big data and cognitive computing

MDPI

Article

The Data Heterogeneity Issue Regarding COVID-19 Lung Imaging in Federated Learning: An Experimental Study

Fatimah Alhafiz * and Abdullah Basuhail *

Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah 21589, Saudi Arabia
* Correspondence: fabdullahalhafiz@stu.kau.edu.sa (F.A.); abasuhail@kau.edu.sa (A.B.)

Abstract: Federated learning (FL) has emerged as a transformative framework for collaborative learning, offering robust model training across institutions while ensuring data privacy. In the context of making a COVID-19 diagnosis using lung imaging, FL enables institutions to collaboratively train a global model without sharing sensitive patient data. A central manager aggregates local model updates to compute global updates, ensuring secure and effective integration. The global model's generalization capability is evaluated using centralized testing data before dissemination to participating nodes, where local assessments facilitate personalized adaptations tailored to diverse datasets. Addressing data heterogeneity, a critical challenge in medical imaging, is essential for improving both global performance and local personalization in FL systems. This study emphasizes the importance of recognizing real-world data variability before proposing solutions to tackle non-independent and non-identically distributed (non-IID) data. We investigate the impact of data heterogeneity on FL performance in COVID-19 lung imaging across seven distinct heterogeneity settings. By comprehensively evaluating models using generalization and personalization metrics, we highlight challenges and opportunities for optimizing FL frameworks. The findings provide valuable insights that can guide future research toward achieving a balance between global generalization and local adaptation, ultimately enhancing diagnostic accuracy and patient outcomes in COVID-19 lung imaging.

Keywords: federated learning; data heterogeneity; non-IID; global model; skew types; generalization metric; personalization metric

check for updates

Academic Editor: Alberto Abelló

Received: 13 December 2024
Revised: 8 January 2025
Accepted: 9 January 2025
Published: 14 January 2025

Citation: Alhafiz, F.; Basuhail, A. The Data Heterogeneity Issue Regarding COVID-19 Lung Imaging in Federated Learning: An Experimental Study. *Big Data Cogn. Comput.* **2025**, *9*, 11. https://doi.org/10.3390/bdcc9010011

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

The COVID-19 pandemic has underscored the urgent need for effective diagnostic tools, particularly in medical imaging. Variability in clinical observations and the testing of virus-related and patient-specific cases across diverse populations has led to ambiguous and inconsistent information [1]. Radiologists have relied on imaging and scanning technologies to analyze the causative virus's behavior and its effects on the lungs. However, the variety of imaging equipment used, and a lack of standardized acquisition formats have resulted in disorganized and fragmented datasets worldwide [2]. Moreover, privacy concerns surrounding patient data further complicate the integration of these heterogeneous datasets, leading to biased outcomes when deep learning models process such data locally [3].

To address these challenges, federated learning (FL) has been employed as a promising solution, enabling collaborative model training while preserving patient data privacy [4]. FL represents the next generation of artificial intelligence (AI), providing a privacy-preserving approach to machine learning (ML) model development [5]. By facilitating model training across distributed healthcare institutions without requiring data centralization, FL ensures

secure, remote, and parallel learning. This innovative framework encourages researchers to implement collaborative learning systems that comply with patient privacy regulations [6].

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 [7]. However, despite its effectiveness in analyzing distributed data, FL faces significant challenges related to data heterogeneity, which can negatively affect overall model performance.

The authors of many studies have evaluated non-independent and non-identically distributed (non-IID) data by redistributing multiple resources into one or more non-IID settings. These studies have yielded varying results, with some demonstrating consistent global model performance [8]. In contrast, Nguyen et al. [9] noted that performance can degrade by up to 50% due to non-IID data. Such discrepancies often arise from differences in the type of skewness within distributed data. Each imaging dataset presents unique characteristics, including patient demographics, imaging equipment, acquisition protocols, and storage formats. These variations must be carefully considered during the partitioning process in FL simulations to accurately interpret results under non-IID conditions.

This study focuses on a horizontal FL architecture, which provides enhanced control over data privacy, security, and participant evaluation. In this architecture, two key metrics are used to assess model performance: The generalization metric evaluates the global model's performance on new, unseen data, particularly an external testing sample [10]. The personalization metric assesses how well the updated model aligns with the specific characteristics of the local data used during training, ensuring it meets the unique requirements of each participating institution [11]. While several studies have concentrated on enhancing the generalization metric at the central node [2,12,13], others have explored improvements to the personalization metric [11]. Achieving an optimal balance between these metrics is essential for robust FL model performance. This paper shows the challenges posed by data heterogeneity in FL by defining real-world data distributions for COVID-19 lung imaging and analyzing their impact under different types of skewness. The primary contributions of this study are given below.

- Defining data heterogeneity: We provide mathematical definitions and illustrate the effects of various types of skewness on both generalization and personalization metrics.
- Interpreting results: We highlight the implications of real-world data heterogeneity for FL model performance across all participating institutions and evaluation metrics.
- Identifying research opportunities: We outline areas requiring further investigation for each type of skewness to guide future research in optimizing FL systems for medical imaging.

By addressing these critical aspects, this study provides a comprehensive analysis of 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.

The remainder of this paper is structured as follows. We discuss related work in Section 2. We present an FL algorithm and the different types of non-IID data in Section 3. Our implementation scenario is revealed in Section 4. Then, we present our non-IID data partition strategies and evaluation metrics in Section 5. Section 6 presents the experimental results, followed by a discussion in Section 7, and conclusions and future work are discussed in Section 8.

2. Related Work

The COVID-19 pandemic facilitated the creation of extensive medical imaging datasets using modalities such as computed tomography (CT) scans and X-ray and ultrasound images. These resources have motivated researchers to apply federated learning (FL) frameworks to COVID-19 imaging datasets for various medical purposes, including lung disease classification [8,14], COVID-19 diagnosis [12,15], severity identification [16,17], lung damage segmentation [11,18], and oxygen-need prediction [2]. Despite this broad applicability, in this study, we focus on COVID-19 imaging datasets for diagnostic purposes, aiming to understand how different non-IID data scenarios impact FL performance.

The existing studies have addressed the challenges regarding data heterogeneity in FL through various approaches. For instance, it has been found that preprocessing methods ensure uniformity in data size and format [2,11], adaptive hyperparameter tuning enhances model performance [19,20], and aggregative strategies can optimize global model updates [21–23]. These approaches have been applied to specific types of data skewness to evaluate and improve global and personalized models. However, comprehensive evaluations of both global and personalized models across all common skewness types in horizontal FL (HFL) remain limited [24].

2.1. Evaluation-Metric-Based Studies

Researchers have employed evaluation metrics to assess FL performance in non-IID scenarios, focusing on both generalization and personalization. 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. [7] 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. [14] achieved through P2P communication without central evaluation. However, Peng et al. [15] 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 [25], while others rely on external testing data to assess personalization [12,15]. Zhou et al. [11] proposed strategies for mitigating the degradation of personalization metrics, while Dou et al. [26] evaluated both generalization and personalization metrics, focusing solely on acquisition skew.

2.2. Skewness-Study-Based Research

Studies exploring data heterogeneity often focus on specific types of skewness, such as label distribution, quantity, or feature skew. The studies [8,27] considered quantity which the simplest skewness type. However, label distribution skew is the most frequently examined [23,28–30]. Adhikari et al. [31] explored feature skewness in novel settings, while others have investigated acquisition skew in clustered architectures using ultrasound and X-ray modalities [32] or combined CT and X-ray modalities within FL frameworks [33]. Despite these efforts, there has not been a comprehensive evaluation of all common skewness types in HFL using both generalization and personalization metrics.

To address the gaps in the literature, in this study, we evaluate FL performance across all major types of non-IID data distributions, providing a comprehensive interpretation of the analyzed models' behavior in relation to distributed training sites. By simulating real-world data heterogeneity scenarios in COVID-19 lung imaging, this research highlights the challenges and opportunities for improving FL systems. The findings can bridge the

gap between technical results and practical implementation, offering insights that can help FL designers and managers identify, mitigate, and plan for worst-case scenarios (Table 1).

Table 1. A comparison of related studies that evaluated non-IID types along with their measured metrics.

Ref.	Heterogeneity Type	Evaluated Factor	Findings
[1]	Acquisition skew	Personalization metric	Focused on a personalized metric fixed by an adaptive local epoch is an effective method.
[2]	Acquisition skew	Generalization and personalization metrics	Distributed data used for training were made visible by pretrained CNN models.
[3–6]	Quantity and label distribution skew	Generalization metric	Maximizing the size of data is an effective way of improving the generalization metric.
[7]	Extreme label skew	Generalization metric	Skew can be managed via hyperparameter settings.
[8]	Feature skew	Generalization metric	Non-IID data had a significant negative impact.
[9]	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.

3. Background

In this study, we used FL to study different data heterogeneity settings, as briefly described in the following subsections.

3.1. Federated Learning

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 without considering the heterogeneity of data from different resources.

Assume that N is the number of hospitals or medical institutions that participate in an FL system, all of which are connected to server S. Each participating site i receives an initial model M_0^g to train with a learning rate of λ_i parallelly for a given number of rounds, T. In a situation involving real medical data, the system considers the issue of non-IID, where each site i uses a given data sample D_i to train the local model and update the weights of the received model θ_i and produce a new version of the local model M_{0+1}^l. All N sites l send back the local models to the server and aggregate them by using the classic optimization global model FedAvg [34], as shown in Equation (1):

$$M_{t+1}^g = \sum_{i=1}^{N} \left(M_{0+1}^l / D_i \right) \cdot M_t^g \quad (1)$$

3.2. Data Heterogeneity

The heterogeneity of medical imaging data extends beyond the differences in dataset sizes among hospitals. The scope of heterogeneity in medical imaging data includes

variations in imaging equipment, protocols for annotation and labeling, and the distribution number of labels. The danger lies in the averaging of local models generated from training distributed data. FL updates the global model over a number of rounds. Each local model trains its own data to construct an optimal model objective, which converges the model's weight to capture the useful features of the local data. Therefore, each local model undergoes a distinct convergence process in which unique model weights are computed to achieve a given optimization goal based on local image data. This leads to the divergences of the models being averaged after the training round, and it could make the global model less accurate over time, harming the generalizability metric [35].

On the other hand, after each round, the hospital receives the latest update of the global model to train the local data further, allowing for adjustments based on the latest global weights. The goal of this iterative process is to improve the local models, with a trade-off between local specificity and global coherence. This will improve the federated learning system's overall performance when data distributions are harmonized. However, in the case of heterogeneous data, the challenge intensifies, as variations in local datasets can lead to significant discrepancies in model performance across different sites, degrading the personalization metric. This could also pose a fairness problem for sites that contribute effectively but do not benefit from the collective knowledge of the global model [36].

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

3.2.1. Quantity Skew/Label Distribution Skew

Quantity skew occurs when 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. For instance, 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.

3.2.2. Extreme Label Skew

Extreme label skew arises when one or more labels are completely absent from the local datasets used for FL rounds. For instance, during the early phases of the COVID-19 pandemic [37], some medical facilities were overwhelmed by a high volume of patients and had limited resources, so images were only labeled as "COVID-19-positive" or not labeled at all for other conditions. This was often the case when using testing kits that could only archive and label images for confirmed COVID-19 cases due to restricted storage and patient volumes [38]. Such a scenario can make it difficult to differentiate between COVID-19 pathology and other diseases like pneumonia or SARS, which share similar symptoms. The lack of diversity in training data can severely affect a model's ability to accurately classify images with ambiguous conditions, leading to poor performance for both global and local models in an FL framework.

3.2.3. 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. These differences can introduce variations in image size, resolution, color intensity, noise levels, and even the background in the images, all of which can negatively impact the performance of FL models. The corresponding model may struggle to generalize across these diverse imaging modalities, and local models may fail to account for the unique characteristics of each site's data.

3.2.4. 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. Modality skew is particularly challenging because each modality has unique properties (regarding, e.g., image size, resolution, and content), requiring a model to learn to adapt to these differences across different sites and local datasets.

In providing these specific examples, we aim to offer a clearer understanding of the types of data skew encountered in medical imaging for COVID-19 and how each may affect the performance of FL models. These cases highlight the importance of addressing data heterogeneity and implementing strategies for mitigating the negative impact on both global and personalized model performance in federated learning systems. In this study, we evaluate each type of skewness as a simple type of skewness or a mix of more than one type of skewness, as described in Section 5.1.1.

4. System Design

As described in the previous Section, the proposed system was designed to study the non-IID issue pertaining to medical imaging in the FL framework. Figure 1 provides an overview of the system's design, comprising four components, each with distinct roles, as outlined below:

- Central server

The central server is considered a trusted node containing a central manager module to control communication between distributed sites and fine-tune the aggregation of local models. Central data serve as a crucial component of the server, enabling the measurement of the global model's generalization based on out-sampled data.

- Hospital nodes

The participating hospitals have three tasks: training the local data, validating the local model, and, finally, testing the received global model.

- Models

All the models in the proposed system share similar architectures, with the only difference being that the weights update according to the local data of the hospital node. The model employs a convolutional neural network (CNN) 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.

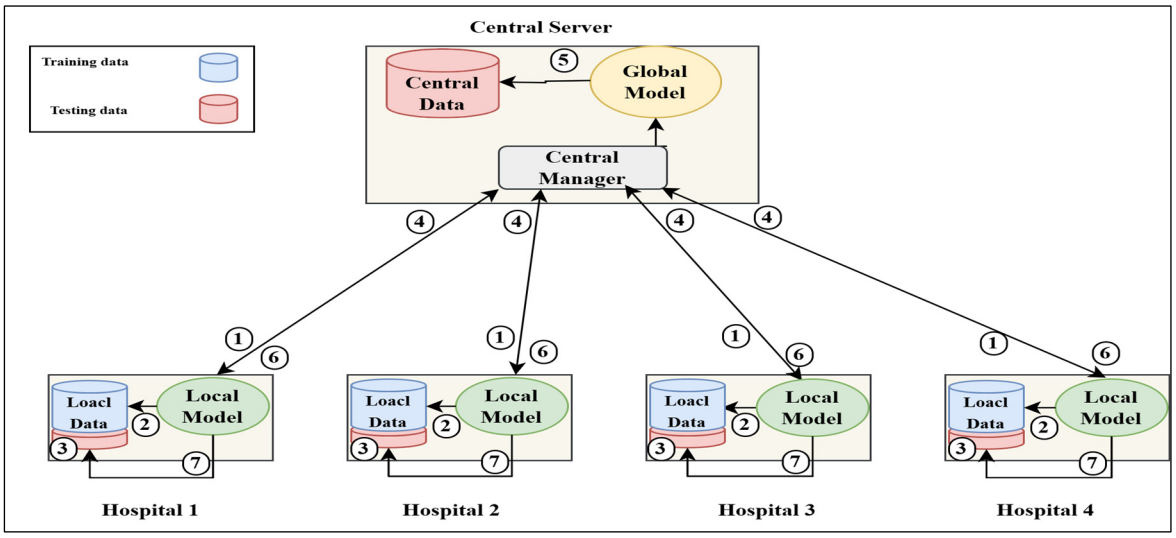


Figure 1. Overview of the proposed FL system architecture.

In this study, we employed a simple CNN, as depicted in Figure 2, to explore the behavior of both local and global models with respect to non-IID data acquired through federated learning. 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.

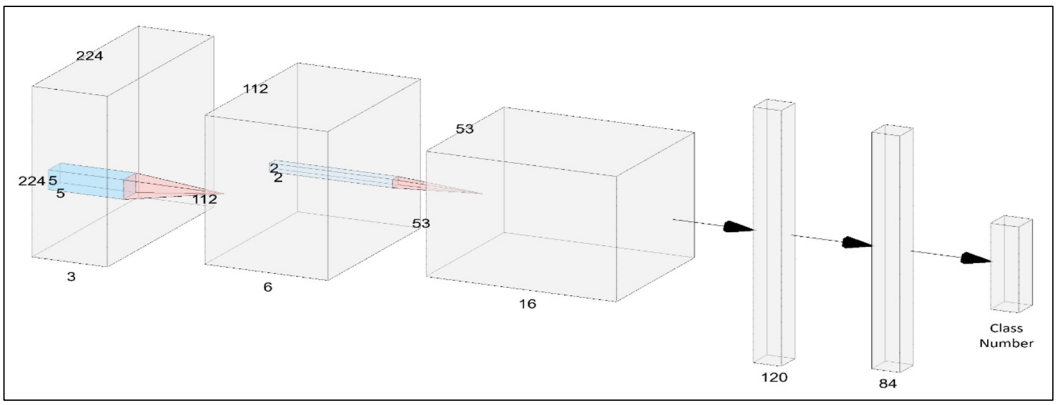


Figure 2. CNN model used for training distributed hospital datasets in FL framework.

- 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.

In this study, we leveraged a combination of seven datasets to evaluate the FL framework, comprising five X-ray and two CT datasets. Among the X-ray datasets, the COVID-19

Radiography Dataset [39,40] contains 21,165 images across four classes: COVID-19 (3616 images), normal (10,192 images), viral pneumonia (1345 images), and lung opacity (6012 images). The COVID-19 X-Ray Dataset [41] features 1512 images divided equally among three classes, namely, COVID-19, pneumonia, and normal, each containing 504 images. We also used a variation of the COVID-19 X-Ray Dataset that includes 3555 images spanning five classes: COVID-19 (711 images), bacterial pneumonia (711 images), viral pneumonia (711 images), lung opacity (711 images), and normal (711 images). The COVID-QU-Ex Dataset [42] comprises 33,920 X-ray images, categorized into COVID-19 (11,956 images), pneumonia (11,263 images), and normal (10,701 images). Additionally, we used a dataset of chest X-rays from Pakistani patients [43], including 450 images, with 390 images of COVID-19-positive patients and 60 images of normal cases.

The CT datasets include the Large COVID-19 CT Dataset [44], which consists of 17,104 images divided into three classes: COVID-19 (7593 images from 466 patients), Normal (6893 images from 604 patients), and CAP (2618 images from 60 patients). The MedRxiv Dataset [45] provides 4,964 CT scans, including COVID-19-positive scans (1252 images), normal scans (1230 images), and additional scans for COVID-19 pathology (2482 images). All the datasets used in this study are publicly available and were accessed via the Kaggle repository API.

The proposed system could achieve its functional requirements based on the following scenario, as illustrated by the numbered steps in Figure 1.

1. The system's scenario begins with sending 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 gauge 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 FedAvg 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. Then, we increase the round number by one, starting a new training session from step 2.

5. Methodology

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 Section outlines the phases of FL implementation. It describes the datasets used to simulate FL using the FLOWER framework, discusses the experiments conducted to evaluate FL performance in different non-IID cases and the variables considered during the

execution of the proposed work across different experiments, and defines the evaluation metrics required to evaluate the local and global models.

5.1. Experiment Settings

5.1.1. Data Settings

To evaluate the performance of the proposed federated learning (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: (1) IID settings for benchmark comparison, distributing data equally across four hospital nodes; (2) simple non-IID scenarios, including data quantity skew and extreme label skew to simulate real-world hospital variations; and (3) hyper-skewness scenarios with acquisition and modality skew to assess generalization in heterogeneous conditions. The central testing datasets remained consistent across all the experiments, with additional external datasets used for hyper-skewness evaluations. In all the experimental configurations, no patient overlap was allowed between hospital subsets to ensure data authenticity.

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, as depicted in Figure 3, to assess the global model's performance. The training data were divided almost equally among the four hospitals, with each hospital receiving a balanced number of images for each class, as shown in Figure 4. The dataset was manually distributed across the nodes, ensuring that each node contained fewer than 100 images per class, as per the distribution guidelines [24].

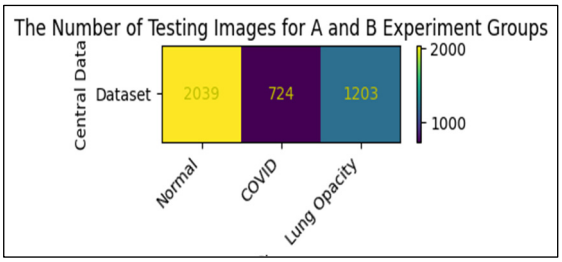


Figure 3. Testing set assigned in central node, experiment A and B.

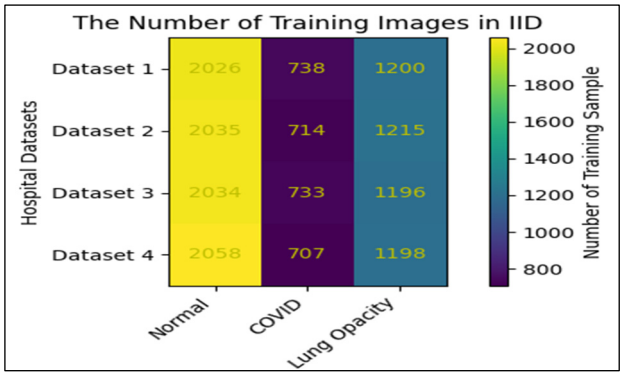


Figure 4. IID data setting, experiment A.

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 validation set was used to test the model's ability to generalize locally at each hospital site. The primary

objective of this experiment was to establish a benchmark for the FL framework's performance, allowing for comparisons with subsequent experiments where data distributions were skewed or non-IID. Because of its use of a uniform distribution, 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, as shown in Figure 5. This distribution reflects real-world scenarios, as shown in Figure 5, 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 [46]. 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.

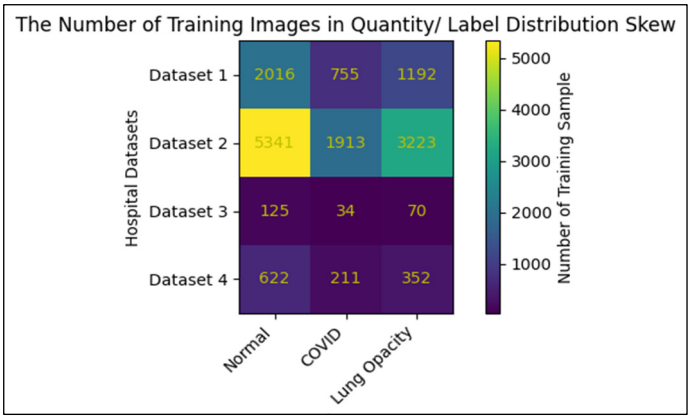


Figure 5. Quantity/label distribution skew data setting, experiment B (1).

2. 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 6. 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 imbalance across distributed datasets in a real-world medical scenario.

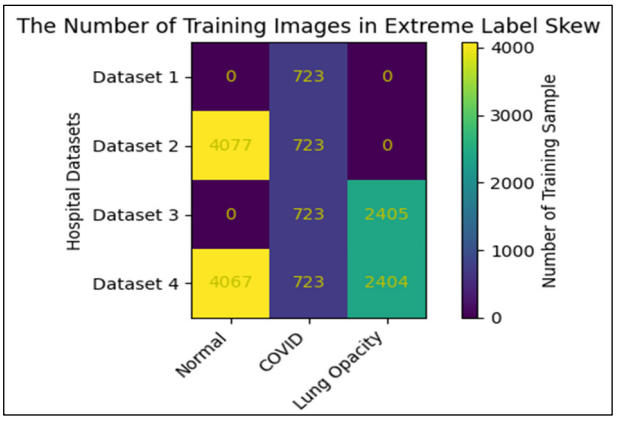


Figure 6. Data settings for extreme label skew, experiment B (2).

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 evaluated 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.

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 7. For central testing, an external dataset of COVID-19 Pakistani patients, comprising two labels, was utilized, as illustrated in Figure 8.

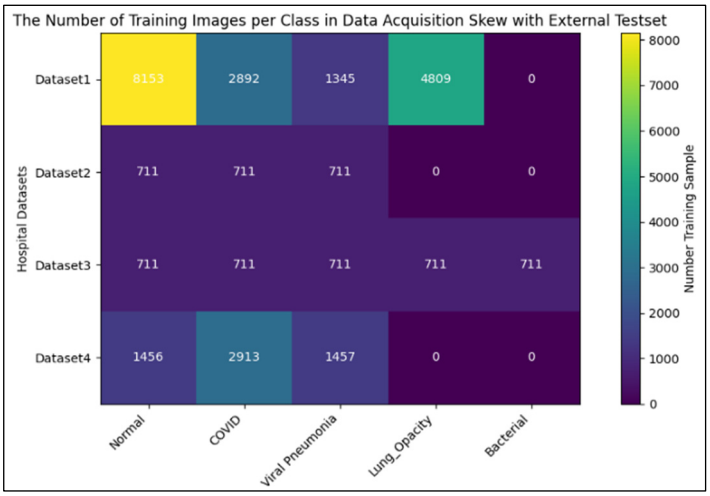


Figure 7. Training data settings used in data acquisition, experiment C (1.a).

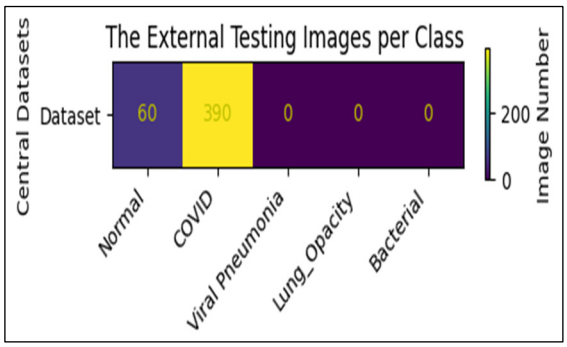


Figure 8. External data for testing the central model, experiment C (1.a).

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 Figures 9 and 10.

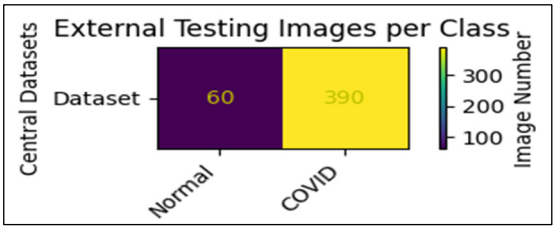


Figure 9. External testing data in the central node, experiment C (1.b).

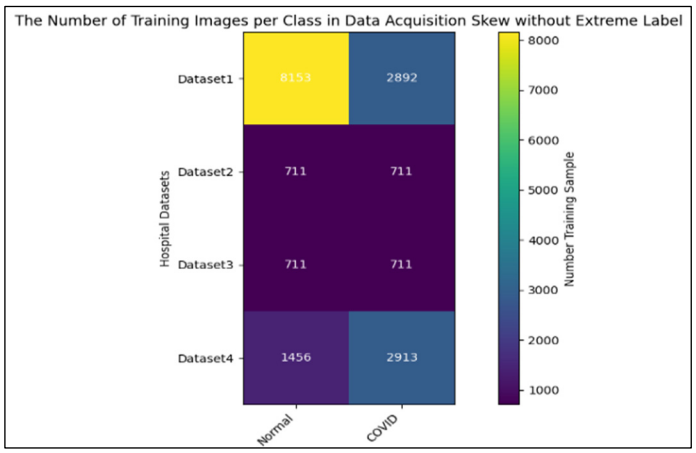


Figure 10. Training data settings for data acquisition, which have the same modality without extreme label skew, experiment C (1.b).

This comprehensive setup simulates diverse real-world scenarios, providing insights into the robustness and adaptability of the proposed FL framework.

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. [31], 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 11. 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 12.



Figure 11. Modality Skew experiment with internal testing set, experiment C (2.a).

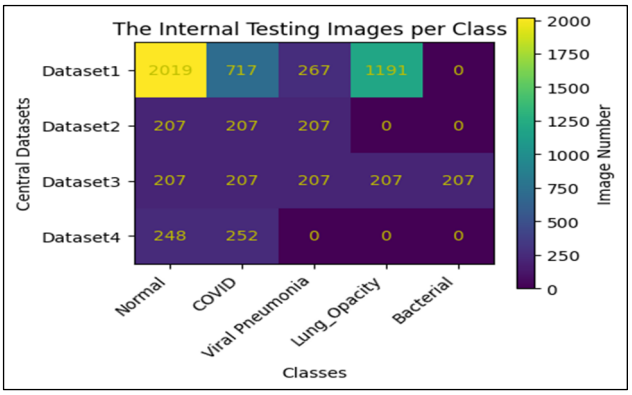


Figure 12. Central testing data for experiment C (2.a).

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 13. External datasets—comprising images obtained via both X-ray and CT modalities—were then used for testing at the central server, as illustrated in Figure 14. This approach evaluated the model's ability to generalize across modalities, with the global model being tested on external data sources.

This experiment setup was designed to assess the robustness of the FL framework in handling data heterogeneity, specifically when imaging modalities vary across distributed sites. It provides valuable insights into how the framework performs when faced with modality skew and external testing conditions.

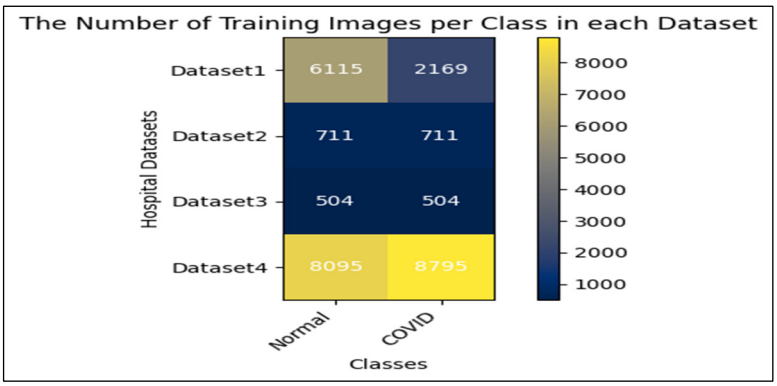


Figure 13. Modality skew with external data test, experiment C (2.b).

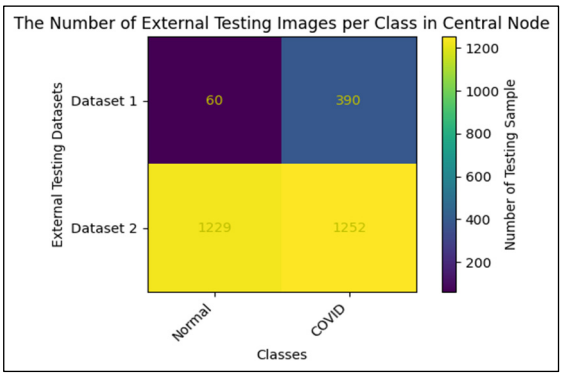


Figure 14. External central test data with modality skew x-ray and CT datasets, experiment C (2.b).

5.1.2. 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 [47], such as random rotations, flips, and scaling, to enhance the diversity of the training data.

The hyperparameters were the same for all experiments, as summarized in Table 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 [48]. 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 [49]. In order to prevent randomness in sampling [2], 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 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 ($\hat{\boldsymbol{\lambda}}$)	0.001

Table 2. *Cont.*

Parameters	Values
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.

5.2. 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 [50]. 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 (2):

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{2}$$

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 [7]. We used a multiple cross-entropy function, as shown in Equation (3), 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.

$$H(y, p) = -\sum_{i=1}^{C} y_i \cdot log(p_i) \tag{3}$$

6. Results

We categorized the framework implementation results based on the experiments discussed in the Methodology Section. As mentioned earlier, our goal was not to improve the performance of the FL system. This Section presents a comparative analysis of the global model's generalization performance and localization performance and the personalization performance of the updated models across all training sites in each experiment, organizing them into the following result groups.

6.1. Group a Results (IID Setting)

The local model evaluations were similar for all hospitals (85–87), as shown in Figure 15. 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.

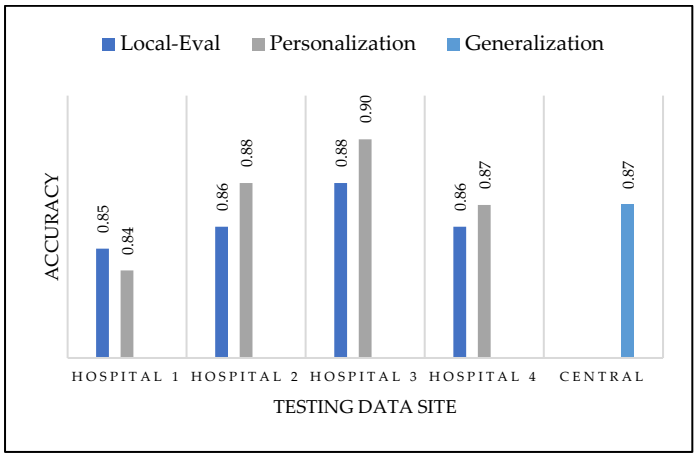


Figure 15. Results regarding IID distribution.

6.2. Group B Results (Simple Skew)

6.2.1. Results Regarding Quantity/Label Distribution Skew

The evaluations of the local model vary depending on the specific set of hospital data. Figure 16 shows that most hospital sites (0, 1, and 3) achieved a higher level of personalization, while Hospital 3 experienced 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, namely, 91. This suggests that the global model is more suitable for sites with numerous and varied data, such as Hospitals 1, 2, and 3.

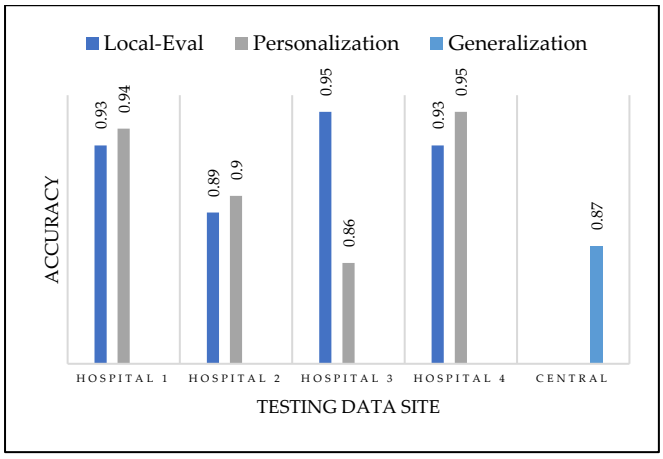


Figure 16. Results regarding non-IID distribution with quantity label distribution skew.

This finding indicates that the reported generalization metric, with a quantity/label distribution skew, yields performance similar to that under IID settings, while a personal-

ization 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.

6.2.2. Results Regarding Extreme Label Skew

As shown in Figure 17, 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.

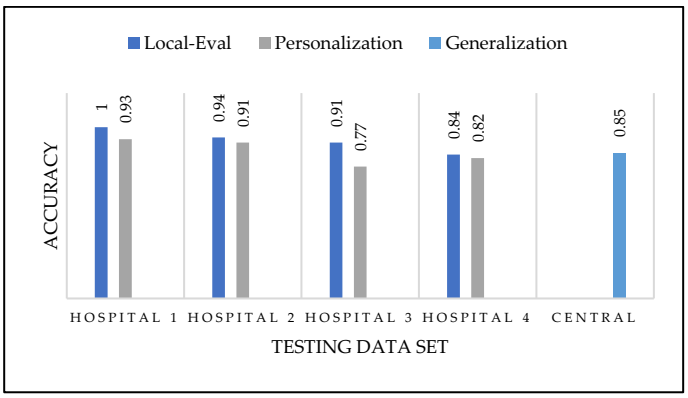


Figure 17. Results regarding non-IID distribution with extreme label skew.

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 at each site, both before and after aggregation.

6.3. Group C Results (Hyper-Skewness)

6.3.1. 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 in the case of removing the extreme label skew from the FL framework, as shown in Figure 18. In the case of ELS, the prediction process for external data decreased to 1% (as shown by the blue curve in Figure 18 for rounds 7 and 15) due to distractions in the training process caused by variant labels across distributed data with insufficient numbers, which were often available in one place and absent in another. Moreover, Figure 19 shows a comparison between the loss prediction

values for the central model and the average of the personal model in all training sites with and without ELS, revealing lower loss prediction in the internal data locally and centrally in the external data without ELS. Also, in this case, ELS 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.

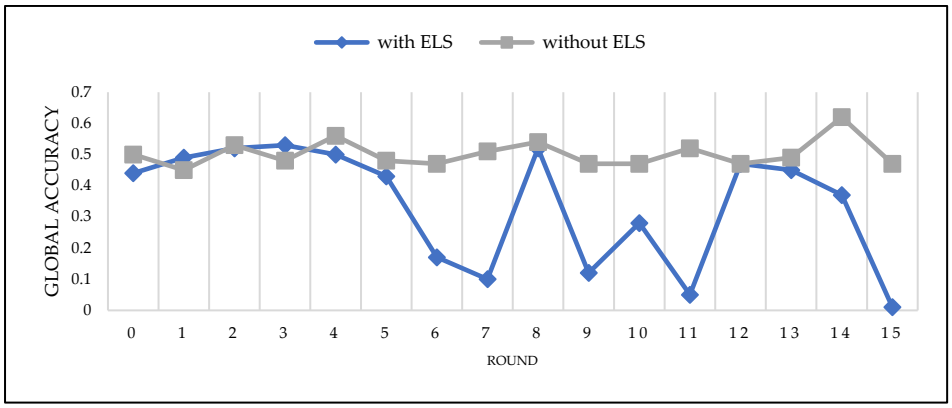


Figure 18. Generalization accuracy with and without extreme label skew (ELS).

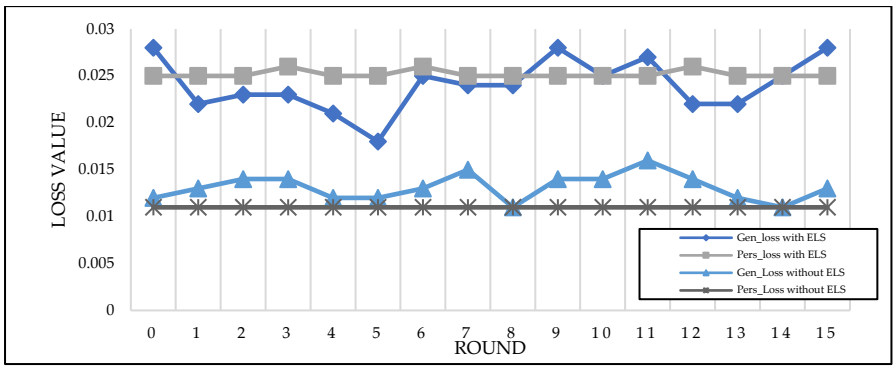


Figure 19. Generalization and personalization loss values with and without ELS.

6.3.2. Results Regarding Modality Skew with Internal and External Data

In Figure 20, 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. Also, 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.

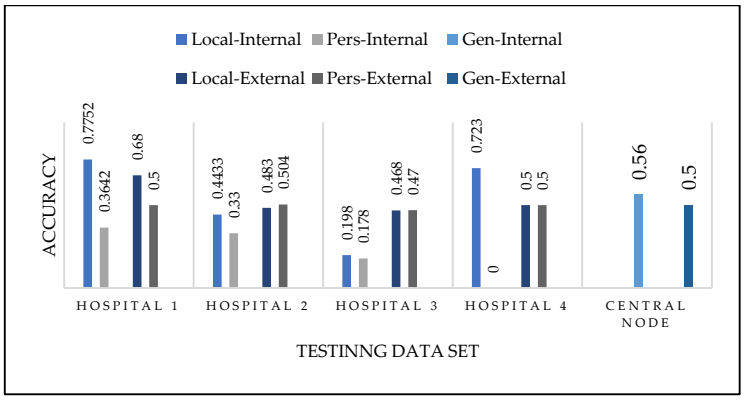


Figure 20. Accuracy results regarding modality skew testing for internal vs. external data.

Figure 21 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 ELS led to higher degradation in the performance accuracy of FL, especially for the personalization metric. Figure 22 shows the confusion matrix of the internal testing data; it predicts five labels, and the image 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 the aggregative model could not generalize features that appear at the hospital site. On the other hand, the confusion matrix for the external testing data revealed a higher true-positive prediction rate for the CVOID-19 label, as Figure 23 shows. Therefore, unique label prediction in FL is an unsolved issue that requires further investigation.

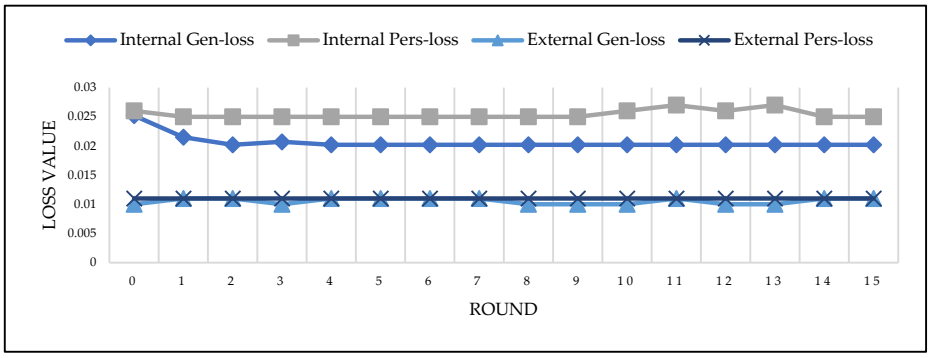


Figure 21. Results regarding generalization and personalization loss for internal and external datasets.

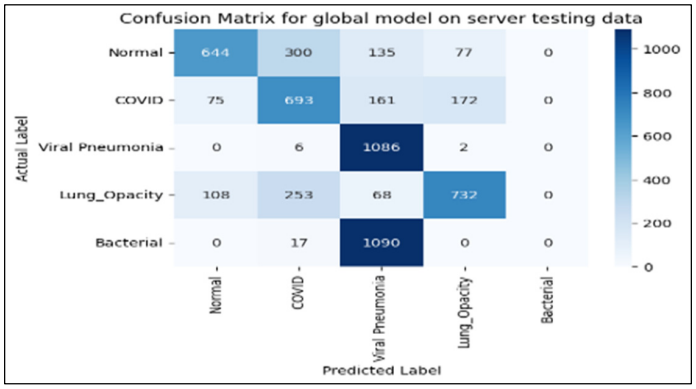


Figure 22. Confusion matrix of modality skew for testing using external data.

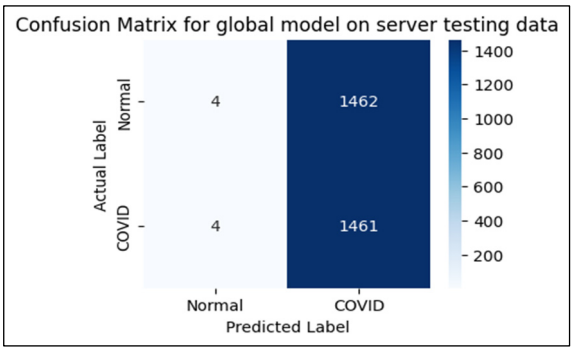


Figure 23. Confusion matrix for modality skew regarding testing on the internal dataset.

7. Discussion

When federated learning 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 1, 2, and 3, all of which were conducted using internal central data. This issue primarily affected the personalization metric, potentially leading to biased results for certain sites. By the end of the experiments on groups A and B, there were infinite settings between both skews; to avoid a trivial variant between the two skews, we followed the same method for limiting the skew ratio, as mentioned in [24]. In contrast, to transition from ELS to label distribution skew, the number of images per label in each local datum was 0: 30 images, as in experiment 2.

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 question; 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 [2].

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 [51].

Another example pertains to label distribution skew, as in the experiment on group B, where there was an imbalanced number of images per label. Differences in label distribution can significantly affect model outcomes. A model trained on Hospital A's dataset, dominated by COVID-19 cases, might become overly specialized in detecting this condition and struggle with other diagnoses, such as lung opacity. This imbalance can precipitate biased predictions, especially in hospitals with a more varied or balanced case distribution. Ensuring a model performs well across all hospitals requires addressing these skews in label representation during training [52].

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 [32].

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 infrared experiment with a clustering architecture [32] or by integrating more than one modality at a training site [33]. 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 metric for group C's outcomes.

Furthermore, training with hyper skew types with extreme label skew led to significant degradation in the personalization metric. The results of experiments 4 and 5 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 ELS without preprocessing to unify the prediction labels.

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.

8. Conclusions and Future Work

Federated learning is a technique that provides an efficient and secure framework for training distributed and scattered data across hospitals and medical institutions. However, the variety of modalities, dimensions, and characteristics, as well as factors like acquisition variations, medical device company, or area demographics within a certain protocol, lead to heterogeneous data or non-IID in an FL framework. Categorizing and classifying data heterogeneity types via extensive experiments based on real situations was the main aim of this study. Another aim was to investigate the effect of common essential performance metrics in FL on all distributed sites in depth. We found that using hyper-skewed data in training, but not simple skew cases, challenges FL performance. While the variable factors of data increased, the contributions of distributed training are only slightly apparent. Overcoming data issues by proposing methods of rendering data uniform or keeping track of metrics across rounds based on the skew type requires further investigation.

Author Contributions: Conceptualization, F.A. and A.B.; methodology, F.A.; software, F.A.; validation, A.B.; formal analysis, F.A.; investigation, F.A.; resources, F.A.; data curation, F.A.; writing—original draft preparation, F.A.; writing—review and editing, A.B.; visualization, A.B.; supervision, A.B.; project administration, A.B.; funding acquisition, not applicable. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The data from the present study are available from the corresponding author for private use only.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Halawa, S.; Pullamsetti, S.S.; Bangham, C.R.M.; Stenmark, K.R.; Dorfmüller, P.; Frid, M.G.; Butrous, G.; Morrell, N.W.; de Jesus Perez, V.A.; Stuart, D.I.; et al. Potential Long-Term Effects of SARS-CoV-2 Infection on the Pulmonary Vasculature: A Global Perspective. *Nat. Rev. Cardiol.* **2022**, *19*, 314–331. [CrossRef]
2. Dayan, I.; Roth, H.R.; Zhong, A.; Harouni, A.; Gentili, A.; Abidin, A.Z.; Liu, A.; Costa, A.B.; Wood, B.J.; Tsai, C.S.; et al. Federated Learning for Predicting Clinical Outcomes in Patients with COVID-19. *Nat. Med.* **2021**, *27*, 1735–1743. [CrossRef] [PubMed]
3. Hryniewska, W.; Bombiński, P.; Szatkowski, P.; Tomaszewska, P.; Przelaskowski, A.; Biecek, P. Checklist for Responsible Deep Learning Modeling of Medical Images Based on COVID-19 Detection Studies. *Pattern Recognit.* **2021**, *118*, 108035. [CrossRef]
4. Mondal, M.R.H.; Bharati, S.; Podder, P.; Kamruzzaman, J. Deep Learning and Federated Learning for Screening COVID-19: A Review. *BioMedInformatics* **2023**, *3*, 691–713. [CrossRef]
5. Kaissis, G.A.; Makowski, M.R.; Rückert, D.; Braren, R.F. Secure Privacy-Preserving and Federated Machine Learning in Medical Imaging. *Nat. Mach. Intell.* **2020**, *2*, 305–311. [CrossRef]
6. Thompson, P.M.; Stein, J.L.; Medland, S.E.; Hibar, D.P.; Vasquez, A.A.; Renteria, M.E.; Toro, R.; Jahanshad, N.; Schumann, G.; Franke, B.; et al. The ENIGMA Consortium: Large-Scale Collaborative Analyses of Neuroimaging and Genetic Data. *Brain Imaging Behav.* **2014**, *8*, 153–182. [CrossRef] [PubMed]
7. Florescu, L.M.; Streba, C.T.; Şerbănescu, M.S.; Mămuleanu, M.; Florescu, D.N.; Teică, R.V.; Nica, R.E.; Gheonea, I.A. Federated Learning Approach with Pre-Trained Deep Learning Models for COVID-19 Detection from Unsegmented CT Images. *Life* **2022**, *12*, 958. [CrossRef] [PubMed]
8. Feki, I.; Ammar, S.; Kessentini, Y.; Muhammad, K. Federated Learning for COVID-19 Screening from Chest X-Ray Images. *Appl. Soft Comput.* **2020**, *106*, 107330. [CrossRef] [PubMed]
9. Nguyen, D.C.; Ding, M.; Pathirana, P.N.; Jin, Y. Federated Learning for COVID-19 Detection with Generative Adversarial Networks in Edge Cloud Computing. *IEEE Internet Things J.* **2020**, *9*, 10257–10271. [CrossRef]
10. Kaissis, G.; Ziller, A.; Passerat-Palmbach, J.; Ryffel, T.; Usynin, D.; Trask, A.; Lima, I.; Mancuso, J.; Jungmann, F.; Steinborn, M.M.; et al. End-to-End Privacy-Preserving Deep Learning on Multi-Institutional Medical Imaging. *Nat. Mach. Intell.* **2021**, *3*, 473–484. [CrossRef]
11. Zhou, J.; Zhou, L.; Wang, D.; Xu, X.; Li, H.; Chu, Y.; Han, W.; Gao, X. Personalized and Privacy-Preserving Federated Heterogeneous Medical Image Analysis with PPPML-HMI. *Comput. Biol. Med.* **2024**, *169*, 107861. [CrossRef]

12. Bai, X.; Wang, H.; Ma, L.; Xu, Y.; Gan, J.; Fan, Z.; Yang, F.; Ma, K.; Yang, J.; Bai, S.; et al. Advancing COVID-19 Diagnosis with Privacy-Preserving Collaboration in Artificial Intelligence. *Nat. Mach. Intell.* **2021**, *3*, 1081–1089. [CrossRef]
13. Siddique, A.A.; Talha, S.U.; Aamir, M.; Algarni, A.D.; Soliman, N.F.; El-Shafai, W. COVID-19 Classification from X-Ray Images: An Approach to Implement Federated Learning on Decentralized Dataset. *Comput. Mater. Contin.* **2023**, *75*, 3883–3901.
14. Bhattacharya, A.; Gawali, M.; Seth, J.; Kulkarni, V. Application of Federated Learning in Building a Robust COVID-19 Chest X-Ray Classification Model. *arXiv* **2022**, arXiv:2204.10505.
15. Peng, L.; Luo, G.; Walker, A.; Zaiman, Z.; Jones, E.K.; Gupta, H.; Kersten, K.; Burns, J.L.; Harle, C.A.; Magoc, T.; et al. Evaluation of Federated Learning Variations for COVID-19 Diagnosis Using Chest Radiographs from 42 US and European Hospitals. *J. Am. Med. Inform. Assoc.* **2023**, *30*, 54–63. [CrossRef]
16. Kumar, R.; Kumar, J.; Aman, A.; Ali, H.; Bernard, C.M.; Ullah, R.; Zeng, S. Blockchain and Homomorphic Encryption Based Privacy-Preserving Model Aggregation for Medical Images. *Comput. Med. Imaging Graph.* **2022**, *102*, 102139. [CrossRef]
17. Xu, Y.; Ma, L.; Yang, F.; Chen, Y.; Ma, K.; Yang, J.; Yang, X.; Chen, Y.; Shu, C.; Fan, Z.; et al. A Collaborative Online AI Engine for CT-Based COVID-19 Diagnosis. *medRxiv* **2020**. [CrossRef]
18. Dong, N.; Voiculescu, I. Federated Contrastive Learning for Decentralized Unlabeled Medical Images. In *Proceedings of the Medical Image Computing and Computer Assisted Intervention–MICCAI 2021, Strasbourg, France, 27 September–October 2021*; Lecture Notes in Computer Science; Springer: Berlin/Heidelberg, Germany, 2021; Volume 12903, pp. 378–387. [CrossRef]
19. Yang, D.; Xu, Z.; Li, W.; Myronenko, A.; Roth, H.R.; Harmon, S.; Xu, S.; Turkbey, B.; Turkbey, E.; Wang, X.; et al. Federated Semi-Supervised Learning for COVID Region Segmentation in Chest CT Using Multi-National Data from China, Italy, Japan R. *Med. Image Anal.* **2021**, *70*, 101992. [CrossRef] [PubMed]
20. Yan, B.; Wang, J.; Cheng, J.; Zhou, Y.; Zhang, Y.; Yang, Y.; Liu, L.; Zhao, H.; Wang, C.; Liu, B. Experiments of Federated Learning for COVID-19 Chest X-Ray Images. *Commun. Comput. Inf. Sci.* **2021**, *1423*, 41–53. [CrossRef]
21. Duan, M.; Liu, D.; Ji, X.; Liu, R.; Liang, L.; Chen, X.; Tan, Y. FedGroup: Efficient Clustered Federated Learning via Decomposed Data-Driven Measure. In Proceedings of the 2021 IEEE Intl Conf on Parallel & Distributed Processing with Applications, Big Data & Cloud Computing, Sustainable Computing & Communications, Social Computing & Networking (ISPA/BDCloud/SocialCom/SustainCom), New York, NY, USA, 30 September–3 October 2021. [CrossRef]
22. Li, X.; Jiang, M.; Zhang, X.; Kamp, M.; Dou, Q. FedBN: Federated Learning on Non-IID Features via Local Batch Normalization. *arXiv* **2021**, arXiv:2102.07623.
23. Zhang, L.; Shen, B.; Barnawi, A.; Xi, S.; Kumar, N.; Wu, Y. FedDPGAN: Federated Differentially Private Generative Adversarial Networks Framework for the Detection of COVID-19 Pneumonia. *Inf. Syst. Front.* **2021**, *23*, 1403–1415. [CrossRef]
24. Prayitno; Shyu, C.R.; Putra, K.T.; Chen, H.C.; Tsai, Y.Y.; Tozammel Hossain, K.S.M.; Jiang, W.; Shae, Z.Y.; Hossain, K.S.M.T.; Jiang, W.; et al. A Systematic Review of Federated Learning in the Healthcare Area: From the Perspective of Data Properties and Applications. *Appl. Sci.* **2021**, *11*, 11191. [CrossRef]
25. Aich, S.; Sinai, N.K.; Kumar, S.; Ali, M.; Choi, Y.R.; Joo, M., II; Kim, H.C. Protecting Personal Healthcare Record Using Blockchain Federated Learning Technologies. In Proceedings of the International Conference on Advanced Communication Technology (ICACT), PyeongChang, Republic of Korea, 13–16 February 2021; pp. 109–112. [CrossRef]
26. Dou, Q.; So, T.Y.; Jiang, M.; Liu, Q.; Vardhanabhuti, V.; Kaissis, G.; Li, Z.; Si, W.; Lee, H.H.C.; Yu, K.; et al. Federated Deep Learning for Detecting COVID-19 Lung Abnormalities in CT: A Privacy-Preserving Multinational Validation Study. *NPJ Digit. Med.* **2012**, *4*, 60. [CrossRef]
27. Ho, T.T.; Tran, K.D.; Huang, Y.; Differential, L.; Using, P.; Images, C.X.; Information, S. FedSGDCOVID: Federated SGD COVID-19 Detection under Local Differential Privacy Using Chest X-Ray Images and Symptom Information. *Sensors* **2022**, *22*, 3728. [CrossRef]
28. Lo, S.K.; Liu, Y.; Lu, Q.; Wang, C.; Xu, X.; Paik, H.-Y.; Zhu, L. Blockchain-Based Trustworthy Federated Learning Architecture. *arXiv* **2021**, arXiv:2108.06912.
29. Jabłecki, P.; Ślazyk, F.; Malawski, M. Federated Learning in the Cloud for Analysis of Medical Images–Experience with Open Source Frameworks. In *Proceedings of the Clinical Image-Based Procedures, Distributed and Collaborative Learning, Artificial Intelligence for Com-bating COVID-19 and Secure and Privacy-Preserving Machine Learning. (DCL 2021, PPML 2021, LL-COVID19 2021, CLIP 2021), Strasbourg, France, 27 September–1 October 2021*; Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics); Springer: Berlin/Heidelberg, Germany, 2021; Volume 12969, pp. 111–119. [CrossRef]
30. Malik, H.; Naeem, A.; Naqvi, R.A.; Loh, W.K. DMFL_Net: A Federated Learning-Based Framework for the Classification of COVID-19 from Multiple Chest Diseases Using X-Rays. *Sensors* **2023**, *23*, 743. [CrossRef] [PubMed]
31. Adhikari, R.; Settles, C. Secure Federated Learning Approaches to Diagnosing COVID-19. *arXiv* **2024**, arXiv:2401.12438.
32. Qayyum, A.; Ahmad, K.; Ahsan, M.A.; Al-Fuqaha, A.; Qadir, J. Collaborative Federated Learning for Healthcare: Multi-Modal COVID-19 Diagnosis at the Edge. *IEEE Open J. Comput. Soc.* **2022**, *3*, 172–184. [CrossRef]

33. Zhang, W.; Zhou, T.; Lu, Q.; Wang, X.; Zhu, C. Dynamic Fusion-Based Federated Learning for COVID-19 Detection. *IEEE Internet Things J.* **2021**, *8*, 15884–15891. [CrossRef]
34. McMahan, B.; Moore, E.; Ramage, D.; Hampson, S.; y Arcas, B.A. Communication-Efficient Learning of Deep Networks from Decentralized Data. In Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, Fort Lauderdale, FL, USA, 20–22 April 2017; Volume 54, p. 10.
35. Zhou, T.; Lin, Z.; Zhang, J.; Tsang, D.H.K. Understanding and Improving Model Averaging in Federated Learning on Heterogeneous Data. *IEEE Trans. Mob. Comput.* **2024**, *23*, 12131–12145. [CrossRef]
36. Ma, T.; Chen, J.; Hoang, T.N. Federated Learning of Models Pretrained on Different Features with Consensus Graphs. *Springer Optim. Its Appl.* **2023**, *213*, 289–319. [CrossRef]
37. Loddo, A.; Pili, F.; di Ruberto, C. Deep Learning for COVID-19 Diagnosis from CT Images. *Appl. Sci.* **2021**, *11*, 8227. [CrossRef]
38. Naz, S.; Phan, K.T.; Chen, Y.P.P. A Comprehensive Review of Federated Learning for COVID-19 Detection. *Int. J. Intell. Syst.* **2022**, *37*, 2371–2392. [CrossRef] [PubMed]
39. Chowdhury, M.E.H.; Rahman, T.; Khandakar, A.; Mazhar, R.; Kadir, M.A.; Mahbub, Z.B.; Islam, K.R.; Khan, M.S.; Iqbal, A.; Emadi, N.A.; et al. Can AI Help in Screening Viral and COVID-19 Pneumonia? *IEEE Access* **2020**, *8*, 132665–132676. [CrossRef]
40. Rahman, T.; Khandakar, A.; Qiblawey, Y.; Tahir, A.; Kiranyaz, S.; Kashem, S.B.A.; Islam, M.T.; Al Maadeed, S.; Zughaier, S.M.; Khan, M.S.; et al. Exploring the Effect of Image Enhancement Techniques on COVID-19 Detection Using Chest X-Ray Images. *Comput. Biol. Med.* **2021**, *132*, 104319. [CrossRef]
41. Vantaggiato, E.; Paladini, E.; Bougourzi, F.; Distante, C.; Hadid, A.; Taleb-Ahmed, A. COVID-19 Recognition Using Ensemble-Cnns in Two New Chest X-Ray Databases. *Sensors* **2021**, *21*, 1742. [CrossRef] [PubMed]
42. Tahir, A.M.; Chowdhury, M.E.H.; Khandakar, A.; Qiblawey, Y.; Khurshid, U.; Kiranyaz, S.; Ibtehaz, N.; Rahman, M.S.; Al-Madeed, S.; Mahmud, S.; et al. COVID-QU-Ex. Kaggle. 2021. Available online: https://www.kaggle.com/datasets/anasmohammedtahir/covidqu (accessed on 10 January 2025).
43. Umair, M.; Khan, M.S.; Ahmed, F.; Baothman, F.; Alqahtani, F.; Alian, M.; Ahmad, J. Detection of COVID-19 Using Transfer Learning and Grad-Cam Visualization on Indigenously Collected X-Ray Dataset. *Sensors* **2021**, *21*, 5813. [CrossRef] [PubMed]
44. Maftouni, M.; Law AC, C.; Shen, B.; Grado ZJ, K.; Zhou, Y.; Yazdi, N.A. 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, 22–25 May 2021.
45. Soares, E.; Angelov, P.; Biaso, S.; Froes, M.H.; Abe, D.K. SARS-CoV-2 CT-Scan Dataset: A Large Dataset of Real Patients CT Scans for SARS-CoV-2 Identification. *medRxiv* **2020**. [CrossRef]
46. Ng, D.; Lan, X.; Yao, M.M.S.; Chan, W.P.; Feng, M. Federated Learning: A Collaborative Effort to Achieve Better Medical Imaging Models for Individual Sites That Have Small Labelled Datasets. *Quant. Imaging Med. Surg.* **2021**, *11*, 852–857. [CrossRef] [PubMed]
47. Hernandez-cruz, N.; Saha, P.; Sarker, M.K.; Noble, J.A. Review of Federated Learning and Machine Learning-Based Methods for Medical Image Analysis. *Big Data Cogn. Comput.* **2024**, *8*, 99. [CrossRef]
48. Li, Q.; Diao, Y.; Chen, Q.; He, B. Federated Learning on Non-IID Data Silos: An Experimental Study. In Proceedings of the 2022 IEEE 38th International Conference on Data Engineering (ICDE), Kuala Lumpur, Malaysia, 9–12 May 2022; pp. 965–978. [CrossRef]
49. Abdul, M.; Id, S.; Taha, S.; Ramadan, M. COVID-19 Detection Using Federated Machine Learning. *PLoS ONE* **2021**, *16*, e0252573. [CrossRef]
50. Kumar, R.; Khan, A.A.; Kumar, J.; Zakria; Golilarz, N.A.; Zhang, S.; Ting, Y.; Zheng, C.; Wang, W.; Zakria; et al. Blockchain-Federated-Learning and Deep Learning Models for COVID-19 Detection Using CT Imaging. *IEEE Sens. J.* **2021**, *21*, 16301–16314. [CrossRef] [PubMed]
51. Rao, A.; Wang, X.; Wen, Y. Challenges in Medical Imaging Analysis with Heterogeneous Datasets. *Med. Image Anal.* **2021**, *72*, 102101.
52. Alhafiz, F.S.; Basuhail, A.A. Non-IID Medical Imaging Data on COVID-19 in the Federated Learning Framework: Impact and Directions. *COVID* **2024**, *4*, 1985–2016. [CrossRef]

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