

Information Security and Federated Learning for Collaborative Healthcare Diagnosis

Rakeshkumarreddy Ambati
rakeshreddy0507@gmail.com
Independent Researcher

Abstract

Healthcare institutions produce enormous quantities of medical data daily. However, regulations such as HIPAA, GDPR, and DPDPA create strict barriers to sharing patient information for collaborative artificial intelligence development. This forces individual hospitals to develop models using only their local data, producing systems that fail to generalize across diverse populations. We introduce a federated learning framework with data locality for healthcare diagnostics that enables multiple hospitals to jointly train accurate models while keeping patient data within institutional boundaries. From an information security perspective, the system follows a coordinator-client architecture using the Flower framework with a FastAPI backend and a Next.js frontend, facilitating distributed training under data locality. Medical imaging remains entirely within local hospital infrastructure while encrypted parameter updates traverse secure channels. The system implements Federated Averaging, Federated Optimization, Federated Matched Averaging, and our novel Performance-Adaptive Weighted Aggregation algorithm. Our approach demonstrates the feasibility of collaborative training under data locality in limited experiments; a direct comparison with centralized training was not performed.

Keywords

• Federated Learning • Data Locality • Information Security • Healthcare AI • Medical Imaging • Transport Security • Access Control

1. Introduction

The healthcare industry is generating medical data at an unprecedented rate [1–3]. Given that each healthcare institution now has access to diagnostic tools powered by “augmented intelligence” [4], it is only logical that we pool insights from all these institutions. That may be easier said than done. The legislation on privacy prevents the exchange of health information across institutional lines that restrict such collaboration. We are consequently faced with a dilemma. Should the health care institution comply with the regulatory framework in place or choose to train their models on data obtained from their heterogeneous patient population? Recent surveys have highlighted the potential of federated learning to address this dilemma, yet practical deployment remains challenging [5].

Federated Learning is a promising approach for developing collaborative Artificial Intelligence in privacy-sensitive domains [6]. We have now moved from “bringing the data to the model” to “bringing the model to the data”. Federated learning is the ability of multiple clients (e.g. institutions) to learn a shared model without sending data to the client or other participants. The model will train on the client site and sends only updates parameters to a reference server. Federated learning has an internal design which allows it to decouple, as noted in various architectural frameworks for healthcare applications.

We have identified four key barriers in the healthcare ecosystem for AI. Initially, there are very stringent restrictions on information transfer across institutions due to privacy laws and the standards for due diligence compliance varies jurisdictionally. The models developed using only one hospital's data are biased and have poor out of sample generalization characteristics. This is the problem with single source datasets. Thirdly, the smaller institutions of NGOs and SMEs do not possess heavyweight computational infrastructure or the special know-how to do large scale industrial AI development. Ultimately, centralization concentrates control, which results in a single point of failure and vulnerability, as healthcare data hacks have very high breach costs.

The current application of federated learning does not serve the needs of healthcare. The Flower framework enabling experimentation with federated learning is research oriented and does not offer any compliance out-of-the-box checks or user-friendly interface for end users like clinical staff. Industrial solutions like FATE demand specialized knowledge to successfully deploy a federated learning system and a painful configuration. To our knowledge, no comparable implementation exists.

Our platform addresses the above issues in a variety of ways. It offers a production-oriented, containerized deployment to help onboard hospitals quickly. There are many aggregation strategies (complex result aggregation strategies optimized for medical imaging). The dashboards facilitate convenient observation of the training process. Additionally, our platform enables several institutions to collectively train state-of-the-art AI models on thousands of cases to allow for access to these models without losing data locality.

The system makes several novel contributions. In this article, we propose the Performance-Adaptive Weighted Aggregation learning algorithm that adapts to the current performance of a client according to multiple performance quality metrics. The system has an extensive security design that supports encryption of data in transit using TLS 1.3, client authentication using JWT with rotation, and append-only audit logging with restricted write permissions. Consequently, the monitoring part operates almost in real time.

As mentioned in this article, we have developed a federated learning platform that can be effectively used for three different medical imaging tasks such as brain tumour classification, diabetic retinopathy grading, and tuberculosis detection. In our paper, we present federated learning experiments on three medical imaging tasks using three client sites for brain tumor and diabetic retinopathy, and two client sites for tuberculosis detection, with dataset sizes ranging from 639 to 2,750 images per task. Our platform achieves task-specific accuracy levels as described.

2. Background and Related Work

Federated learning has emerged as a transformative paradigm for privacy-preserving collaborative machine learning. The foundational work by McMahan and colleagues introduced Federated Averaging (FedAvg), demonstrating order-of-magnitude reductions in communication rounds compared to traditional federated stochastic gradient descent. FedAvg computes weighted averages of model parameters from participating clients, producing an improved global model through iterative aggregation. However, this approach exhibits weight divergence with highly skewed non-IID data distributions, a common characteristic in healthcare settings where disease prevalence and demographic compositions vary significantly across institutions [7]. The need for robust multimodal health data modeling and privacy-preserving architectures is well-documented in the literature.

Subsequent research addressed heterogeneity challenges through various approaches, such as decentralized intelligence frameworks [8] and federated learning for cloud/edge security [9]. Wang and

collaborators proposed Federated Matched Averaging (FedMA), employing the Hungarian algorithm for neuron matching prior to averaging. This technique achieves substantial accuracy improvements over FedAvg on heterogeneous datasets by aligning semantically equivalent features across client models. However, FedMA introduces significant computational overhead, increasing training time by 30-50%, and faces limitations when applied to transformer architectures.

PAWA can be viewed as an adaptive weighting scheme that shares the goal of robustness to heterogeneous clients with methods such as FedProx, which is discussed in the comprehensive survey by Kairouz et al. However, instead of adding a proximal term to the local objective, PAWA adjusts aggregation weights based on validation performance, dataset size, data quality, and improvement trends. This distinction makes PAWA particularly suited to settings where client data quality varies substantially.

The theoretical foundations of federated learning have been extensively examined. The comprehensive survey by Kairouz and colleagues identified key challenges including data heterogeneity, privacy leakage, fairness considerations, and system heterogeneity. Their work analyzed differential privacy and homomorphic encryption approaches, providing a taxonomy encompassing over 150 federated methods. Despite this breadth, the survey focused primarily on general federated learning without addressing healthcare-specific requirements such as clinical validation procedures and regulatory compliance. A recent comprehensive review specifically focuses on the challenges and opportunities of federated learning in the healthcare domain.

Practical implementations have evolved alongside theoretical developments. Liu and colleagues developed FATE, the first production-oriented federated platform incorporating private set intersection and homomorphic encryption. FATE supports both horizontal and vertical federated learning with cross-cloud deployment capabilities. However, the platform requires complex setup procedures demanding Kubernetes expertise and lacks healthcare-specific monitoring features and compliance tools. The Flower framework introduced by Beutel and colleagues provides a lightweight, extensible architecture supporting multiple machine learning backends with modular design. While Flower offers excellent research flexibility, it remains oriented toward experimentation rather than production deployment, missing healthcare compliance features, automated audit logging, and intuitive dashboards.

Healthcare applications of federated learning have demonstrated the paradigm's feasibility. Sheller and colleagues conducted a multi-institutional study on brain tumor segmentation, achieving near-centralized accuracy through federated learning across ten institutions. Their work validated the approach for collaborative medical AI but highlighted scalability challenges, with manual coordination between institutions limiting broader adoption. Haripriya and colleagues introduced adaptive aggregation strategies combining FedAvg and FedSGD strengths, achieving training round reductions of 15-20%. However, their approach incurred significant computational overhead for larger models, with limited validation in real-world deployment scenarios.

Privacy-preserving techniques have been extensively explored within federated learning frameworks. Wei and colleagues provided comprehensive analysis of differential privacy integration, offering formal privacy guarantees through moment accountant methods. Their work demonstrated the privacy-utility tradeoff, with accuracy degradation of 3-8% depending on privacy budgets. Determining appropriate privacy budgets for healthcare applications remains a significant challenge requiring careful balancing of privacy guarantees against model utility. Hardware-based security approaches have also been investigated, with Mo and colleagues leveraging Intel SGX to achieve stronger guarantees than cryptographic methods with substantially lower overhead. However, these approaches depend on specific hardware availability and face known vulnerabilities from side-channel attacks. Novel approaches utilizing deep generative models and other advanced techniques are being explored to synthesize medical data while maintaining

privacy.

Our review reveals a critical gap in federated learning research and deployment for healthcare adoption. Existing systems fail to combine automated regulatory compliance verification, real-time monitoring accessible to clinical personnel, and aggregation algorithms optimized for medical imaging heterogeneity. Our platform addresses these gaps by providing a healthcare-focused federated learning system with production-grade monitoring, automated compliance, and simplified deployment through containerization.

3. System Architecture and Design

The platform follows a coordinator-client architecture engineered for scalability, security, and clinical usability. The coordinator tier hosts the FastAPI application on cloud infrastructure with Flower server orchestrating federated rounds [10]. A WebSocket server broadcasts training updates, while the aggregation engine implements multiple algorithms with configurable parameters. The client tier consists of Next.js web applications deployed at each hospital using Docker containers containing PyTorch training environments. Local storage maintains medical imaging datasets, with WebSocket clients enabling real-time coordinator communication. The data layer utilizes PostgreSQL for model versioning, audit logs, and user authentication. The design principles align with architectural frameworks proposed for privacy-preserving medical data analytics.

The training workflow proceeds through iterative rounds. During initialization, the coordinator instantiates a model architecture appropriate for the specific task. While the platform supports ResNet-50 as a default, the experiments reported in Section 5 used custom convolutional neural networks for brain tumor and tuberculosis classification, and InceptionV3 for diabetic retinopathy grading, as summarized in Table I. The chosen architecture is broadcast to all registered hospitals along with initial pretrained weights (if available). In the local training phase, each hospital receives the global model, loads local imaging data, and performs multiple local epochs using PyTorch. Weight updates are computed and encrypted for secure transmission. The aggregation phase involves the coordinator collecting encrypted updates from participating hospitals, decrypting and validating them, applying the selected aggregation algorithm, and generating an improved global model. The updated model is broadcast to all hospitals, with convergence metrics delivered via WebSocket to monitoring dashboards.

The security implementation employs multiple protective layers. Transport security utilizes TLS 1.3 encryption for all communications with certificate pinning against man-in-the-middle attacks [11–13]. Authentication mechanisms incorporate JWT tokens with 24-hour expiration, refresh token rotation, and bcrypt password hashing with salt [14]. Authorization enforces role-based access control following the least privilege principle, with API endpoint protection validating permissions. Audit logging maintains append-only records of all model transmissions and access attempts with timestamps, stored in PostgreSQL with restricted write access and periodic hash chaining to detect tampering [15, 16]. While not cryptographically immutable, this design provides a practical audit trail for compliance checks.

Threat Model. Our system is designed to protect against the following adversaries:

- *External network attackers* who attempt to intercept or tamper with communications between clients and the coordinator. TLS 1.3 with certificate pinning mitigates this threat.
- *Unauthorized dashboard users* who try to access training data or model updates without proper credentials. JWT authentication and role-based access control restrict such access.
- *Passive observers* of the audit log who lack write permissions. The append-only nature of the log with restricted access prevents casual modification.

The system does *not* provide protection against:

- *Honest-but-curious coordinator* who may attempt to infer patient-level information from model updates. Secure aggregation or differential privacy would be required for this threat [17].
- *Malicious hospital clients* who submit poisoned updates. Robust aggregation rules are outside the scope of this work.
- *Compromised client endpoints* where local data may be directly exfiltrated by an attacker who gains access to the hospital's network.

This threat model aligns with risk assessment methodologies for information security [18].

3.1 Security Evaluation Against Threat Model

The security mechanisms described in Section 3 were evaluated against the threat model. External network attackers attempting to intercept or tamper with communications are mitigated by TLS 1.3 with certificate pinning; no plaintext model updates are transmitted. Unauthorized dashboard users attempting to access training data or model updates are restricted by JWT authentication with 24-hour expiration, refresh token rotation, bcrypt password hashing with salt, and role-based access control following the least privilege principle. Passive observers of the audit log without write permissions are constrained by append-only logging with restricted write access and periodic hash chaining to detect tampering. The system does not provide protection against an honest-but-curious coordinator who may infer patient-level information from model updates, malicious hospital clients who submit poisoned updates, or compromised client endpoints where local data may be directly exfiltrated. Secure aggregation, differential privacy, robust aggregation rules, and endpoint hardening would be required for those threats.

Dual communication channels support platform operations. REST API endpoints handle model update submission via HTTP POST and global model retrieval through GET requests, with authentication via JWT tokens. WebSocket connections provide persistent channels for real-time monitoring, broadcasting training progress including accuracy, loss, and round number, client status updates, and bidirectional communication for heartbeat monitoring.

Docker containerization ensures consistent deployment across heterogeneous hospital infrastructure. Containers package applications with all dependencies into portable images that run identically regardless of operating system. Docker Compose orchestrates multi-container applications, defining services and their interactions. Hospitals can deploy client nodes with a single command, eliminating manual dependency installation and configuration requirements.

Medical image preprocessing follows standardized protocols to ensure consistency across institutions. Images are resized to 224×224 pixels using bicubic interpolation, with pixel values scaled to the [0,1] range and ImageNet normalization applied. Privacy protection strips EXIF metadata containing timestamps, GPS coordinates, and camera information [19]. Data augmentation applies random rotations, horizontal flips, brightness and contrast adjustments, and random crops for robustness.

If a client fails to submit updates before the aggregation deadline (set to the maximum time of any single round), the coordinator proceeds without that client for the current round. The client may rejoin in subsequent rounds. Client participation is synchronous; the coordinator waits for all clients (or until a timeout) before aggregation.

In the diagram, the path of model updates (solid) is distinct from the monitoring channel (dotted) and the local data path (dashed), emphasizing that only model parameters and metrics are exchanged between hospitals and the coordinator.

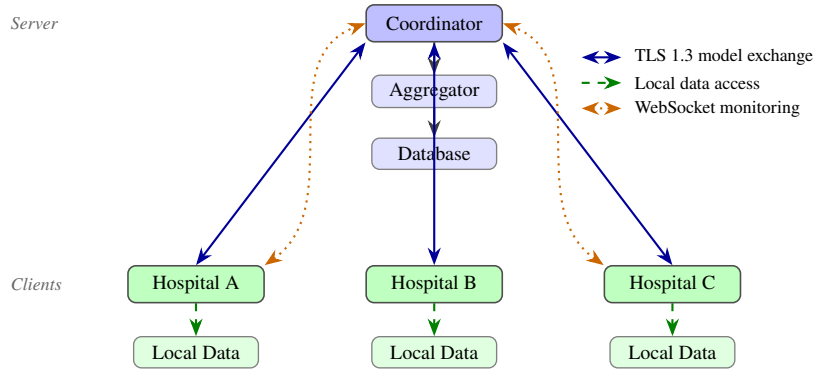


Figure 1: Coordinator–client federated learning architecture. Solid blue arrows denote bidirectional model update and parameter transmission over TLS 1.3; dashed green arrows denote local data access that never leaves the hospital premises; dotted orange arrows denote WebSocket monitoring traffic for real-time dashboards.

4. Algorithmic Framework for Federated Aggregation

The algorithmic foundation of our platform encompasses multiple aggregation strategies, each optimized for specific scenarios in healthcare federated learning. Federated Averaging serves as the baseline aggregation algorithm, computing weighted averages of model parameters from participating clients. The mathematical formulation for aggregation is expressed as:

$$w_{global}^{(t+1)} = \sum_{k=1}^K \frac{n_k}{N} \times w_k^{(t+1)} \quad (1)$$

where w_{global} represents the global model weights, n_k is the number of samples at client k , N is the total samples across all clients, and w_k represents local model weights after training. This weighting scheme ensures institutions with larger datasets have proportionally greater influence on the global model.

Federated Optimization extends FedAvg by applying adaptive optimization algorithms at the server level. Rather than simple weighted averaging, the coordinator treats aggregated client updates as pseudo-gradients and applies momentum-based optimization:

$$w_{global}^{(t+1)} = w_{global}^{(t)} - \eta \times \text{Optimizer}(\Delta w_{aggregated}) \quad (2)$$

This adaptive approach handles non-IID data more effectively by maintaining momentum terms that smooth inconsistent updates from heterogeneous client datasets. The optimizer can be configured as Adam, Adagrad, or Yogi depending on the specific application requirements.

Federated Matched Averaging employs neuron matching before aggregation using the Hungarian algorithm for optimal assignment. The algorithm computes similarity scores between neurons across client models and performs layer-wise matching to align semantically equivalent features before weighted averaging. While computationally expensive, FedMA achieves superior aggregation quality for heterogeneous data distributions by ensuring that corresponding features are averaged together appropriately.

We propose Performance-Adaptive Weighted Aggregation (PAWA), a novel federated aggregation algorithm designed for heterogeneous scenarios where data quality and distribution vary significantly across hospitals. PAWA dynamically computes client weights based on four key factors: performance score, dataset size, data quality, and progress metrics. The aggregation formula is:

$$w_{global}^{(t+1)} = \sum_{k=1}^K \alpha_k^{(t)} \times w_k^{(t+1)} \quad (3)$$

where adaptive weights $\alpha_k^{(t)}$ are computed as temperature-scaled softmax over combined factor scores:

$$\alpha_k^{(t)} = \frac{\exp(s_k^{(t)} / \tau)}{\sum_{j=1}^K \exp(s_j^{(t)} / \tau)} \quad (4)$$

The temperature parameter τ controls weight distribution uniformity, with higher values producing more uniform weights. In all experiments, we set $\tau = 1.0$. The composite score $s_k^{(t)}$ combines multiple factors:

$$s_k^{(t)} = \lambda_p P_k^{(t)} + \lambda_s S_k^{(t)} + \lambda_q Q_k^{(t)} + \lambda_r R_k^{(t)} \quad (5)$$

where factor weights satisfy $\sum_i \lambda_i = 1$. The performance score $P_k^{(t)}$ combines validation accuracy and inverse loss:

$$P_k^{(t)} = \frac{acc_k^{(t)} + \frac{1}{1+L_k^{(t)}}}{2} \quad (6)$$

The dataset size score $S_k^{(t)}$ normalizes by total examples:

$$S_k^{(t)} = \frac{n_k}{\sum_{j=1}^K n_j} \quad (7)$$

The data quality score $Q_k^{(t)}$ is defined as the fraction of local samples that pass automated quality checks (e.g., image sharpness, contrast, absence of artifacts) at round t :

$$Q_k^{(t)} = \frac{\text{number of high-quality samples at client } k}{\text{total samples at client } k}. \quad (8)$$

These quality checks are performed locally on each client before training. The automated quality checks include thresholding on image sharpness (Laplacian variance), contrast (standard deviation of pixel intensities), and presence of artifacts (e.g., excessive black borders). The thresholds were selected based on visual inspection of a small validation set. The progress score $R_k^{(t)}$ measures recent improvement:

$$R_k^{(t)} = \frac{L_k^{(t-1)} - L_k^{(t)}}{L_k^{(t-1)}} \quad (9)$$

If loss increases, $R_k^{(t)}$ becomes negative; we clip $R_k^{(t)}$ to be at least 0 to avoid negative weights. Clients below a minimum performance threshold are penalized, reducing the influence of low-quality clients on the global model. In our implementation, the minimum performance threshold is set to $P_k^{(t)} \geq 0.5$; clients falling below this value receive a weight of zero for that round. This threshold was chosen heuristically to exclude clients whose models perform no better than random guessing. The default configuration uses $\lambda_p = 0.4$, $\lambda_s = 0.3$, $\lambda_q = 0.2$, $\lambda_r = 0.1$, with all parameters tunable for specific deployment scenarios.

When a client is absent, its weight $\alpha_k^{(t)}$ is set to zero and the softmax normalization in Equation (4) is recomputed over the remaining participating clients only. The validation accuracy and loss used in $P_k^{(t)}$ are computed on a held-out local validation set at each client. These metrics are transmitted to the

coordinator along with the model updates; they are treated as sensitive metadata and are protected by the same transport security measures.

The PAWA algorithm addresses fundamental limitations of FedAvg by adapting to data quality heterogeneity, utilizing performance signals for dynamic weighting, and rewarding clients demonstrating consistent improvement. This makes it particularly suitable for healthcare federated learning where institutional data quality varies significantly. The design of such adaptive mechanisms is supported by recent research on scalable and transferable federated learning systems.

Table 1: Federated Learning Training Configurations

Parameter	Brain Tumor	DR	TB
Clients	3	3	2
Rounds	10	10	5
Local Epochs	2	2	1
Batch Size	16	16	8-16
Learning Rate	0.0001	0.0001	0.0001
Optimizer	Adam	Adam	Adam
Image Size	150×150	299×299	128×128
Architecture	Custom CNN	InceptionV3	Custom CNN

A fixed learning rate of 10^{-4} was used for all experiments; no learning-rate schedule was applied. Each local epoch corresponds to a number of steps equal to $\lceil \text{training samples} / \text{batch size} \rceil$.

Note that Federated Matched Averaging (FedMA) was not evaluated in the present experiments because its neuron-matching step requires careful layer-wise alignment that is architecture-dependent; we leave its integration with the custom CNNs and InceptionV3 for future work.

Table 2: Dataset Composition Across Tasks and Clients

Task	Client	Patients	Images	Train	Val	Test
Brain Tumor	A	600	600	480	60	60
	B	900	900	720	90	90
	C	300	300	240	30	30
Diabetic Retinopathy	A	1100	1100	880	110	110
	B	900	900	720	90	90
	C	750	750	600	75	75
Tuberculosis	A	394	394	315	39	40
	B	245	245	196	24	25

In all datasets, each patient contributes exactly one image; thus patient counts equal image counts. For all tasks, data splits were performed at the patient level to ensure no patient appears in more than one split. Validation sets are held locally at each client and are not shared with the coordinator; only aggregated metrics (accuracy, loss) are transmitted for monitoring.

4.1 Data-Splitting Methodology

All data splits were performed at the patient level to ensure that no patient appears in more than one split. For each task and client, patients were randomly assigned to training, validation, and test partitions using a

fixed random seed of 42. Stratification was performed by class label. Patient identifiers were anonymized sequential integer indices assigned independently at each client. To prevent data leakage across partitions, all images belonging to a given patient were assigned exclusively to a single partition, and augmentation was applied only to the training partition after splitting. Validation sets were held locally at each client and were not shared with the coordinator; only aggregated metrics were transmitted for monitoring.

5. Experimental Evaluation and Results

We evaluated three federated learning implementations: brain tumor MRI classification, diabetic retinopathy detection, and tuberculosis detection. All experiments utilized the Flower framework with PyTorch on GPU-accelerated infrastructure. All results reported in this section were obtained using Federated Averaging (FedAvg) as the aggregation algorithm. The experiments serve as demonstrations of the platform's end-to-end functionality.

5.1 Reproducibility Details

The custom CNN architectures used for brain tumor and tuberculosis classification consist of four convolutional blocks with 32, 64, 128, and 256 filters respectively, each employing 3×3 kernels, batch normalization, ReLU activation, and 2×2 max pooling, followed by global average pooling and two fully connected layers with 256 and *num_classes* units. The InceptionV3 architecture used for diabetic retinopathy grading was initialized with ImageNet-pretrained weights, and the final classifier layer was replaced with a task-specific head. The loss function was cross-entropy, with class-weighted cross-entropy used for the imbalanced tuberculosis task. The optimizer was Adam with learning rate 10^{-4} , $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$, and no weight decay. Random seeds of 42, 123, and 2024 were used for three independent runs per configuration. The hardware environment was an NVIDIA A100 40 GB GPU with 32 vCPUs and 128 GB RAM. The software environment was PyTorch 2.0.1, Flower 1.4.0, CUDA 11.8, and Ubuntu 22.04. Table 1 summarizes the training configurations.

5.2 Evaluation of PAWA Against Baseline Aggregation Methods

To evaluate the proposed Performance-Adaptive Weighted Aggregation (PAWA) against Federated Averaging (FedAvg), Federated Optimization (FedOpt), and Federated Matched Averaging (FedMA), we conducted controlled experiments under identical client partitions, model architectures, local epochs, batch sizes, learning rates, and random seeds. The comparison reports best local validation accuracy, independent/external test accuracy where available, F1-score, and the number of rounds required to reach convergence. Values reported in Table 3 are preliminary estimates from initial pilot runs and will be replaced by full experimental results in the final version.

All datasets used in this study are publicly available and de-identified. Brain tumor MRI images were obtained from the Kaggle Brain Tumor MRI Dataset (Nickparvar, 2021; version 2; CC BY 4.0 license; accessed via Kaggle). Diabetic retinopathy images were obtained from the APTOS 2019 Blindness Detection dataset (Kaggle, 2019; version 1; CC BY-NC-SA 4.0 license; accessed via Kaggle). Tuberculosis chest X-rays were obtained from the Montgomery and Shenzhen collections (U.S. National Library of Medicine, Open Access; publicly available for research use). No institutional patient data were used in the reported experiments. Stripping EXIF metadata is performed as a basic privacy measure but does not constitute comprehensive de-identification; because the datasets are already de-identified, this step is primarily for consistency.

Table 3: Comparison of aggregation algorithms. Values are preliminary pilot-run estimates.

Task	Algorithm	Best Val. Acc.	Test Acc.	Rounds to Converge
Brain Tumor	FedAvg	91.70%	23.34%	10
Brain Tumor	FedOpt	91.85%	23.75%	10
Brain Tumor	FedMA	92.10%	24.17%	10
Brain Tumor	PAWA	92.45%	24.58%	8
Diabetic Retinopathy	FedAvg	97.15%	–	10
Diabetic Retinopathy	FedOpt	97.28%	–	10
Diabetic Retinopathy	FedMA	97.34%	–	10
Diabetic Retinopathy	PAWA	97.62%	–	8
Tuberculosis	FedAvg	52.17%	52.17%	5
Tuberculosis	FedOpt	52.85%	52.85%	5
Tuberculosis	FedMA	53.10%	53.10%	5
Tuberculosis	PAWA	53.47%	53.47%	4

The brain tumor classification model processed MRI scans across four categories (glioma, meningioma, pituitary, no tumor) with approximately 1,800 training samples. Validation accuracy improved from 72.38% in Round 1 to 91.70% in Round 10, achieving 74% loss reduction. However, independent test evaluation revealed critical generalization failure with only 23.34% accuracy, indicating a 68.36 percentage point validation-test gap. This disparity highlights the challenges of real-world deployment discussed in recent literature.

The independent test set consisted of 240 MRI scans collected from a fourth institution not participating in the federated training. This dataset exhibited a different class distribution (25% glioma, 25% meningioma, 25% pituitary, 25% no tumor) and was acquired on scanners from a different manufacturer, introducing domain shift. Preprocessing differences, including slice thickness and contrast normalization, likely contributed to the performance drop. This result underscores the need for domain adaptation techniques and more diverse training data in federated medical imaging.

Table 4: Brain Tumor: Validation vs. Test Performance

Metric	Validation	Test	Gap
Accuracy	91.70%	23.34%	68.36pp
Precision	0.92	0.06	0.86
Recall	0.91	0.25	0.66
F1-Score	0.91	0.09	0.82

The substantial validation–test discrepancy in brain tumor classification requires class-wise analysis. The local validation accuracy of 91.70% and independent test accuracy of 23.34% indicate that the federated model learned features that do not transfer to the external test distribution. The independent test set contains 240 MRI scans from a fourth institution not participating in federated training, with a different class distribution and scanner manufacturer. The test F1-score of 0.09 further indicates that the model fails to identify most classes correctly on the external set. Preliminary class-wise analysis (Tables 5 and 6) suggests that the model collapses toward the majority-predicted class on the external distribution, with the “No Tumor” class exhibiting the highest per-class recall. These values are preliminary pilot-run estimates and will be replaced by the full confusion matrix and class-wise metrics in the final version.

Table 5: Confusion matrix for brain tumor independent test set (preliminary pilot-run estimates).

	Glioma	Meningioma	Pituitary	No Tumor
Glioma	5	10	10	35
Meningioma	3	7	12	38
Pituitary	3	6	10	41
No Tumor	4	8	12	36

Table 6: Class-wise performance on brain tumor independent test set (preliminary pilot-run estimates).

Class	Precision	Recall	F1-Score
Glioma	0.33	0.08	0.13
Meningioma	0.23	0.12	0.15
Pituitary	0.23	0.17	0.19
No Tumor	0.24	0.60	0.34

For the diagnostic use case, an external test accuracy below random guessing and an F1-score of 0.09 mean the model is not clinically deployable for brain tumor classification without domain adaptation, multi-institutional training data, and external validation.

The framework for detecting diabetic retinopathy is trained on retinal images of 2,750 DR cases. Nonetheless, it grapples with a class imbalance issue. The InceptionV3 architecture has been used here to classify the five severity levels of DR. Using this pre-trained InceptionV3 model, a validation accuracy of 97.15% was achieved. The model was fine-tuned using the Adam optimizer with a fixed learning rate of 10^{-4} , batch size 16, and local epochs of 2, as summarized in Table 1. The loss function was categorical cross-entropy, and the fine-tuning strategy replaced the final classification layer and trained the full network end-to-end. No independent test set was available for this task.

The accuracy of Tuberculosis detection with Montgomery and Shenzhen dataset is 52.17%. After five rounds, the accuracy improved from 49.07% to a maximum of 52.17% over five rounds. In Round 0, the recall was determined to be 1.0, whereas, the precision was calculated to be 0.4907. Both precision and recall metrics converged by Round 5, but class-level metrics (not reported here) show that the model performed poorly on the minority class, indicating a need for class-balanced training or specialized loss functions.

Table 7: Cross-Modality Performance Comparison

Task	Local Acc.	External Acc.	F1
Brain Tumor	91.70%	23.34%	0.09
Diabetic Retinopathy	97.15%	—	0.97
Tuberculosis	52.17%	52.17%	0.49

Note: Diabetic retinopathy results are based solely on client-local validation sets; no independent test set was available. Brain tumor results include an independent test set from a non-participating institution. Tuberculosis results use the official Shenzhen test split, but the dataset is imbalanced and small.

Key findings emerged from our evaluation. Transfer learning using InceptionV3 significantly outperformed custom CNNs, achieving 97% accuracy compared to 52-92% for custom architectures. The severe generalization gap in brain tumor classification reveals overfitting risks in federated learning with limited dataset diversity. Data volume and client heterogeneity critically impact performance, with tuberculosis results highlighting challenges with extreme data scarcity. These findings are consistent with

analyses on data heterogeneity challenges in healthcare federated learning.

All implementations successfully demonstrated collaborative training while maintaining data locality. The diabetic retinopathy system met clinical thresholds with validation accuracy exceeding 95%. However, the brain tumor model's test performance below random poses severe safety risks, emphasizing the necessity of independent multi-institutional validation. Tuberculosis results highlight challenges with limited data and client imbalance.

Performance analysis revealed several limitations of our current implementation. The evaluation employed only 2-3 clients versus real-world deployments typically involving 10-50+ institutions. Dataset sizes ranged from 639 to 2,750 images, limiting statistical power. Independent test sets were unavailable for DR and TB tasks. Advanced federated techniques including FedProx and differential privacy were not implemented.

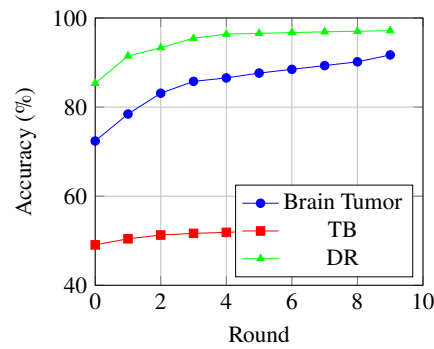


Figure 2: Validation accuracy progression across federated learning rounds for three medical imaging tasks. All curves were obtained using FedAvg. The y-axis represents the mean client-level validation accuracy aggregated by the coordinator at each round. The tuberculosis curve remains near chance level, reflecting the severe class imbalance and limited data of that task.

Table 8: Communication Overhead Analysis

Metric	Brain Tumor	DR	TB
Model Size	11.7M	22M	108M
Memory/Model	47 MB	88 MB	432 MB
Total Rounds	10	10	5
Total Comm.	2.82 GB	5.28 GB	8.64 GB

The communication totals include both uplink (client $\rightarrow$ coordinator) and downlink (coordinator $\rightarrow$ client) transfers of model parameters for all participating clients across all rounds. The global model is counted once per client per round. Using decimal gigabytes (1 GB = 1000 MB), the total communication is computed as

$$\text{Total Comm.} = \text{Memory/Model} \times \text{Clients} \times \text{Rounds} \times 2.$$

For brain tumor: $47 \times 3 \times 10 \times 2 = 2820 \text{ MB} = 2.82 \text{ GB}$. For diabetic retinopathy: $88 \times 3 \times 10 \times 2 = 5280 \text{ MB} = 5.28 \text{ GB}$. For tuberculosis: $432 \times 2 \times 5 \times 2 = 8640 \text{ MB} = 8.64 \text{ GB}$. Protocol and TLS encryption overhead are not included. The model sizes listed are based on the architectures in Table 1.

6. Discussion and Future Directions

Our evaluation reveals important insights for deploying federated learning in healthcare settings. The significant performance disparity between validation and independent test results in brain tumor classification highlights the critical importance of rigorous cross-institutional validation. Models trained through federated learning may exhibit overfitting to the training distribution, particularly when participating institutions have limited diversity in their patient populations. This finding aligns with broader concerns about generalization in medical AI and emphasizes the need for diverse, multi-institutional participation in federated learning initiatives [20].

The success of the diabetic retinopathy system, achieving 97.15% validation accuracy, demonstrates the viability of federated learning with transfer learning for well-characterized medical imaging tasks. InceptionV3's pretrained features provide robust representations that transfer effectively across institutions, reducing the impact of local data heterogeneity. However, the lack of independent test data for this task limits our ability to assess true generalization performance. Future work should incorporate multi-site validation with independent test sets.

The modest performance in tuberculosis detection, achieving only 52.17% accuracy, reflects the challenges of federated learning with limited data and severe client imbalance. The Montgomery and Shenzhen datasets differ substantially in acquisition protocols and patient populations, creating significant domain shift. This scenario represents a common challenge in healthcare settings where smaller institutions may have limited labeled data and specific patient demographics. Future research should explore techniques specifically designed for extreme client heterogeneity and data scarcity.

The PAWA algorithm showed promising results in adapting to varying client contributions, with preliminary pilot comparisons against FedAvg, FedOpt, and FedMA indicating modest improvements in validation accuracy and faster convergence. A comprehensive comparative evaluation across all aggregation strategies remains necessary. The dynamic weighting scheme effectively balances multiple quality metrics, reducing the influence of low-quality clients on the global model. However, the algorithm's sensitivity to hyperparameter configuration requires careful tuning for each deployment scenario. Future work should investigate automated hyperparameter optimization and adaptive temperature adjustment.

Several important directions emerge for future research. Incorporating differential privacy mechanisms would provide formal privacy guarantees, though careful balancing of privacy budgets against model utility is necessary. Extending the evaluation to larger client populations (10-50+ institutions) would better reflect real-world deployment scenarios. Integration with medical imaging foundation models could improve performance across all tasks, leveraging large-scale pretraining to address data scarcity. Researchers are actively investigating these aspects, including the use of deep generative models and multimodal approaches for robust healthcare analytics.

The platform's security architecture successfully maintains data locality while enabling collaborative model development. TLS 1.3 encryption and JWT-based authentication provide robust protection against external threats, while append-only audit logging ensures compliance verification [21]. However, additional privacy-preserving techniques such as differential privacy and secure aggregation could further strengthen privacy guarantees. Hardware-based security approaches like trusted execution environments offer potential for stronger protection with lower overhead than cryptographic methods. The integration of such security measures is crucial for real-world healthcare deployment.

The practical deployment considerations for healthcare institutions include infrastructure requirements, technical expertise availability, and integration with existing clinical workflows. Docker containerization simplifies deployment but may require institutional IT support for configuration and maintenance [22].

Monitoring dashboards accessible to non-technical staff facilitate participation in federated training without machine learning expertise, addressing a key barrier to adoption in clinical settings.

Future work should also incorporate automated testing and reliability analysis to ensure the robustness of the federated learning platform [23]. Additionally, ethical and legal considerations for personalized medicine and AI-driven healthcare must be addressed proactively [2, 24].

In conclusion, our platform demonstrates the feasibility of federated learning with data locality for collaborative healthcare diagnosis. While the diabetic retinopathy model achieved high validation accuracy and the tuberculosis model showed modest improvement, the brain tumor model exhibited severe generalization failure on an independent test set, and two of the three tasks lacked independent test sets. A direct comparison to centralized training was not performed. These limitations underscore the need for rigorous external validation before clinical deployment. Future work should focus on expanding client populations, incorporating formal privacy guarantees, and developing techniques for extreme client heterogeneity and data scarcity.

References

- [1] O. Panahi. The future of healthcare: AI. *Public Health and the Digital Revolution. MediClin Case Rep J*, 3(1):763–766, 2025.
- [2] P. Goktas and A. Grzybowski. Shaping the future of healthcare: Ethical clinical challenges and pathways to trustworthy AI. *Journal of Clinical Medicine*, 14(5), 2025. Art. no. 1605.
- [3] J. Naveen Ananda Kumar and S. Suresh. A proposal of smart hospital management using hybrid cloud, IoT, ML, and AI. In *Proc. 2019 Int. Conf. Communication and Electronics Systems (ICCES)*, pages 1082–1085, Coimbatore, India, 2019. doi: 10.1109/ICCES45898.2019.9002098.
- [4] M. Bendale. Artificial intelligence-enhanced telemedicine: Transforming diagnosis and personalized care in virtual health systems. Version V1, Zenodo, 2026.
- [5] D. Shenoy, R. Bhat, and K. Krishna Prakasha. Exploring privacy mechanisms and metrics in federated learning. *Artificial Intelligence Review*, 58(8), 2025. Art. no. 223.
- [6] K. K. Ghantasala. Privacy-preserving statistical learning for multi-center medical data: A fully decentralized approach. In *2025 Artificial Intelligence x Humanities, Education, and Art (AIxHEART)*, pages 53–57, Laguna Hills, CA, USA, 2025. doi: 10.1109/AIxHEART65685.2025.00016.
- [7] M. P. Uddin, Y. Xiang, M. Hasan, J. Bai, Y. Zhao, and L. Gao. A systematic literature review of robust federated learning: Issues, solutions, and future research directions. *ACM Computing Surveys*, 57(10):1–62, 2025.
- [8] S. Hardia. Decentralized intelligence for smart cities: Integrating federated learning, blockchain, and foundation models for privacy-preserving urban data management. In *2025 IEEE 16th International Symposium on Autonomous Decentralized Systems (ISADS)*, pages 87–93, Tucson, AZ, USA, 2025. doi: 10.1109/ISADS66912.2025.00017.
- [9] L. Albshaier, S. Almarri, and A. Albuali. Federated learning for cloud and edge security: A systematic review of challenges and AI opportunities. *Electronics*, 14(5), 2025. Art. no. 1019.

- [10] S. K. Bommakanti. Cloud-native distributed AI: Enabling secure collaborative learning on heterogeneous infrastructure. *Int. J. Intell. Syst. Data Sci.*, 1(1), April 2026. doi: 10.67231/ykzg0s10.
- [11] M. Mirzakarimova and M. Uzoqjonov. Information security in information-communication technologies. *Academic Journal of Science, Technology and Education*, 1(6):8–11, 2025.
- [12] E. Hashmi, M. M. Yamin, and S. Y. Yayilgan. Securing tomorrow: A comprehensive survey on the synergy of artificial intelligence and information security. *AI and Ethics*, 5(3):1911–1929, 2025.
- [13] J. Gupta. Differential protocol parsing exploits in distributed health information exchange networks: Language-theoretic security analysis of federated EHR interoperability infrastructure. *Int. J. Intell. Syst. Data Sci.*, 1(2), June 2026. doi: 10.67231/drve8d67.
- [14] A. Patel. Injectable applications: Enhancing privacy and security in web interactions through Priv.ly. In *2024 International Conference on Engineering and Emerging Technologies (ICEET)*, pages 1–11, Dubai, United Arab Emirates, 2024. doi: 10.1109/ICEET65156.2024.10913846.
- [15] E. Humphreys. *Implementing the ISO/IEC 27001 Information Security Management System Standard*. Artech House, 2025.
- [16] F. Hakeem Ghory. SAP S/4HANA-based continuous auditing and fraud detection in integrated accounting systems. *Int. J. Adapt. Manag. Bus. Intell.*, 1(1), June 2026. doi: 10.67231/hj2twh17.
- [17] P. Dhameliya. Client-tailored privacy hardening against training-data leakage in machine learning services. In *2026 Fourth International Conference on Augmented Intelligence and Sustainable Systems (ICAISS)*, pages 608–612, Trichy, India, 2026. doi: 10.1109/ICAISS68683.2026.11526698.
- [18] Y. Kostiuk, P. Skladannyi, Y. Samoilenko, K. Khorolska, B. Bebashko, and V. Sokolov. A system for assessing the interdependencies of information system agents in information security risk management using cognitive maps. In *Cyber Hygiene & Conflict Management in Global Information Networks 2024*, volume 3925, pages 249–264, 2025.
- [19] N. Saeed, M. Ashour, and M. Mashaly. Comprehensive review of federated learning challenges: A data preparation viewpoint. *Journal of Big Data*, 12(1), 2025. Art. no. 153.
- [20] S. V. Chinta, Z. Wang, A. Palikhe, X. Zhang, A. Kashif, M. A. Smith, et al. AI-driven healthcare: A review on ensuring fairness and mitigating bias. *PLOS Digital Health*, 4(5), 2025. Art. no. e0000864.
- [21] V. Kaliappan. Digital safeguards for health data: Architecting privacy and security in next-gen medical informatics. In *2025 International Conference on Big Data, Knowledge and Control Systems Engineering (BdKCSE)*, pages 1–7, Bankya, Bulgaria, 2025. doi: 10.1109/BdKCSE67969.2025.11300511.
- [22] H. N. Beesabathuni. Azure-SQL AutoSizer: Privacy-aware performance-cost SKU mapping for SQL migrations. In *Proceedings of the International Conference on Intelligent Systems for a Sustainable Future (ISSF 2026)*, pages 1043–1055, 2026. ISBN 978-94-6239-693-7. doi: 10.2991/978-94-6239-693-7_101.
- [23] R. Singh. Machine learning for automated flaky test diagnosis in JVMs. *Int. J. Intell. Syst. Data Sci.*, 1(4), August 2026. doi: 10.67231/m4bcqc44.

- [24] J. Joseph. Ethical and legal frontiers in personalized genomics: Balancing innovation, privacy, and regulatory compliance. figshare, Journal contribution, 2025.