learning for COVID-19 lung imaging. These works highlight various strategies, applications, and limitations in addressing this critical challenge.

Table A.1: A summary of proposed FL framework for COVID-19 lung medical imaging.

Study	**Aim**	**Application**	**Contributions**	**Limitations**
Xu et al., 2020 [63]	Compares FL model accuracy with radiologists in diagnosis	Diagnosing CT lung images (COVID-19, viral pneumonia, bacterial pneumonia, healthy)	Comparable sensitivity-specificity with radiologists; real FL experiments with three hospitals in Wuhan	Trade-off between performance and communication; 16 hours for 200 rounds
Zhang et al., 2021 [73]	Improves communication efficiency using dynamic fusion FL	Diagnosing X-ray and CT lung images (COVID-19, pneumonia, healthy)	Reduced communication overhead (model scaled to 1/16 time of galaxy FL)	Did not consider reverse engineering threats

Study	Aim	Application	Contributions	Limitations
Feki et al., 2021 [14]	Investigates FL settings with non-IID and unbalanced data	Diagnosing chest X-ray images (COVID-19, healthy)	More rounds improved accuracy; more participants led to fast convergence; labeled distribution skew worsened performance	Small dataset (108 COVID-19 images)
Liu et al., 2021 [24]	Compares FL performance of four DL models	Diagnosing X-ray lung images (COVID-19, pneumonia, healthy)	ResNeXt performed best on COVID-19 images	Models trained on 2% COVID-19 labels; ignored non-IID issues
Jablecki et al., 2021 [77]	Measures non-IID impact on FL accuracy	Diagnosing chest X-ray images (COVID-19, pneumonia, healthy)	More local epochs increased GPU time but not accuracy; Non-IID reduced accuracy from 0.923 to 0.39	First round had longest execution time due to TensorFlow; ignored GPU resource limits
Dou et al., 2021 [29]	Improves FL generalizability using multi-hospital data	Quantifying lesions from COVID-19 CT images	Larger datasets reduced model bias and improved generalizability	40ms per test round; ignored reverse engineering threats
Qayyum et al., 2022 [74]	Addresses imaging modality heterogeneity	Diagnosing chest X-rays and ultrasound images (COVID-19, normal)	Shared models across modalities improved generalizability	Did not specify data distribution across clients; privacy not guaranteed
Yang et al., 2021 [31]	Evaluates FL with heterogeneous data and unlabeled data	Segmenting and annotating COVID-19 lung lesions (CT images)	Data augmentation improved model generalizability; trade-off needed for aggregation frequency	Did not address non-IID impact or bias detection

Study	Aim	Application	Contributions	Limitations
Bai et al., 2021 [62]	Tests FL with highly heterogeneous data	Diagnosing chest X-ray images (COVID-19, pneumonia, healthy)	Reported FLOPS for different models	Lacked bias analysis; ignored participant dropout effects
Kumar et al., 2021 [71]	Overcomes centralization with blockchain and HE	COVID-19 segmentation and classification	Introduced dataset (34,006 CT scans, 89 patients); FL model accuracy 84.21%	Did not report blockchain latency or cost efficiency
Abdul et al., 2021 [25]	Studies FL hyperparameter impact on accuracy/loss	Diagnosing chest X-ray images (COVID-19, pneumonia, healthy)	Softmax + SGD gave best accuracy/loss	Results limited due to incompatibility with other studies
Zhang et al., 2021 [65]	Uses GANs and DP to improve FL	Diagnosing chest X-ray images (COVID-19, pneumonia, healthy)	Fake images improved accuracy (0.84%) and reduced loss (3.0%)	Considered only label distribution skew, ignoring other skew types
Dong et al., 2021 [64]	Labels unlabeled data using contrastive learning	COVID-19 diagnosis (FL with metadata transfer)	Reduced annotation costs; improved performance over FedAvg	No privacy-preserving methods applied
Dayan et al., 2021 [32]	Uses FL for predicting oxygen needs	Predicting mechanical ventilation or death	FL performed well with 25% weight updates; diversity improved generalizability (38%)	Did not report computational cost/time
Nguyen et al., 2021 [92]	Uses GANs to generate synthetic X-ray images	Diagnosing chest X-rays (COVID-19, pneumonia, healthy)	Improved generalizability of models	No privacy-preserving methods applied

Study	Aim	Application	Contributions	Limitations
Lo et al., 2021 [66]	Improves FL fairness using blockchain	Diagnosing chest X-rays (COVID-19, pneumonia, opacity, healthy)	Blockchain contracts improved accountability	No privacy measures implemented
Bhattacharya et al., 2022 [59]	Maintains non-IID nature using three data sources	Diagnosing chest X-rays (COVID-19, healthy)	Personalized client models improved local performance	Ignored privacy attacks; unclear system configuration
Ho et al., 2022 [69]	Enhances privacy and accuracy with FL	Diagnosing chest X-rays (COVID-19, pneumonia, healthy)	SPP-CNN achieved best accuracy; DP noise reduced accuracy by 0.17%	Dataset imbalance (3616 COVID-19 vs. 10,192 normal images)
Kumar et al., 2022 [17]	Uses blockchain-ledger HE to improve FL	Diagnosing chest X-rays (COVID-19, pneumonia, healthy)	Ensured model quality; HE reduced accuracy drop	Blockchain latency caused delays
Wang et al., 2022 [30]	Reduces FL third-party dependence using blockchain	Diagnosing CT lung images (COVID-19, healthy)	Asynchronous FL performed well even with non-IID	Ensuring local model quality was difficult
Kandati and Gadekallu, 2023 [98]	Reduces communication cost using swarm optimization	Diagnosing X-ray images (Normal, COVID-19, pneumonia)	Effective for small datasets with few participants	High global model conversion time; large search space

Glossary

3D-ResNet101 A deep 3D convolutional residual network with 101 layers, designed for spatiotemporal data processing such as medical imaging and video analysis. 100

BIMCV Biomedical Imaging Center of Valencia is a medical research institution in Spain that provides open-access medical imaging datasets, particularly for research in radiology and artificial intelligence. 104

CAD Computer-aided diagnosis system embedded in most imaging devices. 2

COVID-19 stands for Coronavirus Disease 2019. It is caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) . 1

CT A CT scan (Computed Tomography scan) is an advanced medical imaging technique that uses X-rays and computer processing to generate detailed cross-sectional images of the body. It provides more detailed information than regular X-ray images, making it particularly useful for diagnosing diseases affecting soft tissues, bones, and organs. 24

DICOM Digital Imaging and Communications in Medicine is an international standard for storing, transmitting, and managing medical imaging data. 24, 98, 104

ENIGMA project for Enhancing Neuro Imaging Genetics through Meta-analysis. 15

GitHub A web-based platform for version control and collaborative software development using Git, enabling developers to manage, track, and collaborate on code projects. 103

Google Colab Google Colaboratory is a cloud-based platform that allows users to write and execute Python code in a Jupyter Notebook environment. It is widely used for machine learning, deep learning, and data science projects due to its free access to GPUs and TPUs. Google Colab supports libraries like TensorFlow and PyTorch, making it ideal for AI research. However, the free and paid versions have limitations, such as session timeouts (maximum 24 hours), restricted RAM and GPU availability, and execution limits, which can impact large-scale training tasks. [98]

HFL Horizontal Federated Learning works in collection of data records from distributed datasets. [8]

Kaggle An online platform for data science and machine learning competitions, offering datasets, notebooks, and cloud-based tools for collaborative AI development. [103]

ResNet Residual Network is a deep neural network architecture introduced by Microsoft Research in 2015, designed to train very deep networks effectively. It won the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) in 2015 by a significant margin and has since become a foundational model in computer vision tasks. [100]

RGB (Red-Green-Blue) is a color model where colors are formed by combining different intensities of red, green, and blue light. [78]

RSNA Radiological Society of North America is a non-profit professional organization dedicated to advancing radiology through education, research, and innovation. RSNA hosts one of the largest annual conferences in medical imaging and provides open-access datasets, such as the RSNA Pneumonia Detection Challenge dataset and the RSNA COVID-19 Dataset, which are widely used in medical AI research. [104]

RT-PCR manual reverse transcription-polymerase chain reaction tests. [2]

SGD Stochastic Gradient Descent is an optimization algorithm used to minimize loss functions in machine learning by updating model parameters iteratively. It updates weights using a small batch or single sample at a time, making it efficient for large-scale datasets and deep learning tasks.. [78]

Viral RNA Sequence The genetic material of an RNA virus, composed of ribonucleic acid (RNA), which encodes the information required for viral replication and infection. RNA viruses can be single-stranded (ssRNA) or double-stranded (dsRNA). 105

مشكلة البيانات الغير متجانسة لصور الرئة المتعلقة بمرضى كوفيد –١٩ في التعلم المتحد

المستخلص

برز التعلم المتحد في تدريب البيانات الطبية لما يقدمه من ضمان الحفاظ على خصوصية بيانات المرضى. التعلم المتحد هو إطار عمل يندرج تحت تعليم الآلة التعاوني حيث يتيح لعدد من الجهات الموزعة بتدريب بياناتها محليا دون الحاجة لمشاركة هذه البيانات أو الإفصاح عنها وتقوم هذه الجهات فقط بمشاركة أوزان النموذج الذي تم تدريبه. يتم احتساب اوزان جديده من نماذج التعلم التي تم تدريبها محليا ويسمى بالنموذج العالمي.

بحيث يكون النموذج العالمي أكثر قدرة على تعميم النتائج وأقل تأثرًا بالانحرافات في البيانات المحلية نظراً لتحقق شرطي نجاح التعلم العميق وهي سعة البيانات الكبيرة والتنوع في هذه البيانات الخاضعة للتدريب.

تعد مشكلة عدم تجانس البيانات الطبية تحدياً جوهرياً في نتائج التعلم المتحد وظهرت هذه المشكلة في جانحة كوفيد -١٩ بشكل جليّ. حيث اختلفت شدة المرض وأعراضه بين أعمار مختلفة من المصابين في عدد من التوزيعات السكانية حول العالم أدى ذلك لظهور مشكلة البيانات غير المتجانسة في المستشفيات والمراكز الصحية الموزعة.

يناقش هذا البحث مشكلة عدم تجانس البيانات في صور الرئة الملتقطة لمرضى كوفيد -١٩، وتتمثل الإضافة الجوهرية في هذه الدراسة في اقتراح أنواع جديدة من تغاير البيانات الطبية بما في ذلك الانحراف الحاد في التصنيفات وانحراف الخصائص.

يقسم هذا البحث أنواع التباين أو التغير في البيانات بناء على نوع الانحرافات المتعلقة بهذه البيانات. كما يقدم وصفاً رياضياً لستة أنواع من الانحرافات مستندة إلى الأحوال الواقعية التي تشهدها البيانات الطبية في المراكز الموزعة.

ومن خلال تقييم هذه الأنواع على أداء التعلم المتحد ضمن سيناريوهات مختلفة من توزيع البيانات حيث بدأت بتوزيع مثالي ومن ثم يزداد تعقيداً لقياس قدرة النموذج على قراءة نتائج صحيحة في ظروف تغاير حادة، ظهرت من خلالها انحراف التوزيع الكمي وتوزيع التصنيفات بنتائج مشابهة للتوزيع المثالي بينما أظهر إنحراف التصنيفات الحاد وانحراف أنواع التصوير كأسوائها أثراً على أداء التعلم المتحدة، وأظهر إنحراف طرق جمع البيانات بنتائج مرضية مما قد يحقق الأهداف المرجوة من تعاون الأطراف الموزعة.

من خلال التحليل العميق لأداء التعلم المُتحد في ظل التباين الواقعي للبيانات ، تؤكد هذه الدراسة على التحديات والفرص المحورية لتطوير أُطر تعلم تعاوني أكثر قوة ، وأكثر قابلية للتعميم ، وتتمتع بقدرة أعلى على التخصيص دون المساس بمبادئ الخصوصية والموثوقية في التطبيقات الطبية .

الكلمات المفتاحية:

التعلم المتُحد؛ تغاير أو عدم تجانس البيانات؛ بيانات غير مستقلة وغير متوزعة بشكل متماثل؛ مقياس التعميم المركزي؛ مقياس التخصيص المحلي.

إهداء

الحمد لله الذي لا ينبغي الشكر إلا له الحمد لله الذي أعان ويسر وبسط وقدم وأخر.. الحمد لله حمداً متصلاً متواتراً ومتعاقباً متوالياً لا انقطاع له.. الحمد لله الذي بلغنا ملذات الوصول ...

وبعد حمده وإعادة الفضل له جل جلاله...

أمد يدي شكري وأمتناني لوالدي الحبيب.. غرست.. وسقيت وقد حان وقت الحصاد هنيئا لك بما قدمت يداك..

ثم الشكر لوالدتي الغالية: التي ما فتئت أكف الضراعة تطرق باب السماء كلما لمحت هماً عالقاً في صدري.. فطوبى لمن ضمت اسمه دعواتك الحنونة..

زوجي.. وقناديل الضياء عبد الهادي، علي، سعيد، والوردة الجميلة ريما.. شكراً لدعمكم المتواصل وصبركم الجميل.. لا حرمكم الله الأجر.. وأقر العين بأفراحكم ...

أخواني وأخواتي ... معاول الطريق وجبر الأيام الثقال..

شكراً لهتافاتكم التي تشعل لهيب الإقدام في صدري.

مشرفي المتقاعد.. البروفسور عبد الله با سهيل.. من كان لك الفضل بعد الله في اكتمال هذا العمل.. قدمت الكثير.. وجهت وقومت، ودعمت وبذلت الوقت والجهد..

شكر الله سعيك وجزاك المولى خير ما جزى معلماً لطلابه..

مشرفي الحالي البروفيسور كمال جنبي.. شاكرة ومقدرة جميل استقبالكم لإخراج جهدي المتواضع للنور.. وسعيكم الممزوج بالإحسان.. رفع الله قدركم ولا حرمكم الأجر والمثوبة..

والشكر موصول للصديقات والأساتذة الأفاضل في الجامعة ممن أعانوا ويسروا وبذلوا..

كلماتي ممزوجة بالعجز عن الوفاء والامتنان لكم جميعاً.. كتبت بعضها هنا وما يجوب الفؤاد ضاقت به الأوراق..

أهدي هذا العمل المتواضع، عرفاناً وامتناناً لما قدمتموه لي من دعم لا يُقدر بثمن..

مشكلة البيانات الغير متجانسة لصور الرئة المتعلقة بمرضى كوفيد –١٩ في التعلم المتحد

فاطمة سعيد آل حافظ

بحث مقدم لنيل درجة الدكتوراة في علوم الحاسبات

إشراف

البروفيسور. كمال منصور جمبي

كلية الحاسبات وتقنية المعلومات

جامعة الملك عبد العزيز

جدة ، المملكة العربية السعودية

ذو القعدة ١٤٤٦هـ - مايو ٢٠٢٥م