Scalable Federated Learning for Massive Medical Image Classification: Tackling Noisy and Imbalanced Data

Hadjir Zemmouri
MISC Laboratory, Abdelhamid Mehri University, Constantine, Algeria
zemmouri.hadjir@univ-constantine2.dz (corresponding author)

Akram Kout
Setif Ferhat Abbas University, Setif, Algeria | MISC Laboratory, Abdelhamid Mehri University, Constantine, Algeria
akram-kout@univ-setif.dz

Said Labed
MISC Laboratory, Abdelhamid Mehri University, Constantine, Algeria
said.labed@univ-constantine2.dz

Received: 10 May 2025 | Revised: 7 June 2025, 29 June 2025, and 3 July 2025 | Accepted: 5 July 2025

 DOI: https://doi.org/10.48084/etasr.12040

ABSTRACT

The proliferation of medical imaging data, coupled with stringent privacy regulations, necessitates scalable and reliable classification methods to address the challenges of big data. Motivated by the critical need for accurate pneumonia detection from Chest X-rays (CXR) across diverse clinical settings, this research confronts the five fundamental challenges of big data: volume, velocity, variety, value, and veracity. In such scenarios, traditional centralized methods are often constrained by data heterogeneity, computational bottlenecks, and privacy risks. To overcome these constraints, we propose a Federated Learning (FL) system that distributes model training across multiple clients, thereby ensuring data privacy while efficiently managing large-scale datasets. Our method leverages transfer learning by fine-tuning a pre-trained VGG11 model and employs FedProx regularization to mitigate client drift arising from non-Independent and non-Identically Distributed (non-IID) data distributions. Furthermore, we introduce an innovative data partitioning technique that simulates real-world conditions by generating imbalanced label distributions with a Dirichlet process and injecting Gaussian noise to mimic image quality variations. By enabling distributed local training and dynamic learning rate adjustments, our approach effectively manages high-volume, high-velocity data while preserving data privacy. Experimental results demonstrate that our proposed method efficiently aggregates diverse and noisy client updates while achieving competitive performance in pneumonia classification.

Keywords-federated learning; big data classification; distributed training; class imbalance; pneumonia

I. INTRODUCTION

The availability of improved imaging technologies has enabled hospitals and research institutes to generate huge amounts of high-resolution medical images daily [1, 2]. While this rapid data expansion offers unprecedented opportunities for improving diagnostic accuracy and clinical outcomes [3], it also introduces substantial challenges concerning the five V's of big data: volume, velocity, variety, value, and veracity [4]. Moreover, consolidating data from multiple sources into a centralized repository is computationally intensive, presents privacy and legal constraints [5], and is complicated by disparities in data quality, label distributions, and noise, factors that often hinder the effectiveness of traditional Machine Learning (ML) and centralized learning approaches [6].

Federated Learning (FL) has emerged as a promising framework to tackle these challenges. By enabling decentralized model training across multiple clients (e.g., hospitals or imaging centers), FL preserves data privacy while reducing computational overhead [7]. In a typical FL setup, each client trains a local model using its private data and transmits only model updates to a central server, where the updates are aggregated to form a global model. However, due to the inherently non-Independent and non-Identically Distributed (non-IID) nature of medical data across institutions, client drift may occur, resulting in model updates

diverging due to heterogeneous and noisy data distributions [8].

In recent years, FL has been widely applied in the medical imaging domain to address privacy and regulatory challenges, while facilitating collaborative learning across institutions. For instance, the FL approach in [9] demonstrated effective pneumonia detection from chest radiographs across multi-site healthcare datasets. Similarly, the ensemble FL approach in [10] combined models trained across different data silos, showing improvements in diagnostic accuracy. In another line of work, authors in [11] introduced an FL framework for harmonizing distributed datasets with varied tasks, offering valuable insights into managing client drift and data heterogeneity through adaptive aggregation mechanisms. These studies underline the potential of FL in managing the unique challenges of multi-institutional healthcare data, particularly in heterogeneous settings.

In parallel, advances in Deep Learning (DL) have facilitated the development of powerful Computer-Aided Diagnosis (CAD) systems for Chest X-ray (CXR) analysis [12]. For example, the system in [1] demonstrated that deep Convolutional Neural Networks (CNNs) can effectively extract clinically relevant features, enabling automatic and accurate disease detection. The CXray-EffDet model [13] exemplifies the application of efficient network architectures for classifying various chest disorders, further highlighting the utility of transfer learning in this domain. Moreover, pre-trained networks such as VGG11, originally trained on large-scale datasets like ImageNet, have been successfully fine-tuned for medical imaging tasks, including pneumonia classification [14], reducing training time while improving diagnostic accuracy. Table I provides a comparison of key approaches in pneumonia detection from CXR images, highlighting model architectures, performance metrics, and limitations. The reviewed literature indicates a growing trend toward combining FL and DL to build scalable, privacy-preserving, and accurate diagnostic models for CXR analysis.

This study proposes a novel FL-based architecture designed for the classification of large-scale medical imaging datasets, with a specific focus on pneumonia detection from CXR images. The core of our approach involves fine-tuning a pre-trained VGG11 model using transfer learning, thereby leveraging rich feature representations acquired from large-scale natural image datasets. This enables more effective feature extraction and improved generalization across diverse clinical data. To mitigate client drift, we incorporate FedProx regularization [15], which penalizes significant deviations from the global model during local training, thereby stabilizing convergence in the presence of heterogeneous data distributions. In addition, we introduce a realistic data partitioning strategy to simulate client-level heterogeneity. This strategy includes two components: i) label imbalance is introduced using a Dirichlet distribution, reflecting the unequal class distributions commonly observed in real-world medical datasets; and ii) image-level noise is added via Gaussian perturbations, simulating variability in imaging quality and acquisition protocols across different institutions. These augmentations address the veracity and variety dimensions of big data, contributing to the development of a more robust and generalized global model. Building on prior work, which demonstrates that controlled noise injection and synthetic label imbalance can improve model robustness and generalization [16], our framework integrates these augmentation techniques directly into the FL pipeline.

TABLE I. COMPARISON OF APPROACHES FOR PNEUMONIA DETECTION

Ref.	Model Used	Performance Area Under Curve (AUC) / Accuracy	Limitations
[1]	Custom deep CNN architecture	~ 90% Accuracy	Centralized design; privacy concerns and scalability issues when applied to multi-institutional scenarios.
[9]	Deep CNN with advanced FL aggregation (FedAvg/FedProx)	AUC ~ 0.90	High communication overhead; sensitive to data heterogeneity across clients.
[10]	Ensemble of CNNs within a federated framework	~ 88% Accuracy	Increased computational complexity; requires careful ensemble tuning and integration.
[11]	Multi-task CNNs with specialized aggregation strategies	AUC ~ 0.87	Aggregation complexity; may be limited when tasks or data features are too divergent across clients.
[13]	EfficientDet-based model	AUC ~ 0.92	High computational requirements; may need further adaptation for federated scenarios or environments with strict resource constraints.
[14]	DenseNet-based architecture (CheXNet)	AUC ~ 0.88 (indicative)	Centralized approach; requires large-scale annotated data and significant computational resources; may lack privacy preservation in multi-institutional use
[13, 17]	CNNs (e.g., Visual Geometry Group (VGG), ResNet, InceptionResNetV2)	~ 95% Accuracy	Centralized learning approach, class imbalance issues

II. METHODOLOGY

This section presents the proposed FL system for large-scale medical image classification, with a particular emphasis on pneumonia detection from CXR images, as illustrated in Figure 1. The approach addresses the core challenges associated with big data through an integrated framework that combines data preprocessing, heterogeneous data partitioning with noise injection, federated optimization using FedProx, and transfer learning based on a pre-trained VGG11 model. This methodology is designed to ensure model robustness against heterogeneous, non-IID, and noisy data, while simultaneously maintaining data privacy and computational efficiency.

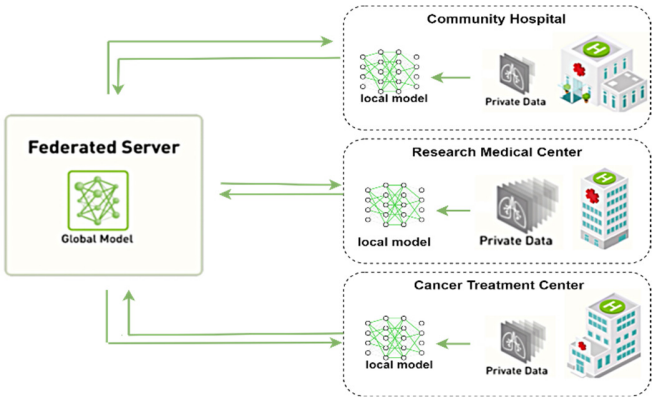

Fig. 1. FL system for large medical image classification.

A. Data Partitioning and Noise Injection

To simulate real-world scenarios characterized by data heterogeneity and quality variations, we employ two key techniques:

- Data partitioning via Dirichlet distribution: The CXR dataset is partitioned among multiple clients using a Dirichlet distribution [18], which introduces label imbalance by assigning varying class proportions to each client. As a result, some clients predominantly possess pneumonia cases, while others contain mostly normal instances. This strategy reflects the inherent data distribution disparities commonly observed across institutions in practical medical settings.
- Gaussian noise injection: To model variations in image quality stemming from different imaging devices or acquisition conditions, a controlled level of Gaussian noise is injected into the images for selected clients. The noise levels vary across clients, mimicking inconsistent image quality. This augmentation is essential for training a model that is robust to noise and capable of generalizing effectively to real-world medical imaging data.

B. Data Preprocessing

Data preprocessing plays a crucial role in ensuring consistency, quality, and suitability of input images for effective model training. Our preprocessing pipeline leverages established tools and frameworks, including Python's Pillow, OpenCV, and TorchVision from the PyTorch library. The steps include:

- Resizing: All CXR images are resized to a fixed resolution of 224 × 224 pixels to meet the input requirements of the pre-trained VGG11 model. This standardization facilitates efficient batch processing and ensures architectural compatibility.
- Normalization: Each image is normalized using the mean and standard deviation values from the ImageNet dataset, aligning the input pixel distribution with that of the data used during the VGG11 model's initial training. This alignment enhances training stability and accelerates convergence.
- Data Augmentation: To improve model generalization and reduce overfitting, various augmentation techniques are applied during training, including random horizontal flips, rotations, affine transformations, and random cropping/resizing. During validation and testing, only resizing and normalization are performed to ensure consistent and unbiased evaluation.
- Quality Control: Images with severe artifacts, corrupt content, or incorrect or missing labels are excluded from the dataset to maintain data integrity and prevent negative impacts on model training.

C. Federated Learning (FL) Framework

With the datasets partitioned and preprocessed, we implement an FL system to facilitate decentralized training while ensuring data privacy. The overall process is illustrated in Figure 2.

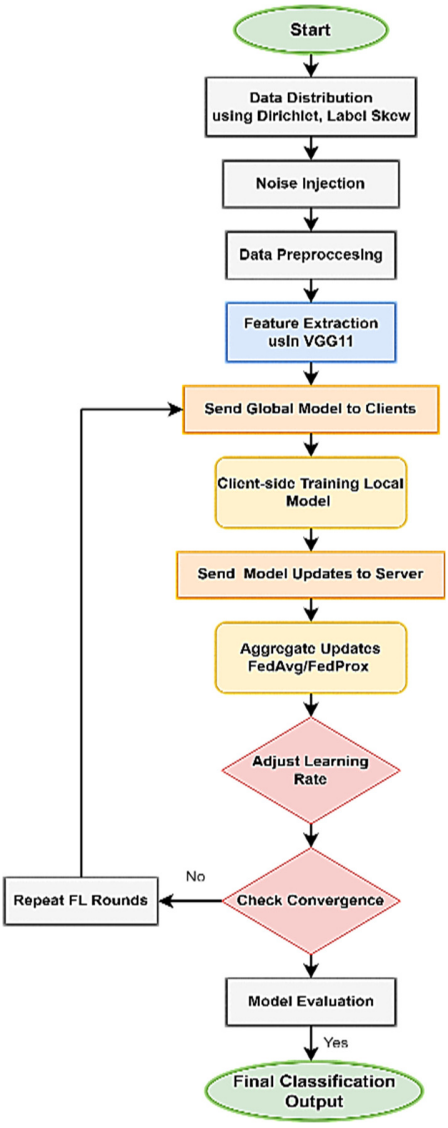

Fig. 2. Flowchart of the proposed approach.

1) Global Model Initialization and Distribution

The central server initializes a global model with a pre-trained VGG11 architecture. The global model is thereafter distributed securely to a subset of participating clients. Each round employs random client selection to ensure diverse data representation and equitable participation.

2) Local Training with FedProx Regularization

Each selected client trains the received model on its local dataset for a fixed number of epochs. To handle non-IID data and reduce client drift, the local training incorporates FedProx regularization into the loss function:

$$L = L_{CE} + (\mu/2)\,\|w_{local} - w_{global}\|^2 \quad (1)$$

where L is the overall loss function used for training in the FL setting, L_{CE} is the cross-entropy loss, μ is the regularization coefficient, w_{local} and w_{global} are the global and local model weights, respectively. This proximal term penalizes large deviations from the global model, reducing client drift while still allowing each client to adapt to its local data.

3) Aggregation of Local Updates

After training, clients transmit only their model weight updates to the central server. The server aggregates these updates using a weighted averaging scheme:

$$w_{global}^{t+1} = \sum_{i=1}^{K} \frac{n_i}{N} w_{local}^{i} \quad (2)$$

where: w_{global}^{t+1} is the new global model weights after aggregating updates from all participating clients at round $t + 1$, K is the number of clients that participated in the current training round, w_{local}^{i} is the local model weight of client i after training on its dataset, n_i is the number of training samples at client i, and N is the total number of training samples across all participating clients in round t.

4) Dynamic Learning Rate Adaptation and Redistribution

To improve convergence and stability in the presence of heterogeneous updates, we introduce Dynamic Learning Rate Adaptation (DLRA). The global learning rate $\eta^{(t+1)}$ is adjusted based on changes in model performance (e.g., validation accuracy or loss) between rounds:

$$\eta^{(t+1)} = \eta^{t} \cdot f(\Delta\mathcal{M}^{t}) \quad (3)$$

where $\Delta\mathcal{M}^{t}$ is the performance change at round t, and $f(\Delta\mathcal{M}^{t})$ is a function that determines how the learning rate should be modified based on the performance change. Specifically:

$$f(\Delta\mathcal{M}^{t}) = \gamma_{decay}, \text{if } \Delta\mathcal{M}^{t} > \varepsilon \quad (4)$$

$$f(\Delta\mathcal{M}^{t}) = \gamma_{increase}, \text{if } \Delta\mathcal{M}^{t} < \varepsilon \quad (5)$$

where ε is a small threshold (e.g., 0.001) to ignore minor fluctuations.

The DLRA mechanism is formalized in Algorithm 1. It reduces the learning rate when improvements are consistent (to support fine-tuning) and slightly increases it if performance stagnates, encouraging further exploration. After each update, the refined model is redistributed to clients for the next training round. To maintain communication efficiency, only model updates are exchanged, not raw data. This approach ensures continuous global model improvement while preserving data privacy and minimizing network overhead through secure update protocols.

Algorithm 1: DLRA
Input: Initial learning rate η^0, decay factor γ_{decay}, increase factor $\gamma_{increase}$
for each global round t do
    Train local models with η^t
    Aggregate global model w_{global}^{t}
    Evaluate validation performance: Compute $\Delta\mathcal{M}^t$
    if($\Delta\mathcal{M}^t$> 0) then /Improved accuracy or lower loss/
        $\eta^{(t+1)}=\eta^t$ * γ_{decay}
    else if ($\Delta\mathcal{M}^t$ ~ 0 or worsens) then
        $\eta^{(t+1)}=\eta^t$ * $\gamma_{increase}$
    end if
end for
Output: Optimized learning rate $\eta^{(t+1)}$

D. Model Architecture and Transfer Learning

Leveraging the power of transfer learning, our model is built upon the robust VGG11 architecture, which has proven effective in FL contexts. VGG networks are highly structured and sequential, making them particularly suitable for feature extraction tasks [15]. Each convolutional layer builds upon established hierarchical patterns, facilitating effective learning even in transfer scenarios (Figure 3).

- Feature extraction: The convolutional base of a pre-trained VGG11 model (originally trained on ImageNet) is employed to extract high-level features from CXR images. Despite domain differences, ImageNet pretraining offers a strong foundation for capturing relevant visual patterns.
- Custom classifier adaptation: We have replaced the original VGG11 classifier with a fully connected network for binary classification (pneumonia versus normal) to handle different spatial dimensions. Our proposed classifier comprises: Adaptive Average Pooling → Fully Connected (128 neurons, ReLU) → Output Layer (SoftMax for 2 classes).
- Integration with FL: This modified architecture is seamlessly integrated into the FL framework. Local training is performed using FedProx regularization, ensuring robustness to data heterogeneity by discouraging significant divergence from the global model while allowing local adaptability.

III. EXPERIMENTS AND RESULTS

A. Dataset

The dataset used in our FL experiments is the Chest X-Ray Dataset for Pneumonia Detection [19], a widely adopted

benchmark in medical image classification research. It comprises chest radiographs labeled as either pneumonia (diseased lungs) or normal (healthy lungs), with 4,280 pneumonia and 1,583 normal cases. The images were sourced from multiple medical institutions and are publicly available on the Kaggle platform. To better suit our experimental requirements, the dataset was repartitioned from its original configuration. Specifically, we employed:

- A training set of 4,198 images (3,216 pneumonia, 982 normal).
- A validation set of 1,000 images (500 pneumonia, 500 normal).
- A test set of 665 images (564 pneumonia, 101 normal).

Figure 4 presents 8 example images of the dataset.

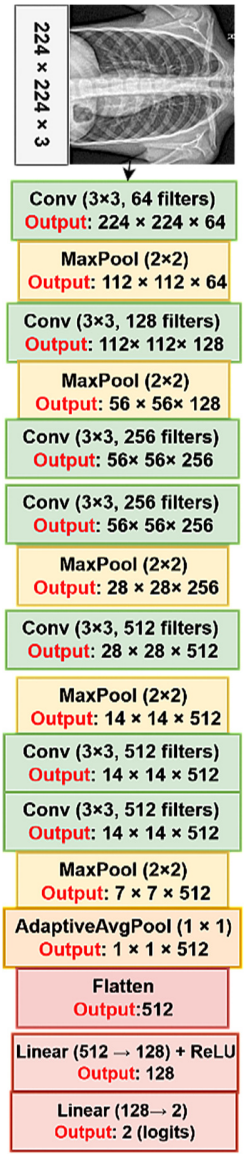

Fig. 3. The model architecture.

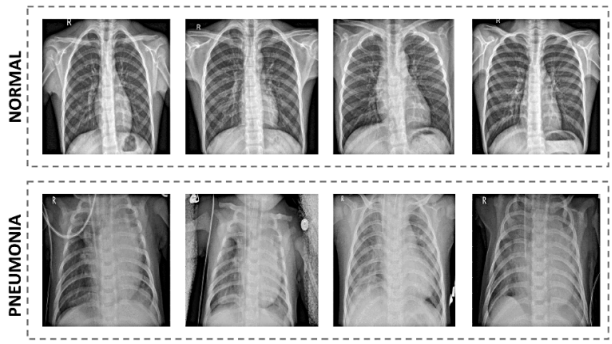

Fig. 4. Example of the two categories of images under study (pneumonia and normal).

1) Experimental Setup

We conducted our experiments using a high-performance FL environment simulating 60 clients, with 6 clients randomly selected per global round to participate in training. Each client performed local updates on its data and transmitted only model parameters to a central server, equipped with an AMD Ryzen 9 CPU and an NVIDIA RTX 3090 GPU.

A VGG11-based model was fine-tuned using FedProx regularization ($\mu = 1.0 \times 10^{-2}$) to handle non-IID data. Key hyperparameters included 200 global rounds, 5 local epochs per client, adaptive learning rate, and Stochastic Gradient Descent (SGD) optimizer with momentum (0.9) and weight decay (1.0×10^{-5}).

2) Validation Metrics

To evaluate the performance and robustness of our federated approach, we employed the following metrics:

- Area Under the Receiver Operating Characteristic Curve (AUC-ROC): Captures the model's ability to distinguish between pneumonia and normal cases across varying classification thresholds. This metric is particularly suitable for imbalanced datasets where accuracy may be misleading.
- Recall: Measures the model's sensitivity, its ability to correctly identify pneumonia-positive cases. High recall indicates fewer false negatives, which is crucial in medical diagnostics.
- F1-score: The harmonic mean of precision and recall, offering a balanced metric that reflects both false positives and false negatives. It is particularly valuable for evaluating performance on imbalanced data.
- Training time per round: Quantifies the computational efficiency of the FL framework by measuring the time required to complete one round of local training and global aggregation.

B. Results and Analysis

The outcomes of our FL experiments over 200 communication rounds are presented in Table II, with performance trends of AUC-ROC and recall with F1-score illustrated in Figures 5 and 6, respectively.

TABLE II. PERFORMANCE METRICS AT DIFFERENT TRAINING ROUNDS

Round	AUC-ROC (%)	Learning rate	Recall (%)	F1-score (%)	Time (s)
2	50.00	1.0×10^{-4}	50.00	52.01	21.32
10	95.75	1.0×10^{-8}	95.75	90.16	38.84
50	75.00	1.0×10^{-7}	75.20	75.00	49.98
100	81.25	1.0×10^{-8}	81.25	80.57	52.04
150	78.75	1.0×10^{-8}	78.75	77.11	53.12
200	82.25	1.0×10^{-8}	82.25	80.57	48.42

During the initial 10 rounds, the model exhibited rapid improvement, with AUC-ROC increasing from 50.00% to 95.75%, recall reaching 95.75%, and F1-score peaking at 90.16%. This sharp rise underscores the effectiveness of transfer learning with the pre-trained VGG11 architecture, which enabled robust feature extraction from the outset. Additionally, the significant drop in learning rate from 1.0×10^{-4} to 1.0×10^{-8} contributed to training stability by preventing abrupt weight updates, supporting fast and stable convergence.

Between rounds 50 and 100, AUC-ROC values stabilized between 75% and 81% on both test and validation sets. This phase of consistent performance reflects the impact of FedProx regularization, which mitigated divergence among client updates. The use of weighted federated aggregation helped maintain balanced contributions from heterogeneous clients. A slight dip in AUC-ROC at round 50 (75.00%) suggests potential noise in local updates from specific clients. Future work could incorporate adaptive client weighting to reduce such instability.

Performance fluctuations observed during later rounds are likely due to non-IID client drift, driven by disparities in disease prevalence, image quality, and noise across client datasets. These inconsistencies led to variable update contributions, occasionally destabilizing global model performance.

In parallel, training time remained stable across rounds, averaging around 50 seconds per round, confirming the computational scalability of the proposed FL framework. round 2 was the fastest (21.32 s), attributed to limited early updates. The longest round, round 150 (53.12 s), may reflect increased computational load due to more complex model refinements and gradient computations in later stages.

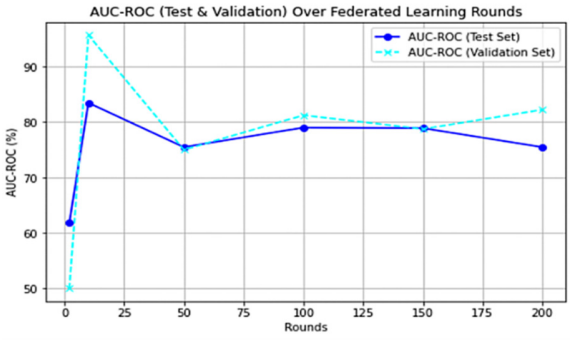

Fig. 5. AUC-ROC over FL rounds (test vs. validation).

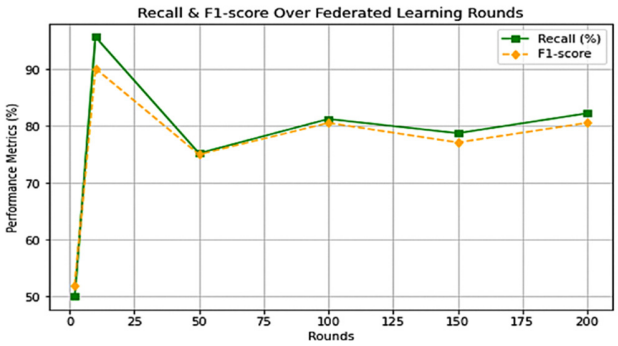

Fig. 6. Recall and F1-score trends over rounds (validation set).

The confusion matrix (Figure 7) shows that the model successfully identified the vast majority of pneumonia cases with minimal false negatives (only 20). Although there are some false positives (35 normal cases predicted as pneumonia), the high recall indicates that the model is highly effective at detecting actual disease cases, which is critical in medical diagnostics.

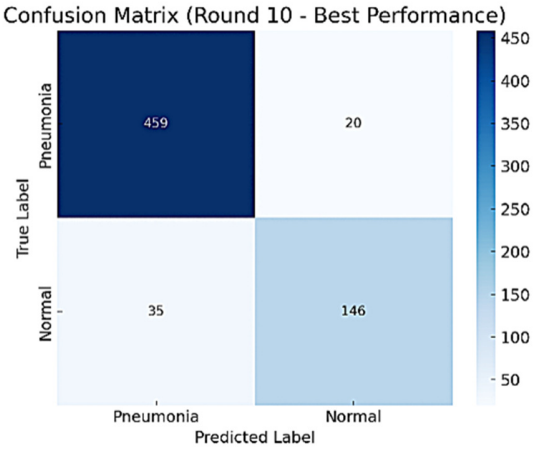

Fig. 7. Confusion matrix: pneumonia detection (test set).

To evaluate the effectiveness of our proposed methodology, we compared its performance with state-of-the-art pneumonia classification techniques, including both centralized DL models and alternative FL frameworks. Table III summarizes the comparison based on the AUC-ROC metric, along with each approach's noted limitations.

Traditional FL methods such as FedAvg + ResNet50 show degraded performance in non-IID settings due to a lack of robust regularization, resulting in lower AUC-ROC scores (88.90%). Centralized models such as CheXNet (93.40%) and the CNN variants used in [12] (95.07%) perform competitively but lack privacy preservation, making them unsuitable for multi-institutional real-world deployments where data cannot be shared. In contrast, our approach offers comparable or superior performance (95.75%) in a privacy-preserving federated setting. Moreover, ensemble-based FL techniques (e.g., in [10]) involve training and maintaining multiple models, leading to significant communication and

computational overhead. Our single-model approach is computationally more efficient, requiring just ~50 seconds per round, and is better suited for scalable deployments in distributed healthcare environments.

TABLE III. COMPARISON OF DIFFERENT APPROACHES FOR PNEUMO DETECTION

Study	Model Used	AUC-ROC (%)	Limitations
[9]	FedAvg + ResNet50	88.90	Lacks robust regularization against client drift
[10]	CNN + FL ensemble	90.10	Requires multiple models, increasing communication overhead
[11]	FL with custom aggregation	91.50	High dependency on client data distributions
[12]	CNN-based-model (VGG16, VGG19, InceptionResNetV2)	95.07	Computationally expensive; limited generalization
[14]	CheXNet (DenseNet-121)	93.40	Requires centralized training; privacy concerns
Proposed approach	FedProx + VGG11	**95.75**	Late-stage fluctuations due to non-IID client drift

IV. CONCLUSION AND FUTURE WORKS

This study introduced a scalable Federated Learning (FL) framework for pneumonia detection from Chest X-ray (CXR) images, designed to address the challenges of large-scale, decentralized medical datasets. The proposed approach achieved a peak validation Area Under the Receiver Operating Characteristic Curve (AUC-ROC) of 95.75% and test AUC-ROC of 83.49%, with a recall of 95.75%, demonstrating strong diagnostic performance. Key contributions include the use of FedProx regularization to mitigate client drift, dynamic learning rate adaptation for training stability, and a novel data partitioning strategy that integrates Dirichlet-based label imbalance and Gaussian noise injection to simulate real-world heterogeneity. Compared to existing methods, our framework offers a privacy-preserving, computationally efficient, and highly scalable solution for decentralized medical Artificial Intelligence (AI). The system maintained consistent training times (~50 seconds per round), supporting its applicability in distributed hospital networks and large-scale federated deployments.

In summary, the proposed FL architecture effectively handles the volume, velocity, variety, value, and veracity challenges of medical big data while ensuring patient privacy. Future research will focus on enhancing model robustness, supporting real-world clinical integration, and extending the framework to multimodal data sources to further advance the capabilities of federated healthcare AI.

REFERENCES

[1] G. Gana *et al.*, "Development and performance testing of a deep learning computer-aided diagnosis system for chest X-rays," *European Respiratory Journal*, vol. 60, no. suppl. 66, 2022, https://doi.org/10.1183/13993003.congress-2022.3085.

[2] P. Chakraborty and C. Tharini, "Pneumonia and Eye Disease Detection using Convolutional Neural Networks," *Engineering, Technology & Applied Science Research*, vol. 10, no. 3, pp. 5769–5774, Jun. 2020, https://doi.org/10.48084/etasr.3503.

[3] N. Kumar, A. Hashmi, M. Gupta, and A. Kundu, "Automatic Diagnosis of Covid-19 Related Pneumonia from CXR and CT-Scan Images," *Engineering, Technology & Applied Science Research*, vol. 12, no. 1, pp. 7993–7997, Feb. 2022, https://doi.org/10.48084/etasr.4613.

[4] H. Zemmouri, S. Labed, and A. Kout, "A survey of parallel clustering algorithms based on vertical scaling platforms for big data," in *2022 4th International Conference on Pattern Analysis and Intelligent Systems (PAIS)*, Oum El Bouaghi, Algeria, Oct. 2022, pp. 1–8, https://doi.org/10.1109/pais56586.2022.9946663.

[5] A. Z. Tan, H. Yu, L. Cui, and Q. Yang, "Towards Personalized Federated Learning," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 34, no. 12, pp. 9587–9603, Dec. 2023, https://doi.org/10.1109/tnnls.2022.3160699.

[6] Y. Wang, M. M. Rosli, N. Musa, and F. Li, "Multi-Class Imbalanced Data Classification: A Systematic Mapping Study," *Engineering, Technology & Applied Science Research*, vol. 14, no. 3, pp. 14183–14190, Jun. 2024, https://doi.org/10.48084/etasr.7206.

[7] S. Ram, Y. N. Kiran, A. Bhute, and T. Khare, "Federated Learning for Accurate Labeling of Chest X-Ray Scans," in *2024 36th Conference of Open Innovations Association (FRUCT)*, Lappeenranta, Finland, Oct. 2024, pp. 649–654, https://doi.org/10.23919/fruct64283.2024.10749865.

[8] H. Zhu, J. Xu, S. Liu, and Y. Jin, "Federated learning on non-IID data: A survey," *Neurocomputing*, vol. 465, pp. 371–390, Nov. 2021, https://doi.org/10.1016/j.neucom.2021.07.098.

[9] S. Sharma, K. Guleria, and A. Dogra, "FedPneu: Federated Learning for Pneumonia Detection across Multiclient Cross-Silo Healthcare Datasets," *Current Medical Imaging Reviews*, vol. 21, Mar. 2025, https://doi.org/10.2174/0115734056333970241212132150.

[10] A. Mabrouk, R. P. D. Redondo, M. A. Elaziz, and M. Kayed, "Ensemble Federated Learning: An approach for collaborative pneumonia diagnosis," *Applied Soft Computing*, vol. 144, Sep. 2023, Art. no. 110500, https://doi.org/10.1016/j.asoc.2023.110500.

[11] P. Kulkarni, A. Kanhere, P. H. Yi, and V. S. Parekh, "From Isolation to Collaboration: Federated Class-Heterogeneous Learning for Chest X-Ray Classification." *arXiv*, Nov. 15, 2024, https://doi.org/10.48550/arXiv.2301.06683.

[12] P. R. Kaur, A. Sharma, I. Singh, and R. Malhotra, "Deep Learning-Based Pneumonia Recognition from Chest X-Ray Images," *International Journal of Performability Engineering*, vol. 18, no. 5, 2022, Art. no. 380, https://doi.org/10.23940/ijpe.22.05.p8.380386.

[13] M. Nawaz, T. Nazir, J. Baili, M. A. Khan, Y. J. Kim, and J.-H. Cha, "CXray-EffDet: Chest Disease Detection and Classification from X-ray Images Using the EfficientDet Model," *Diagnostics*, vol. 13, no. 2, Jan. 2023, Art. no. 248, https://doi.org/10.3390/diagnostics13020248.

[14] P. Rajpurkar *et al.*, "CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning." *arXiv*, 2017, https://doi.org/10.48550/ARXIV.1711.05225.

[15] V. Iglovikov and A. Shvets, "TernausNet: U-Net with VGG11 Encoder Pre-Trained on ImageNet for Image Segmentation." *arXiv*, 2018, https://doi.org/10.48550/ARXIV.1801.05746.

[16] C. Shorten and T. M. Khoshgoftaar, "A survey on Image Data Augmentation for Deep Learning," *Journal of Big Data*, vol. 6, no. 1, Dec. 2019, https://doi.org/10.1186/s40537-019-0197-0.

[17] N. Kumar, J. Manzar, Shivani, and S. Garg, "Underwater Image Enhancement using Deep Learning," *Multimedia Tools and Applications*, vol. 82, no. 30, pp. 46789–46809, Dec. 2023, https://doi.org/10.1007/s11042-023-15525-4.

[18] J. Lin, "On The Dirichlet Distribution," Department of Mathematics and Statistics, Queens University, 2016.

[19] *Chest X-Ray Images (Pneumonia).* (2018), Kaggle. [Online]. Available: https://www.kaggle.com/paultimothymooney/chest-xray-pneumonia.