open access

Published online on: https://journal.uii.ac.id/jurnalsnati/

Check for updates

DOI: 10.20885/snati.v5.i1.44597

JURNAL SNATI
Sains, Nalar, dan Aplikasi Teknologi Informasi

Jurnal Sains, Nalar, dan Aplikasi Teknologi Informasi

Vol. 5 No. 1 (2026) 32 - 44 ISSN Media Electronic: 2807-5935

Comparative Evaluation of Federated Learning Algorithms in Dirichlet Non-IID Medical Imaging

Michael Angello Qadosy Riyadi[1]*, Adinda Mariasti Dewi[2], Zahid Abdullah Nur Mukhlishin[3], Zalsabilah Rezky Amelia Arep[4]
[1,2]Department of Information Technology, Telkom University, Surabaya Campus, Surabaya, Indonesia
[3,4]Department of Data Science, Telkom University, Surabaya Campus, Surabaya, Indonesia
[1]michaelangello@student.telkomuniversity.ac.id, [2]adindamariasti@student.telkomuniversity.ac.id,
[3]zahidabdullahnur@student.telkomuniversity.ac.id, [4]zalsabilahrezky@student.telkomuniversity.ac.id

Abstract

Machine learning has achieved diagnostic performance comparable to clinical experts on medical imaging, yet centralized training paradigms necessitate patient data aggregation, risking violations of privacy regulations such as GDPR and HIPAA. In 2023, 1,853 healthcare data breaches were reported in the United States, compromising over 133 million medical records, rendering raw inter-institutional data exchange increasingly unsustainable. Federated Learning (FL) offers a viable solution by enabling collaborative model training without data transfer. However, prior studies predominantly evaluate single algorithms and often neglect non-IID Dirichlet-distributed conditions and probabilistic calibration metrics like log-loss. This study rigorously compares FedAvg, FedProx, FedSVRG, and FedAtt across three MedMNIST v2 datasets—PneumoniaMNIST (binary), DermaMNIST, and BloodMNIST (multi-class)—using three clients under non-IID Dirichlet partitioning (α=0.1) over 50 communication rounds. FedProx demonstrates the most consistent performance and stability, achieving accuracy of 0.9521 and log-loss of 0.1850 on PneumoniaMNIST; 0.8595 and 0.4066 on BloodMNIST; and 0.5747 and 1.5996 on DermaMNIST. It also exhibits fastest convergence and superior probability calibration. Thus, FedProx's proximal regularization enhances FL robustness against extreme clinical heterogeneity, establishing it as a scalable, privacy-preserving framework for cross-institutional medical image diagnostics.

Keywords: Aggregation; healthcare; federated learning; data transfer; privacy-preserving

1. Introduction

Machine learning has achieved diagnostic performance on par with clinical experts across diverse medical image analysis tasks, including thoracic abnormality detection and skin lesion classification, as synthesized in recent comprehensive reviews [1], [2], [3], [4]. However, centralized training paradigms necessitate cross-institutional aggregation of patient data, potentially violating stringent privacy regulations such as the European Union's General Data Protection Regulation (GDPR) and the United States' Health Insurance Portability and Accountability Act (HIPAA) [5], [6], [7]. In 2023 alone, 1,853 healthcare data breaches were reported in the United States, compromising over 133 million patient records, underscoring that direct exchange of raw data is no longer sustainable in modern healthcare ecosystems [8], [9], [10], [11].

Federated Learning (FL) offers an alternative paradigm by enabling collaborative model training without relocating sensitive data from its origin [12], [13], [14]. Each medical institution—acting as a client—trains a local model on its internal dataset, transmitting only parameter updates to a central server for global aggregation. In the medical domain, inter-institutional data heterogeneity is not a simulated artifact but a true reflection of clinical specialization: dermatology oncology centers predominantly manage melanoma cases, whereas general hospitals handle markedly lower proportions [15], [16]. Such variation induces highly non-independent and identically distributed (non-IID) data distributions, severely degrading the efficacy of conventional FL algorithms [17]. Empirical studies demonstrate that Federated Averaging (FedAvg)—the foundational method—suffers accuracy degradation of up to 25.7% under extreme non-IID conditions compared to IID settings [18]. This decline is exacerbated by extreme class imbalance, where minority classes—such as rare diseases—may constitute less than 1% of the global data, rendering their diagnostic signals vulnerable to suppression during aggregation [19], [20], [21]. This phenomenon not only impairs generalization but also amplifies predictive bias, particularly in clinically critical classes [22].

Received: 13-11-2025 | Revised: 10-01-2026 | Accepted: 15-01-2026 | Published: 31-01-2026

Although recent literature has explored FL applications in medicine, most studies evaluate a single algorithm without systematic comparison, neglect the Dirichlet distribution as a standard for simulating non-IID heterogeneity, and omit comprehensive evaluation metrics, including log-loss for probability calibration: Author et al. tested only FedEst-NIID on tabular MIMIC-IV data without images, Dirichlet partitioning, or log-loss [23]; Author et al. evaluated FedAvg+DP on a single COVID-19 cohort with departmental heterogeneity, excluding log-loss and probabilistic simulation [24]; Author et al. used FedAvg as a proof-of-concept on skin lesions across three sites with real clinical non-IID data [25]; and Author et al. implemented SCAFFOLD on a single hematology dataset with fixed labels, with neither algorithm comparison nor log-loss [26]. The collective absence of multi-algorithm comparative evaluation, representative Dirichlet-based non-IID simulation, and calibration metrics such as log-loss constitute a significant methodological gap, limiting deep understanding of FL robustness and reliability in heterogeneous clinical environments.

To address these challenges, several advanced FL variants have been proposed with adaptive mechanisms. Federated Proximal (FedProx) introduces proximal regularization to constrain local model divergence from the global model [27], [28]; Federated Stochastic Variance Reduced Gradient (FedSVRG) leverages variance control to mitigate gradient oscillation in heterogeneous data [29]; whereas Federated Attentive Aggregation (FedAtt) employs cosine similarity-based attention to prioritize informative updates [30]. Despite their theoretical promise, systematic comparative evaluations—particularly on medical imaging datasets with Dirichlet-based non-IID partitioning that probabilistically replicates inter-institutional variation—remain scarce. Moreover, comprehensive evaluation metrics—including accuracy, precision, recall, F1-score, ROC-AUC, and log-loss (specifically assessing probability calibration yet rarely utilized in medical FL literature)—have not been thoroughly explored to reveal convergence dynamics and long-term stability.

This study aims to bridge these methodological gaps by rigorously evaluating the performance of FedAvg, FedProx, FedSVRG, and FedAtt on three medical imaging datasets from the MedMNIST v2 collection [15]: PneumoniaMNIST (5,856 samples, binary classification), DermaMNIST (10,015 samples, 7 classes), and BloodMNIST (17,092 samples, 8 classes). Data are partitioned across three heterogeneous clients using a low-concentration Dirichlet distribution, yielding statistically significant inter-client class distribution disparities. This approach enables in-depth analysis of algorithmic resilience to both statistical and system heterogeneity while providing critical insights for developing reliable FL systems in real-world clinical settings [31].

2. Research Methods

Figure 1 illustrates the overall research workflow. The dataset is initially loaded and partitioned into three heterogeneous clients using a Dirichlet distribution to simulate non-IID conditions. Each client subsequently performs pixel intensity normalization. Next, 5-fold cross-validation based on StratifiedKFold is applied to each client. In each fold, every client trains a local CNN model and then transmits the local model weights to the center [32]. The center performs global aggregation separately to produce four global models: Federated Averaging (FedAvg), Federated Proximal (FedProx), Federated Stochastic Variance Reduced Gradient (FedSVRG), and Federated Attentive Aggregation (FedAtt). Evaluation is conducted on the local models using each client's test set, as well as on the global models using all three client test sets, with computation of accuracy, precision, recall, F1-score, ROC-AUC, and log-loss. The processes of local training, weight transmission, aggregation, and evaluation are repeated for 50 communication rounds.

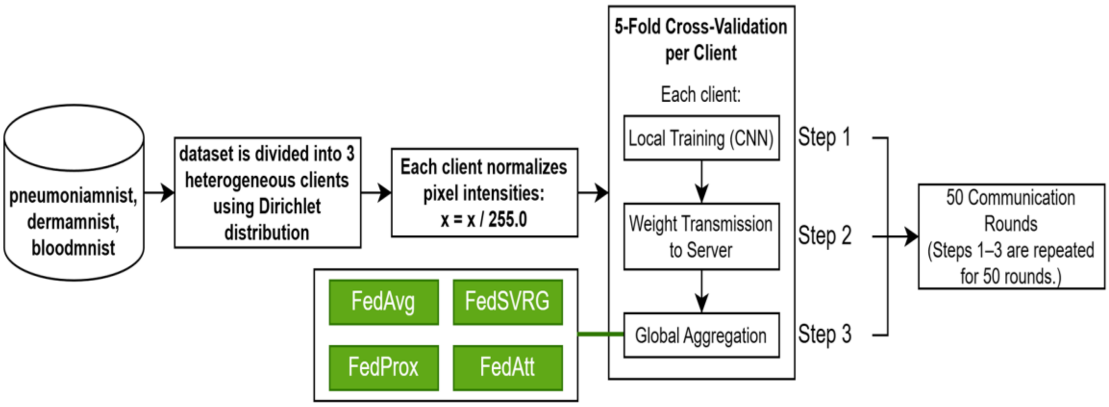

pneumoniamnist, dermamnist, bloodmnist
dataset is divided into 3 heterogeneous clients using Dirichlet distribution
Each client normalizes pixel intensities: x = x / 255.0
5-Fold Cross-Validation per Client
Each client:
Local Training (CNN)
Weight Transmission to Server
Global Aggregation
Step 1
Step 2
Step 3
50 Communication Rounds (Steps 1–3 are repeated for 50 rounds.)
FedAvg
FedSVRG
FedProx
FedAtt

Figure 1. Research Flow

2.1. Dataset

This study utilizes three datasets, namely PneumoniaMNIST, DermaMNIST, and BloodMNIST, which encompass variations in classification types, including both binary and multiclass tasks [33]. These datasets were selected to represent diverse data characteristics within the context of evaluating federated learning algorithms. A comprehensive summary of each dataset is presented in Table 1, while Figure 2 displays example samples from each class across the respective datasets.

The datasets are subsequently partitioned into three heterogeneous clients to simulate non-IID conditions commonly encountered in healthcare federated learning applications. The partitioning is performed using a Dirichlet distribution to probabilistically control class proportions across clients, thereby reflecting real-world variations in patient populations across different medical institutions [31].

The class proportion on client k for class c is determined via a random vector drawn from the Dirichlet distribution, as shown in Equation 1.

$$\overrightarrow{\pi_c} \sim \text{Dirichlet}\left(\alpha \cdot \overrightarrow{1_K}\right),\ c\ \in \{0,1\},\ k\ =\ 1,2,3 \qquad (1)$$

Where α is the concentration parameter that controls the degree of heterogeneity (a smaller α yields a more imbalanced distribution), and $\overrightarrow{1_K}$ is a one-vector of dimension $K\ =\ 3$. Equation 1 produces a proportion $\overrightarrow{\pi_c} = \left(\pi_{c,1}, \pi_{c,2}, \pi_{c,3}\right),\ \sum_{k=1}^{3} \pi_{c,k} = 1$.

The parameter α in the Dirichlet distribution was set to 0.1 to increase data imbalance across clients. The partitioning ensures each sample is assigned to one client without overlapping. This configuration preserves the integrity of the non-IID simulation, common in distributed healthcare systems. The design supports evaluating method robustness against data diversity in real-world institutions. This setting enables realistic evaluation and shows its generalization across clients. The distribution of the datasets across the three clients is presented in Tables 2–4, while the results of the Kolmogorov-Smirnov test between clients are shown in Table 5.

Based on Tables 2–5, the class distribution and mean intensity across clients differ significantly in all datasets ($p < 0.05$). The Chi-square test was employed to assess whether the class distribution across clients differs statistically, while the Kolmogorov-Smirnov (KS) test was used to compare the distribution of mean pixel intensity across clients [34], [35]. The significant results from both tests confirm that each dataset is genuinely heterogeneous, making it suitable for simulating federated learning scenarios, where models are frequently confronted with imbalanced and heterogeneous data.

2.2. Federated Learning (FL)

Federated Learning (FL) is a distributed machine learning paradigm that enables collaborative model training across multiple clients without exchanging raw data, preserving privacy and reducing

Table 1. Medical Image Dataset Summary

Dataset Name	Classification Type	Total Samples	Class Distribution (All Data)
PneumoniaMNIST	Binary	5,856	Class 0 (Normal): 1,583 (27.03%) Class 1 (Pneumonia): 4,273 (72.97%)
DermaMNIST	Multiclass	10,015	Class 0 (akiec): 327 (3.27%) Class 1 (bcc): 514 (5.13%) Class 2 (bkl): 1,099 (10.97%) Class 3 (df): 115 (1.15%) Class 4 (mel): 1,113 (11.11%) Class 5 (nv): 6,705 (66.95%) Class 6 (vasc): 142 (1.42%)
BloodMNIST	Multiclass	17,092	Class 0 (basophil): 1,218 (7.13%) Class 1 (eosinophil): 3,117 (18.24%) Class 2 (erythroblast): 1,551 (9.07%) Class 3 (immature granulocytes): 2,895 (16.94%) Class 4 (lymphocyte): 1,214 (7.10%) Class 5 (monocyte): 1,420 (8.31%) Class 6 (neutrophil): 3,329 (19.48%) Class 7 (platelet): 2,348 (13.74%)

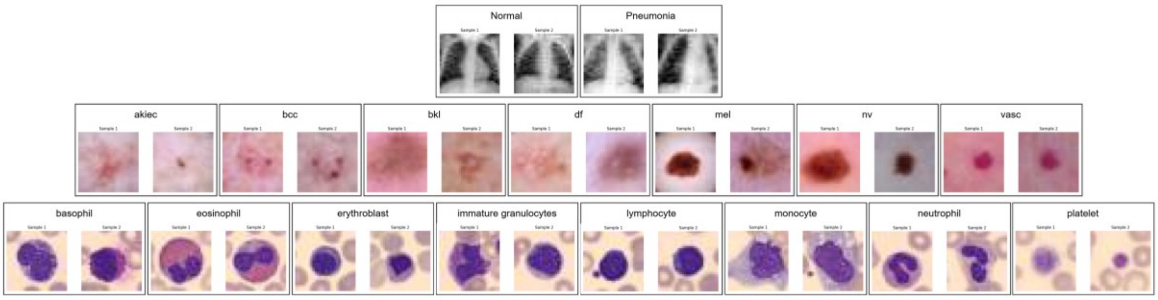


Figure 2. Example Images per Class for Datasets PneumoniaMNIST, DermaMNIST, and BloodMNIST

Table 2. PneumoniaMNIST

Total Samples	Client 1 Class Distribution	Client 2 Class Distribution	Client 3 Class Distribution	Chi-square	p-value
5,360	Class 0 (Normal): 1,043 (39.4%) Class 1 (Pneumonia): 1,602 (60.6%)	Class 0 (Normal): 75 (4.0%) Class 1 (Pneumonia): 1,815 (96.0%)	Class 0 (Normal): 465 (56.4%) Class 1 (Pneumonia): 360 (43.6%)	1,003.47	1.25e-218

Table 3. DermaMNIST

Total Samples	Client 1 Class Distribution	Client 2 Class Distribution	Client 3 Class Distribution	Chi-square	p-value
3,569	Class 0 (akiec): 183 (5.1%) Class 1 (bcc): 20 (0.6%) Class 2 (bkl): 550 (15.4%) Class 3 (df): 21 (0.6%) Class 4 (mel): 28 (0.8%) Class 5 (nv): 2,717 (76.1%) Class 6 (vasc): 50 (1.4%)	Class 0 (akiec): 84 (5.2%) Class 1 (bcc): 35 (2.2%) Class 2 (bkl): 319 (19.8%) Class 3 (df): 39 (2.4%) Class 4 (mel): 1,065 (65.9%) Class 5 (nv): 35 (2.2%) Class 6 (vasc): 38 (2.4%)	Class 0 (akiec): 60 (1.2%) Class 1 (bcc): 459 (9.5%) Class 2 (bkl): 230 (4.8%) Class 3 (df): 55 (1.1%) Class 4 (mel): 20 (0.4%) Class 5 (nv): 3,953 (81.8%) Class 6 (vasc): 54 (1.1%)	7,288.59	0.00e+00

Table 4. BloodMNIST

Total Samples	Client 1 Class Distribution	Client 2 Class Distribution	Client 3 Class Distribution	Chi-square	p-value
8,289	Class 0 (basophil): 353 (4.3%) Class 1 (eosinophil): 141 (1.7%) Class 2 (erythroblast): 863 (10.4%) Class 3 (immature granulocytes): 339 (4.1%) Class 4 (lymphocyte): 1,134 (13.7%) Class 5 (monocyte): 821 (9.9%) Class 6 (neutrophil): 2,370 (28.6%) Class 7 (platelet): 2,268 (27.4%)	Class 0 (basophil): 258 (5.0%) Class 1 (eosinophil): 2,926 (56.4%) Class 2 (erythroblast): 638 (12.3%) Class 3 (immature granulocytes): 362 (7.0%) Class 4 (lymphocyte): 30 (0.6%) Class 5 (monocyte): 549 (10.6%) Class 6 (neutrophil): 394 (7.6%) Class 7 (platelet): 30 (0.6%)	Class 0 (basophil): 607 (16.8%) Class 1 (eosinophil): 50 (1.4%) Class 2 (erythroblast): 50 (1.4%) Class 3 (immature granulocytes): 2,194 (60.7%) Class 4 (lymphocyte): 50 (1.4%) Class 5 (monocyte): 50 (1.4%) Class 6 (neutrophil): 565 (15.6%) Class 7 (platelet): 50 (1.4%)	16,242.23	0.00e+00

Table 5. Kolmogorov-Smirnov Test Results (Average Intensity Distribution)

Dataset	Scenario	p-value
PneumoniaMNIST	client_1 vs client_2	4.68e-12
PneumoniaMNIST	client_1 vs client_3	2.03e-02
PneumoniaMNIST	client_2 vs client_3	9.10e-14
DermaMNIST	client_1 vs client_2	5.65e-24
DermaMNIST	client_1 vs client_3	1.22e-02
DermaMNIST	client_2 vs client_3	7.01e-33
BloodMNIST	client_1 vs client_2	1.15e-320
BloodMNIST	client_1 vs client_3	9.01e-321
BloodMNIST	client_2 vs client_3	2.35e-31

communication costs [36]. Each client independently trains a local model using its private dataset and periodically sends the model parameters (weights) to a central server for global aggregation [37]. The server then updates and redistributes the global model to all clients, allowing continuous improvement through multiple communication rounds [38]. This iterative process ensures that knowledge from diverse data sources is integrated without compromising data confidentiality. In this study, FL addresses data heterogeneity and privacy concerns across medical institutions. The overall workflow of FL implemented in this research is illustrated in Figure 3, providing a clearer overview of the collaborative training process. This framework also allows evaluating model performance under non-IID data distributions and highlights robustness when deployed across diverse clients.

2.3. Federated Averaging (FedAvg)

Federated Averaging (FedAvg) is the foundational algorithm in FL that performs aggregation of local model parameters via a weighted average based on the proportion of each client's data size relative to the total data [39]. At communication round t, the server initializes the global model $w^{(t)}$, after which each

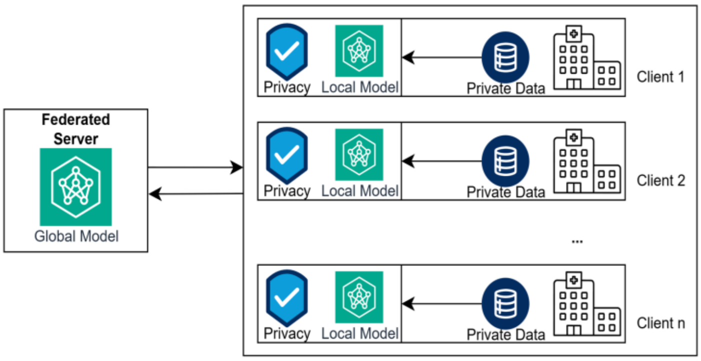


Figure 3. Workflow of Federated Learning

client k performs local updates over several epochs to produce $w_k^{(t+1)}$. Global aggregation is performed according to Equation 2.

$$w^{(t+1)} = \sum_{k=1}^{K} \frac{n_k}{n} w_k^{(t+1)} \tag{2}$$

Where n_k is the number of samples at client k, $n = \sum_{k=1}^{K} n_k$ is the total number of samples, and K is the number of clients. Equation 2 ensures that each client's contribution to the global model is proportional to its dataset size, thereby preventing dominance by clients with large datasets and maintaining overall data representation balance.

2.4. Federated Proximal (FedProx)

FedProx is an enhancement over FedAvg that incorporates a proximal term to address system and data heterogeneity [27]. The algorithm constrains the deviation of local models from the global model during local training. At client k, the local objective function is modified to Equation 3.

$$\min_w f_k(w) + \frac{\mu}{2}|w - w^{(t)}|^2 \tag{3}$$

Where $f_k(w)$ is the local loss function, $w^{(t)}$ is the global model at round t, and μ is the proximal regularization coefficient. Equation 3 adds a quadratic penalty on the distance between local and global parameters. After local optimization, aggregation follows the same weighted averaging as in Equation 2.

2.5. Federated Stochastic Variance Reduced Gradient (FedSVRG)

FedSVRG, in this implementation based on the SCAFFOLD framework, uses control variates to reduce gradient variance and accelerate convergence on non-IID data [40]. The algorithm maintains two control variables: c (global) and c_k (local per client). The local gradient update at client k is modified according to Equation 4.

$$\widetilde{g_k} = \nabla f_k(w_k) - c_k + c \tag{4}$$

Where $\nabla f_k(w_k)$ is the local stochastic gradient. Equation 4 corrects the local gradient by subtracting the client-specific variance component (c_k) and adding the global average variance (c). After local training, control variations are updated based on parameter differences, $\Delta w_k = w_k - w^{(t)}$, then $c \leftarrow c + \frac{1}{K}\sum \Delta w_k$ and $c_k \leftarrow c$.

2.6. Federated Attentive Aggregation (FedAtt)

FedAtt introduces an attention mechanism to assign adaptive aggregation weights based on cosine similarity between local and global model parameters [30]. At each round, the server computes a similarity score s_k for each client k using Equation 5.

$$s_k = \sum_l \text{cosine\_similarity}\left(w_l^{(t)}, w_{k,l}^{(t+1)}\right) \tag{5}$$

Where index l iterates over all parameter layers. Equation 5 measures the extent to which client k's local update aligns with the current global model. Then, attention weights α_k are computed via softmax according to Equation 6.

$$\alpha_k = \frac{\exp(s_k)}{\sum_j \exp(s_j)} \tag{6}$$

and global aggregation is performed according to Equation (7).

$$w^{(t+1)} = \sum_{k=1}^{K} \alpha_k w_k^{(t+1)} \tag{7}$$

Equation 6 ensures that $\sum \alpha_k = 1$, distributing weights probabilistically. Equation 7 enables the server to prioritize clients with more informative or consistent updates.

2.7. Local Model Architecture: Convolutional Neural Network (CNN)

The local model employed by each client is a Convolutional Neural Network (CNN). The architecture consists of two convolutional layers with ReLU activation functions, one max pooling layer for spatial dimension reduction, dropout for regularization, and two fully connected layers at the end [41]. This design enables hierarchical feature extraction from image data. For binary classification tasks, the local model architecture is summarized in Table 6.

Table 6. CNN Architecture for Binary Classification

Layer	Configuration	Output Shape	Parameter
Input	Citra grayscale / RGB	1 × H × W / 3 × H × W	–
conv1	Conv2d(1 → 16, kernel 3×3, stride 1, padding 1)	16 × H × W	160
ReLU	–	16 × H × W	0
conv2	Conv2d(16 → 32, kernel 3×3, stride 1, padding 1)	32 × H × W	4.640
ReLU	–	32 × H × W	0
pool	MaxPool2d(kernel 2×2, stride 2)	32 × (H/2) × (W/2)	0
dropout	Dropout(p=0.25)	32 × (H/2) × (W/2)	0
flatten	–	1568	0
fc1	Linear(1568 → 64)	64	100.416
ReLU	–	64	0
fc2	Linear(64 → 1)	1	65

For multiclass classification tasks, a similar architecture is used with adjustments to the input channels and final output dimension, as shown in Table 7.

Table 7. CNN Architecture for Multiclass Classification

Layer	Configuration	Output Shape	Parameter
Input	Citra grayscale / RGB	1 × H × W / 3 × H × W	–
conv1	Conv2d(3 → 16, kernel 3×3, stride 1, padding 1)	16 × H × W	448
ReLU	–	16 × H × W	0
conv2	Conv2d(16 → 32, kernel 3×3, stride 1, padding 1)	32 × H × W	4.640
ReLU	–	32 × H × W	0
pool	MaxPool2d(kernel 2×2, stride 2)	32 × (H/2) × (W/2)	0
dropout	Dropout(p=0.25)	32 × (H/2) × (W/2)	0
flatten	–	1568	0
fc1	Linear(1568 → 64)	64	100.416
ReLU	–	64	0
fc2	Linear(64 → 7)	7	455

2.7. Experiments and Evaluation

Experiments were conducted to evaluate the performance of four federated learning algorithms—Federated Averaging (FedAvg), Federated Proximal (FedProx), Federated Stochastic Variance Reduced Gradient (FedSVRG), and Federated Attentive Aggregation (FedAtt). The entire process—from data loading, non-IID partitioning, initialization of 5-fold stratified cross-validation, local training, global aggregation, to metric evaluation structured in Algorithm 1.

Algorithm 1. Experimental Flow of Federated Learning with 5-Fold Stratified Cross-Validation

```
Input: Global dataset D, number of clients K, Dirichlet parameter α,
number of communication rounds T
Output: Average local and global performance metrics for each
algorithm
Steps:
1. Load dataset D.
   Normalize pixel intensities of each sample to the range [0, 1].

2. Partition dataset D into K non-IID subsets using the Dirichlet(α)
   distribution.
   Each class is allocated to clients proportionally based on a random
   vector sampled from the Dirichlet distribution.

3. For each client k = 1, ..., K:
     Apply StratifiedKFold (5-fold) on subset D_k.
     Generate 5 pairs of datasets (train_data_i^k, test_data_i^k), for i
      = 1, ..., 5.

4. Initialize the global model w^(0) for four algorithms:
     FedAvg, FedProx, FedSVRG, and FedAtt using a CNN
     architecture.
     Initialize control variates c^(0) and c_k^(0) = 0 (for FedSVRG
     only).

5. For each communication round t = 1 to T:
     For each fold i = 1 to 5:
       Local Training Phase:
         For each client k = 1, ..., K:
           Copy global model: w_k^(t) = w^(t−1).
           Perform local updates using the loss function:
              - CrossEntropyLoss (for multi-class) or
              - BCEWithLogitsLoss (for binary).
           Add a proximal term to the objective function (FedProx
           only).
           Apply gradient correction with control variates
           (FedSVRG only).

       Global Aggregation Phase (performed separately for each
       algorithm):
         FedAvg: w^(t) = Σ_(k=1)^K (n_k / n) * w_k^(t)
         FedProx: w^(t) = Σ_(k=1)^K (n_k / n) * w_k^(t) after local
         optimization with the proximal term
         FedSVRG: Update c and c_k based on Δw_k, then w^(t) =
         (1 / K) Σ w_k^(t)
         FedAtt: Compute attention scores s_k, normalize via
         softmax to obtain α_k, then w^(t) = Σ α_k * w_k^(t)

       Evaluation Phase:
         Local model evaluation:
           For each client k = 1, ..., K:
              Evaluate w_k^(t) on test_data_i^k → metric_k.
           Average local metric = (1 / K) Σ metric_k.

         Global model evaluation:
           For each client k = 1, ..., K:
              Evaluate w_alg^(t) on test_data_i^k → metric^k.
           Average global metric = (1 / K) Σ metric^k.

         Save all metrics for each round and each fold.

6. Compute the final average of all metrics over T rounds and 5
   folds for each algorithm.

7. Return the average local and global performance for FedAvg,
   FedProx, FedSVRG, and FedAtt.
```

Performance evaluation was carried out using six classification metrics: accuracy, precision, recall, F1-score, ROC-AUC, and log-loss [42]. Each metric was computed based on predictions from both local and global models on test data. The definitions and formulations of each metric are presented sequentially.

The first metric is accuracy, which measures the proportion of correct predictions relative to the total number of samples, given by Equation 8.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \tag{8}$$

Where TP is true positive, TN is true negative, FP is false positive, and FN is false negative.

The second metric is precision, which measures the proportion of positive predictions that are positive, defined in Equation 9.

$$\text{Precision} = \frac{TP}{TP+FP} \tag{9}$$

Equation 9 is critical in applications where false positives carry high consequences, such as disease detection.

The third metric is recall (or sensitivity), which measures the proportion of actual positive cases successfully detected, given by Equation 10.

$$\text{Recall} = \frac{TP}{TP+FN} \tag{10}$$

Equation 10 emphasizes the model's ability to identify all positive instances, which is crucial when false negatives are costly.

The fourth metric is the F1-score, the harmonic mean of precision and recall, formulated in Equation 11.

$$\text{F1-score} = 2 \times \frac{\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}} \tag{11}$$

Equation 11 provides a balanced measure between precision and recall, making it suitable as a single metric for imbalanced data.

The fifth metric is ROC-AUC (Area Under the Receiver Operating Characteristic Curve), which assesses the model's ability to distinguish between positive and negative classes across various thresholds, computed as the integral of the ROC curve. An AUC of 1 indicates perfect separation, while AUC = 0.5 corresponds to random guessing.

The sixth metric is log-loss (or cross-entropy loss), which evaluates the quality of predicted probabilities. For binary classification, it is defined in Equation 12.

$$\text{Log-loss} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(\widehat{p_i}) + (1-y_i)\log(1-\widehat{p_i})] \tag{12}$$

For the multiclass case, log-loss is given by Equation 13.

$$\text{Log-loss} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} y_{i,c}\log\left(\widehat{p_{i,c}}\right) \tag{13}$$

Where y_i is the true label, $\widehat{p_i}$ is the predicted probability, N is the number of samples, and C is the number of classes. Equations 12 and 13 more heavily penalize confident incorrect predictions, thus encouraging well-calibrated probabilities.

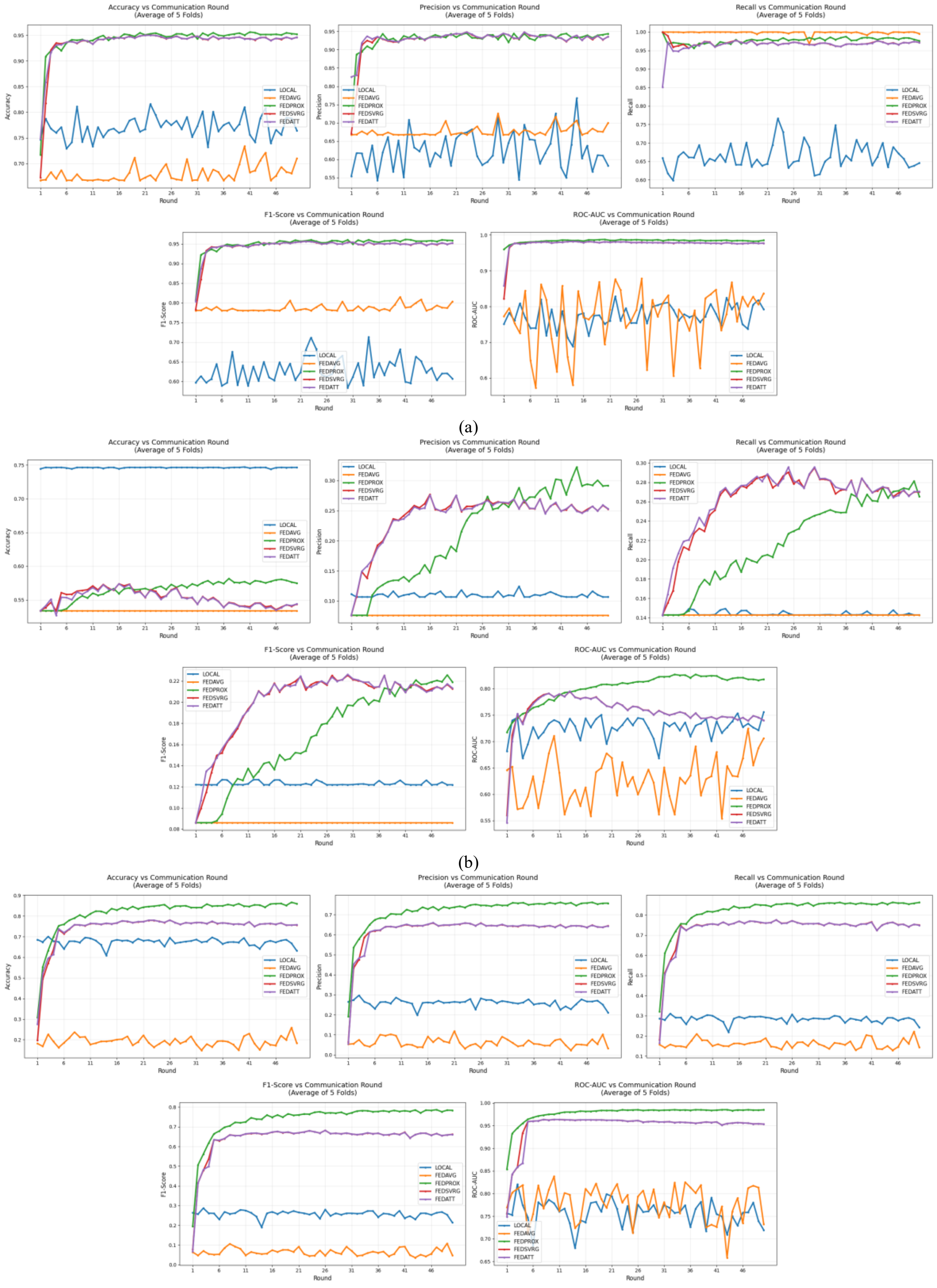
3. Results and Discussions

This study evaluates the performance of four federated learning algorithms—FedAvg, FedProx, FedSVRG, and FedAtt—alongside local client models on three medical imaging datasets partitioned in a non-IID manner using a Dirichlet distribution with a low concentration parameter to simulate real-world heterogeneity across healthcare institutions. The experimental protocol employs 5-fold stratified cross-validation per client and conducts training over 50 communication rounds, with evaluation metrics including accuracy, precision, recall, F1-score, ROC-AUC, and log-loss, computed as averages across all folds and clients. Table 8 presents the mean performance at three critical points—Round 1 (initialization), Round 25 (mid-convergence), and Round 50 (final stabilization)—for the PneumoniaMNIST, DermaMNIST, and BloodMNIST datasets, while Figure 4 illustrates the complete trends from Round 1 to Round 50 across all six metrics, providing a longitudinal view of convergence speed, stability, and response to data heterogeneity.

On the PneumoniaMNIST dataset, a binary classification task with moderate imbalance, FedProx consistently demonstrates the best performance throughout the communication rounds. At Round 50, it achieves accuracy of 0.9521, precision of 0.9439, recall of 0.9763, F1-score of 0.9589, ROC-AUC of 0.9852, and log-loss of 0.1850—significantly outperforming the local model (accuracy 0.7642) and FedAvg (accuracy 0.7097). A sharp improvement is observed by Round 25, with accuracy reaching 0.9470 and log-loss dropping to 0.1521, indicating that the proximal term effectively constrains local model deviation from the global model, thereby accelerating convergence and enhancing generalization under non-IID conditions. FedSVRG and FedAtt achieve nearly identical values at Round 50, with accuracy around 0.9461, F1-score of 0.9525, and ROC-AUC near 0.977, but log-loss increases from approximately 0.177 at Round 25 to 0.2687, suggesting degradation in probability calibration in later stages. FedAvg maintains high recall above 0.995 from Round 1 but keeps precision below 0.700 through Round 50, reflecting a strong bias toward predicting the majority class. The local model exhibits fluctuations, peaking at accuracy 0.7836 at Round 25 before declining to 0.7642 at Round 50, underscoring the limitations of isolated training due to constrained data size and diversity per client.

Table 8. Mean Performance (5-Fold) at Rounds 1, 25, and 50 Across All Datasets

Dataset	Model	Round	Accuracy	Precision	Recall	F1-Score	ROC-AUC	Log-Loss
PneumoniaMNIST	Lokal	1	0,7462	0,5538	0,6594	0,5968	0,7509	0,4550
		25	0,7836	0,6085	0,6323	0,6166	0,7538	0,4123
		50 (Final Model)	0,7642	0,5828	0,6456	0,6068	0,7921	0,4267
	FedAvg	1	0,6672	0,6672	1,0000	0,7804	0,7726	0,6541
		25	0,6992	0,6903	0,9961	0,7968	0,7603	0,6189
		50 (Final Model)	0,7097	0,6998	0,9955	0,8035	0,8367	0,6130
	FedProx	1	0,7170	0,7018	0,9994	0,8071	0,9600	0,5623
		25	0,9470	0,9322	0,9772	0,9531	0,9861	0,1521
		50 (Final Model)	0,9521	0,9439	0,9763	0,9589	0,9852	0,1850
	FedSVRG	1	0,6724	0,6703	0,9998	0,7828	0,8219	0,6523
		25	0,9434	0,9339	0,9691	0,9500	0,9790	0,1774
		50 (Final Model)	0,9461	0,9360	0,9716	0,9525	0,9772	0,2687
	FedAtt	1	0,7482	0,8257	0,8511	0,8037	0,8576	0,5575
		25	0,9438	0,9349	0,9687	0,9504	0,9790	0,1766
		50 (Final Model)	0,9461	0,9360	0,9716	0,9525	0,9773	0,2687
DermaMNIST	Lokal	1	0,7444	0,1113	0,1425	0,1220	0,6814	0,8444
		25	0,7455	0,1104	0,1445	0,1246	0,7308	0,7851
		50 (Final Model)	0,7464	0,1066	0,1429	0,1219	0,7556	0,7892
	FedAvg	1	0,5339	0,0763	0,1429	0,0861	0,6458	1,4891
		25	0,5339	0,0763	0,1429	0,0861	0,6331	1,4996
		50 (Final Model)	0,5339	0,0763	0,1429	0,0861	0,7061	1,4642
	FedProx	1	0,5339	0,0763	0,1429	0,0861	0,7172	1,4723
		25	0,5653	0,2460	0,2269	0,1790	0,8114	1,5342
		50 (Final Model)	0,5747	0,2919	0,2652	0,2190	0,8178	1,5996
	FedSVRG	1	0,5339	0,0763	0,1429	0,0861	0,5598	1,5901
		25	0,5559	0,2594	0,2905	0,2194	0,7664	1,8269
		50 (Final Model)	0,5438	0,2528	0,2700	0,2125	0,7397	2,4578
	FedAtt	1	0,5339	0,0763	0,1429	0,0861	0,5462	1,6044
		25	0,5557	0,2593	0,2957	0,2201	0,7662	1,8113
		50 (Final Model)	0,5439	0,2532	0,2705	0,2133	0,7396	2,4540
BloodMNIST	Lokal	1	0,6853	0,2653	0,2862	0,2638	0,7564	0,9850
		25	0,6518	0,2271	0,2605	0,2340	0,7150	1,0638
		50 (Final Model)	0,6315	0,2114	0,2426	0,2140	0,7194	1,0999
	FedAvg	1	0,1816	0,0530	0,1568	0,0640	0,7690	2,0899
		25	0,1952	0,0555	0,1731	0,0727	0,7133	1,9878
		50 (Final Model)	0,1831	0,0320	0,1439	0,0481	0,7319	1,9868
	FedProx	1	0,3079	0,1921	0,3200	0,1944	0,8535	1,8511
		25	0,8554	0,7511	0,8570	0,7764	0,9846	0,3938
		50 (Final Model)	0,8595	0,7571	0,8619	0,7834	0,9848	0,4066
	FedSVRG	1	0,1965	0,0646	0,1808	0,0771	0,7544	1,9529
		25	0,7683	0,6446	0,7602	0,6681	0,9602	0,8231
		50 (Final Model)	0,7568	0,6441	0,7501	0,6618	0,9534	1,0513
	FedAtt	1	0,2756	0,0543	0,1643	0,0679	0,7488	1,8887
		25	0,7688	0,6448	0,7595	0,6680	0,9602	0,8227
		50 (Final Model)	0,7569	0,6434	0,7487	0,6607	0,9534	1,0509



(a)

(b)

(c)

Figure 4. Average 5-Fold Performance Convergence Trends from Rounds 1 to 50 across All Datasets: (a) PneumoniaMNIST, (b) DermaMNIST, and (c) BloodMNIST

The DermaMNIST dataset, involving multiclass classification with extreme imbalance, reveals global aggregation failure across nearly all federated learning algorithms. FedAvg, FedSVRG, and FedAtt remain stagnant at accuracy 0.5339 and F1-score 0.0861 from Round 1 to Round 50, equivalent to majority-class guessing with no meaningful learning of minority classes. FedProx shows gradual progress, improving to accuracy 0.5747 and F1-score 0.2190 by Round 50, yet these values remain far below clinically acceptable performance, while log-loss stays high above 1.4642 for FedAvg and exceeds 2.45 for FedSVRG and FedAtt in the final phase. Conversely, the local model maintains stable accuracy around 0.746 and improves ROC-AUC from 0.6814 at Round 1 to 0.7556 at Round 50, confirming that when one class dominates over 66% of global data and up to 81% on certain clients, local training outperforms federated collaboration, which dilutes rare class signals during aggregation.

To mitigate the observed global aggregation failure under severe class imbalance, where minority class signals are diluted during standard weighted averaging, future extensions could incorporate advanced aggregation strategies such as class-wise or class-balanced aggregation (e.g., weighting contributions per class or using distribution-discrepancy-based methods to prioritize minority classes during aggregation) [43], [44], [45]. Additionally, personalized federated learning approaches (e.g., FedPer, pFedMe, Ditto, or Per-FedAvg) offer promising alternatives by allowing client-specific model adaptations, which decouple shared global representations from personalized heads to better preserve local class distributions and improve minority class performance in extreme non-IID scenarios like DermaMNIST [46], [47], [48]. These extensions will be explored in future work to enhance robustness in highly imbalanced medical imaging tasks.

On the BloodMNIST dataset, an eight-class multiclass task with moderate imbalance, FedProx again dominates, achieving accuracy 0.8595, precision 0.7571, recall 0.8619, F1-score 0.7834, ROC-AUC 0.9848, and log-loss 0.4066 at Round 50, with peak performance reached by Round 25 (accuracy 0.8554, log-loss 0.3938). FedSVRG and FedAtt reach accuracy of 0.7568 and 0.7569, respectively, with F1-scores around 0.661 at Round 50, but log-loss rises from approximately 0.823 at Round 25 to over 1.05, indicating probability calibration degradation over iterations. FedAvg experiences a drastic decline to accuracy 0.1831 and F1-score 0.0481 by Round 50, while the local model degrades from accuracy 0.6853 at Round 1 to 0.6315 at Round 50, likely due to overfitting on highly limited and imbalanced per-client subsets. Figure 4 reveals clear convergence patterns: FedProx exhibits rapid and stable improvement across all metrics for PneumoniaMNIST and BloodMNIST, with consistent log-loss reduction, whereas FedSVRG and FedAtt show oscillations in precision and recall before plateauing, and all federated learning curves flatten at low levels for DermaMNIST while the local model sustains gradual ROC-AUC improvement.

These findings conclude that FedProx is the most effective and robust federated learning algorithm for medical image classification under non-IID heterogeneity, particularly for binary and moderately imbalanced multiclass tasks, due to its ability to balance local contributions with global consistency via proximal regularization. In contrast, FedAvg, FedSVRG, and FedAtt fail completely under extreme imbalance conditions as seen in DermaMNIST. The main contributions of this study include providing a comprehensive benchmark using per-client 5-fold stratified cross-validation on non-IID partitioned MedMNIST datasets, identifying global aggregation failure points when a single class dominates over 65% of the data, and analyzing convergence trends over 50 communication rounds using six evaluation metrics covering both discriminative and probabilistic performance.

However, this study has several limitations that warrant careful consideration. First, the experiments were conducted using only three clients to simulate inter-institutional heterogeneity in a controlled and computationally feasible manner. This minimal number of clients effectively highlights the impact of extreme non-IID conditions (Dirichlet α=0.1) on algorithmic robustness, as evidenced by the complete failure of global aggregation in severely imbalanced scenarios such as DermaMNIST. However, such a limited client count does not fully represent large-scale federated learning deployments in real-world clinical settings, where dozens to hundreds of healthcare institutions may participate.

Increasing the number of clients is expected to influence the findings in several ways. In non-IID settings, a larger number of clients typically leads to more stable global aggregation due to the law of large numbers, which enables better averaging of local updates and reduces the dominance of any single highly divergent client. This can mitigate client drift and improve overall convergence speed and stability, particularly for algorithms like FedProx that already incorporate proximal regularization to constrain local divergence. Conversely, extreme heterogeneity across more clients could amplify communication overhead and require more robust handling of statistical and system heterogeneity, potentially exposing additional challenges in probability calibration (log-loss) if minority class signals remain diluted. Recent literature on federated learning in medical imaging supports this observation: studies simulating scalability with varying client counts (e.g., 3–10 or more on chest X-ray or histopathology datasets) demonstrate that performance often improves with scale under moderate non-IID conditions, while severe label skew continues to challenge standard FedAvg but benefits more from proximal or personalized variants [49], [50].

Second, the proximal coefficient μ in FedProx was fixed to a single value throughout the experiments. The primary objective of this study is a rigorous comparative evaluation of multiple federated learning algorithms (FedAvg, FedProx, FedSVRG, and FedAtt) under extreme Dirichlet non-IID conditions in medical imaging datasets. To ensure a fair and unbiased comparison across all algorithms, we deliberately used a fixed μ (as is common in comparative FL studies). Performing a detailed sensitivity analysis alone would extend the scope significantly and introduce unfairness in the algorithmic comparison, as the other methods (e.g., FedSVRG's control variants or FedAtt's attention mechanism) also have tunable hyperparameters that were not explored for the same reason of maintaining comparability.

Third, the local model architecture employed a relatively simple and shallow Convolutional Neural Network (CNN) with only two convolutional layers, selected for its computational efficiency and to maintain focus on algorithmic comparison rather than backbone optimization. While this design suffices for initial benchmarking on small-scale MedMNIST images (28×28), deeper and more powerful backbones such as ResNet (e.g., ResNet-18/50) or Vision Transformers (ViT) could potentially enhance feature extraction capabilities, leading to improved robustness against non-IID heterogeneity, better handling of complex patterns in multiclass tasks (e.g., DermaMNIST and BloodMNIST), and superior convergence or probability calibration in federated settings. Recent studies in medical imaging indicate that ResNet variants often achieve higher accuracy in non-IID federated scenarios (e.g., COVID-19 CXR classification with ResNet50/101), while ViTs excel in capturing global dependencies but may require larger data or pre-training to outperform CNNs on small datasets like MedMNIST [51], [52], [53], [54]. However, deeper models also introduce risks such as increased overfitting on limited per-client data or higher communication costs.

Fourth, while the motivation for privacy preservation through Federated Learning is emphasized in the introduction (no raw data exchange, compliance with GDPR/HIPAA), this study does not implement any formal privacy-enhancing mechanisms beyond the inherent data locality of FL. Mechanisms such as differential privacy (e.g., via DP-SGD or noise addition in local updates) and secure aggregation (e.g., SecAgg protocol for encrypted summation of model updates) are critical for real-world clinical deployment to protect against model inversion attacks, membership inference, and unintended leakage of sensitive patient information. The absence of these mechanisms is due to the current focus on comparative algorithmic performance under non-IID conditions; incorporating them would require additional computational overhead and parameter tuning, which is outside the primary scope of this benchmark study.

To address these limitations, we recommend the following for future work: conducting dedicated scalability experiments by increasing the number of clients to 5–20 (or higher) while maintaining the same Dirichlet α=0.1 partitioning on MedMNIST datasets or transitioning to real multi-institutional medical imaging data; performing hyperparameter sensitivity analyses, including tuning of μ in FedProx (e.g., across a range of values such as 0.01–10.0) alongside similar tuning for other algorithms; integrating personalized federated learning approaches to address extreme imbalance cases; developing class-aware or minority-focused attention-based aggregation mechanisms; evaluating deeper backbones such as ResNet-18/50 or lightweight ViT variants (e.g., DeiT-Small) under the same non-IID conditions to quantify performance gains in robustness and calibration; incorporating formal privacy mechanisms such as differential privacy (DP-SGD, DP-FedAvg) and secure aggregation (SecAgg) to strengthen protection against privacy attacks and better align the framework with clinical practice requirements (e.g., full GDPR/HIPAA compliance); and testing these extensions with real-world medical imaging data from actual institutions. Such rigorous extensions would further quantify the robustness, scalability, and practical deployability of FedProx in large-scale, heterogeneous, and privacy-sensitive clinical environments.

4. Conclusion

Overall, this study provides a comprehensive evaluation of the resilience of four major federated learning algorithms under extreme heterogeneity conditions of medical image data, simulated through non-IID partitioning based on Dirichlet distribution (α=0.1), and demonstrates that FedProx consistently emerges as the most robust method in terms of convergence speed, training stability, and probability calibration quality, both in binary classification scenarios (PneumoniaMNIST) and multiclass tasks with moderate imbalance (BloodMNIST). In contrast, FedAvg, FedSVRG, and FedAtt experience significant performance degradation under extreme imbalance conditions (DermaMNIST), where the global aggregation process dilutes diagnostic signals of minority classes, rendering local training more effective than federated collaboration in such cases. The main scientific contributions of this study include the introduction of a novel benchmark integrating Dirichlet-based non-IID simulation, per-client stratified 5-fold cross-validation, and probabilistic metrics such as log-loss; the identification of critical global aggregation failure points when a single class dominates over 65% of the data; and in-depth analysis of convergence trends over 50 communication rounds across six evaluation metrics. These findings imply that FedProx offers a scalable, privacy-preserving framework for cross-institutional medical image diagnostics, with potential applications in real-world clinical settings to enhance diagnostic accuracy for rare diseases while mitigating data privacy risks. Speculatively, this robustness could extend to broader healthcare AI systems, though further validation is needed. For future research, we advise exploring scalability with more clients, hyperparameter tuning, personalized FL mechanisms, deeper model backbones, and formal privacy integrations to align more closely with clinical deployment requirements.

Reference

[1] K. Bhavsar, A. Abugabah, J. Singla, A. Ali, A. Bashir, and Nikita, “A Comprehensive Review on Medical Diagnosis Using Machine Learning,” Comput. Mater. Contin., vol. 67, pp. 1997–2014, Dec. 2020, doi: 10.32604/cmc.2021.014943.

[2] S. K. Zhou et al., “A Review of Deep Learning in Medical Imaging: Imaging Traits, Technology Trends, Case Studies With Progress Highlights, and Future Promises,” Proc. IEEE, vol. 109, no. 5, pp. 820–838, 2021, doi: 10.1109/JPROC.2021.3054390.

[3] G. Litjens et al., “A Survey on Deep Learning in Medical Image Analysis,” Med. Image Anal., vol. 42, pp. 60–88, 2017, doi: 10.1016/j.media.2017.07.005.

[4] A. Garg and V. Mago, “Role of machine learning in medical research: A survey,” Comput. Sci. Rev., vol. 40, p. 100370, May 2021, doi: 10.1016/j.cosrev.2021.100370.

[5] C. Prince, N. Omrani, and F. Schiavone, “Online privacy literacy and users’ information privacy empowerment: the case of GDPR in Europe,” Inf. Technol. People, vol. 37, no. 8, pp. 1–24, Jan. 2024, doi: 10.1108/ITP-05-2023-0467.

[6] M. Taufiq and A. S. Kenyo, “The Legal Protection of Personal Data in the Digital Era: A Comparative Study of Indonesian Law and the GDPR,” Int. J. Bus. Law Educ., vol. 6, no. 2, pp. 1260–1268, Aug. 2025, doi: 10.56442/ijble.v6i2.1178.

[7] K. Theodos and S. Sittig, “Health Information Privacy Laws in the Digital Age: HIPAA Doesn’t Apply,” Perspect. Health Inf. Manag., vol. 18, no. Winter, p. 1l, 2020.

[8] K. Li, A. Lohachab, M. Dumontier, and others, “Privacy preservation in blockchain-based healthcare data sharing: A systematic review,” Peer--Peer Netw. Appl., vol. 18, p. 302, 2025, doi: 10.1007/s12083-025-02148-9.

[9] M. M. Ahmed, O. J. Okesanya, M. Oweidat, Z. K. Othman, S. S. Musa, and D. E. Lucero-Prisno III, “The ethics of data mining in healthcare: challenges, frameworks, and future directions,” BioData Min., vol. 18, no. 1, p. 47, 2025, doi: 10.1186/s13040-025-00461-w.

[10] K. Anyaso and V. Okoye, “The Impact of Big Data and Predictive Analytics on U.S. Healthcare Delivery: Opportunities, Challenges, and Future Directions,” World J. Adv. Res. Rev., vol. 24, no. 1, pp. 2275–2287, 2024, doi: 10.30574/wjarr.2024.24.1.3266.

[11] L. C. Bell and E. Shimron, “Sharing Data Is Essential for the Future of AI in Medical Imaging,” Radiol. Artif. Intell., vol. 6, no. 1, p. e230337, 2024, doi: 10.1148/ryai.230337.

[12] G. Choi, W. C. Cha, S. U. Lee, and S.-Y. Shin, “Survey of medical applications of federated learning,” Healthc. Inform. Res., vol. 30, no. 1, pp. 3–15, 2024, doi: 10.4258/hir.2024.30.1.3.

[13] C. Zhang, Y. Xie, H. Bai, B. Yu, W. Li, and Y. Gao, “A survey on federated learning,” Knowl.-Based Syst., vol. 216, p. 106775, 2021, doi: https://doi.org/10.1016/j.knosys.2021.106775.

[14] Q. Yang, Y. Liu, T. Chen, and Y. Tong, “Federated Machine Learning: Concept and Applications,” ACM Trans Intell Syst Technol, vol. 10, no. 2, Jan. 2019, doi: 10.1145/3298981.

[15] H. Zhu, J. Xu, S. Liu, and Y. Jin, “Federated learning on non-IID data: A survey,” Neurocomputing, vol. 465, Sep. 2021, doi: 10.1016/j.neucom.2021.07.098.

[16] L. Kwak and H. Bai, “The Role of Federated Learning Models in Medical Imaging,” Radiol. Artif. Intell., vol. 5, no. 3, p. e230136, 2023, doi: 10.1148/ryai.230136.

[17] Y. Zhao, M. Li, L. Lai, N. Suda, D. Civin, and V. Chandra, “Federated Learning with Non-IID Data,” 2018, doi: 10.48550/ARXIV.1806.00582.

[18] J. Liu, J. Huang, Y. Zhou, and others, “From distributed machine learning to federated learning: a survey,” Knowl. Inf. Syst., vol. 64, pp. 885–917, 2022, doi: 10.1007/s10115-022-01664-x.

[19] F. Sufi, “Addressing Data Scarcity in the Medical Domain: A GPT-Based Approach for Synthetic Data Generation and Feature Extraction,” Information, vol. 15, no. 5, p. 264, 2024, doi: 10.3390/info15050264.

[20] R. Schäfer, T. Nicke, H. Höfener, and others, “Overcoming data scarcity in biomedical imaging with a foundational multi-task model,” Nat. Comput. Sci., vol. 4, pp. 495–509, 2024, doi: 10.1038/s43588-024-00662-z.

[21] J. Gehrmann, E. Herczog, S. Decker, and others, “What prevents us from reusing medical real-world data in research,” Sci. Data, vol. 10, p. 459, 2023, doi: 10.1038/s41597-023-02361-2.

[22] I. Araf, A. Idri, and I. Chairi, “Cost-sensitive learning for imbalanced medical data: a review,” Artif. Intell. Rev., vol. 57, p. 80, 2024, doi: 10.1007/s10462-023-10652-8.

[23] K. Y. Chen, C. R. Shyu, Y. Y. Tsai, and others, “Effective Non-IID Degree Estimation for Robust Federated Learning in Healthcare Datasets,” J. Healthc. Inform. Res., vol. 9, pp. 437–464, 2025, doi: 10.1007/s41666-025-00195-8.

[24] H. Zhao, D. Sui, Y. Wang, L. Ma, and L. Wang, “Privacy-Preserving Federated Learning Framework for Multi-Source Electronic Health Records Prognosis Prediction,” Sensors, vol. 25, no. 8, p. 2374, 2025, doi: 10.3390/s25082374.

[25] Z. Deng, Y. Yang, and K. Suzuki, “Federated Active Learning Framework for Efficient Annotation Strategy in Skin-Lesion Classification,” J. Invest. Dermatol., vol. 145, no. 2, pp. 303–311, 2025, doi: 10.1016/j.jid.2024.05.023.

[26] D. Ryu, T. Bak, D. Ahn, and others, “Deep learning-based label-free hematology analysis framework using optical diffraction tomography,” Heliyon, vol. 9, no. 8, p. e18297, 2023, doi: 10.1016/j.heliyon.2023.e18297.

[27] J. Cui, Y. Li, Q. Zhang, Z. He, and S. Zhao, “A Federated Learning Framework Using FedProx Algorithm for Privacy-Preserving Palmprint Recognition,” in Biometric Recognition, S. Yu, W. Jia, X. Shu, X. Yuan, J. Gui, J. Tang, C. Shan, and Q. Liu, Eds., Singapore: Springer Nature Singapore, 2025, pp. 187–196. doi: https://doi.org/10.1007/978-981-96-1071-6_17.

[28] C. Mathew and P. Asha, “FedProx: FedSplit algorithm based federated learning for statistical and system heterogeneity in medical data communication,” J Internet Serv Inf Secur, vol. 14, no. 3, pp. 353–370, 2024, doi: 10.58346/JISIS.2024.I3.021.

[29] M. R. Rostami and S. S. Kia, “Federated Learning Using Variance Reduced Stochastic Gradient for Probabilistically Activated Agents.” 2023. [Online]. Available: https://arxiv.org/abs/2210.14362

[30] C. M. Thwal, Y. L. Tun, K. Kim, S.-B. Park, and C. S. Hong, “Transformers with Attentive Federated Aggregation for Time Series Stock Forecasting,” in 2023 International Conference on Information Networking (ICOIN), IEEE, Jan. 2023, pp. 499–504. doi: 10.1109/icoin56518.2023.10048928.

[31] M. Chaudhary, L. Gaur, and A. Chakrabarti, “Detecting the Employee Satisfaction in Retail: A Latent Dirichlet Allocation and Machine Learning approach,” in 2022 3rd International Conference on Computation, Automation and Knowledge Management (ICCAKM), 2022, pp. 1–6. doi: 10.1109/ICCAKM54721.2022.9990186.

[32] S. Bates, T. Hastie, and R. Tibshirani, “Cross-Validation: What Does It Estimate and How Well Does It Do It?,” J. Am. Stat. Assoc., vol. 119, no. 546, pp. 1434–1445, 2024, doi: 10.1080/01621459.2023.2197686.

[33] J. Yang, R. Shi, and B. Ni, “MedMNIST Classification Decathlon: A Lightweight AutoML Benchmark for Medical Image Analysis,” in 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI), 2021, pp. 191–195. doi: 10.1109/ISBI48211.2021.9434062.

[34] A. T. Hutcheson and K. G. Brown, “Chi-Square,” in Statistics for Psychology Research, Cham: Palgrave Macmillan, 2024. doi: 10.1007/978-3-031-60970-1_10.

[35] A. Zeimbekakis, E. D. Schifano, and J. Yan, “On Misuses of the Kolmogorov–Smirnov Test for One-Sample Goodness-of-Fit,” Am. Stat., vol. 78, no. 4, pp. 481–487, 2024, doi: 10.1080/00031305.2024.2356095.

[36] T. Li, A. K. Sahu, A. Talwalkar, and V. Smith, “Federated Learning: Challenges, Methods, and Future Directions,” IEEE Signal Process. Mag., vol. 37, no. 3, pp. 50–60, 2020, doi: 10.1109/MSP.2020.2975749.

[37] J. Reyes, L. Di Jorio, C. Low-Kam, and M. Kersten-Oertel, “Precision-Weighted Federated Learning,” Comput. Intell., vol. 41, no. 6, p. e70150, 2025, doi: https://doi.org/10.1111/coin.70150.

[38] T. Kołodziej and P. Rościszewski, “Towards Scalable Simulation of Federated Learning,” in International Conference on Neural Information Processing, Springer, 2021, pp. 248–256.

[39] H. Reguieg, M. E. Hanjri, M. E. Kamili, and A. Kobbane, “A Comparative Evaluation of FedAvg and Per-FedAvg Algorithms for Dirichlet Distributed Heterogeneous Data,” in 2023 10th International Conference on Wireless Networks and Mobile Communications (WINCOM), 2023, pp. 1–6. doi: 10.1109/WINCOM59760.2023.10322899.

[40] S. P. Karimireddy, S. Kale, M. Mohri, S. Reddi, S. Stich, and A. T. Suresh, “SCAFFOLD: Stochastic Controlled Averaging for Federated Learning,” in Proceedings of the 37th International Conference on Machine Learning, H. D. III and A. Singh, Eds., in Proceedings of Machine Learning Research, vol. 119. PMLR, Jul. 2020, pp. 5132–5143. [Online]. Available: https://proceedings.mlr.press/v119/karimireddy20a.html

[41] S. M. Anwar, M. Majid, A. Qayyum, and others, “Medical Image Analysis using Convolutional Neural Networks: A Review,” J. Med. Syst., vol. 42, no. 11, p. 226, 2018, doi: 10.1007/s10916-018-1088-1.

[42] G. Naidu, T. Zuva, and E. M. Sibanda, “A Review of Evaluation Metrics in Machine Learning Algorithms,” in Artificial Intelligence Application in Networks and Systems, R. Silhavy and P. Silhavy, Eds., Cham: Springer International Publishing, 2023, pp. 15–25. doi: https://doi.org/10.1007/978-3-031-35314-7_2.

[43] A. Xiong and others, “A Multi-Task Based Clustering Personalized Federated Learning Method,” Big Data Min. Anal., vol. 7, no. 4, pp. 1017–1030, Dec. 2024, doi: 10.26599/BDMA.2024.9020001.

[44] C. Smestad and J. Li, “A Systematic Literature Review on Client Selection in Federated Learning,” in Proceedings of the 27th International Conference on Evaluation and Assessment in Software Engineering, in EASE ’23. New York, NY, USA: Association for Computing Machinery, 2023, pp. 2–11. doi: 10.1145/3593434.3593438.

[45] Y. Shanmugarasa, H. Paik, S. S. Kanhere, and L. Zhu, “A systematic review of federated learning from clients’ perspective: challenges and solutions,” Artif. Intell. Rev., vol. 56, no. Suppl 2, pp. 1773–1827, 2023, doi: 10.1007/s10462-023-10563-8.

[46] H. Lin, X. Hu, S. Bai, and Y. Liu, “Personalized Federated Learning Algorithm Based on User Grouping and Group Signatures,” in Information and Communications Security, J. Han, Y. Xiang, G. Cheng, W. Susilo, and L. Chen, Eds., in Lecture Notes in Computer Science, vol. 16218. Springer, Singapore, 2026. doi: 10.1007/978-981-95-3543-9_24.

[47] Y. Liu, S. Li, W. Li, H. Qian, and H. Xia, “A Personalized Federated Learning Algorithm Based on Dynamic Weight Allocation,” Electronics, vol. 14, no. 3, p. 484, 2025, doi: 10.3390/electronics14030484.

[48] H. Yang, J. Li, M. Hao, and others, “An efficient personalized federated learning approach in heterogeneous environments: a reinforcement learning perspective,” Sci. Rep., vol. 14, p. 28877, 2024, doi: 10.1038/s41598-024-80048-3.

[49] M. Adnan, S. Kalra, J. C. Cresswell, G. W. Taylor, and H. R. Tizhoosh, “Federated learning and differential privacy for medical image analysis,” Sci. Rep., vol. 12, no. 1, p. 1953, 2022, doi: 10.1038/s41598-022-05539-7.

[50] M. Mohammadi, M. Vejdanihemmat, M. Lotfinia, and others, “Differential privacy for medical deep learning: methods, tradeoffs, and deployment implications,” Npj Digit. Med., 2026, doi: 10.1038/s41746-025-02280-z.

[51] X. Yang, J. Luo, and M. N. Mohammed, “Federation Learning of Optimized Convolutional Neural Network Structure for Intrusion Detection,” in 2022 Second International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT), Bhilai, India, 2022, pp. 1–7. doi: 10.1109/ICAECT54875.2022.9807964.

[52] Divya, N. Anand, and G. Sharma, “Convolutional neural network (CNN) and federated learning-based privacy preserving approach for skin disease classification,” J. Supercomput., vol. 80, pp. 24559–24577, 2024, doi: 10.1007/s11227-024-06309-0.

[53] S. Mehta, V. Kukreja, and R. Gupta, “Exploring the Potential of Federated Learning CNN for Interactive Virtual Tours of UNESCO Cultural Heritage Sites: A Case Study,” in 2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT), Delhi, India, 2023, pp. 1–6. doi: 10.1109/ICCCNT56998.2023.10307071.

[54] D. Süer Tümen and M. Nergiz, “Federated Learning-Based CNN Models for Orthodontic Skeletal Classification and Diagnosis,” Diagnostics, vol. 15, no. 7, p. 920, 2025, doi: 10.3390/diagnostics15070920.