Federated Performance-based Averaging (FedPA): A Robust and Selective Learning Framework for Chest X-Ray Classification in Heterogeneous Data Environments

Atif Mahmood[1], Tashin Khan Sadique[2], Saaidal Razalli Azzuhri[3], Roziana Ramli[4], Leila Ismail[5]
Faculty of Data Science and Information Technology,
INTI International University, Nilai, Malaysia, 71800[1,2]
Department of Computer System and Technology-Faculty of Computer Science and Information Technology,
Universiti Malaya, Kuala Lumpur, Malaysia, 50603[3]
Department of Computer and Information Sciences, Northumbria University, Newcastle, UK[4]
Department of Computer Science and Software Engineering, United Arab Emirates University, UAE[5]

***Abstract*—Chest X-ray imaging remains a cornerstone in the diagnosis of thoracic conditions such as COVID-19, pneumonia, and lung opacity. Despite advancements in deep learning, the development of robust and generalizable models is limited by data privacy constraints, as patient data cannot be centralized across institutions. Federated Learning (FL) has emerged as a promising solution by enabling collaborative model training without sharing raw data. However, standard FL algorithms like FedAvg, FedProx, and FedSGD aggregate all client updates without considering their individual quality, making them vulnerable to performance degradation in the presence of data heterogeneity, label noise, or underperforming clients. To address these challenges, this study proposes Federated Performance-Based Averaging (FedPA), a novel selective aggregation strategy that incorporates only those client models that meet a predefined performance threshold during training. By leveraging an accuracy-based filtering mechanism, FedPA ensures that only sufficiently trained and reliable local models contribute to global updates. The method was evaluated on a multi-class, non-IID chest X-ray dataset containing four classes: Normal, COVID-19, Pneumonia, and Lung Opacity. Using DenseNet as the backbone model, experiments were conducted across four federated clients, each biased toward a specific class to simulate real-world data distributions. Results demonstrate that FedPA significantly outperforms baseline federated algorithms across key metrics, achieving a global accuracy of 91.82%, F1-score of 92.48%, and recall of 92.08%. The method also achieved faster convergence, higher stability, and reduced round-to-round accuracy fluctuations. System-level evaluations further show that FedPA offers competitive efficiency in terms of inference time, throughput, CPU usage, and memory footprint, making it suitable for deployment in resource-constrained clinical environments. Overall, FedPA offers a practical and effective advancement in federated learning for medical imaging. By filtering unreliable client contributions, it preserves model quality and privacy, presenting a viable path for clinical deployment in scenarios where data centralization is infeasible due to ethical, legal, or logistical constraints.**

***Keywords*—Public health; industrial growth; federated learning; FedAvg; FedPA; FedSGD**

I. Introduction

Deep learning [1] and Medical imaging has become an necessary tool in current healthcare, playing an important role in the diagnosis, monitoring, and treatment of various diseases [2]. Technologies such as X-rays, computed tomography (CT), magnetic resonance imaging (MRI), and ultrasound, combined with AI-driven analysis (see Fig. 1), offer detailed insights into the human body, enhancing the ability of healthcare professionals to detect abnormalities at an early stage with greater accuracy and efficiency [3]. With rapid advancements in artificial intelligence (AI) and deep learning, medical image analysis has significantly improved [4], offering enhanced accuracy, speed, and efficiency in diagnosing conditions such as cancer [5], pneumonia [6], neurological disorders [7], and cardiovascular diseases [8]. These AI-driven approaches facilitate automated detection, segmentation, and classification of medical images, reducing the burden on radiologists and improving patient outcomes [9], [10].

The collection and management of medical imaging datasets for AI applications face significant challenges due to strict privacy regulations and ethical concerns. Federated Learning (FL) [11] has emerged as an innovative, privacy-preserving solution that enables collaborative model training across healthcare institutions without sharing raw patient data (see Fig. 1). By training models locally and only sharing model updates, FL ensures data privacy while leveraging distributed datasets [12]. Federated averaging (FedAvg) is a common aggregation procedure in federated contexts. FedAvg, however, experiences convergence issues, especially when there is significant diversity in the data distributions among clients [13].

In the next section, we will examine the role of federated learning in healthcare, highlighting its current applications and limitations. This discussion will also identify the existing research gaps, thereby framing the specific aim and contribution of this paper in addressing those challenges.

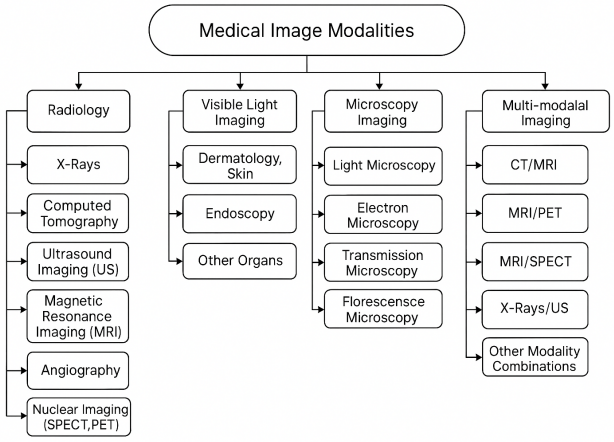

Fig. 1. Medical imaging modalities classification [3].

A. Federated Learning in Healthcare

In a standard FL setup, each client trains a local model using its own dataset and transmits model updates to a central server, which aggregates these updates to improve the global model. This iterative process continues until the model reaches convergence (see Fig. 2). However, traditional FL algorithms such as FedAvg struggle with non-IID data distributions, client heterogeneity, and communication overhead, particularly in healthcare applications where data availability and quality vary significantly across institutions.

Federated Learning has gained significant traction in the healthcare domain due to its ability to enable collaborative model training while ensuring compliance with data privacy regulations such as HIPAA (Health Insurance Portability and Accountability Act) and GDPR (General Data Protection Regulation) [14]. In healthcare, FL is particularly useful for applications such as medical imaging analysis, predictive diagnostics, personalized treatment planning, and disease progression monitoring.

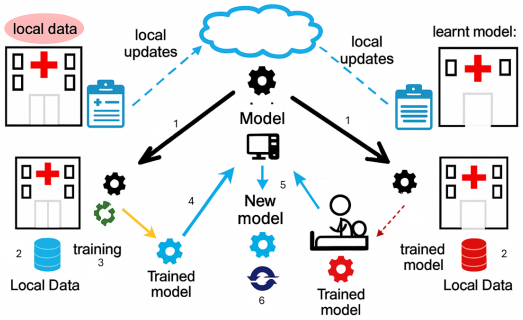

Fig. 2. Federated learning step by step process.

Recent studies have introduced various aggregation techniques to enhance federated learning, with Federated Averaging (FedAvg) being one of the most widely adopted due to its simplicity and efficiency [15]. However, FedAvg assumes equal participation from all clients, making it vulnerable to unreliable or delayed clients, particularly in real-world federated networks where client availability is inconsistent due to factors such as network failures, power constraints, and computational limitations. This inconsistency can lead to several challenges, including unstable global updates, where a high dropout rate among clients causes model updates to become erratic and degrade overall performance; inefficient optimization, as the server lacks control over client participation, making it difficult to streamline the learning process; and reduced model quality, since the absence of high-performing clients in certain rounds may lead to lower-quality updates.

To address some of these limitations, several advanced aggregation techniques have been proposed. Federated Learning with Proximal Optimization (FedProx) [16] improves FedAvg by mitigating the negative effects of client data heterogeneity. However, tuning its additional hyperparameter is challenging, and its default values may not generalize well across different datasets. Federated Multi-Task Learning via Multi-Task Association (FedMA) [17] optimizes both global and local models jointly, which enhances convergence, but its complexity makes it difficult to implement in some federated learning systems. Quantization-based Federated Learning (QFFL) [18] reduces communication overhead by compressing model updates through quantization, but this approach introduces quantization errors, which can impair convergence and final model accuracy.

Despite advancements in Federated Learning (FL), existing methods still face major challenges in medical imaging. High communication overhead, caused by frequent model updates, burdens low-bandwidth healthcare networks. Non-IID data due to variations in demographics, imaging tools, and protocols limits model generalizability.

Client participation is often inefficient, particularly from resource-constrained or outdated devices, contributing low-quality updates and hindering convergence. Data heterogeneity and noise further destabilize training. Scalability is also limited by diverse hardware capabilities across institutions.

Moreover, FL systems remain vulnerable to privacy and security threats such as model inversion and adversarial attacks. Unreliable client connectivity common in mobile or remote setups adds to participation inconsistency, complicating the learning process.

To address these limitations, we propose Performance-Based Federated Averaging (FedPA). A novel aggregation technique specifically designed for medical imaging applications. Unlike the traditional FedAvg method, which aggregates updates from all participating clients regardless of their quality, FedPA selectively incorporates only those updates that meet a predefined accuracy threshold (e.g., 80%). This selective inclusion significantly reduces noise from poorly performing clients, accelerates convergence, and enhances overall model robustness.

In addition to performance-aware selection, FedPA allows low-performing clients to download the updated global model parameters, enabling them to improve their local models over time. Once their performance meets the participation threshold, they are reintegrated into the aggregation process. This strategy promotes progressive learning and inclusivity while maintaining model quality and minimizing communication overhead.

To evaluate the effectiveness of FedPA, we applied it to a multi-class chest X-ray dataset for thoracic disease detection (Normal, COVID-19, Lung Opacity, Pneumonia). Each client was assigned a specific class, simulating a realistic non-IID setting. Experimental results demonstrate that FedPA outperforms traditional FL methods in terms of accuracy, F1-score, robustness, and communication efficiency highlighting its potential as a performance-aware and scalable FL approach for healthcare AI applications

The remainder of this paper is structured as follows. Section 2 presents a literature review of existing FL techniques in medical imaging, discussing the limitations of traditional aggregation methods. Section 3 details the methodology, including an overview of FL, the architecture of the proposed FedPA framework, model configurations, and dataset preprocessing. Section 4 provides experimental results and discussion, comparing FedAvg, traditional FedPA, and Performance-Based FedPA in terms of accuracy, loss, convergence time, and efficiency. Section 5 discusses the broader implications of the findings, highlighting advantages, challenges, and potential real-world applications of FedPA. Section 6 concludes the paper with a summary of findings and future research directions, focusing on further optimizations and the integration of secure aggregation techniques for enhanced privacy in FL.

II. Aggregation Strategies in Federated Learning

1) Federated Averaging (FedAvg): Federated Averaging (FedAvg) is the most commonly used FL aggregation strategy, introduced by [11]. In FedAvg, each participating client trains a local model for several epochs and sends the updated model weights to the central server. The global model is updated using the weighted average of all client models, formulated as:

$$w_{t+1} = \sum_{k=1}^{K} \frac{n_k}{N} w_k^t \tag{1}$$

where w_k^t is the local model update from client k, n_k is the number of samples on client k, and $N = \sum_{k=1}^{K} n_k$ represents the total number of data samples across all clients.

While FedAvg is effective in IID (independent and identically distributed) settings, it faces challenges in heterogeneous and non-IID environments, where certain clients may contribute low-quality updates, slowing down convergence.

Federated learning studies in medical imaging commonly use FedAvg as the standard baseline, often comparing it with optimization-enhanced and personalized alternatives. For instance, [19] report an accuracy approximately 97% for FedAvg under IID conditions, while personalized approaches in [20] and [21] show AUROC values of 0.95 for FedAvg versus 0.94 for local models. These studies cover diverse imaging modalities such as MRI, CT, X-ray, OCT and dataset sizes ranging from 2,100 to over 84,000 images. Methods tailored to handle data heterogeneity, such as FedOpt, often demonstrate improved accuracy and robustness under non-IID conditions. However, most of these approaches (six out of seven) incur higher client-side computational costs, and communication overhead typically remains moderate to high.

The [22] study shows that using EfficientNet-B0 with the FedAvg algorithm in a federated learning framework improves privacy and diagnostic accuracy for MRI brain tumor detection. EfficientNet-B0 outperforms ResNet in handling data heterogeneity, emphasizing the potential of federated learning for robust medical image analysis.

Referred by [23] SecureFed, a secure federated learning aggregation method, outperforms FedAvg, FedMGDA+, and FedRAD in lung abnormality analysis using chest X-rays. It demonstrates superior robustness and fairness on COVID-19 datasets and is adaptable for multimodal medical data analysis.

The [24] study introduces Auto-FedAvg, a data-driven federated learning approach that dynamically adjusts aggregation weights based on data distribution and training progress. It outperforms existing methods on heterogeneous CIFAR-10 and proves effective in two medical imaging tasks: COVID-19 lesion segmentation in chest CT and pancreas segmentation in abdominal CT.

While FedAvg is simple and effective, it suffers from inefficiencies when dealing with non-IID data and heterogeneous client performance, leading to slower convergence and suboptimal accuracy.

2) Federated Proximal (FedProx): To address the limitations of FedAvg, FedProx [16] introduced a proximal term to stabilize local updates, especially when clients have heterogeneous computing power or varying data distributions. The modified objective function is:

$$w_k^{t+1} = \arg\min_w \left(F_k(w) + \frac{\mu}{2} \|w - w_t\|^2 \right)$$

where u controls how much a client update deviates from the global model. FedProx helps mitigate drift in non-IID settings but does not directly address client selection based on performance.

The author in [25] presents a comparative analysis of federated aggregation algorithms for binary classification of X-ray images across a limited number of hospitals with varying data heterogeneity. Among the evaluated methods, FedProx consistently outperforms others, making it the most effective approach in handling statistical variation in distributed medical imaging settings.

Referred by [26], FedProx outperforms FedAvg in federated learning for medical image analysis, demonstrating superior performance in cancer classification despite heterogeneous client data. However, increasing the number of communication rounds between the server and clients degrades model performance, affecting convergence and accuracy.

The author in [27] introduces a hybrid federated learning algorithm, FedProx:FedSplit, designed to tackle both statistical and system heterogeneity in medical data communication. This approach enhances model convergence and improves the overall accuracy of the global model in distributed healthcare settings.

This [28] study explores brain tumor detection using MRI images in a federated learning (FL) setup to preserve data privacy. A VGG19-based model was trained across four clients with non-IID data, incorporating Grad-CAM for explainability. Using a centralized test set for fair evaluation, the model achieved high accuracy of 97.18% using FedAvg, 98.24% using FedProx, and 98.45% using Scaffold. It demonstrating the effectiveness of FL in privacy-preserving, real-world medical imaging scenarios.

3) Federated Proximal (FedBN): Federated Batch Normalization (FedBN) [29]is a variant of federated learning designed to tackle non-IID (non-identically distributed) data across clients by adjusting how batch normalization layers are handled. In standard federated learning (e.g. FedAvg), all model parameters (including batch normalization statistics) are averaged across clients, which can be problematic when clients have very different feature distributions (a situation known as feature shift). FedBN addresses this by keeping the batch normalization parameters local to each client. In other words it does not include the BN layers running mean and variance in the global model aggregation [30].

Medical imaging federated learning often faces site-specific data distributions for example, MRI scans from different hospitals may have varying intensity distributions due to different scanners or protocols, even if the underlying task (e.g. tumor classification) is the same [29]. FedBN is well-suited to such settings: by not forcing a single global normalization, it allows each hospitals model to adjust to its own imaging characteristics while still contributing to a shared global model. This approach has been applied in a number of medical imaging tasks to improve generalization across institutions. For instance, researchers have evaluated FedBN on multi-center datasets for diagnostic classification and segmentation problems, observing better performance compared to vanilla federated training when data distributions differ [31]

Notably, FedBN has been used as a state-of-the-art baseline in: Histopathology Image Classification Cell Nuclei Segmentation MRI Segmentation The method's success in these peer-reviewed studies demonstrates how personalized normalization can boost performance in privacy-preserving medical AI collaborations.

While FedBN was introduced to address the issue of batch normalization (BN) layers in a non-IID setting. In standard BN, statistics are computed from local mini-batches that might not generalize properly across clients. FedBN sidesteps BN layers during aggregation, allowing the client-specific normalization parameters while aggregating the remaining ones. This modification improves generalization, thus preventing the collapse in performance due to misaligned feature statistics [13].

In FedBN, the model parameters are separated into two components:

$$w = \{w_{\text{shared}}, w_{\text{BN}}\}$$

where:

- w_{shared} are the weights of shared layers (e.g., convolutional layers),
- $w_{\text{BN}} = \{\gamma, \beta, \mu, \sigma^2\}$ are the batch normalization parameters: scale, shift, running mean, and running variance.

The FedBN update rules are then:

$$w_{\text{shared}}^{(t+1)} = \sum_{k=1}^{K} \frac{n_k}{n} w_{\text{shared},k}^{(t)} \tag{2}$$

$$w_{\text{BN},k}^{(t+1)} = w_{\text{BN},k}^{(t)} \quad \text{(not aggregated)} \tag{3}$$

This means batch normalization parameters remain local to each client to better handle feature distribution heterogeneity.

A. Federated Stochastic Gradient Descent (FedSGD)

FedSGD is a foundational algorithm in federated learning that updates the global model by aggregating gradients from multiple clients rather than full model weights. Unlike FedAvg, which averages model parameters after multiple local updates, FedSGD performs only one gradient computation per client per round, reducing local computation but increasing communication cost. The global objective is to minimize the weighted empirical risk:

$$\min_{\theta} f(\theta) = \sum_{k=1}^{K} \frac{n_k}{n} f_k(\theta)$$

Each client k computes its local gradient using the current global model θ^t:

$$g_k^t = \nabla f_k(\theta^t)$$

The server aggregates gradients from the selected clients $\mathcal{S}_t$ and updates the global model as:

$$\theta^{t+1} = \theta^t - \eta \sum_{k \in \mathcal{S}_t} \frac{n_k}{n_{\mathcal{S}_t}} g_k^t$$

If uniform weighting is used instead of sample-size-based weighting:

$$\theta^{t+1} = \theta^t - \eta \cdot \frac{1}{|\mathcal{S}_t|} \sum_{k \in \mathcal{S}_t} g_k^t$$

This approach ensures lightweight local computation and is suitable for bandwidth-constrained distributed settings.

FedAc [32] can achieve a linear speedup with fewer rounds of synchronization compared to FedAvg, improving communication efficiency. FedAc has stronger theoretical guarantees than FedAvg, particularly for functions that are third-order smooth. The authors developed a novel analysis approach and a strategic tradeoff between acceleration and stability to derive the FedAc algorithm.

The key idea of the FedSel framework [33] is to privately select the top-k most important dimensions in each iteration of

federated SGD, in order to reduce the noise injected into the gradients under local differential privacy. The authors claim that their FedSel framework outperforms the state-of-the-art solutions in terms of privacy, accuracy, and time complexity.Experiments on real-world and synthetic datasets verify the effectiveness and efficiency of the FedSel framework.

The paper [34] analyzes the convergence of local descent methods like Federated Stochastic Gradient Descent (FedSGD) for solving nonconvex optimization problems in federated learning with heterogeneous data.

Despite notable advances in federated aggregation strategies such as FedAvg, FedProx, FedBN, and FedSGD, existing approaches remain limited in their ability to adaptively account for client performance quality under highly heterogeneous and non-IID medical data. Most methods either assume uniform participation (FedAvg), add complexity with sensitive hyperparameters (FedProx), or address only feature shift without considering contribution reliability (FedBN), while lightweight approaches like FedSGD incur high communication costs without performance-based filtering. Furthermore, recent personalized and optimization-driven extensions improve convergence but often impose higher computational and communication overhead, making them less practical for resource-constrained healthcare environments. A clear gap therefore exists in the literature for a performance-aware, lightweight, and selective aggregation mechanism that not only filters out low-quality updates but also ensures inclusivity for weaker clients to progressively improve an area where FedPA positions itself as a novel and practical contribution.

III. Federated Learning Performance-based Averaging (FedPA)

Federated Averaging (FedAvg) is one of the most widely adopted aggregation methods in federated learning due to its simplicity and communication efficiency [15]. However, it assumes uniform participation and contribution quality from all clients, which rarely holds in real-world settings. Client availability often varies due to network instability, power constraints, and hardware limitations, leading to unreliable or delayed updates. This inconsistency results in unstable global model updates, inefficient optimization, and reduced model quality especially when high-performing clients are absent from certain rounds. Furthermore, FedAvg lacks mechanisms to differentiate between high- and low-quality updates, allowing poor local models to degrade overall performance, particularly when data quality, size, or computational resources vary significantly across clients.

To overcome this limitation, Performance-Based FedPA is introduced. This approach integrates a client contribution evaluation mechanism that selectively aggregates model updates from high-performing clients while still allowing lower-performing clients to participate and improve. This strategy ensures that the global model prioritizes high-quality updates, leading to improved overall performance and stability.

Step 1: Initialization

- Initialize the global model w_0.
- Set the performance threshold τ (e.g., 80% accuracy).
- Define all participating clients K.

Step 2: Federated Learning Process

For each global round $t = 1, 2, \ldots, T$:

- Server selects a subset of clients $S_t \subseteq K$.
- Each client $k \in S_t$:
  - Trains its local model w_k^t using its dataset.
  - Evaluates its local model accuracy:

$$\text{Accuracy}(w_k^t)$$

  - If $\text{Accuracy}(w_k^t) \geq \tau$, the client is added to the selected clients set C_{selected}:

$$C_{\text{selected}} = \{k \mid \text{Accuracy}(w_k^t) \geq \tau\}$$

  - If $\text{Accuracy}(w_k^t) < \tau$:
    - The client still receives the global model update but does not contribute to aggregation.
    - This ensures continuous learning for weaker clients.

Step 3: Optimization and Weight Computation

For each selected client $k \in C_{\text{selected}}$:

- Compute the contribution weight α_k based on accuracy:

$$\alpha_k = \frac{\text{Accuracy}(w_k^t)}{\sum_{j \in C_{\text{selected}}} \text{Accuracy}(w_j^t)}$$

- Compute the local model update Δw_k^t:

$$\Delta w_k^t = w_k^t - w_t$$

Step 4: Aggregation Using FedAvg

Once all selected clients send their updates to the server, the global model is updated using FedAvg.

Using the weighted averaging formula:

$$w_{t+1} = \sum_{k \in C_{\text{selected}}} \frac{\alpha_k}{N} w_k^t$$

where:

- $N = |C_{\text{selected}}|$, the number of selected clients.
- Stable and effective for non-IID data.

- Reduces variance in updates.
- Slower convergence but better final accuracy.

Step 5: Updating the Global Model

- Server broadcasts the updated global model w_{t+1} to all clients.
- The process repeats for T rounds.
- At the end, the final model w_T is returned.

A. Advantages of Performance-Based FedPA

1) Improved global model accuracy: By prioritizing contributions from high-accuracy clients, the global model avoids the negative impact of poor updates, leading to faster and more stable convergence. Ensures that well-trained models influence learning while still allowing weaker models to improve.

2) Faster convergence: Since only high-performing models significantly influence the global update, the model learns more effectively per round, reducing the number of communication rounds required for convergence. Avoids instability caused by noisy updates from clients with poor local training.

3) Adaptive participation for weaker clients: Clients not meeting the accuracy threshold are still included in the federated learning process, receiving global updates and improving over multiple rounds. This encourages progressive learning, allowing weaker models to gradually enhance their accuracy and contribute more effectively in future rounds.

4) Robustness to Non-IID data: Real-world federated learning scenarios, such as multi-institutional medical imaging, often involve non-IID (heterogeneous) datasets. Performance-Based FedPA naturally adapts to these conditions by giving higher importance to clients with better model performance, which mitigates the impact of data heterogeneity.

IV. Methodology

A. Data Acquisition and Preprocessing

This study is based on dataset images of the chest radiograph classifying images into four classes-Normal, COVID-19, Viral Pneumonia, and Lung Opacity. Data from this collection used for the study is from publicly available clinical repositories to make sure the patient's demographics and radiological conditions are as diverse as needed. Such diversity will strengthen the generalization capability of the trained models for better and more effective application in real-world clinical conditions (Fig. 3).

Each image was preprocessed through a similar standard pipeline that was used across federated clients. All images were resized to a consistent input resolution of 224x224 pixels, which is widely adopted in convolutional neural network (CNN) models. Afterward, the pixel intensities were normalized by scaling the values into the range from [0, 1]. This definition of normalization was further adjusted per channel by mean and standard deviation values calculated on the training set to standardize input data of different images.

Various data augmentation techniques were utilized during training to reduce overfitting and improve the generalization abilities of the model. These augmentation methods included random horizontal flipping, small-angle rotations, random cropping, and intensity rescaling. The augmentations mimic image acquisition variability across the various institutions and imaging equipment, which is one of the main requirements for building robust models in medical image analysis. These techniques improve model robustness by preparing a model that performs better in real data differing from the acquisition conditions and quality.

All the changed images, label and metadata have been stored in an organized NumPy archive (covid_dataset.npz) to directly load and access such data during the training operation. This format is readily available for any federated client to train on such dataset providing a facility to have a smooth Read-Write process in distributed learning in healthcare AI systems.

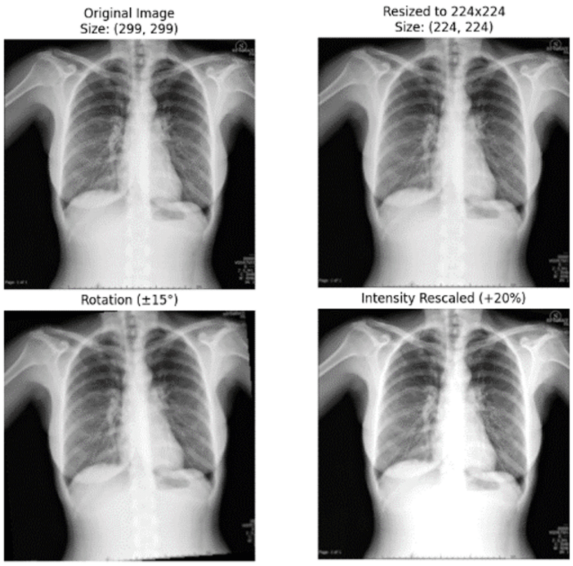

Fig. 3. This figure shows the preprocessing pipeline, the original image is resized into 224 x 224 size. This is further rotated and has its intensities increased for better feature recognition and generalization.

B. Data Partitioning and Distribution Strategy

The partitioning of the preprocessed chest X-ray dataset into four virtual clients simulating different hospitals or settings under federated learning, therefore representing three thoracic diseases-COVID-19, Pneumonia, and Lung Opacity-and one Normal class, would provide insight into these instances' diverse medical diagnosis as seen in real-world clinical practice. A truly non-IID data partitioning approach for federated learning algorithms is justified because, in practice, medical data can vary from one institution to another depending on specialization, as well as geographical attractiveness for certain diseases. Now, the data of each client was divided into a training set (70%), a validation set (15%), and a global test set (15%). The training and validation data were kept by the local class distributions, whereby the model of each client's training learned from an equal balance of images to prevent any biases

that could have arisen from unequal class proportions. The global test set, which holds 15% of the total dataset, was evenly distributed across all classes, allowing it to be completely separate to ensure an unbiased evaluation of the aggregated model's performance. This partitioning strategy exploits natural statistical variability that different clients have for the evaluation of federated learning methods under decentralized, imbalanced, and real-world clinical data conditions. Such a mechanism becomes necessary to allow for the generalization of the trained models over dissimilar clinical terrains, while addressing nature issues like data heterogeneity and class imbalance.

C. Model Architecture

Proposed model works on deep learning algorithms to classify chest X-ray images under four categories: Normal, COVID-19, Lung Opacity, and Pneumonia. The architecture promises extraction of multi-level spatial features with a reduction in computational complexity, thereby making it an ideal flaunt for federated learning settings.

1) The input layer: images of size 224 x 224 in grayscale feed into the model such that all the images could also undergo pre-processing in the same way for all federated clients.

2) Convolutional layer: Three convolutional blocks are required to capture the most relevant features such as edges, textures, patterns related to abnormalities in the lungs, and consist of a convolutional layer each followed by ReLU activation and max pooling.

3) Flatten layer: This layer converts 2-D feature maps made from the convolutional layers into single one-dimensional vectors for subsequent processing.

4) Dense layer: This is a dense layer built with 128 neurons and by ReLU activation to model very complex behaviours. A dropout layer is included at 30% rate to reduce overfitting and thus enhance generalization across clients.

5) Output layer: There are 4 neurons that correspond to classes using the SoftMax activation function for multi-class classification.

The model is optimized with the categorical cross-entropy loss for multi-class classification and is best suited for fast training by the Adam optimizer. It is a lightweight model that affords on-the-fly training on pretty cheap machines on the client-side in a federated learning system.

D. Global and Local Model Training in Federated Learning

In federated learning, model training involves two main phases: local model training and global model aggregation. These phases are iteratively repeated to make the global model better while ensuring data privacy.

1) Client model training: The local model for a specific client will be trained based on its local private data and the updates to be sent to the global server will only be that specific model updates such as weights or gradients. The training procedure involves:

a) Model initialization: Clients receive the global model weights at the beginning of the training round.

b) Local epochs: Clients train their model for a set number of epochs on local data.

c) Update sharing: After training, each client computes the updates (model weights or gradients) before sending them back to the global server.

This decentralized structure ensures that client sensitive data never leaves clients.

2) Global model training: The global server aggregates all local updates to improve the global model after all the local updates have been collected. This will include averaging the weights of all models from participating clients after the process. The most common aggregation algorithms include the following:

a) FedAvg: This is basically the first federated learning algorithm where the server computes a weighted factor average of model updates to carry out its aggregation. This would be especially useful when the distribution of data was relatively similar among the clients.

b) FedProx: This adds a proximal term in a local objective with respect to deviation from the global model for local models. This is usually good where very much dissimilar data exist across clients.

c) FedBN: This is a specific method designed for very different data distributions across client bases, as might occur in medical imaging tasks. FedBN keeps different batch normalization statistics for each client to handle domain shifts, thereby completely solving the problem of tasks.

d) FedSGD: Unlike FedAvg, which uses model weights, FedSGD is used to aggregate gradients instead of model parameters. It thus minimizes data exchange between clients and the server, thus optimally suited under low-latency communication networks.

e) FedPA (Federated Performance Averaging): With this new method, performance-based client selection is introduced. So in FedPA, a client must surpass a certain performance threshold to add its model updates to the global model.

Ensure that the federated learning algorithms harness the global model through the knowledge gained from each client while having local data private. They are especially concerned with healthcare and medical imaging since they leverage data diversity to deal with privacy alongside efficient model aggregation (Fig. 4).

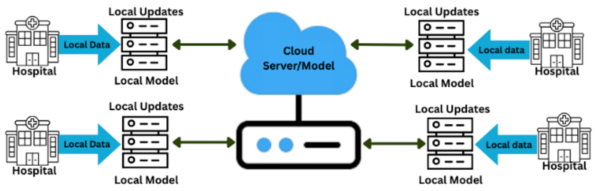

Fig. 4. Federated learning environment whereby hospitals learn and update local models on local patient data and use these to periodically update a shared central server for collaborative model updates without patient-data sharing.

V. EXPERIMENTS

A. Hardware and Software Requirements

The experiments were conducted on a workstation running Windows 11, equipped with an Intel Core i7 processor, 16 GB of RAM, and an NVIDIA RTX 4050 GPU with 12 GB of VRAM. The software environment was configured with Python 3.10, PyTorch 2.0.1, Torchvision 0.15.2, NumPy 1.24, scikit-learn 1.2, tqdm 4.65, and the CUDA Toolkit version 11.7. Model training was accelerated using CUDA where applicable, leveraging the available GPU resources. All models were implemented in PyTorch and executed in a single-machine, single-GP

B. Hyperparameter Settings

A careful selection of hyperparameters is critical in federated learning, as it significantly influences convergence speed, model generalization, and robustness to non-IID medical data distributions. This aligns with prior research in federated medical imaging and deep learning optimization. To ensure fair evaluation across all methods, the hyperparameters were standardized, drawing on empirical validation and insights from earlier studies addressing model regularization, optimizer behavior, and communication efficiency in federated settings.

Throughout all experiments, consistent hyperparameter settings were maintained. The AdamW optimizer was chosen for its improved handling of weight decay over standard Adam, with a stable initial learning rate of 0.001. A cosine annealing learning rate scheduler was applied at every epoch to adjust the learning rate dynamically. A weight decay of 0.01 was used to mitigate overfitting. The loss function employed was cross-entropy loss, a standard choice for classification tasks. A batch size of 16 was selected to balance memory usage and gradient estimation accuracy. Each client performed three local epochs per round, and all federated learning variants (FedAvg, FedProx, FedBN, and FedPA) were trained over 14 communication rounds. Data augmentation techniques, including random horizontal flips and rotations, were applied during training to enhance generalization. A DenseNet-based convolutional neural network served as the common base architecture across all approaches.

These hyperparameter values were fixed for all experiments to ensure both comparability and reproducibility, in line with best practices recommended in recent federated learning studies focused on medical imaging.

C. Client Configuration and Performance-Based Participation

As viewed from this research perspective, the federated learning setup imitates the real-life medical setting where the data distributions are inherently non-IID and client-specific. Four clients were set up to represent the four separate clinical categories from the chest X-ray dataset: Normal, COVID-19, Lung Opacity, and Pneumonia. Each of the four clients had access exclusively to its own set of labeled images, creating heterogeneity in class distribution and simulating isolated data silos of hospitals or diagnostic centers.

In the training process, each client was intended to keep optimizing its copy of the global model independently for five epochs per round using local data. This practice seems to fit well with reality since in federated health care, local computation is generally chosen over frequent communication due to privacy and latency issues. Nevertheless, global performance could be compromised if all clients participated uniformly, as when some clients perform really poorly due to poor data quality or too few training samples.

Consequently, the proposed FedPA framework implemented a performance-based client selection mechanism. At the end of every round, the clients assessed their local model(s) on a validation split and saved the best checkpoint. Only clients that achieved at least 85% validation accuracy were eligible to take part in the global aggregation. Generalization and robustness in federated learning can benefit from contribution-based client selection, as earlier studies have pointed out..

This participation filter remained in effect throughout the 15 communication rounds. FedPA, by removing low-quality model updates, fosters stable learning and reduces the risk of drift in the central model, which is often due to unreliable model updates by clients; a challenge posed in federated scenarios by heterogeneous and imbalanced datasets.

D. Global Aggregation Strategy

For every modification of the global model, there was an aggregation of local models by the participating clients at the end of each communication round, though significantly different regarding the mechanism of aggregation. In FedAvg, aggregation used to be done by the average of all client updates uniformly, presuming the reliability of all participants is equal. On a somewhat similar note, FedProx did not introduce any method of aggregation but provided a proximal term to recursively impose on local training so that students would penalize divergence from the global model within the locality, indirectly affecting aggregation quality. For another communication-intensive approach, FedSGD used gradients from each client, not the model updates but only required synchronous updates and higher communication overhead. FedBN performed normal averages on parameters but excluded all batch normalization layers from the averaging process such that the client held domain-specific statistics on normalization-an important feature in non-IID settings.

By contrast, the proposed FedPA approach redefined aggregation through a selection mechanism based on performance: It's not the fact that all clients contributed updates that matters, but instead, FedPA had just included clients whose local model outperformed a given validation accuracy threshold-all the way in this case, at or above the bar of 85 percent. Clients not meeting this criterion are not included in that specific round's aggregation. The selective strategy ensured that noisy and inferior updates were filtered out and only the well-trained local model participated in the global parameters. Thus, FedPA offered a greater quality enhancement under client performance and data heterogeneity-influencing factors concerning simple convergence and robustness, primarily in complicated evidence-dependent contexts, for example, medical imaging..

E. Evaluation Metrics

The performance of the federated learning models was evaluated based on the following main metrics

- Accuracy: It is the metric that indicates the overall performance of the model at classifying chest X-ray images correctly.

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \quad (4)$$

- Precision: Indicates the ability of the model to reduce false positives by correctly identifying only true positive instances.

$$Precision = \frac{TP}{TP + FP} \quad (5)$$

- Recall : Reflects the model's ability to correctly identify all positive instances, minimizing false negatives.

$$Sensitivity = \frac{TP}{FN + TP} \quad (6)$$

- F1 Score: It combines precision and recall into one measure and therefore sets a balance between the two by calculating their harmonic mean.

$$F1\ score = 2 * \frac{precision * recall}{precision + recall}. \quad (7)$$

- Inference time (s): Time needed to make predictions based on an input offered to the model by the user Additional measurements, such as CPU usage, memory usage, average inference time ,total inference time, throughput , were examined to determine the computational efficiency of the models.

VI. Result and Analysis

This section portrays a comparative analysis of the five federated learning models such as FedAvg, FedBN, FedPA, FedProx, and FedSGD. Evaluated on a common test set under exactly identical hardware and software conditions. The effect of the models is analyzed in determining the best among them in terms of the trade-off between performance in classification and efficiency at the system level for real-world deployment in almost all medical applications, such as chest X-ray analysis (Table I).

TABLE I. Performance Metrics Comparison

Method	Accuracy	Precision	Recall	F1 Score
FedAvg	0.9140	0.9259	0.9157	0.9206
FedBN	0.9124	0.9326	0.9053	0.9180
FedPA	0.9182	0.9295	0.9208	0.9248
FedProx	0.9113	0.9295	0.9071	0.9175
FedSGD	0.9129	0.9296	0.9102	0.9192

The standard evaluation metrics chosen for models are accuracy, precision, recall, and F1 score; each of these has its own clinical implications, especially in tasks related to disease detection.

1) Accuracy: FedPA obtained an accuracy of 91.82%, which is greater than FedAvg with 91.39% and FedSGD with 91.29%. This was achieved by performing client selection based on performance, excluding undertrained or noisy updates, which tended to minimize global model variance and aid generalization. FedAvg and FedSGD combine all clients equally, thus muting the effect of high-performing clients and slowing down convergence.

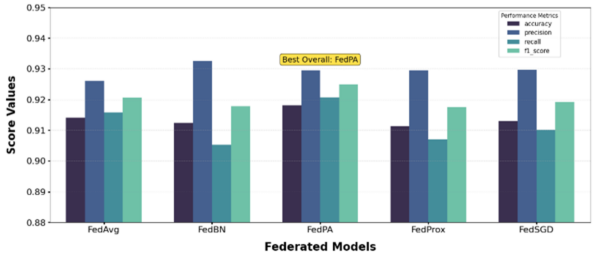

Fig. 5. Comparison of predictive performance metrics (Accuracy, Precision, Recall, F1 Score) of various federated models.

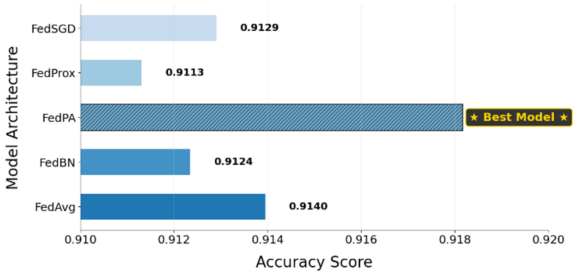

Fig. 6. Comparison of Accuracy across multiple federated models.

Precision FedBN leads in precision at 93.26% due to the use of client-specific batch normalization, thereby helping the model adapt to non-IID data distributions across clients. FedPA, close behind at 92.95%, strikes a good balance in precision without sacrificing too much recall. The implication is that FedBN really tries not to miss negative samples while FedPA has a better clinical trade-off by reducing the risk of under-detecting actual cases.

2) Memory: With 92.08%, FedPA achieves the highest recall performance, which indicates that it captures a lot of true positives, especially in lives-critical contexts like conversations about the identification of COVID-19 or pneumonia. This selective aggregation from FedPA contributes to this recall because high-performers among clients amplify the visibility of their presence to the global model and, hence, facilitate better identification of minority or diffused class patterns, even in the presence of data heterogeneity.

F1-ScoreFedPA has again captured the F1 score of 92.48%, thereby making it possible for it to prove its efficiency with respect to false positive or false negative prediction. With FedSGD and FedAvg coiled very closely behind, neither has the robustness of FedPA because it does not filter clients. The harmonic property of the F1 score underlines how selective aggregation according to FedPA effects compromise between sensitivity and specificity, both very important in terms of clinical reliability (see Fig. 5 and 6).

A confusion matrix is a table used to evaluate the performance of a classification model by comparing predicted labels with true labels. It is especially useful for understanding the types of errors a model makes. Here's the general structure for a binary classification confusion matrix (see Fig. 7 to 11):

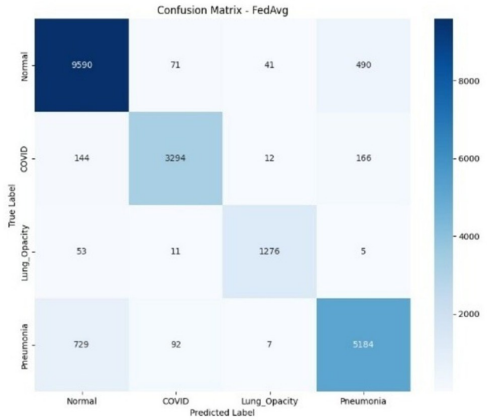

Fig. 7. Confusion Matrix of FedAvg.

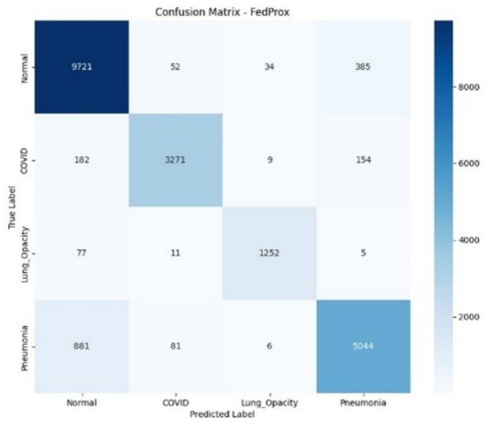

Fig. 8. Confusion matrix of FedProx.

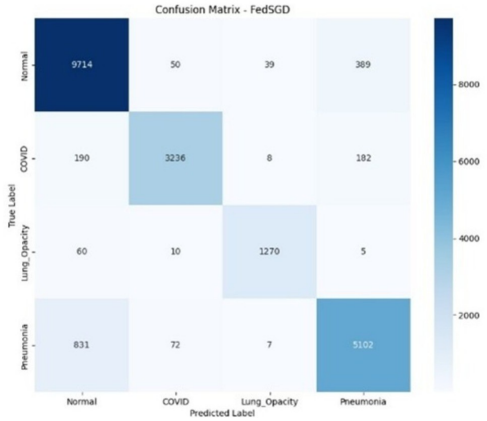

Fig. 9. Confusion matrix of FedSGD.

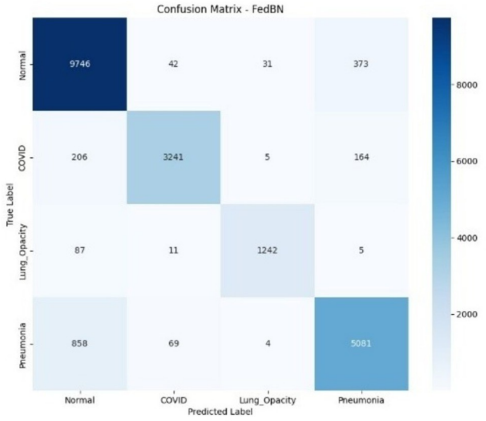

Fig. 10. Confusion matrix of FedBN.

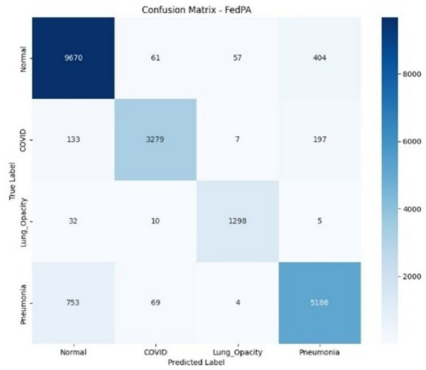

Fig. 11. Confusion matrix of FedPA: Proposed method.

A. System-Level Evaluation Metrics

This is now an assessment by which each model is compared for efficiency mainly due to the inference time, throughput, CPU usage, and memory footprint that are necessary for real-time medical AI deployment (Table II).

TABLE II. System Metrics Comparison

Method	Average Inference Time (ms)	Throughput	CPU Usage (%)	Memory Usage (MB)
FedAvg	0.003427	9328	13.15	12976
FedBN	0.002838	11266	13.11	13013
FedPA	0.002810	11377	12.85	13028
FedProx	0.002794	11442	12.98	13040
FedSGD	0.002812	11368	12.13	13053

1) Inference time: In terms of inference time, FedProx and FedPA are the fastest (2.8 ms/image), achieved by lightweight parameter updates and subsequently reduced model divergence during training. FedAvg comes in at 3.4 ms, likely due to model instability because of noisy client aggregation.

2) Throughput: FedProx attains the highest throughput (11,442 images/s) closely followed by FedPA (11,376) and FedSGD (11,368). These models benefit from efficient forward pass computation; besides, FedPA benefits from stabilized model updates that prevent computational spikes during testing. FedAvg, with a throughput of 9338, is thus lower due to its slow inference time and an increase in computational load.

3) CPU usage: FedSGD is the lightest on the CPU (12.13%), followed by FedPA (12.85%). Their minimal effort on steep client-side operations and sparse updates reduces compute requirements, thus lending themselves to edge environments. The higher CPU usage of FedAvg (13.15%) is likely due to its dense parameter averaging additionally worsened by client drift in non-IID settings.

4) Memory usage: Models cluster around 13 GB, with FedSGD (13.05 GB) and FedPA (13.11 GB) being somewhat lighter on memory. Since the overhead of memory and CPU power is manageable, these systems may also be deployed on a resource-limited scenario like mobile clinics (Fig. 12).

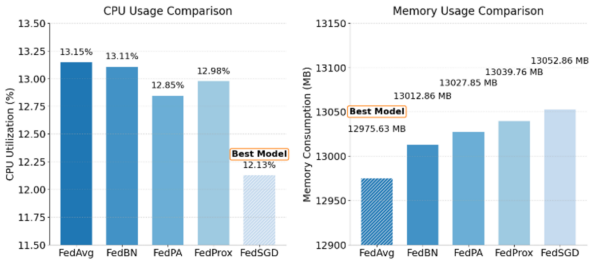

Fig. 12. Relative comparison of system resource usage between models with CPU usage and memory usage set side by side to emphasize the computational efficiency of the models.

B. Summarized Key Insights

We combine performance measures and systems metrics to identify the best model, lightest model, and most balanced model.

1) Best performing model: FedPA is characterized by high accuracy, recall, and F1 scores with low inference time and moderate resource consumption. Its inclusion strategy includes only reliable updates in the global model computation, thereby increasing the robustness of the FedPA in non-IID settings.

2) Most lightweight model: The FedSGD is by far the most lightweight in terms of CPU and memory resource consumption while still keeping throughput at a high level. It does have slightly less accuracy in the classification realm, but it is best suited for scenarios where resource constraints apply.

3) Most balanced model: FedPA, once again, sets the mark by balancing state-of-the-art accuracy with reasonable performance metrics. FedPA considers both efficiency and accuracy in diagnosis, a crucial capability for healthcare AI. In contrast, FedProx sacrifices accuracy for throughput (Fig. 13).

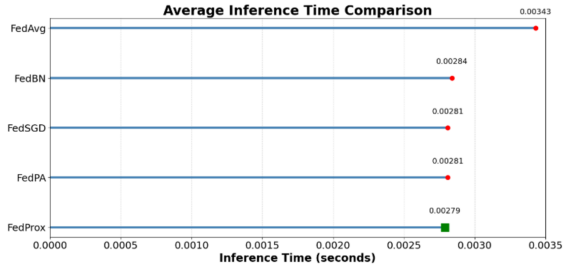

Fig. 13. Spider chart comparison of all eight test measures across all the models with a focus on the general predominance of FedPA regarding predictive performance, system efficiency, and consumption of resources.

C. Strengths of FedPA in Federated Learning

It is performance-aware aggregation that FedPA relies on for its superiority, where applying the client changes under a local accuracy condition of 85% would filter out updates and speed convergence, all while not falling into the trap of overfitting for the dominant client distribution.

Unlike FedAvg and FedProx, which accept all updates equally, FedPA selectively amplifies clients not by each, but only by those contributing meaningfully to learning and thus reducing negative transfer-the well-known problem of non-IID federated settings. It is both economical of communication bandwidth use, and doesn't waste computation cycles on low-quality updates; it is hence scalable and robust.

High recall and F1-score metrics further justify the clinical relevance of FedPA, especially in contexts where increased false negatives can lead to fatalities. Moreover, it is extremely stable across rounds and generalizes strongly; hence, it is earmarked for federated medical applications in which the parameters of accuracy and reliability are critical.

VII. Conclusion

This study introduced Federated Performance-Based Averaging (FedPA), a performance-aware federated learning approach designed to address challenges of data heterogeneity, variable client performance, and privacy concerns in medical imaging. Unlike traditional aggregation strategies such as FedAvg, FedProx, and FedSGD, which combine all client updates indiscriminately. FedPA selectively incorporates only high-performing client models based on a predefined accuracy threshold. This strategy helps to mitigate the negative impact of undertrained or noisy clients on global model performance.

In the context of COVID-19 chest X-ray classification, FedPA consistently outperformed other federated algorithms across all major evaluation metrics, achieving the highest accuracy, recall, and F1-score. It also demonstrated faster convergence and greater training stability, with reduced performance fluctuations and better retention across both strong and weak clients.

Looking ahead, several enhancements can further strengthen FedPA. These include adaptive thresholding that dynamically adjusts performance criteria across training rounds, and techniques to handle label noise through confidence-based filtering. Incorporating federated hyperparameter optimization methods may improve

convergence and fairness, while expanding to multi-modal data such as CT, MRI, and electronic health records could broaden the model's clinical utility. Real-world validation across geographically diverse hospital networks would also be essential to assess FedPA's scalability, generalization, and equity in healthcare delivery.

Overall, FedPA represents a significant advancement for privacy-preserving federated learning in healthcare. Its strong diagnostic performance, stability, and system efficiency make it a practical candidate for deployment in real-world clinical environments where centralized data access is limited.

A. Future Work

In this study, the evaluation of the proposed FedPA framework primarily focused on accuracy as the performance threshold for client selection. While this provided promising improvements in robustness and convergence, accuracy alone may not sufficiently capture the impact of class imbalance, which is a common challenge in medical imaging tasks. As part of future work, we aim to extend the thresholding mechanism by incorporating the F1-score alongside accuracy, thereby ensuring that both precision and recall are adequately considered in performance-based aggregation. This dual-metric approach will allow the framework to better handle skewed class distributions and minority class detection. Additionally, we plan to design adaptive selection strategies that dynamically choose the most suitable performance metric (e.g., accuracy, F1, precision, recall, or AUC) depending on the dataset characteristics and clinical context. Beyond this, ablation experiments will be conducted to systematically evaluate the contribution of each component in the FedPA pipeline, enabling us to identify the most effective configurations for stability and generalization. Ultimately, these enhancements aim to finalize FedPA into a robust, performance-aware federated learning algorithm that is well-suited for real-world commercial deployment in healthcare AI applications.

Author Contributions

Atif Mahmood and Tashin Khan Sadique are the main contributors to the conceptualization, methodology, and manuscript preparation. The remaining authors contributed to data analysis, validation, and review of the manuscript.

Data Availability Statement

The study utilized publicly available datasets. The dataset and corresponding code will be made available upon reasonable request.

References

[1] F. Fatoni, T. B. Kurniawan, D. A. Dewi, M. Z. Zakaria, and A. M. M. Muhayeddin, "Fake vs real image detection using deep learning algorithm," *Journal of Applied Data Sciences*, vol. 6, no. 1, pp. 366–376, 2025.

[2] D. S. Ting, Y. Liu, P. Burlina, X. Xu, N. M. Bressler, and T. Y. Wong, "Ai for medical imaging goes deep," *Nature medicine*, vol. 24, no. 5, pp. 539–540, 2018.

[3] S. Iqbal, A. N. Qureshi, J. Li, and T. Mahmood, "On the analyses of medical images using traditional machine learning techniques and convolutional neural networks," *Archives of Computational Methods in Engineering*, vol. 30, no. 5, pp. 3173–3233, 2023.

[4] A. S. Panayides, A. Amini, N. D. Filipovic, A. Sharma, S. A. Tsaftaris, A. Young, D. Foran, N. Do, S. Golemati, T. Kurc *et al.*, "Ai in medical imaging informatics: current challenges and future directions," *IEEE journal of biomedical and health informatics*, vol. 24, no. 7, pp. 1837–1857, 2020.

[5] K. Kourou, T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis, and D. I. Fotiadis, "Machine learning applications in cancer prognosis and prediction," *Computational and structural biotechnology journal*, vol. 13, pp. 8–17, 2015.

[6] C. A. Gao, N. S. Markov, T. Stoeger, A. Pawlowski, M. Kang, P. Nannapaneni, R. A. Grant, C. Pickens, J. M. Walter, J. M. Kruser *et al.*, "Machine learning links unresolving secondary pneumonia to mortality in patients with severe pneumonia, including covid-19," *The Journal of Clinical Investigation*, vol. 133, no. 12, 2023.

[7] P. Khan, M. F. Kader, S. R. Islam, A. B. Rahman, M. S. Kamal, M. U. Toha, and K.-S. Kwak, "Machine learning and deep learning approaches for brain disease diagnosis: principles and recent advances," *Ieee Access*, vol. 9, pp. 37 622–37 655, 2021.

[8] A. Kilic, "Artificial intelligence and machine learning in cardiovascular health care," *The Annals of thoracic surgery*, vol. 109, no. 5, pp. 1323–1329, 2020.

[9] A. Barragán-Montero, U. Javaid, G. Valdés, D. Nguyen, P. Desbordes, B. Macq, S. Willems, L. Vandewinckele, M. Holmström, F. Löfman *et al.*, "Artificial intelligence and machine learning for medical imaging: A technology review," *Physica Medica*, vol. 83, pp. 242–256, 2021.

[10] W. Salah Eldin and A. Kaboudan, "Ai-driven medical imaging platform: Advancements in image analysis and healthcare diagnosis," *Journal of the ACS Advances in Computer Science*, vol. 14, no. 1, 2023.

[11] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Artificial intelligence and statistics*. PMLR, 2017, pp. 1273–1282.

[12] M. Adnan, S. Kalra, J. C. Cresswell, G. W. Taylor, and H. R. Tizhoosh, "Federated learning and differential privacy for medical image analysis," *Scientific reports*, vol. 12, no. 1, p. 1953, 2022.

[13] M. N. Hossen, K. Ahmed, F. M. Bui, and L. Chen, "Fedrsmax: An effective aggregation technique for federated learning with medical images," *2023 IEEE Canadian Conference on Electrical and Computer Engineering (CCECE)*, pp. 229–234, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:264521437

[14] L. Kwak and H. Bai, "The role of federated learning models in medical imaging," *Radiology: Artificial Intelligence*, vol. 5, no. 3, p. e230136, 2023.

[15] T. Li, M. Sanjabi, A. Beirami, and V. Smith, "Fair resource allocation in federated learning," *arXiv preprint arXiv:1905.10497*, 2019.

[16] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith, "Federated optimization in heterogeneous networks," *Proceedings of Machine learning and systems*, vol. 2, pp. 429–450, 2020.

[17] V. Smith, C.-K. Chiang, M. Sanjabi, and A. S. Talwalkar, "Federated multi-task learning," *Advances in neural information processing systems*, vol. 30, 2017.

[18] H. Zhao, "Non-iid quantum federated learning with one-shot communication complexity," *Quantum Machine Intelligence*, vol. 5, no. 1, p. 3, 2023.

[19] S. Amgain, P. Shrestha, S. Bano, I. del Valle Torres, M. Cunniffe, V. Hernandez, P. Beales, and B. Bhattarai, "Investigation of federated learning algorithms for retinal optical coherence tomography image classification with statistical heterogeneity," *ArXiv*, vol. abs/2402.10035, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:267681948

[20] H. Guan, P.-T. Yap, A. Bozoki, and M. Liu, "Federated learning for medical image analysis: A survey," *Pattern Recognition*, p. 110424, 2024.

[21] 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 diagnosis using chest radiographs from 42 us and european hospitals," *Journal of the American Medical Informatics Association*, vol. 30, no. 1, pp. 54–63, 2023.

[22] L. Zhou, M. Wang, and N. Zhou, "Distributed federated learning-based deep learning model for privacy mri brain tumor detection," *arXiv preprint arXiv:2404.10026*, 2024.

[23] A. Makkar and K. Santosh, "Securefed: federated learning empowered medical imaging technique to analyze lung abnormalities in chest x-rays," *International Journal of Machine Learning and Cybernetics*, vol. 14, no. 8, pp. 2659–2670, 2023.

[24] Y. Xia, D. Yang, W. Li, A. Myronenko, D. Xu, H. Obinata, H. Mori, P. An, S. A. Harmon, E. B. Turkbey, B. I. Turkbey, B. J. Wood, F. Patella, E. Stellato, G. Carrafiello, A. M. Ierardi, A. L. Yuille, and H. R. Roth, "Auto-fedavg: Learnable federated averaging for multi-institutional medical image segmentation," *ArXiv*, vol. abs/2104.10195, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:233324290

[25] M. Delehouzée, X. Lessage, T. Reginster, and S. Mahmoudi, "Performance analysis of aggregation algorithms in cross-silo federated learning for non-iid data," *2024 4th International Conference on Embedded & Distributed Systems (EDiS)*, pp. 74–79, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:274708806

[26] M. Subramanian, V. Rajasekar, S. VE, K. Shanmugavadivel, and P. Nandhini, "Effectiveness of decentralized federated learning algorithms in healthcare: a case study on cancer classification," *Electronics*, vol. 11, no. 24, p. 4117, 2022.

[27] C. Mathew and P. Asha, "Fedprox: Fedsplit algorithm based federated learning for statistical and system heterogeneity in medical data communication," *Journal of Internet Services and Information Security*, vol. 14, no. 3, pp. 353–370, 2024.

[28] M. Muntaqim and T. A. Smrity, "Federated learning framework for brain tumor detection using mri images in non-iid data distributions," *Journal of Imaging Informatics in Medicine*, pp. 1–19, 2025.

[29] Z. Zhou, G. Luo, M. Chen, Z. Weng, and Y. Zhu, "Federated learning for medical image classification: A comprehensive benchmark," *arXiv preprint arXiv:2504.05238*, 2025.

[30] N. Koutsoubis, A. Waqas, Y. Yilmaz, R. P. Ramachandran, M. Schabath, and G. Rasool, "Future-proofing medical imaging with privacy-preserving federated learning and uncertainty quantification: A review," *arXiv preprint arXiv:2409.16340*, 2024.

[31] 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*, vol. 36, no. 1, 2022, pp. 1087–1095.

[32] H. Yuan and T. Ma, "Federated accelerated stochastic gradient descent," *ArXiv*, vol. abs/2006.08950, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:219708203

[33] R. Liu, Y. Cao, M. Yoshikawa, and H. Chen, "Fedsel: Federated sgd under local differential privacy with top-k dimension selection," in *Database Systems for Advanced Applications: 25th International Conference, DASFAA 2020, Jeju, South Korea, September 24–27, 2020, Proceedings, Part I 25*. Springer, 2020, pp. 485–501.

[34] F. Haddadpour and M. Mahdavi, "On the convergence of local descent methods in federated learning," *ArXiv*, vol. abs/1910.14425, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:207757958