

Auditable Federated Learning for Multimodal Brain Tumor Classification under Label Skew and Client Dropout

Thiruvenkadam T

Vellore Institute of Technology

Sivakumar N

Marwadi University

Emerson Raja

`emerson.raja@mmu.edu.my`

Multimedia University

Article

Keywords: Federated learning, brain tumour classification, multimodal medical imaging, FedProx, non-IID data, client dropout, weight auditing, trustworthy AI, healthcare

Posted Date: October 9th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10305486/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Auditable Federated Learning for Multimodal Brain Tumor Classification under Label Skew and Client Dropout

^{1,2}T. Thiruvankadam, ³N. Sivakumar, ^{4,*}J. Emerson Raja

¹Department of Information Security, School of Computer Science and Engineering, Vellore Institute of Technology, Vellore, India.

²Postdoctoral Research Fellow, Faculty of Engineering and Technology, Multimedia University, Jalan Ayer Keroh lama, 75450 Melaka, Malaysia.

mailone.thiru@gmail.com

³Department of Computer Engineering, Marwadi University, Rajkot, Gujarat.
drsivakumar.nadarajan@gmail.com

⁴Faculty of Engineering and Technology, Multimedia University, Jalan Ayer Keroh lama, 75450 Melaka, Malaysia.

emerson.raja@mmu.edu.my

Abstract: Federated learning (FL) allows hospitals to collaboratively train deep learning models without exchanging patient images, but reported FL studies in medical imaging often assume independent and identically distributed (IID) client data and full participation in every communication round both conditions seldom hold in real clinical networks. Furthermore, although the entire premise of FL rests on the iterative averaging of client weight updates, the optimisation dynamics of those weights are rarely audited at the round level, limiting clinicians' and regulators' ability to inspect how the global model evolves under stress. This study presents an auditable federated framework for binary brain-tumour classification from a combined multimodal dataset of 4618 axial computed-tomography (CT) and 5000 magnetic-resonance imaging (MRI) slices, totalling 9618 labelled images. A lightweight three-block convolutional neural network is trained across five simulated clients using FedProx with proximal coefficient $\mu = 0.01$ and FedAvg aggregation, under two contrasting client partitions: an IID scenario in which each client receives a balanced sample of healthy and tumour cases, and a deliberately aggressive non-IID scenario in which each client is assigned an 80 % majority class. Clinically realistic stress is added by dropping a randomly chosen client in approximately 30 % of rounds. At every round, for every active client, the framework records three weight summaries initial (received from the server), updated (after local training), and aggregated (new global model) together with per-client accuracy, loss, precision, recall, F1-score, communication payload, wall-clock round time, and process memory. After ten communication rounds, the global model attains 96.05 % accuracy and an F1-score of 0.9641 on the held-out test set under the IID partition, and 94.96 % accuracy with F1 0.9547 under the 80 % label-skew non-IID partition, corresponding to an IID-to-non-IID degradation of only 1.09 percentage points despite the severe class skew and 30 % per-round client dropout. Inspection of the weight-audit trace reveals a consistent and measurable shift in the output-layer mean under non-IID training that is invisible in conventional accuracy curves, suggesting that round-level weight auditing can serve as a lightweight forensic signal of distributional drift in federated medical AI.

Keywords: Federated learning, brain tumour classification, multimodal medical imaging, FedProx, non-IID data, client dropout, weight auditing, trustworthy AI, healthcare.

1. INTRODUCTION

Brain tumours represent one of the most critical neurological disorders and a major cause of death from cancer worldwide [1,2]. Precise and timely diagnosis is critical for treatment planning, prognosis estimation and improved patient outcome [3]. The diagnosis of brain tumours is predominantly conducted using magnetic resonance imaging (MRI) because of its exceptional soft-tissue contrast. However, computed tomography (CT) has the advantage of rapid imaging and is widely used in clinical emergency settings [4,5]. Current developments in deep learning have enhanced the automated detection and classification of

tumours in medical images. Recently, CNNs have demonstrated excellent performance in feature extraction and classification tasks of brain MRI and CT data. [6,7]. Furthermore, multimodal imaging techniques that utilise both MRI and CT information are known to be more robust and diagnostically reliable than single modality approaches [8].

Centralised deep learning approaches can achieve high predictive performance but necessitate large-scale datasets collected from multiple institutions. Privacy regulations, ethical concerns and institutional governance policies [9] often limit the sharing of medical images across hospitals. Federated Learning (FL) is a new distributed machine learning framework that allows collaborative training of models while maintaining the privacy of raw patient data [10]. The Federated Averaging (FedAvg) algorithm laid out the basis of decentralised learning and has been widely applied in healthcare applications [11]. In practice, however, data heterogeneity is often encountered, since hospitals differ in patient demographics, imaging protocols, disease prevalence and scanner configurations [12]. Non-IID data distribution might be a challenge to model convergence and prediction performance. [13].

Many federated studies also assume that clients fully participate in every round of communication. In practical healthcare settings, hospitals may face network disruptions, maintenance windows, computation constraints, or policy restrictions that result in sporadic client participation [14]. Also, the majority of the research largely focus on measures like as accuracy and loss, and provide little insight into the change of model parameters over training. Therefore, recent works have highlighted the need for transparency, trustworthiness and auditability of healthcare AI systems [15].

In this study, we present a federated framework for binary brain-tumour classification from a mixed multimodal dataset of CT and MRI images that simultaneously solves these three gaps. The framework uses a purposely lightweight 3-block convolutional neural network trained with FedProx to reduce client drift, and simulates ~ 30 % client dropout per round, and compares a well-balanced IID partition with a purposely aggressive 80 % label-skew non-IID partition. We make three key contributions: we report three explicit weight summaries for each active client at each round: initial, updated and aggregated weight summaries, along with conventional performance metrics, communication and memory cost, and round time. The resulting trace can be examined after training to confirm that updates from each institution were within expected statistical bounds and that the aggregated global model evolved smoothly.

1.1 Contributions

The major contributions of this study are summarised below: (i) We provide an auditable FL protocol that tracks the initial, updated and aggregated weight statistics such as mean, standard deviation, minimum and maximum for each active client in each round, in addition to the usual per-round performance metrics. The trace is light-weight, does not require any additional model parameters and is fully compatible with privacy-preserving aggregation. (ii) We test the framework on a real multimodal brain-tumour dataset of 9618 CT and MRI images in two contrasting partitions, IID and 80% label-skew non-IID, and under about 30% per-round client dropout, closer to clinical deployment than IID full-participation studies. (iii) We present a full set of deployment-aware metrics, such as round time, communication payload and process memory, in addition to usual performance metrics, enabling the framework to be assessed on both predictive quality and practical cost. (iv) We empirically demonstrate that the weight-audit trace shows a clear and consistent bias signature in the output layer during non-IID training, which is hidden by the standard

accuracy curves, proposing round-level weight inspection as a potential low-cost integrity check for federated medical AI.

1.2 Paper Organisation

The rest of the paper is structured as follows. Section 2 reviews FL foundations, FL applications to medical imaging, brain-tumour detection from CT and MRI, and transparency-oriented FL. Section 3 describes the dataset, preprocessing pipeline, model architecture, federated optimisation, client partitioning, dropout protocol, weight-auditing logging, and evaluation. Section 4 summarises the experimental setup. Section 5 presents results, including per-round convergence curves for both scenarios, per-client metrics, the weight-audit trace, dropout effects, and deployment indicators. Section 6 provides comparative analysis with prior reported results on similar datasets and federated medical-imaging studies. Section 7 discusses the findings, limitations, and outlines future work. The conclusion is given in section 8.

2. RELATED WORK

2.1 Federated Learning Foundations

Federated averaging (FedAvg) introduced the canonical FL training loop, which is still the baseline aggregator for most subsequent work [16]. The non-IID problem, statistical heterogeneity, was already identified early as a main obstacle to convergence. This gave rise to a series of optimisers such as FedProx that offers convergence guarantees in the presence of partial participation and bounded heterogeneity [17]. SCAFFOLD then addressed client drift through control variates and demonstrated better convergence in heterogeneous data distributions [18].

Subsequent work examined personalised federated learning, optimisation with limited communication and robust aggregation schemes that can cope with adversarial or unreliable clients. The impact of non-IID data on the federated convergency was shown [19], communication efficiency and personalisation were highlighted as important research directions in comprehensive surveys [20]. Aggregation methods robust to Byzantine attacks such as Krum and Byzantine-robust distributed learning have increased reliability in the presence of adversaries [21,22]. Further studies have uncovered hidden vulnerabilities of distributed optimisation and have motivated the development of more robust aggregation frameworks [23]. Later security-focused investigations demonstrated the feasibility of backdoor attacks and model-poisoning attacks against federated systems [24,25]. More recent solutions such as FLTrust and FLDetector added trust-aware aggregation and malicious-client detection to improve robustness [26,27].

2.2 Federated Learning in Medical Imaging

Federated learning is an effective method for various healthcare imaging tasks, including classification of chest X-ray, detection of retinal diseases, COVID-19 diagnosis using radiographs and CT images, analysis of histopathological images, and prediction of neurological diseases [28,29]. Large-scale multi-institutional studies have confirmed that federated models are better than centralised learning while ensuring local data ownership and regulatory compliance [30].

Federated learning is a concept identified by as a promising model for collaborative AI for healthcare and challenges related to data diversity, efficiency in communication, protection of privacy, and deployment

scalability have been identified [31,32]. However, many existing studies focus on predictive performance with limited analysis of client heterogeneity, intermittent participation and optimisation dynamics.

2.3 Detection of Brain Tumour Using CT and MRI

Deep convolutional neural networks are being commonly utilised for classification of brain tumours from MRI images, and the reported accuracies have often been more than 95% on benchmark datasets [33]. Transfer-learning methods using VGG, ResNet, DenseNet, and EfficientNet architectures [34,35] further improve classification performance and generalisation capability. CT-based deep learning models have also demonstrated promising results in emergency and rapid screening settings, although MRI is still the main method for diagnosing brain tumours [36]. Multimodal CT+MRI approaches are still relatively scarce, but have shown promising results by using complementary anatomical and structural information from both modalities [37]. Federated variants of multimodal brain tumour classification are still relatively scarce, where most of the existing studies focus on privacy preservation and predictive accuracy, rather than robustness under client dropout and severe non-IID conditions [38].

2.4 Federated Learning Transparency and Auditability

In the federated learning, auditability has been traditionally limited to privacy guarantees, secure aggregation, differential privacy and blockchain-assisted logging of update transactions [39]. Recent studies have advocated for more transparent federated evaluation practices including reproducibility of experimentation, heterogeneity stress testing, and comprehensive reporting of optimisation dynamics [40]. The round-level inspection of weight statistics as a mechanism of identifying distributional drift and potential bias is largely unexplored in the literature. Most of the existing works only provide overall standard performance metrics and do not give much insight about the progression of global model parameters. This study fills this gap by proposing a lightweight weight-auditing protocol that can track optimisation behaviour under realistic non-IID and client-dropout conditions. Below table 1 presents the comparison of FL techniques for brain tumour analysis and medical image applications.

Table 1. Comparison of FL techniques for brain tumour analysis and medical image applications

Ref.	Methodology Used	Modality	Non-IID Handling	Client Dropout Handling	Privacy Preservation	Weight Auditing	Key Limitation / Research Gap
[41]	FedAvg-based Federated Learning for Multi-Institutional Medical Imaging	Brain MRI	Not Addressed	No	FL-based privacy through decentralized training	No	Did not consider model interpretability, client failures, or statistical heterogeneity.
[42]	Federated U-Net Brain Tumor Segmentation	Brain MRI	Limited	No	Privacy-preserving FL architecture	No	Segmentation only. No heterogeneity analysis or explainability mechanism.
[43]	Siloed Federated Learning for	Histopathology Images	Partially Addressed	No	Distributed model training	No	Limited evaluation on medical image heterogeneity, and

	Histopathology						lack of privacy enhancement techniques beyond FL.
[44]	FedDyn (Dynamic Regularization) Algorithm	Generic FL Datasets Fully Addressed	Generic FL Datasets Fully Addressed	Partial	No explicit privacy mechanism	No	It balances local and global objectives with a dynamic penalty term, but does not consider explainability or medical imaging specific needs.
[45]	Federated Learning with Differential Privacy	Medical Images	Partial	No	Differential Privacy	No	Privacy improved. Model utility degraded. Explainability decreased.
[46]	FedMix Mixed-Supervised Federated Learning	Medical Image Segmentation	Fully Addressed	No	Federated Learning	No	Less focus on segmentation and no interpretability and dropout robustness
[47]	Self-Supervised Federated Learning	Medical Imaging	Fully Addressed	No	Federated Learning	No	Does not include issues of explainability or client reliability.
[48]	Explainable Federated Learning Framework	Brain MRI	Not Addressed	No	Federated Learning	No	Explainability has improved, but heterogeneity and communication reliability are not addressed.
[49]	Transfer Learning + Federated Learning	Brain MRI	Not Addressed	No	Federated Learning	No	Does not handle Non-IID distributions, interpretability, client failures.
[50]	SelfFed Self-Supervised Federated Learning	Medical Images	Fully Addressed	No	Federated Learning	No	Focus on label scarcity and heterogeneity. No XAI or auditing mechanisms.
[51]	Federated Learning with ResNet-50, Capsule Networks, HGBOA Feature Selection, and Blockchain-based Secure Aggregation	Brain CT	Partially Addressed	No	FL-based decentralized training with blockchain	Yes	No explicit evaluation under severe Non-IID settings, no client-dropout resilience analysis, no communication-efficiency study, and limited explainability/interpretability assessment.

[52]	A Federated Multimodal Framework with Layer Skipping, FedPropSAG Prototype Aggregation, and NSGA-II Multi-Objective Optimisation	Brain MRI (multimodal architecture, empirically tested on MRI only)	Brain MRI Dirichlet-based non-IID partitioning, $\alpha = 0.5$	Not Addressed (identified as future work)	Hybrid Differential Privacy + Secure Aggregation	No	Validated on a single artificially partitioned public dataset with only 5 simulated clients No real multi-institutional validation, scalability testing or robustness to client dropout.
[53]	Causal Explainable FL with Blockchain	Medical Imaging	Partial	No	Blockchain + FL	Partial	Provides trust and explainability but does not fix client dropout resilience.
[54]	Federated Transformer-GNN	Brain MRI (Multimodal)	Not Addressed	No	Privacy-Preserving FL	No	Limited evaluation on heterogeneous clients and real-world deployment scenarios

3. METHODOLOGY

The proposed auditable federated learning approach for brain tumour classification is illustrated in Fig. 1. The framework comprises of four stages: data preparation, partitioning of client data, federated model training, round-wise evaluation and logging of audit. The training data are shared across multiple customers under both the IID and non-IID settings after multimodal data fusion and preprocessing to mimic the various levels of data heterogeneity commonly seen in remote healthcare settings.

In each round of communication, the server sends a broadcast of the global model to the participating clients for local optimisation. Next, the updated client models are combined to produce a new global model. Intermittent client dropout is introduced to simulate realistic participation variability. In parallel, the framework logs model statistics before local training, after local updates and after global aggregation, providing a transparent way to monitor model progress during the federated learning process. The final global model is evaluated at the end of each round using predictive and deployment-oriented metrics, which provides a comprehensive assessment of classification performance and practical deployment feasibility.

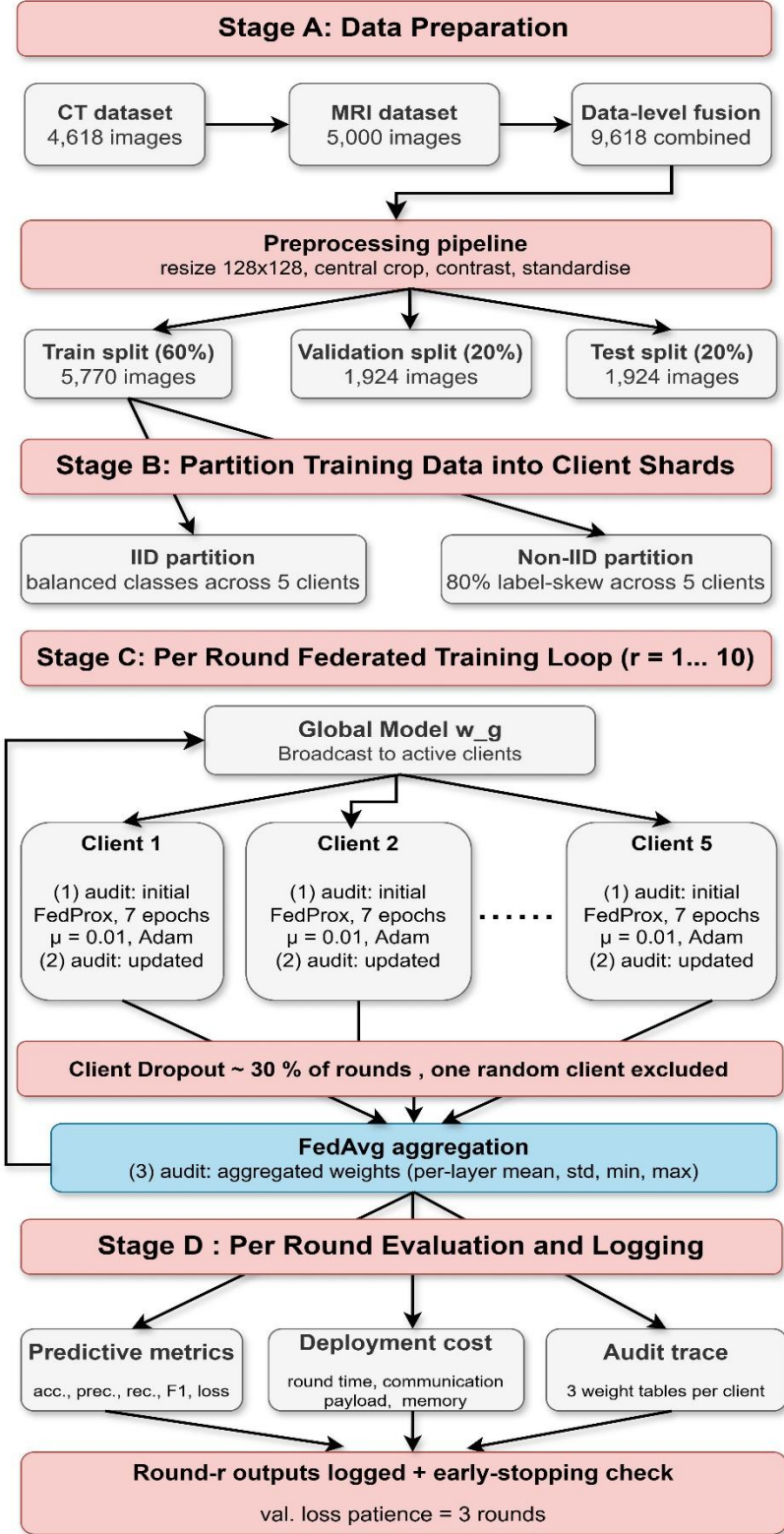


Fig. 1 Workflow of the Proposed Auditable Federated Learning Framework for Brain Tumour Classification

3.1 Dataset

Experiments use a publicly available combined brain-tumour dataset comprising 4618 axial CT slices and 5000 MRI slices, distributed across two classes: 4300 healthy and 5318 tumour cases, for a total of 9618 labelled images [56]. Each image is scaled to 128×128 pixels using three coloured channels, and normalised with respect to the unit interval. CT and MRI samples are merged into one pool, shuffled with a predefined randomly selected seed, and divided into 5770 training, 1924 validation, and 1924 test images following a stratified 60/20/20 partition. Table 2 shows the composition of the combined class.

Table 2. Dataset composition

Subset	Healthy	Tumour	Total
CT	2300	2318	4618
MRI	2000	3000	5000
Combined	4300	5318	9618
Train (60 %)	—	—	5770
Validation (20 %)	—	—	1924
Test (20 %)	—	—	1924

3.2 Preprocessing

Each image undergoes a deterministic preprocessing pipeline implemented in TensorFlow. The pipeline consists of: (a) resizing to 128×128 , (b) central cropping to 95 % of image area followed by re-resizing to 128×128 to focus attention on the cranial region while suppressing peripheral artefacts and borders, (c) contrast adjustment with a factor of 1.2, (d) per-image standardisation (zero mean, unit variance), (e) clipping to the range ± 3.0 standard deviations to suppress extreme intensities, and (f) rescaling back to the unit interval. The same pipeline is applied to CT and MRI samples without modality-specific tuning, supporting reproducibility and avoiding modality-specific tuning that would complicate federated deployment.

3.3 Model Architecture

Each client uses an identical compact convolutional neural network. The framework has three convolutional blocks with 32, 64 and 128 filters corresponding to each other, each with a 2×2 max-pooling. The extracted features are flattened and then fed to a dense layer with 64 units, with ReLU activation followed by a single output unit with activation of the sigmoid for binary classification. Table 3 shows the number of parameters and output shape for each layer.

Table 3. Compact CNN architecture used at every client

Layer	Output shape	Trainable params
Input ($128 \times 128 \times 3$)	$128 \times 128 \times 3$	0
Conv2D-1 (32 filters, 3×3 , ReLU)	$126 \times 126 \times 32$	896
MaxPool 2×2	$63 \times 63 \times 32$	0
Conv2D-2 (64 filters, 3×3 , ReLU)	$61 \times 61 \times 64$	18496
MaxPool 2×2	$30 \times 30 \times 64$	0

Layer	Output shape	Trainable params
Conv2D-3 (128 filters, 3×3 , ReLU)	$28 \times 28 \times 128$	73856
MaxPool 2×2	$14 \times 14 \times 128$	0
Flatten	25088	0
Dense (64, ReLU)	64	1605696
Dense (1, Sigmoid)	1	65

The choice of a deliberately compact backbone reflects the deployment-oriented objective of this work: lightweight clients reduce per-round computation, memory footprint, and communication payload, which is critical in resource-constrained hospital settings. A systematic accuracy-versus-efficiency comparison against deeper backbones such as EfficientNet, DenseNet121, ResNet50, and Vision Transformers within the federated setting is identified as future work.

3.4 Federated Optimisation with FedProx

FedProx Local Objective Function

Local training is conducted at each client using the FedProx optimisation algorithm . FedProx generalises the binary cross-entropy loss function with an extra proximal regularisation parameter in order to restrict the divergence between local and global model parameters. The local objective function for the client i can be expressed as::

$$L_i(w_i) = BCE(y, \hat{y}(w_i)) + \frac{\mu}{2} \|w_i - w_g\|^2 \quad (1)$$

where $L_i(w_i)$ is the local loss function, $BCE(\cdot)$ is the binary cross-entropy loss, w_i and w_g are the local and global model parameters respectively, and μ is the proximal coefficient. In this paper, μ is set to 0.01. The proximal term regularises the local updates to avoid variances from the global model, which helps to improve training stability under non-IID data distributions and minimal client engagement.

FedAvg Global Aggregation

After local training, the updated client models are transferred to the central server and combined using the FedAvg algorithm. at communication round t the update of the global model is given as:

$$w_g^{t+1} = \frac{1}{|S_t|} \sum_{i \in S_t} w_i^t \quad (2)$$

where S_t is the set of active clients participating in round t , $|S_t|$ is the number of active clients, and w_i^t corresponds to the locally updated model parameters received from client i . The aggregated model w_g^{t+1} is then sent to all participating clients for a subsequent phase of communication.

3.5 Client Partitioning

Two contrasting client partitions are evaluated. In the IID partition, the 5770 training samples are distributed round-robin across five clients, yielding approximately balanced label distributions per client. In the non-IID partition, an 80 % label-skew protocol assigns each client a majority class: clients 1 and 3 receive 923 healthy and 361 tumour samples (healthy-majority); clients 2 and 4 receive 230 healthy and 1054 tumour samples (tumour-majority); client 5 receives the remainder (274 healthy and 360 tumour samples). The exact partitions are summarised in Table 4.

Table 4. Client label distribution for the two scenarios

Client	IID Healthy	IID Tumour	Non-IID Healthy	Non-IID Tumour
1	516	638	923	361
2	514	640	230	1054
3	534	620	923	361
4	514	640	230	1054
5	502	652	274	360
Total	2580	3190	2580	3190

The 80 % skew is intentionally aggressive: it is more severe than the moderate Dirichlet skews typically used to benchmark FL optimisers, and is intended as a stress test of the proximal regularisation.

3.6 Client Dropout

To emulate intermittent client availability, approximately 30 % of communication rounds are designated as dropout rounds. In each dropout round, one client is selected uniformly at random and excluded from training and aggregation for that round only. Specifically, in ten rounds, three dropout rounds are sampled at the start of the run. The remaining clients train and submit updates normally; aggregation then averages the surviving client weights. For the IID scenario, rounds 3, 9, and 10 are dropout rounds; for the non-IID scenario, rounds 2, 4, and 6 are dropout rounds. The specific clients dropped in those rounds are recorded in the experimental log.

3.7 Auditable Weight Logging

The core of the proposed framework is the explicit round-level weight auditing protocol. In each communication round r and for each active client i , three weight summaries are recorded:

- Initial weights (w_g^r): Parameters of the global model provided to client i at the start of round r .
- Updated weights (w_i^r): The local model parameters of client i after FedProx training for seven local epochs during round r .
- Aggregated weights (w_g^{r+1}): The server generates the new global model parameters after the FedAvg aggregation of all the active client models in the round r .

Each summary is reported as a per-layer table containing the tensor shape, mean, standard deviation, minimum, and maximum of the layer’s weight values. For the architecture used here, this yields ten layer-rows per summary and three summaries per (round, client) pair. Across both scenarios, the framework therefore produced approximately 300 weight-audit tables (5 clients $\times$ 10 rounds $\times$ 3 stages, less the dropped clients’ entries). One representative example is reproduced in Section 5; full traces are released alongside the source code.

Two properties make the protocol practical for deployment. First, it requires no additional model parameters and adds negligible computation: a small set of summary statistics per layer is computed in milliseconds. Second, it is compatible with privacy-preserving aggregation: only summary statistics of the locally updated weights are logged for audit, not individual gradients or examples. In a deployment scenario where regulators or clinicians must justify a model’s behaviour, this trace provides a lightweight, low-cost

forensic signal indicating that updates from each institution remained within expected statistical bounds and that the aggregated global model evolved smoothly.

3.8 Evaluation Protocol

The global model aggregated after each communication round is assessed on the remaining validation and test partitions. The reported predictive metrics are accuracy, binary cross-entropy loss, precision, recall and F1-score, all measured at a predetermined decision-making threshold value of 0.5. In addition, three deployment-aware indicators are recorded per round: (i) round wall-clock time, in seconds, from the start of local training to the end of aggregation; (ii) total communication payload, computed as the product of the number of float32 weight parameters and 4 bytes, summed over active clients; and (iii) the resident-set memory footprint of the training process, measured via the operating system. Early stopping was activated if validation losses did not get better for three rounds. In the current runs patience was not met and both scenarios ran for the full 10 rounds.

3.9 Algorithm

Input: clients $\{1, \dots, N\}$, dataset partition $\{D_i\}$, global rounds R , local epochs E , learning rate η , proximal coefficient μ , dropout rounds R_d

Output: final global model $\mathbf{w}_g^R$, full audit trace, per-round metric log

Initialization: set global model weights $\mathbf{w}_g^0$

for $r = 1$ to R do

 if $r \in R_d$ then

 drop one random client c from active set S_r

 else

$S_r \leftarrow \{1, \dots, N\}$

 end if

 for each client $i \in S_r$ in parallel do

 Record initial weights table for client i (per-layer summary of $\mathbf{w}_g^{r-1}$)

$\mathbf{w}_i \leftarrow \mathbf{w}_g^{r-1}$

 for epoch $e = 1$ to E do

 Update $\mathbf{w}_i$ using Adam on

$$L_i(\mathbf{w}_i) = BCE(y, \hat{y}) + \frac{\mu}{2} \|\mathbf{w}_i - \mathbf{w}_g^{r-1}\|^2$$

 end for

 Record updated weights table for client i (per-layer summary of $\mathbf{w}_i$)

 Return $\mathbf{w}_i$ to server

 end for

 Aggregate client models

$$\mathbf{w}_g^r = \frac{1}{|S_r|} \sum_{i \in S_r} \mathbf{w}_i$$

Record aggregated weights table for round r (per-layer summary of $\mathbf{w}_g^r$)

Evaluate $\mathbf{w}_g^r$ on validation and test partitions

if no validation improvement for P rounds, break

break

end if

end for

Return final global model $\mathbf{w}_g^R$, full audit trace, and per-round metric log

4. EXPERIMENTAL SETUP

Experimental activities are conducted through a single Google Colab session using GPU accelerator. The implementation uses TensorFlow with the Keras API. Each of the two scenarios (IID and 80 % label-skew non-IID) is trained for 10 communication rounds with 5 clients, 7 local epochs per round, batch size 32, Adam optimiser, and FedProx with $\mu = 0.01$. Within each scenario, three rounds are randomly selected as dropout rounds: rounds 3, 9, and 10 for the IID scenario, and rounds 2, 4, and 6 for the non-IID scenario. Random seeds are fixed where supported. Table 5 is presented with the complete set-up.

Table 5. Experimental configuration

Parameter	Value
Total labelled images	9618 (4618 CT + 5000 MRI)
Class balance	4300 healthy / 5318 tumour
Image size	128 × 128 RGB
Train / Validation / Test	5770 / 1924 / 1924
Number of clients	5
Communication rounds	10
Local epochs per round	7
Batch size	32
Local optimiser	Adam, FedProx $\mu = 0.01$
Aggregation	FedAvg (uniform mean)
Client dropout rate	~30 % of rounds, one client dropped
IID dropout rounds	3, 9, 10
Non-IID dropout rounds	2, 4, 6
Loss	Binary cross-entropy + proximal term
Decision threshold	0.5
Early-stopping patience	3 rounds (not triggered)
Implementation	TensorFlow / Keras, Google Colab GPU

5. RESULTS

5.1 Client Label Distribution

Figure 2 visualises the resulting client label distributions for the two scenarios. The IID partition yields approximately balanced classes per client, with each client holding between 502 and 534 healthy samples and between 620 and 652 tumour samples. The non-IID partition yields strongly skewed clients: clients 2 and 4 hold approximately $4.6 \times$ more tumour than healthy samples, while clients 1 and 3 hold approximately $2.6 \times$ more healthy than tumour samples. The asymmetry between healthy-majority and tumour-majority clients reflects the underlying class imbalance of the combined dataset.

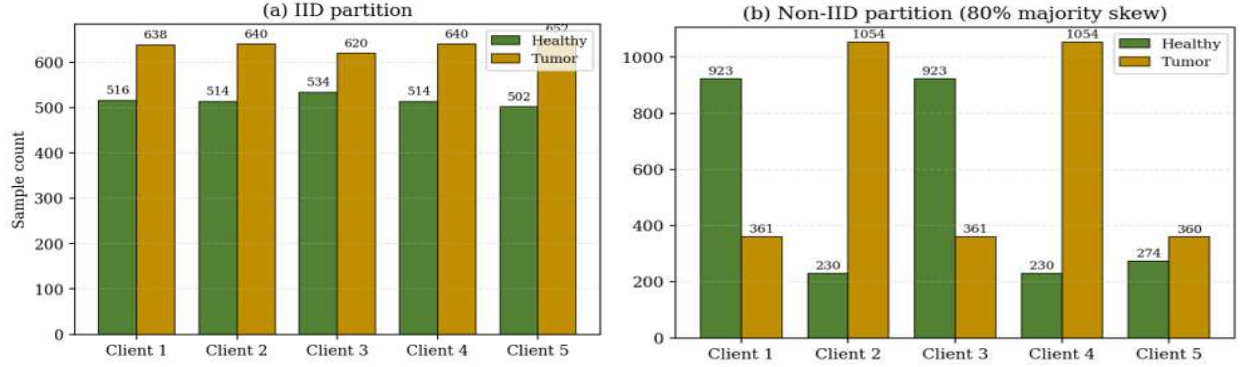


Fig. 2 Client label distributions. (a) IID partition with approximately balanced classes per client. (b) Non-IID partition with 80 % majority-class skew per client.

5.2 Weight-Audit Trace (Illustrative Example)

Table 6 reproduces a representative weight-audit table for client 1 in round 1 of the IID scenario, focusing on the output layer (a 64×1 dense layer). The three rows correspond to the three logged stages. The full trace contains ten such layer-rows per stage per client per round; only the output layer is shown here for space, but the same statistics are recorded for every convolutional and dense layer. As expected, local training perturbs both the mean and the spread of the layer’s weights, while FedAvg aggregation produces a smoothed estimate that lies between the initial and updated values.

Table 6 Representative weight audit — Client 1, Round 1 (IID), output layer (Dense 64×1)

Stage	Mean	Std	Min	Max
Initial (server)	0.009488	0.166480	−0.303559	0.301457
Updated (post-local)	0.007966	0.173970	−0.336037	0.330076
Aggregated (FedAvg)	0.009725	0.173076	−0.317303	0.329405

5.3 Per-Round Convergence

Figure 3 reports the per-round convergence of validation accuracy, validation loss, test accuracy, and test F1-score for both scenarios. In the IID setting, validation accuracy improves monotonically from 0.8649 at round 1 to 0.9621 at round 10; test accuracy improves correspondingly from 0.8602 to 0.9605, with test F1-score reaching 0.9641. In the non-IID setting, validation accuracy increases from 0.8311 at round 1 to 0.9491 at round 10; test accuracy improves from 0.8259 to 0.9496, with test F1-score reaching 0.9547. The complete per-round metric trace for the global model is reproduced in tables 7 and 8 and final test-set performance after 10 communication rounds is presented in table 9.

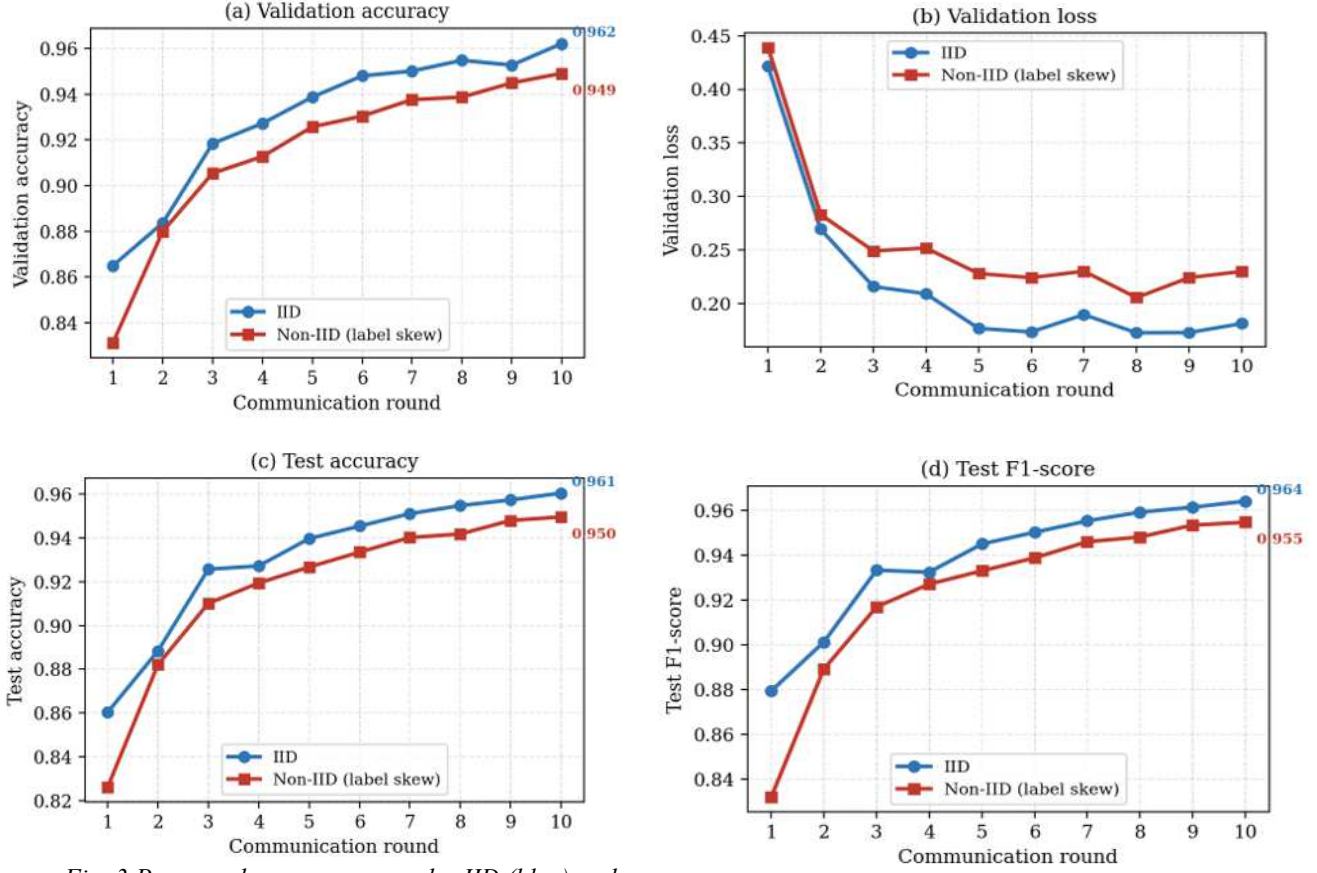


Fig. 3 Per-round convergence under IID (blue) and non-IID (red). Dotted vertical lines mark dropout rounds. (a) Validation accuracy. (b) Validation loss. (c) Test accuracy. (d) Test F1-score.

A central observation is that the gap between IID and non-IID narrows steadily across rounds: the test-accuracy gap of 3.43 percentage points at round 1 contracts to only 1.09 percentage points at round 10. This indicates that the combination of FedProx proximal regularisation and class-balanced majority dynamics across multiple clients absorbs much of the per-client class skew over time.

Table 7 Per-round global metrics — IID scenario

Rd	Val Acc	Val Loss	Test Acc	Test Loss	Test F1	Active
1	0.8649	0.4214	0.8602	0.4192	0.8793	5
2	0.8836	0.2695	0.8883	0.2680	0.9012	5
3	0.9184	0.2160	0.9257	0.2129	0.9333	4*
4	0.9272	0.2093	0.9272	0.2053	0.9324	5
5	0.9387	0.1770	0.9397	0.1850	0.9450	5
6	0.9480	0.1737	0.9454	0.1756	0.9502	5
7	0.9501	0.1898	0.9511	0.1940	0.9553	5
8	0.9548	0.1729	0.9548	0.1730	0.9592	5
9	0.9527	0.1731	0.9574	0.1469	0.9614	4*
10	0.9621	0.1815	0.9605	0.1638	0.9641	4*

Asterisks mark dropout rounds in which one client was excluded.

Table 8 Per-round global metrics — Non-IID scenario

Rd	Val Acc	Val Loss	Test Acc	Test Loss	Test F1	Active
1	0.8311	0.4391	0.8259	0.4386	0.8319	5
2	0.8799	0.2831	0.8820	0.2855	0.8891	4*
3	0.9054	0.2493	0.9101	0.2544	0.9169	5
4	0.9127	0.2519	0.9194	0.2444	0.9271	4*
5	0.9257	0.2281	0.9267	0.2179	0.9330	5
6	0.9304	0.2243	0.9335	0.2280	0.9388	4*
7	0.9376	0.2303	0.9402	0.2372	0.9460	5
8	0.9387	0.2057	0.9418	0.2056	0.9481	5
9	0.9449	0.2243	0.9480	0.2312	0.9534	5
10	0.9491	0.2301	0.9496	0.2331	0.9547	5

Table 9 Final test-set performance after 10 communication rounds

Scenario	Accuracy	Precision	Recall	F1
IID (balanced)	0.9605	0.9687	0.9596	0.9641
Non-IID (80 % skew)	0.9496	0.9489	0.9605	0.9547
Absolute difference	+0.0109	+0.0198	−0.0009	+0.0094

5.4 Per-Client Behaviour

Figure 4 plots the per-client local-training F1-score across rounds for both scenarios. Under the IID partition, all five clients converge to F1 values above 0.99 by round 7, indicating that each client’s local model fits its balanced shard well. Under the non-IID partition, the trajectory is more revealing: the two healthy-majority clients (clients 1 and 3) start at substantially lower F1 (0.7163 and 0.7987 respectively in round 1) than the tumour-majority clients (0.9508 and 0.9516), a direct consequence of the class-balance asymmetry in the combined dataset. By round 4, all clients have converged to F1 above 0.96. Table 10 reports the round-10 per-client local performance for completeness.

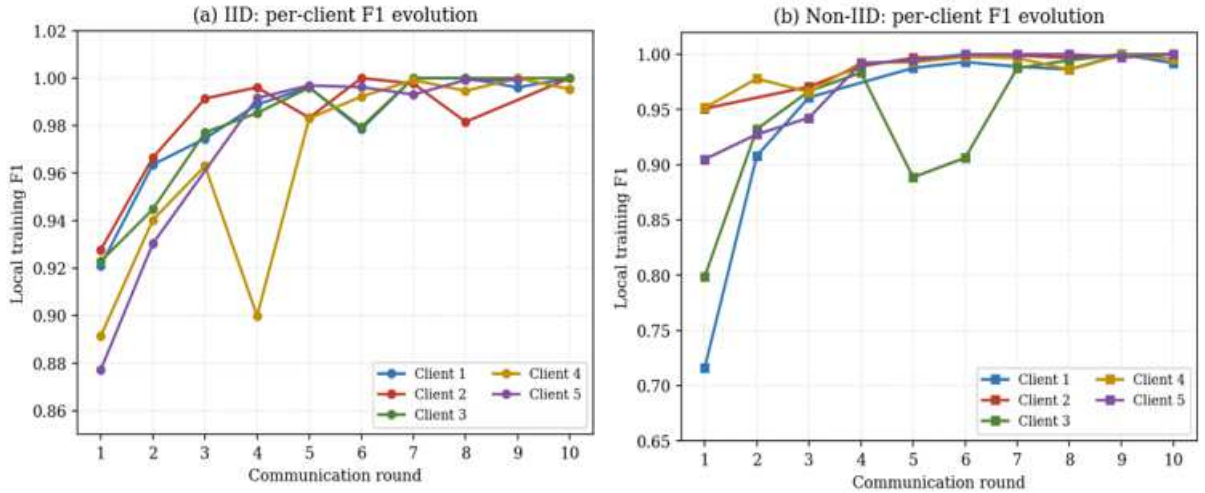


Fig. 4. Per-client local-training F1-score across rounds. (a) IID: all clients converge quickly and tightly. (b) Non-IID: healthy-majority clients (1, 3) start substantially lower, reflecting the asymmetric class-balance of the underlying dataset, before converging by round 4.

Table 10. Per-client local-training metrics at round 10

Client	Scen.	Acc.	Prec.	Recall	F1
1	IID	1.0000	1.0000	1.0000	1.0000
2	IID	1.0000	1.0000	1.0000	1.0000
3	IID	1.0000	1.0000	1.0000	1.0000
4	IID	0.9948	0.9922	0.9984	0.9953
5	IID	dropped	—	—	—
1	Non-IID	0.9953	0.9837	1.0000	0.9918
2	Non-IID	1.0000	1.0000	1.0000	1.0000
3	Non-IID	1.0000	1.0000	1.0000	1.0000
4	Non-IID	0.9930	0.9943	0.9972	0.9957
5	Non-IID	1.0000	1.0000	1.0000	1.0000

5.5 Effect of Client Dropout

In both cases, dropout rounds did not lead to an obvious degradation in the combined global model. In the IID case, round 3 for client 5 is dropped produced the single largest one-round jump in test accuracy of the entire run, from 0.8883 to 0.9257; rounds 9 and 10 (each with one client dropped) continued to improve or maintain accuracy. In the non-IID scenario, dropout rounds 2, 4, and 6 likewise showed continued monotonic improvement. This empirical robustness against approximately 30 % per-round dropout supports the practical viability of the framework under intermittent client availability, though a more systematic dropout sensitivity study varying the dropout fraction and the dropped client’s identity is identified as future work.

5.6 Weight-Audit Diagnostic Signal

Figure 5 plots the aggregated output-layer weights of Layer 8 with a dense of 64×1 by showing the mean and standard deviation across rounds for both scenarios. Two observations emerge. First, the output-layer mean grows monotonically throughout training for both scenarios, but is consistently and substantially higher under non-IID training: the final round-10 mean is 0.0418 for non-IID compared to 0.0205 for IID, a factor of approximately twofold. This persistent shift is a direct trace of the label-skew partition: clients training on majority-tumour data push the decision-layer bias in one direction, and although FedAvg eventually balances the model, the residual bias remains detectable in the aggregated weight signature. Second, the output-layer standard deviation grows monotonically in both scenarios and is slightly lower under non-IID than under IID, consistent with the proximal term constraining the spread of weight values under high client heterogeneity.

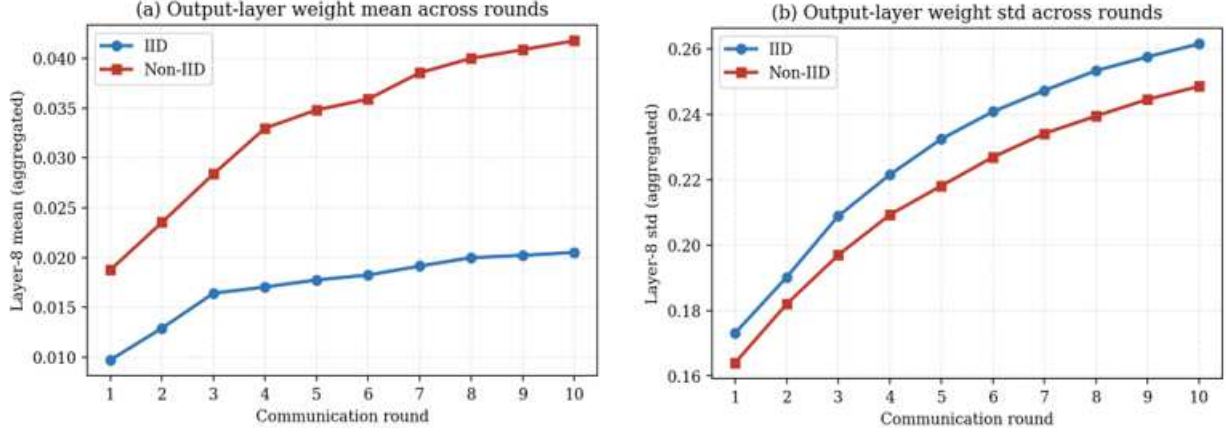


Fig. 5. Aggregated output-layer (Layer 8) weight statistics across communication rounds. (a) Mean: persistently higher under non-IID, reflecting class-skew-induced bias absorbed by the output-layer bias. (b) Standard deviation: slightly lower under non-IID, consistent with proximal-term regularisation.

To extend this observation across the full architecture, Figure 6 compares aggregated layer-wise weight means at round 1 and round 10 for both scenarios. The output layer L8 shows the largest IID/non-IID divergence by a wide margin; the convolutional layers L0–L4 show smaller but consistent differences in direction and magnitude. This is consistent with the well-known property of binary classifiers that the output-layer bias is the most sensitive parameter to class imbalance, and indicates that the weight-audit trace can localise the source of class-skew-induced bias to a specific layer.

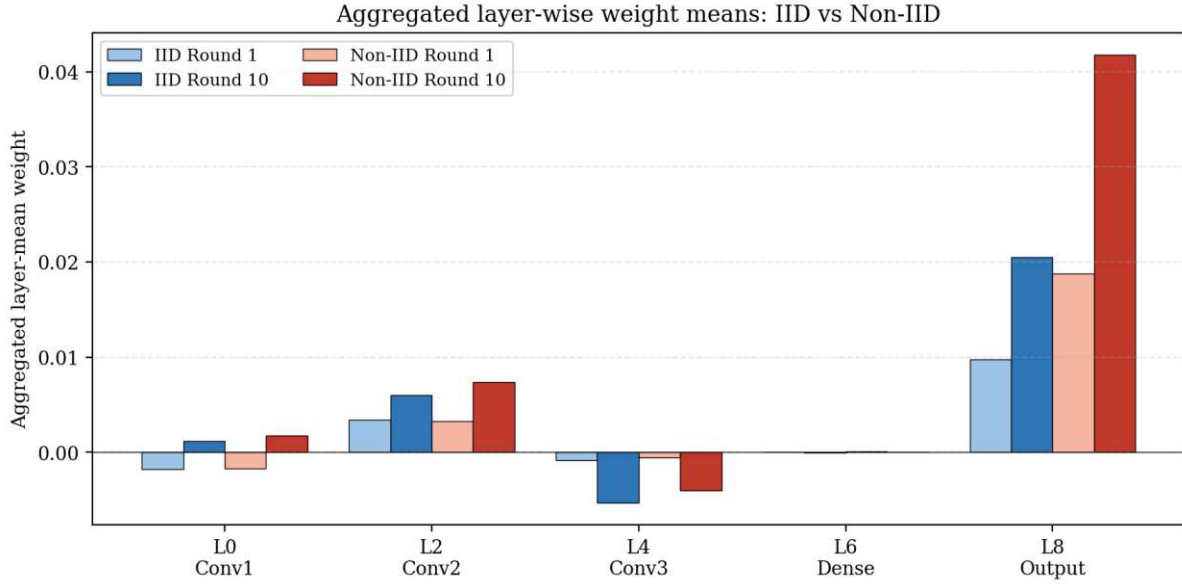


Fig. 6. Aggregated layer-wise weight means at round 1 and round 10 for IID and non-IID scenarios. The output layer (L8) shows the largest IID/non-IID divergence.

Table 11 summarises the round-1 and round-10 means and standard deviations of the output layer for both scenarios.

Table 11 Aggregated output-layer (L8) statistics across rounds

Round	IID Mean	IID Std	Non-IID Mean	Non-IID Std
1	0.009725	0.173076	0.018759	0.163994
3	0.016418	0.208885	0.028386	0.196948
5	0.017754	0.232444	0.034816	0.218120
7	0.019141	0.247299	0.038547	0.234131
10	0.020516	0.261554	0.041761	0.248552
Round-10 ratio	$1.00 \times$	$1.00 \times$	$2.04 \times$	$0.95 \times$

Importantly, the IID/non-IID divergence in the output-layer mean is not visible in the aggregate test accuracy curves of Figure 2, which appear to converge similarly. This is the empirical content of our auditability claim: the weight-audit trace exposes a measurable distributional bias signature that conventional aggregate-metric reporting alone would conceal.

5.7 Deployment-Aware Metrics

Figure 7 reports the three deployment-aware indicators per round: wall-clock round time, communication payload, and process memory. Round time averages 45.30 s (IID) and 45.24 s (non-IID), confirming that the non-IID partition does not substantially increase per-round cost. Communication payload averages 31.94 MB (IID) and 32.30 MB (non-IID), again essentially identical. Process memory grows steadily from approximately 16.5 GB in early rounds to approximately 17.5 GB (IID) and 18.4 GB (non-IID) in later rounds, reflecting the cumulative effect of Python garbage-collection patterns and CUDA memory caching rather than a true monotonic increase in the working set—a known artefact of running multi-client simulation inside a single Python process. Table 12 summarises mean values.

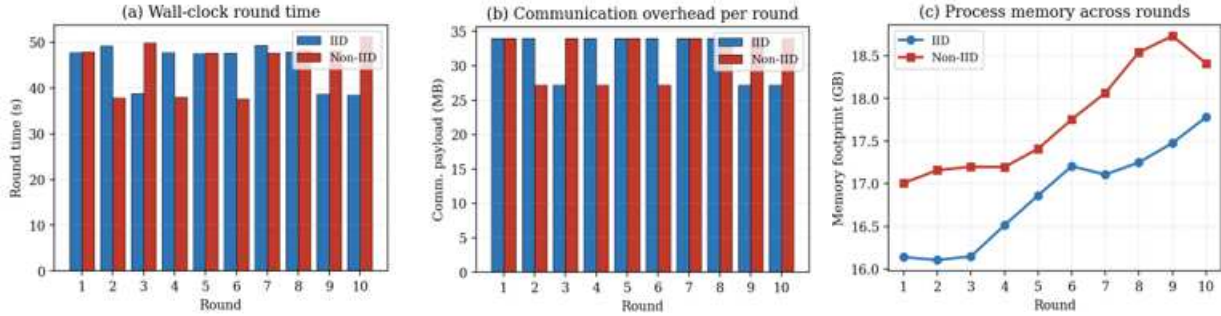


Fig. 7. Per-round deployment metrics for both scenarios. (a) Wall-clock round time. (b) Communication payload per round. (c) Process memory footprint.

Table 12. Mean deployment indicators per round across 10 rounds

Metric	IID	Non-IID
Mean round time (s)	45.30	45.24
Mean communication payload (MB)	31.94	32.30
Mean process memory (GB)	17.35	18.05
Mean active clients per round	4.7	4.7

6. COMPARATIVE ANALYSIS WITH PRIOR WORK

An indicative comparison of the proposed AuditFed-BTC framework with the representative studies on brain tumour classification is shown in Table 13. The accuracies reported in the compared works should be considered indicative and not directly comparable, as the compared works use different datasets, imaging modalities, class distributions and classification tasks.

Table 13 Indicative comparison with prior work

Work	Approach	Dataset	Federated?	Reported Acc.
AuditFed-BTC (Ours, IID)	Compact CNN + FedProx, IID, round-level audit	Combined CT+MRI (9,618 imgs)	Yes	0.9605
AuditFed-BTC (Ours, non-IID 80%)	Compact CNN + FedProx, non-IID 80%, round-level audit	Combined CT+MRI (9,618 imgs)	Yes	0.9496
[55] Saeedi et al., <i>BMC MID</i> (2023) —	Centralised 2D CNN, MRI only	Kaggle MRI 3,264 imgs (4-class)	No	0.9647
[56] Ghanta et al., <i>PeerJ CS</i> (2025)	FL + ResNet50 + FedProx, MRI only, non-IID	Kaggle Brain Tumour MRI (4-class)	Yes	0.9771 IID / 0.8719 non-IID
[57] Muntaqim & Smrity, <i>J. Imaging Inform. Med.</i> (2025)	FL + VGG19 + FedProx/SCAFFOLD + Grad-CAM, non-IID (low heterogeneity)	Kaggle Brain Tumour MRI (4-class)	Yes	0.9718 (FedAvg)

The proposed AuditFed-BTC achieves 96.05% accuracy in the IID setting and 94.96% accuracy in the severe non-IID setting, demonstrating the performance comparable to recent centralised and federated approaches. While some works reach slightly higher peak accuracy, the proposed framework suffers a limited accuracy degradation of only 1.09 percentage points under 80% label skew and 30% client dropout, highlighting its robustness under heterogeneous federated environments.

7. DISCUSSION

7.1 Robustness Under Realistic Stress

The headline empirical finding is that the framework remains both accurate and stable under combined stress from label-skew non-IID data and intermittent client dropout. A final test accuracy of 94.96 % (and F1 of 0.9547) under 80 % label skew and approximately 30 % per-round dropout is competitive with values typically reported under fully IID, full-participation settings, and the gap to the IID upper bound observed in this study is only 1.09 percentage points. The per-client analysis (Figure 3) reveals that this aggregate stability is driven by the rapid convergence of all clients to high local F1 after only a handful of rounds, even when individual clients begin with substantially lower performance due to their skewed class distribution. This supports the practical viability of FedProx for federated brain-tumour detection under conditions closer to clinical deployment than the IID full-participation regime typically reported.

7.2 Auditability as a Practical Trust Signal

The most distinctive observation enabled by the weight-audit trace is the persistent, measurable bias in the output-layer mean under non-IID training. Importantly, this signature is invisible in the conventional aggregate accuracy and loss curves—both scenarios appear to converge similarly when only those curves are inspected—but it is plainly visible in the round-by-round weight statistics. In a deployment scenario where regulators or clinicians must justify a model’s behaviour, such a trace provides a lightweight, low-cost forensic signal indicating that the underlying client distribution may differ from what was assumed. Inspecting the trace requires no access to client data and does not affect any privacy guarantee, and the additional logging cost is negligible. Formalising the qualitative observation into a quantitative integrity test—for example, a hypothesis test on output-layer mean drift across rounds—is a natural direction for future work.

7.3 Why the IID/Non-IID Gap Closes

The narrowing of the IID/non-IID test-accuracy gap from 3.43 to 1.09 percentage points across the run can be understood as the cumulative effect of three mechanisms. First, the FedProx proximal term penalises local drift from the global model, so per-client updates that would otherwise overfit to a single class are pulled back toward a more balanced global solution. Second, the simple uniform FedAvg aggregation averages over five clients whose majority classes are balanced across the federation (two healthy-majority, two tumour-majority, one minority), so even though each individual client is class-skewed, the collective gradient direction is approximately balanced. Third, multiple communication rounds allow the residual per-client bias to dissipate. The persistent shift in output-layer mean despite this dissipation indicates that some bias survives in the model’s parameters even when its test accuracy converges, providing the basis of the auditability finding.

7.4 Limitations

A number of constraints are required to be stated. At first, the model uses the fusion of data level of CT and MRI, and does not compare the fusion of feature level or decision level like comparison is considered as future work. Second, only one non-IID partition (80 % label skew) is examined, Dirichlet-based label and quantity skew at multiple severities would strengthen the empirical evidence. Third, performance metrics are reported from a single held-out test partition under a single random seed repeated trials with confidence intervals and significance testing are not provided here. Fourth, although confusion matrices and per-class metrics are computable from the saved predictions, ROC curves and AUC scores require the threshold-swept probability outputs that we did not retain for this study; future evaluations will retain probability outputs and add full diagnostic analysis. Fifth, although Figure 5 illustrates the weight-audit trace across all layers, a quantitative statistical test on the bias signature has not yet been formalised. Sixth, we did not use explainable-AI processes like Grad-CAM, which will be added in a subsequent extension.

7.5 Future Work

Several extensions follow naturally from this study. (i) Comparing FedProx with FedAvg, SCAFFOLD, and a centralised upper bound on the same data split would quantify the specific contribution of the proximal term under the stress conditions used here. (ii) Comparing the compact CNN against EfficientNet, DenseNet121, ResNet50, and Vision Transformer backbones would map the accuracy-versus-efficiency frontier within the federated setting. (iii) Incorporating Grad-CAM heatmaps and reliability diagrams would

address explainability and calibration concerns. (iv) Formalising the weight-audit trace into a quantitative integrity test—e.g., a control-chart-style hypothesis test on aggregated layer-mean drift—would translate the diagnostic observation reported here into a deployable monitoring tool. (v) Studying the effect of varied dropout rates and adversarial (rather than random) client dropout would test the framework’s resilience against more realistic threat models. (vi) Validating on an independent multi-institutional dataset would address the external-validity question that no single-dataset study can answer.

8. CONCLUSION

In this study, we have proposed an auditable federated learning system for multimodal brain-tumour classification under realistic deployment stress. Using a deliberately lightweight three-block CNN, FedProx optimisation, and a round-level weight-auditing protocol, the framework was evaluated on a combined 9618-image CT and MRI dataset both in IID and 80 % label-skew non-IID partitions with approximately 30 % per-round client dropout. The framework attains 96.05 % test accuracy under IID and 94.96 % under non-IID, with an IID-to-non-IID degradation of only 1.09 percentage points despite the severe class skew. The recorded weight-audit trace reveals a persistent and measurable bias in the output-layer signature under non-IID training that is invisible in conventional accuracy curves, suggesting that round-level weight inspection can serve as a lightweight forensic signal for federated medical AI. We argue that explicit per-round weight reporting should become a more common practice in federated medical-imaging research, both for scientific transparency and for practical clinical trust, and we have shown that this can be done with negligible additional cost using only summary statistics of the local and aggregated model weights.

Funding:

The authors received no financial support for the research of this article.

REFERENCES

- [1] Ostrom, Q. T., Cioffi, G., Waite, K., Kruchko, C., & Barnholtz-Sloan, J. S. (2022). CBTRUS statistical report: Primary brain and other central nervous system tumors diagnosed in the United States in 2015–2019. *Neuro-Oncology*, 24(Suppl. 5), v1–v95. <https://doi.org/10.1093/neuonc/noac202>
- [2] Louis, D. N., Perry, A., Wesseling, P., Brat, D. J., Cree, I. A., Figarella-Branger, D., Hawkins, C., Ng, H. K., Pfister, S. M., Reifenberger, G., Soffietti, R., von Deimling, A., & Ellison, D. W. (2021). The 2021 WHO classification of tumors of the central nervous system: A summary. *Neuro-Oncology*, 23(8), 1231–1251. <https://doi.org/10.1093/neuonc/noab106>
- [3] Weller, M., van den Bent, M., Preusser, M., Le Rhun, E., Tonn, J. C., Minniti, G., Bendszus, M., Balana, C., Chinot, O., Dirven, L., French, P., Hegi, M., Jakola, A., Platten, M., Roth, P., Rudà, R., Soffietti, R., von Deimling, A., Wick, W., & Reifenberger, G. (2021). EANO guidelines on the diagnosis and treatment of diffuse gliomas. *Nature Reviews Clinical Oncology*, 18(3), 170–186. <https://doi.org/10.1038/s41571-020-00447-z>
- [4] Smits, M. (2016). Imaging of oligodendroglioma. *British Journal of Radiology*, 89(1060), 20150857. <https://doi.org/10.1259/bjr.20150857>
- [5] Villanueva-Meyer, J. E., Mabray, M. C., & Cha, S. (2017). Current clinical brain tumor imaging. *Neurosurgery*, 81(3), 397–415. <https://doi.org/10.1093/neuros/nyx103>

- [6] Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., van der Laak, J. A. W. M., van Ginneken, B., & Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42, 60–88. <https://doi.org/10.1016/j.media.2017.07.005>
- [7] Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., Cui, C., Corrado, G., Thrun, S., & Dean, J. (2019). A guide to deep learning in healthcare. *Nature Medicine*, 25(1), 24–29. <https://doi.org/10.1038/s41591-018-0316-z>
- [8] Li, Y., El Habib Daho, M., Conze, P.-H., Zeghlache, R., Le Boité, H., Tadayoni, R., Cochener, B., Lamard, M., & Quéllec, G. (2024). *A review of deep learning-based information fusion techniques for multimodal medical image classification*. *Computers in Biology and Medicine*, 177, 108635. <https://doi.org/10.1016/j.compbio.2024.108635>
- [9] Price, W. N., & Cohen, I. G. (2019). Privacy in the age of medical big data. *Nature Medicine*, 25(1), 37–43. <https://doi.org/10.1038/s41591-018-0272-7>
- [10] Yang, Q., Liu, Y., Chen, T., & Tong, Y. (2019). Federated machine learning: Concept and applications. *ACM Transactions on Intelligent Systems and Technology*, 10(2), 1–19. <https://doi.org/10.1145/3298981>
- [11] Yang, Q., Liu, Y., Cheng, Y., Kang, Y., Chen, T., & Yu, H. (2019). Federated learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 13(3), 1–207. <https://doi.org/10.2200/S00960ED2V01Y201910AIM043>
- [12] Sattler, F., Wiedemann, S., Müller, K. R., & Samek, W. (2020). Robust and communication-efficient federated learning from non-IID data. *IEEE Transactions on Neural Networks and Learning Systems*, 31(9), 3400–3413. <https://doi.org/10.1109/TNNLS.2019.2944481>
- [13] Hsieh, K., Phanishayee, A., Mutlu, O., & Gibbons, P. (2020). The non-IID data quagmire of decentralized machine learning. *Proceedings of the 37th International Conference on Machine Learning Workshop*, 1–8. <https://proceedings.mlr.press/v119/hsieh20a.html>
- [14] Bonawitz, K., Eichner, H., Grieskamp, W., Huba, D., Ingerman, A., Ivanov, V., Kiddon, C., Konečný, J., Mazzocchi, S., McMahan, H. B., Overveldt, T., Petrou, D., Ramage, D., & Roselander, J. (2019). Towards federated learning at scale: System design. *Proceedings of Machine Learning and Systems*, 1, 374–388. <https://doi.org/10.48550/arXiv.1902.01046>
- [15] Topol, E. J. (2019). High-performance medicine: The convergence of human and artificial intelligence. *Nature Medicine*, 25(1), 44–56. <https://doi.org/10.1038/s41591-018-0300-7>
- [16] McMahan, H. B., Moore, E., Ramage, D., Hampson, S., & Arcas, B. A. Y. (2017). Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)* (pp. 1273–1282). <https://proceedings.mlr.press/v54/mcmahan17a.html>
- [17] Li, T., Sahu, A. K., Zaheer, M., Sanjabi, M., Talwalkar, A., & Smith, V. (2020). Federated optimization in heterogeneous networks. *Proceedings of Machine Learning and Systems*, 2, 429–450. <https://doi.org/10.48550/arXiv.1812.06127>
- [18] Karimireddy, S. P., Kale, S., Mohri, M., Reddi, S. J., Stich, S. U., & Suresh, A. T. (2020). SCAFFOLD: Stochastic controlled averaging for federated learning. *Proceedings of the 37th International Conference on Machine Learning*, 5132–5143. <https://proceedings.mlr.press/v119/karimireddy20a.html>
- [19] Zhao, Y., Li, M., Lai, L., Suda, N., Civin, D., & Chandra, V. (2018). Federated learning with non-IID data. *arXiv Preprint arXiv:1806.00582*. <https://doi.org/10.48550/arXiv.1806.00582>
- [20] Sabah, F., Chen, Y., Yang, Z., Raheem, A., Azam, M., Ahmad, N., & Sarwar, R. (2024). Communication optimization techniques in personalized federated learning: Applications, challenges and future directions. *Information Fusion*, 117, 102834. <https://doi.org/10.1016/j.inffus.2024.102834>

- [21] Blanchard, P., El Mhamdi, E. M., Guerraoui, R., & Stainer, J. (2017). Machine learning with adversaries: Byzantine tolerant gradient descent. *Advances in Neural Information Processing Systems*, 30, 119–129. https://proceedings.neurips.cc/paper_files/paper/2017/file/f4b9ec30ad9f68f89b29639786cb62ef-Paper.pdf
- [22] Yin, D., Chen, Y., Kannan, R., & Bartlett, P. (2018). Byzantine-robust distributed learning: Towards optimal statistical rates. *Proceedings of the 35th International Conference on Machine Learning*, 5650–5659. <https://doi.org/10.48550/arXiv.1803.01498>
- [23] El Mhamdi, E. M., Guerraoui, R., & Rouault, S. (2018). The hidden vulnerability of distributed learning in Byzantium. *Proceedings of the 35th International Conference on Machine Learning*, 3521–3530. <https://doi.org/10.48550/arXiv.1802.07927>
- [24] Bagdasaryan, E., Veit, A., Hua, Y., Estrin, D., & Shmatikov, V. (2020). How to backdoor federated learning. *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics*, 2938–2948. <https://doi.org/10.48550/arXiv.1807.00459>
- [25] Fang, M., Cao, X., Jia, J., & Gong, N. Z. (2020). Local model poisoning attacks to Byzantine-robust federated learning. *Proceedings of the 29th USENIX Security Symposium*, 1605–1622. <https://www.usenix.org/conference/usenixsecurity20/presentation/fang>
- [26] Cao, X., Fang, M., Liu, J., & Gong, N. Z. (2021). FLTrust: Byzantine-robust federated learning via trust bootstrapping. *Proceedings of the Network and Distributed System Security Symposium (NDSS)*. <https://doi.org/10.14722/ndss.2021.24434>
- [27] Zhang, Z., Cao, X., Jia, J., & Gong, N. Z. (2022). FLDetector: Defending federated learning against model poisoning attacks via detecting malicious clients. *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2545–2555. <https://doi.org/10.48550/arXiv.2207.09209>
- [28] Rieke, N., Hancox, J., Li, W., Milletari, F., Roth, H. R., Albarqouni, S., ... Cardoso, M. J. (2020). The future of digital health with federated learning. *NPJ Digital Medicine*, 3(1), 119. <https://doi.org/10.1038/s41746-020-00323-1>
- [29] Dayan, I., Roth, H. R., Zhong, A., Harouni, A., Gentili, A., Abidin, A. Z., Liu, A., Costa, A. B., Wood, B. J., Tsai, C. S., et al. (2021). Federated learning for predicting clinical outcomes in patients with COVID-19. *Nature Medicine*, 27(10), 1735–1743. <https://doi.org/10.1038/s41591-021-01506-3>
- [30] Kaissis, G., Makowski, M., Rückert, D., & Braren, R. (2020). Secure, privacy-preserving and federated machine learning in medical imaging. *Nature Machine Intelligence*, 2(6), 305–311. <https://doi.org/10.1038/s42256-020-0186-1>
- [31] Mothukuri, V., Parizi, R. M., Pouriyeh, S., Huang, Y., Dehghantanha, A., & Srivastava, G. (2021). A survey on security and privacy of federated learning. *Future Generation Computer Systems*, 115, 619–640. <https://doi.org/10.1016/j.future.2020.10.007>
- [32] S M Ruhul Kabir Howlader, Lu Liu, Xiao Chen, Haiqin Wu, Bo Yuan, and Atanu Bhattacharjee. 2026. Federated Learning Applications in Healthcare Informatics: A Comprehensive Review. *ACM Comput. Surv.* 58, 12, Article 302 (September 2026), <https://doi.org/10.1145/3809485>.
- [33] Deepak, S., & Ameer, P. M. (2021). Brain tumor classification using deep CNN features via transfer learning. *Computers in Biology and Medicine*, 111, 103345. <https://doi.org/10.1016/j.compbiomed.2019.103345>
- [34] Abiwinanda, N., Hanif, M., Hesaputra, S. T., Handayani, A., & Mengko, T. R. (2019). Brain tumor classification using convolutional neural network. *World Congress on Medical Physics and Biomedical Engineering*, 183–189. https://doi.org/10.1007/978-981-10-9035-6_33

- [35] Swati, Z. N. K., Zhao, Q., Kabir, M., Ali, F., Ali, Z., Ahmed, S., & Lu, J. (2019). Brain tumor classification for MR images using transfer learning and fine-tuning. *Computerized Medical Imaging and Graphics*, 75, 34–46. <https://doi.org/10.1016/j.compmedimag.2019.05.001>
- [36] Badža, M. M., & Barjaktarović, M. Č. (2020). Classification of brain tumors from MRI images using a convolutional neural network. *Applied Sciences*, 10(6), 1999. <https://doi.org/10.3390/app10061999>
- [37] Sharif, M. I., Li, J. P., Amin, J., Sharif, A., & Saleem, M. A. (2020). Active deep neural network features selection for segmentation and recognition of brain tumors using MRI images. *Pattern Recognition Letters*, 129, 181–189. <https://doi.org/10.1016/j.patrec.2019.11.019>
- [38] Sheller, M. J., Reina, G. A., Edwards, B., Martin, J., & Bakas, S. (2019). Multi-institutional deep learning modeling without sharing patient data: A feasibility study on brain tumor segmentation. In A. Crimi et al. (Eds.), *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries* (Lecture Notes in Computer Science, Vol. 11383, pp. 92–104). Springer. https://doi.org/10.1007/978-3-030-11723-8_9
- [39] Lu, Y., Huang, X., Dai, Y., Maharjan, S., & Zhang, Y. (2020). Blockchain and federated learning for privacy-preserved data sharing in industrial IoT. *IEEE Transactions on Industrial Informatics*, 16(6), 4177–4186. <https://doi.org/10.1109/TII.2019.2942190>
- [40] Kairouz, P., McMahan, H. B., & collaborators. (2021). Advances and open problems in federated learning. *Foundations and Trends® in Machine Learning*, 14(1–2), 1–210. <https://doi.org/10.1561/22000000083>
- [41] Sheller, M. J., Edwards, B., Reina, G. A., Martin, J., Pati, S., Kotrotsou, A., Milchenko, M., Xu, W., Marcus, D., Colen, R. R., & Bakas, S. (2020). Federated learning in medicine: Facilitating multi-institutional collaborations without sharing patient data. *Scientific Reports*, 10(1), 12598. <https://doi.org/10.1038/s41598-020-69250-1>
- [42] Li, W., Milletari, F., Xu, D., Rieke, N., Hancox, J., Zhu, W., Baust, M., Cheng, Y., Ourselin, S., Cardoso, M. J., & Feng, A. (2019). Privacy-preserving federated brain tumour segmentation. In H.-I. Suk, M. Liu, P. Yan, & C. Lian (Eds.), *Machine Learning in Medical Imaging* (Lecture Notes in Computer Science, Vol. 11861, pp. 133–141). Springer. https://doi.org/10.1007/978-3-030-32692-0_16
- [43] Andreux, M., du Terrail, J. O., Beguier, C., & Tramel, E. W. (2020). Siloed federated learning for multi-centric histopathology datasets. In S. Albarqouni et al. (Eds.), *Domain Adaptation and Representation Transfer, and Distributed and Collaborative Learning* (Lecture Notes in Computer Science, Vol. 12444, pp. 129–139). Springer. https://doi.org/10.1007/978-3-030-60548-3_13
- [44] Acar, D. A. E., Zhao, Y., Navarro, R. M., Mattina, M., Whatmough, P. N., & Saligrama, V. (2021). Federated learning based on dynamic regularization. *Proceedings of the International Conference on Learning Representations (ICLR 2021)*. <https://doi.org/10.48550/arXiv.2111.04263>
- [45] Adnan, M., Kalra, S., Cresswell, J. C., Taylor, G. W., & Tizhoosh, H. R. (2022). Federated learning and differential privacy for medical image analysis. *Scientific Reports*, 12(1), 1953. <https://doi.org/10.1038/s41598-022-05539-7>
- [46] Wicaksana, J., Yan, Z., Zhang, D., Huang, X., Wu, H., Yang, X., & Cheng, K.-T. (2023). FedMix: Mixed supervised federated learning for medical image segmentation. *IEEE Transactions on Medical Imaging*, 42(7), 1955–1968. DOI: 10.1109/TMI.2022.3233405
- [47] Yan, R., Qu, L., Wei, Q., Huang, S.-C., Shen, L., Rubin, D., Xing, L., & Zhou, Y. (2023). Label-efficient self-supervised federated learning for tackling data heterogeneity in medical imaging. *IEEE Transactions on Medical Imaging*, 42(7), 1932–1943. <https://doi.org/10.1109/TMI.2022.3233574>

- [48] Mastoi, Q.-U.-A., Latif, S., Brohi, S., Ahmad, J., Alqhatani, A., Alshehri, M. S., Al Mazroa, A., & Ullah, R. (2025). Explainable AI in medical imaging: An interpretable and collaborative federated learning model for brain tumor classification. *Frontiers in Oncology*, 15, 1535478. <https://doi.org/10.3389/fonc.2025.1535478>
- [49] Albalawi, E., T. R., M., Thakur, A., Kumar, V. V., Gupta, M., Khan, S. B., & Almusharraf, A. (2024). Integrated approach of federated learning with transfer learning for classification and diagnosis of brain tumor. *BMC Medical Imaging*, 24(1), 110. <https://doi.org/10.1186/s12880-024-01261-0>
- [50] Khowaja, S. A., Dev, K., Anwar, S. M., & Linguraru, M. G. (2025). SelfFed: Self-supervised federated learning for data heterogeneity and label scarcity in medical images. *Expert Systems with Applications*, 261, 125493. <https://doi.org/10.1016/j.eswa.2024.125493>
- [51] Al-Saleh, A., Tejani, G. G., Mishra, S., Sharma, S. K., & Mousavirad, S. J. (2025). A federated learning-based privacy-preserving image processing framework for brain tumor detection from CT scans. *Scientific Reports*, 15, Article 23578. <https://doi.org/10.1038/s41598-025-07807-8>
- [52] Lakshmi Vasanthi, K., Sree Darshne, J., Venkatasubbu, P., & Ramasubramanian, P. (2026). A federated multimodal deep learning framework for brain tumor classification using MRI. *Frontiers in Artificial Intelligence*, 9, 1754000. <https://doi.org/10.3389/frai.2026.1754000>
- [53] Mu, J., Kadoch, M., Yuan, T., Lv, W., Liu, Q., & Li, B. (2024). Explainable federated medical image analysis through causal learning and blockchain. *IEEE Journal of Biomedical and Health Informatics*, 28(6), 3206–3218. <https://doi.org/10.1109/JBHI.2024.3375894>
- [54] Protani, A., Taiello, R., & Molina Van Den Bosch, M. (2025). Federated Transformer-GNN for privacy-preserving brain tumor localization with modality-level explainability. *arXiv preprint arXiv:2601.15042*. <https://doi.org/10.48550/arXiv.2601.15042>
- [55] Saeedi, S., Rezayi, S., Keshavarz, H., & Niakan Kalhori, S. R. (2023). MRI-based brain tumor detection using convolutional deep learning methods and chosen machine learning techniques. *BMC Medical Informatics and Decision Making*, 23(1), 16. <https://doi.org/10.1186/s12911-023-02114-6>
- [56] Ghanta, S., et al. (2025). Federated proximal learning with data augmentation for brain tumor classification under heterogeneous data distributions. *PeerJ Computer Science*, 11, e3396. <https://doi.org/10.7717/peerj-cs.3396>
- [57] Muntaqim, M. Z., & Smrity, T. A. (2025). Federated learning framework for brain tumor detection using MRI images in non-IID data distributions. *Journal of Imaging Informatics in Medicine*, 38, 3909–3927. <https://doi.org/10.1007/s10278-025-01484-9>
- [58] Brain tumor multimodal image (CT & MRI). Kaggle dataset. Available from: <https://www.kaggle.com/datasets/murtozalikhon/brain-tumor-multimodal-image-ct-and-mri>