Iraqi Journal for Computers and Informatics
Available online: https://ijci.uoitc.edu.iq/index.php/ijci
Volume 52, issue 1. 2026

Print ISSN: 2313-190X, Online ISSN: 2520-4912

Explainable Federated Learning for Brain Tumor Classification Using Multi-Source MRI Data

Suhad Muhy Helal [a,b], Belal Al-Khateeb [c]

[a] *Informatics Institute for Postgraduate Studies, University of Information Technology and Communications (UoITC), Baghdad, Iraq.*
[b] *College of Islamic Sciences, University of Baghdad, Baghdad, Iraq.*
[c] *Computer Science Department, College of Computer Science and Information Technology, University of Anbar, Ramadi, Iraq.*

ARTICLEINFO

Keywords:

Brain Tumor Classification, Magnetic Resonance Imaging (MRI), Federated Learning (FL), ResNet18, Non-IID Data, Grad-CAM.

ABSTRACT

Early diagnosis and clinical decision-making depend on accurate brain tumor classification using magnetic resonance imaging (MRI). However, traditional deep learning methods usually rely on centralized medical data, which raises privacy concerns and limits the use of distributed clinical data. This research proposes a privacy-preserving federated learning framework for MRI image-based binary brain tumor classification using a decentralized ResNet-18 architecture that enables collaborative training without sharing raw patient data. To reflect realistic clinical conditions, the framework integrates heterogeneous multi-source datasets in different image formats (PNG and JPG) and evaluates performance under both IID and non-IID settings. Experiments were conducted using the Kaggle Brain Tumor MRI dataset and Mendeley Data distributed across five simulated institutions. Within the evaluated experimental setup, the proposed framework achieved approximately 92% accuracy under IID conditions and 91.5% under non-IID settings, with an F1-score of approximately 0.90. Client-level evaluation demonstrated the model's ability to handle data heterogeneity, while convergence analysis indicated stable training behavior across communication rounds. In addition, Grad-CAM visualization was employed to provide visual interpretability, showing that the model focuses on clinically relevant anatomical regions during prediction. Overall, the results demonstrate that combining federated learning with heterogeneous multi-source MRI data can preserve privacy, maintain robustness and interpretability, and achieve competitive classification performance, highlighting the potential of federated deep learning as a practical and scalable solution for privacy-aware medical image analysis in realistic clinical environments.

1. INTRODUCTION

Brain tumors are considered one of the most serious neurological conditions, and they require accurate diagnosis to improve treatment outcomes and patient survival. Magnetic Resonance Imaging (MRI) is widely used for brain tumor detection due to its accurate assessment and superior soft tissue contrast of brain structures. However, the interpretation of MRI scans by radiologists manually can be time-consuming and prone to human errors, particularly when handling huge volumes of medical imaging. In

E-mail address: ms202420023@uoitc.edu.iq [a]
belal-alkhateeb@uoanbar.edu.iq [c]
Corresponding*: *Suhad Muhy Helal*

Received 21 April 2026, Accepted 09 May 2026

DOI: 10.25195/ijci.v52i1804

addition to being time-consuming, manual MRI scan interpretation is prone to inter-observer variability and diagnostic uncertainty, especially when it comes to identifying minor tumor boundaries and heterogeneous tissue patterns.
For this reason, deep learning-based computer-aided diagnosis systems are being developed to support clinical decision-making [1].
In recent years, deep learning models such as CNNs have been widely used in medical image analysis, including brain tumor classification. Architectures based on ResNet, in particular, are effective at extracting meaningful features from medical images, which helps improve diagnostic accuracy. However, most deep learning approaches rely on large centralized datasets for model training. On the other hand, ethical considerations, privacy concerns, and legal regulations often restrict sharing sensitive patient data across institutions in real clinical environments. The availability of large-scale centralized datasets for deep learning model training is limited by institutional policies, ethical considerations, and stringent privacy regulations such as the European General Data Protection Regulation (GDPR), and the Health Insurance Portability and Accountability Act (HIPAA) of the United States, which impose strict limitations on medical data sharing [2], [3].
To address these challenges, Federated Learning (FL) has emerged as a distributed machine learning approach that allows multiple institutions to train a global model collaboratively without sharing raw data. Each client trains the model locally and sends only model updates to a central server, instead of exchanging sensitive medical data. The knowledge from all participating clients is combined with these updates to build a global model while ensuring data privacy. This framework is mostly useful in healthcare, where it is essential to protect sensitive patient data [3].
In the domain of medical imaging, FL allows hospitals, research centers, and medical institutions to work together to improve diagnostic models while safeguarding sensitive data. Several recent studies have shown that FL can achieve performance comparable to centralized training while preserving data privacy. Because data is often distributed across multiple institutions, FL is especially useful for detecting brain tumor tasks [4].
Existing studies in the literature are broadly classified into centralized deep learning methods and federated learning-based approaches. While centralized approaches often achieve strong predictive performance, they require data sharing, whereas FL-based approaches preserve privacy through decentralized learning. Nevertheless, several important limitations remain in the current literature. Many studies assume independent and identically distributed (IID) data or rely on single-source datasets, which do not adequately represent real clinical environments characterized by heterogeneous multi-source data. In addition, explainability mechanisms are often absent or only partially explored, limiting the interpretability of model predictions in medical applications. Furthermore, few studies evaluate federated frameworks under realistic hospital-like simulations involving multiple institutions and heterogeneous client distributions. These limitations highlight the need for robust, privacy-preserving, and interpretable federated learning frameworks capable of handling realistic distributed medical imaging scenarios.
This paper presents a Federated Learning (FL) model for classifying MRI-based brain tumor. The suggested system replicates a distributed healthcare scenario with multiple clients, which represent various data sources. The Federated Averaging (FedAvg) method is used by the central server to aggregate adjustments to the model while the client locally trains a deep CNN based on the ResNet-18 architecture. The global model might improve progressively as participating clients share their knowledge during the training process. The system was trained for 25 federated rounds with five clients using a distributed dataset configuration for the experimental evaluation.
The main contributions of this research are summarized as follows:

- A five-client federated learning framework for MRI-based binary brain tumor classification evaluated under both IID and Non-IID settings using heterogeneous multi-source datasets.
- Integration of a ResNet18 architecture within a decentralized federated pipeline trained over 25 communication rounds to investigate classification performance and convergence behavior.
- Incorporation of Grad-CAM explainability to visualize clinically relevant regions influencing model predictions and enhance interpretability in medical decision-support settings.

While the proposed model ensures data privacy, it shows that FL can significantly leverage distributed medical datasets, which makes it a viable alternative for practical cooperative healthcare applications.
The remaining sections of this article are arranged as follows. Section 2 describes the review of the research relevant to FL in medical imaging. Section 3 presents a detailed description of the suggested methodology, which contains the FL framework and partitioning data techniques. Section 4 deals with decisions and results. Finally, section 5 concludes and outlines future research possibilities.

2. Related Works

Medical image processing systems have greatly benefited from recent developments in deep learning, especially when it uses MRI scans to diagnose brain cancer. However, the majority of deep learning models utilize centralized training paradigms, which need to compiling medical data from a single source. This method raises concerns about patient privacy, regulatory compliance and data ownership. To faces these constraints, FL has been an effective strategy that enables models to train among institutions collaboratively without needing the sharing of private data.

2.1 FL with CNN-based Models

Various researchers are working on the integration of convolutional neural networks (CNNs) within federated learning frameworks for brain tumor classification. For instance, in [5], the authors proposed a FL based DL model that leveraged convolutional neural networks CNNs for automated and accurate brain tumor classification. The model utilized a modified VGG16 architecture and was trained on data distributed across multiple institutions. The model showed strong classification performance, which reached an overall accuracy of about 98%, indicating the effectiveness of integrating transfer learning with FL in the medical imaging field. Despite the advantages of FL in healthcare applications, data biases and diversity, complexity of FL, and computational resources, are remaining several limitations and challenges in this domain.

Another research [6] investigated the use of multiple pre-trained CNN models were evaluated, and the best three models (DenseNet121, VGG19, and InceptionV3) were integrated using an ensemble approach, which was then used as the global model within an FL framework for MRI-based brain tumor classification. In this framework, multiple CNN architectures were trained locally on distributed MRI datasets across multiple clients, and their learned parameters were aggregated at the server level. With an accuracy of about 91.05% while keeping patient data privacy, the results showed that FL can reach performance comparable to centralized models. While this research achieved improved accuracy, there were several limitations, such as an imbalance of class distribution, the use of standard federated average (e.g., FedAvg), a limited number of clients, and the existence of low-quality images. Similarly, [7] examined the use of EfficientNet architectures within an FL framework for classifying MRI-based tumor. This research reached an accuracy of about (80.17%), in addition, FL can effectively deal with distributed medical datasets while achieving strong classification performance and keeping data confidentiality across different institutions. However, limitations related to data heterogeneity and model interpretability were not fully addressed, which suggested the need for more powerful optimization strategies. To further strengthen privacy-preserving collaborative learning, [8] developed CNN based FL framework for brain tumor and molecular subtype classification, enabling multi-institution training without sharing raw patient data. The models gave an accuracy of 99.8%. Despite its privacy advantages, the study's validation was restricted to a relatively small dataset. Recent studies have further explored the application of federated learning in brain tumor detection to address privacy concerns and enable collaborative medical data analysis. For instance, [9] proposed a distributed and privacy-preserving horizontal federated learning framework for malignant glioma detection using MRI images. Their approach evaluated both IID and non-IID data distributions with different numbers of participating clients, including configurations with five and ten clients. The proposed framework utilized a pre-trained MobileNetV2 backbone within a federated training environment based on the Federated Averaging (FedAvg) algorithm. Experimental results demonstrated very high classification performance, achieving 99.76% accuracy under IID data distribution and 99.71% under non-IID settings with five clients. Although these promising results, there are limitations that can be observed, such as being dependent on a single dataset, did not test cross-dataset generalization, which is important for assessing the robustness of the model, and data heterogeneity. In addition, this study relied on a fixed model architecture without exploring a hybrid, or alternative deep learning strategy, limiting its adaptability in complex conditions.

On the other hand, [10] introduced an FL based DL approach for classifying brain tumors. The framework used the ResNet-18 model, which was used to develop the system, and was trained on the MRI scans from the Kaggle Brain Tumor MRI dataset. In addition, the use of FedAvg for aggregate model updates across different clients. The results of this research achieved 98% of accuracy, while keeping privacy across participating clients. However, this study did not provide a detailed analysis of non-IID data distributions, and it only limited evaluation of scalability with diverse data characteristics across clients. In brain tumor classification using federated medical imaging settings, studies [11] and [12] developed FL with the effectiveness of combining transfer learning with FedAvg and FedProx. This integration achieved accuracies of (97.19%) and (98.4%), respectively. Although this performance, their applicability may be limited because of the use a small number of clients, and issues with associated computational and communication costs.

More recent work [13] proposed a federated learning framework for MRI-based brain tumor detection using VGG19 under heterogeneous non-IID settings. The study evaluated multiple aggregation strategies, including FedAvg, FedProx, and Scaffold, while incorporating Grad-CAM to improve interpretability. The experimental results demonstrated strong classification performance, achieving 97.18% with FedAvg, 98.24% with FedProx, and 98.45% with Scaffold. However, the framework remained dependent on a specific dataset and required broader clinical validation to assess its robustness across diverse medical environments.

2.2 FL with Explainability (FL + XAI)

To improve transparency in medical decision-making, authors recommended combining explainable artificial intelligence (XAI) with FL. [14], illustrated a cooperative FL model using GoogLeNet, which included explainability methods like Grad-CAM and saliency maps. The model was trained across a number of distributed clients and achieved an overall accuracy of approximately 94% while providing visual explanations that help clinicians better understand the model's predictions. On the other hand, this study has some drawbacks, including the assumption of uniform resources availability among clients, variances in computational capacity, and potential privacy problems. To this family of study [15] used explainable AI (XAI) with (FL) for brain cancer classification utilizing MRI data from multi-medical centers. The proposed used a modified ResNet-18 with FedAvg and

differential privacy (DP). It attained 92% centralized training accuracy, 90% in FL without DP, and 88% with DP. Grad-CAM and SHAP was used in this research, as explainability techniques, which highlight relevant tumor boundaries clinically. However, this paper did not examine different data distribution scenarios, relied on simulated institutions, also depending on a single CNN mode.

2.3 FL with Privacy, Security, and Optimization

Beyond traditional FL frameworks, modern studies have focused on improving model performance and communication costs. In [16], the authors proposed an FL approach that leverages knowledge distillation to address data heterogeneity among clients while reducing communication overhead. The method achieved classification accuracies of 94,38% under IID settings and 93.34% under non-IID settings. In addition, this works improved performance on distributed brain tumors for images using MRI, while maintaining data privacy during the training process. In addition, this approach can be further optimized by employing different strategies, tuning key parameters, and using different pretrained models.

Later studies aimed to improve unsolved problems, such as system security and model robustness, based on the FL adoption. For example, in [17], a model called Aniso-ResCapHGBO-Net was introduced, which combined ResNet-50, Capsule Networks, and HGBOA optimization, and Blockchain for secure brain tumor detection using CT images. The model achieved an accuracy 99.07%. Despite the framework's enhanced security and architecture expressiveness, additional challenges were introduced, including high computational complexity, increased communication overhead, and limited large-scale clinical validation.

Besides, [18] provided a systematic review considering over 250 papers to introduce FL methodologies, which are based on privacy-preserving, collaborative learning in domains such as healthcare and IoT. The research gaps including data heterogeneity, security concerns, communication overhead, aggregation techniques, and differential privacy, that covered in the review. However, it did not particularly address explainable FL or its use multi-source MRI data to classify brain tumors.

2.4 Research Gap Analysis

Despite significant advances in federated learning for medical imaging, there are still several important limitations in the current literature. First, many studies rely on single-source datasets, which limits cross-domain generalization and reduces robustness under heterogeneous clinical conditions. Second, most of the existing works focus on standard FL settings without incorporating stronger privacy-enhancing mechanisms or analyzing realistic privacy constraints, while privacy preservation is a core objective of the federated learning. Third, more research on explainability is still needed, where few studies have applied XAI methods such as Grad-CAM or SHAP to facilitate the clinical interpretability. Moreover, many previous works either assume IID data distributions or offer limited evaluation under realistic Non-IID and multi-institutional conditions. These limitations motivate the need for interpretable, privacy-aware and robust FL frameworks for handling heterogeneous multi-source MRI environments.

Table 1 *Summarizes And Compares Of The Key Characteristics Of The Reviewed Studies*.

Study	Dataset	Year	FL Method	Explainability	Privacy	Accuracy	Limitation
[5]	MRI	2024	VGG16 + FL	No	No	98%	Non-IID challenges, computational cost, limited data diversity
[6]	MRI	2023	CNN + FL	No	No	91.05% (FL), 96.68% (Base model)	Class imbalance, scalability issues, potential overfitting
[7]	MRI	2024	EfficientNet +FL	No	No	92.2	Data heterogeneity and limited generalization
[8]	MRI	2025	CNN + FL	No	No	99.8	Limited scalability and adaptability to diverse tumor patterns
[9]	MRI	2025	MobileNetV2 + FedAvg	No	No	99.76% (IID), 99.71% (Non-IID)	Binary-only evaluation and limited multi-class applicability
[10]	MRI	2025	ResNet-18 + FedAvg	Grad-CAM visualization	No	98	Limited explainability evaluation, as Grad-CAM was primarily used for visualization without quantitative validation.

Study	Dataset	Year	FL Method	Explainability	Privacy	Accuracy	Limitation
[11]	MRI	2025	CNN + FedAvg + FedProx	No	No	97.19%	Model instability under heterogeneous client distributions
[12]	MRI	2025	VGG-based deep CNN with FedAvg	No	No	98.4%	Data heterogeneity, communication cost, aggregation challenges
[13]	MRI	2025	VGG19 with FL	Grad-CAM	No	FedAVG: (97.18%) FedProx: (98.24%) Scaffold: (98.45%)	required broader clinical validation to assess its robustness across diverse medical environments
[14]	MRI	2025	GoogLeNet + FL	Grad-CAM + Saliency Maps	No	94.24%	Fixed client assumptions and communication instability
[15]	MRI	2026	ResNet-18 + FedAvg	Grad-CAM + SHAP	Yes	92% (Centralized), 90% (FL), 88% (FL+DP)	Limited Non-IID evaluation and reduced FL performance
[16]	MRI	2024	VGGNet16 + Knowledge Distillation (KD)	No	Yes	94.38% (IID), 93.34% (Non-IID)	Computational complexity and architecture dependency
[17]	CT	2025	ResNet-50 + Capsule Networks + HGBOA optimization + Blockchain	No	Yes	99.07%	High computation, communication overhead, limited clinical validation
[18]	-	2025	Review Study	-	-	-	Does not address explainable FL with multi-source MRI data

The research studies that combine FL with CNN-based architectures for medical image classification are the most relevant of those evaluated. However, the most of these strategies rely on single-source datasets, or assume ideal IID data distributions, which limits their use in real-world clinical settings with dispersed and data heterogeneity. In addition, very few studies combine explainability methods with FL, and even fewer deal with both interpretability and data heterogeneity at the same time. This highlights a significant research gap in creating reliable, interpretable, and privacy-preserving FL frameworks that can manage multi-source non-IID medical data.

In this work, we proposed a federated learning framework for binary brain tumor classification using MRI images. Unlike several previous studies that rely on complex architectures or centralized preprocessing pipelines, our approach focuses on a lightweight convolutional architecture integrated within a federated training environment. The Federated Averaging (FedAvg) technique is used by the system's multiple distributed clients to aggregate model updates during communication rounds. Additionally, Grad-CAM is integrated to promote clinical interpretability and improve transparency by offering visual explanations of the model's predictions.

3. Proposed Methodology

The suggested FL approach for privacy-preserving MRI image-based brain tumor classification is discussed in this section. Without exchanging raw medical data. The system simulates a distributed medical setting in which several clients work together to develop a global deep learning model. A central server and multiple distributed clients, each containing a private subset of the MRI dataset, make up the overall architecture. Figure 1 illustrates the general design of the suggested framework.

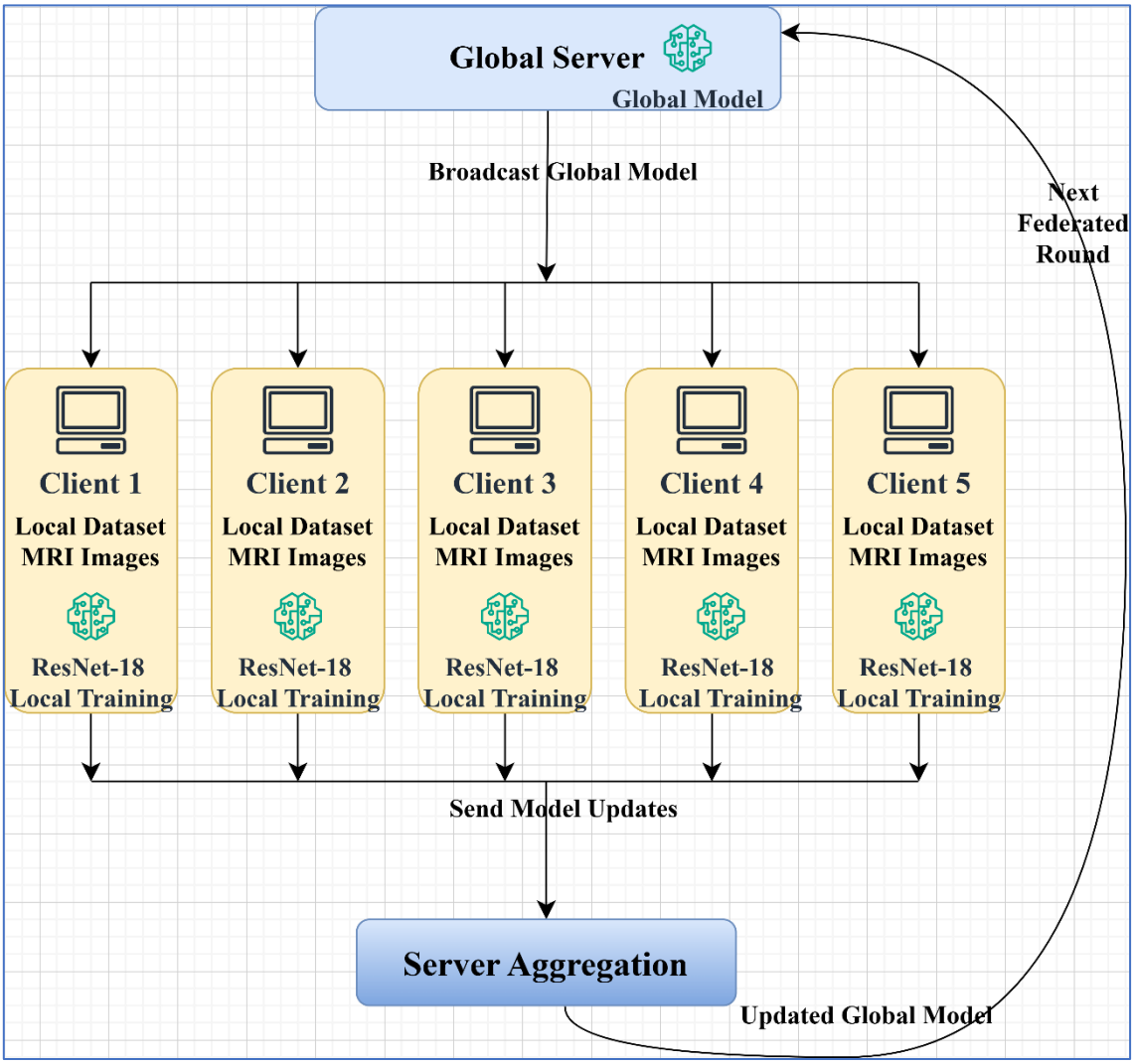

Figure 1: *Proposed Methodology*

3.1 Dataset Preparation and Preprocessing

To produce a dataset that is suitable for FL research, brain MRI images were collected from two publicly available datasets. The datasets include the Brain Lesion MRI Dataset on Mendeley Data[2] and the Brain Tumor MRI Dataset on Kaggle[3]. It is critical to mention that the domain variability occurs when datasets from multiple sources (Kaggle and Mendeley) are combined due to differences in image formats, acquisition procedures, and preprocessing pipelines. Although this unpredictability strengthens the model, it also makes learning more challenging.

A. Dataset Description

The Brain Tumor MRI dataset from Kaggle, is the first dataset, which comprises images aggregated from the Figshare, SARTAJ datasets, and Br35H datasets. The dataset is organized into training and testing subsets. The training set contains 5712 MRI images distributed across four tumor classes: glioma (1321 images), meningioma (1339 images), no tumor (1595 images), and pituitary (1457 images). The testing set consists of 1311 MRI images categorized into the same four classes: 300 glioma, 306 meningioma, 405 no tumor, and 300 pituitary images. All images are provided in JPG format. The second dataset is a Brain lesion MRI and co-related MRS Spectroscopy Dataset from Mendeley Data, which is significant because it combines conventional imaging (MRI) and expert diagnostic information, and includes both MRI and MRS images, as well as expert diagnosis, allowing for exceptional reuse in medical and diagnostic research. However, for the purpose of this research, only two categories were selected from the Kaggle dataset: glioma and no-tumor. The glioma images were treated

[2] Riyadh, Sura (2024), "Brain lesion MRI and co-related MRS Spectroscopy Dataset", Mendeley Data, V1, doi: 10.17632/v3gwhkyjsg.1

[3] https://www.kaggle.com/datasets/masoudnickparvar/brain-tumor-mri-dataset/data

as tumor samples, while the no-tumor images were used as normal brain samples. This selection allowed the construction of a binary classification dataset that focuses on detecting the presence or absence of brain tumors. To further increase the number of tumor samples and improve dataset diversity, additional MRI images were incorporated from the Brain Tumor MRI Dataset repository available through Mendeley Data. Unlike the Kaggle dataset, which contains multiple tumor types and normal images, the Mendeley dataset consists exclusively of brain tumor MRI scans. Therefore, all images from the Mendeley dataset were considered as tumor samples and merged into the tumor class of the final dataset.

By combining tumor images from both datasets and normal brain images from the Kaggle dataset, a binary classification dataset was constructed consisting of two classes: (Tumor, No Tumor). The model can learn more generalized tumor features from MRI images collected from different sources according to this integration method, which increases the variability of tumor appearances.

Table 2 presents the final class composition after merging the Kaggle and Mendeley datasets. The resulting binary dataset consists of tumor and no-tumor categories with an imbalanced class distribution.

Table 2: *Final Dataset Composition After Dataset Merging.*

Class	*Number of Samples*
Tumor	1659
No Tumor	2000
Total	3659

Before federated distribution, the merged dataset was separated into training and testing subsets. The test set was kept independent and was not used during client-level training or data partitioning. Only the training subset was distributed among the federated clients for IID and Non-IID experiments. The Kaggle and Mendeley datasets originate from independent repositories, and no known patient overlap exists between the two data sources. To prevent potential data leakage, all slices extracted from the same MRI volume were kept within the same client partition and were not shared across different clients or across training and testing subsets.

B. Load and Preprocessing Data

Because of using two datasets, there are two preprocessing phases:

1. Brain Tumor MRI Dataset from Kaggle: all MRI images were resized to a fixed resolution of 224 × 224 pixels to match the input requirements of the ResNet-18 architecture. The images were normalized using the standard ImageNet mean and standard deviation to achieve consistent intensity scaling and stable optimization across samples.
2. Brain lesion MRI and co-related MRS Spectroscopy Dataset from Mendeley Data: This dataset originally provides MRI scans in NIfTI (.nii /.nii.gz) format representing volumetric medical imaging data. Since deep learning models used in this study operate on 2D images, a preprocessing pipeline was implemented to convert the volumetric scans into 2D slices. The conversion process was implemented using Python with the libraries: Nibabel for reading NIfTI medical imaging files, NumPy for numerical processing, and Pillow (PIL) for image generation. The middle slices along the axial dimension were extracted to capture the most informative brain regions for each MRI volume. To guarantee consistency across patients, a fixed number of slices was selected from the central region of the volume. Each extracted slice was produced as a grayscale PNG image after being normalized to the range [0,255]. The images were arranged using the tumor labels included in the dataset metadata file, after conversion. Only the high-grade and low-grade tumor cases were selected for this study, because the Mendeley dataset includes multiple diagnostic categories (high-grade tumor, low-grade tumor, and benign lesion). The benign lesion category was excluded, as it represents tumor-like abnormalities rather than confirmed tumor cases. Therefore, only the selected high-grade and low-grade samples were categorized as tumor class in the final dataset.

For the training set, light data augmentation techniques were used in order to decrease overfitting and enhance generalization. These included random horizontal flipping with a probability of 0.5, and small-angle rotations (±10 degrees). On the other hand, only deterministic preprocessing techniques, such as resizing and normalization, were used for validation and test sets to guarantee unbiased evaluation. Class weights are dynamically computed from each local dataset and incorporated into the loss function during training to address class imbalance across clients. During the learning process, these steps help ensure that minority classes are fairly represented.

3.2 Data Distribution for Federated Learning

In order to create a realistic federated learning environment, the dataset was distributed among several clients, each of which represented a different medical institution. By using a custom Python process, the data was organized and divided into client-specific folders. The training data was divided between five clients, with each client having its own training and validation sets. Two categories of data distribution were examined including:

- IID: this category use to maintain class balance, and stratified sampling is used to spread data evenly among clients.
- Non-IID: this category simulates real-world variation in data across institutions, which is reflected in this configuration. Data distribution differs between clients, some clients having more tumor samples and others clients having more normal cases.

3.3 Federated Learning Framework

In this paper, a centralized (FL) system is developed to facilitate cooperative training of a DL model for brain tumor classification without exchanging raw clinical data. The system uses a client-server architecture, in which a central server uses the Federated Averaging (FedAvg) [1] algorithm to aggregate the model parameters while several clients train local models on their own private datasets. The local training objective (FedProx) includes a proximal regularization term in addition to the standard FedAvg aggregation. This term constrains local model updates to remain close to the global model parameters, improving training stability and reducing client drift caused by data heterogeneity in non-IID settings. Python and PyTorch were used to construct the implementation, and socket-based messaging with serialized data transport is used for client-server communication. **Algorithm 1** shows how the proposed federated learning framework is implemented step-by-step, highlighting the communication rounds, local training procedure, and global model aggregation.

Algorithm 1: Federated Learning Training Procedure

Input: *Client updates* $\left\{\left(W_k^{(t)}, n_k\right)\right\}_{k=1}^{K}$, *T Number of communication rounds*

Output: *Optimized global model* $W^{(t+1)}$

Step1: *Initialize global model* $W^{(0)}$ *at the server*
Step2: *Establish connection with all clients K*
Step3: *for each round t=1 to T do*
Step3.1: *Server broadcasts current global model* $W^{(t)}$ *to all clients.*
Step3.2: *For each client* $k \in \{1, \ldots, K\}$ *do:*
- *Load global weights* $W^{(t)}$.
- *Train local model using private dataset.*
- *Compute evaluation metrics (Accuracy, Precision, Recall, F1-score)*
- *Send updated weights* $W_k^{(t)}$ *and sample size* n_k *to server.*

Step3.3: *Server aggregates client updates using FedAvg*

$$W^{t+1} = \sum_{k \in S_t} \frac{n_k}{\sum_{j \epsilon S_t} n_j} W_k^{(t)} \quad (1)$$

Step3.4: *Evaluate updated global model on test dataset:*
- *Compute (Accuracy, Precision, Recall, F1-score)*
- *Generate confusion matrix.*

Step3.5: *Save best global model based on F1-score.*
Step3.6: *Store results (metrics, curves, confusion matrix)*
End For
Step4: *Return final optimized global model*

The F1-score which provides a balanced evaluation of precision and recall, is used to determines which global model is the best. In medical classification tasks, where both false positives and false negatives must be carefully considered, this is very crucial.
To improve training stability and convergence across communication rounds, a learning rate decay strategy was adopted. The initial learning rate is set to a moderate value (0.0003), to ensure effective learning in early stages, while gradual decay is applied, to reduce update fluctuations and prevent overshooting during later rounds.
While a smaller learning rate in later stages, helps fine-tuning the model parameters and achieving stable convergence, a comparatively higher learning rate at the beginning enables the model to learn more quickly. This strategy is particularly important in FL, where the distributions of data heterogeneity may lead to unstable updates across clients.

3.4 Model Architecture Design

The model architecture adopted within the proposed FL framework for brain tumor classification, was described in this section. The ResNet18 convolutional neural network was selected as the primary deep learning model, to ensure a strong yet computationally feasible backbone suitable for federated and privacy-preserving training.

- **ResNet18**

In this paper, a well-known convolutional neural network was using of ResNet-18, with residual learning to enhance the capabilities of model's learning. Training of deep neural networks is challenging due to the vanishing gradient problem. ResNet uses shortcut connections to learn residual mappings while preserving important features across network layers. ResNet-18 provides a balance between model complexity and computational efficiency for deep feature extraction in brain tumor classification while retaining model stability. The ResNet-18 architecture has demonstrated strong performance in different medical imaging tasks; and is ideal in classifying similar tumor types including glioma, meningioma, and pituitary tumors, as well as distinguishing non-tumor cases [2]. The detailed configuration of the model is summarized in Table 3.

Table 3: *RESNET18-BASED MODEL ARCHITECTURE CONFIGURATION.*

Component	Description
Base Model	ResNet18 pre-trained on ImageNet
Input Size	224 × 224 × 3 (RGB MRI images)
Feature Extractor	Convolutional layers of ResNet18 (excluding final FC layer)
Frozen Layers	Early layers are frozen in initial rounds and later unfrozen
Trainable Layers	Final layers and classifier head are fine-tuned
Fine-Tuning Strategy	Dynamic fine-tuning (freeze → unfreeze across rounds)
Pooling	Adaptive Average Pooling (inherent in ResNet18 architecture)
Flattening	Output feature maps are flattened before classification
Classifier Head	Fully Connected (FC) layer modified to match binary classification task
Activation Function	ReLU used in hidden layers
Output Layer	Fully Connected layer with 2 neurons (tumor / no-tumor)
Loss Function	CrossEntropyLoss with label smoothing (0.1) and class weights
Regularization	FedProx proximal term added to loss
Optimizer	Adam optimizer with weight decay (L2 regularization)
Learning Rate	Initial learning rate is set to 0.0003 and gradually decayed across communication rounds.
Gradient Clipping	Applied to stabilize training and prevent exploding gradients
Learning Rate Strategy	Exponential decay applied per communication round

In addition, to improve generalization performance and reduce model overconfidence, label smoothing with a factor of 0.1 was applied in the CrossEntropyLoss function. The ResNet18 pre-trained model serves as a foundation to leverage rich feature representations obtained from large-scale datasets. To adapt to binary classification (tumor vs. no tumor), the final fully connected layer was modified. To increase model generalization with medical imaging data, fine-tuning was used. To preserve pre-trained representations, the ResNet18's feature extraction layers are frozen during the initial communication rounds. After several rounds, all layers are unfrozen, which allow for full fine-tuning and adaptation to the medical dataset.

The hyperparameter setting of this work was selected based on preliminary empirical tuning, the constraints of federated learning, and common practices in the deep learning literature. The Adam optimizer was selected, because of its adaptive optimization ability and stable convergence behavior in medical image classification tasks. To balance the convergence speed and training stability in distributed optimization, the initial learning rate was set to 0.0003. We used twenty-five communication rounds to achieve sufficient collaborative learning with acceptable computational and communication overhead for federated clients. Additionally, we applied label smoothing with a factor of 0.1 to reduce model overconfidence and enhance the generalization performance in the heterogeneous data scenario.

In this study, the ResNet-18 model was chosen since it strikes a reasonable compromise between predictive performance, computational efficiency, and communication overhead, which makes it suitable for federated learning scenarios. Compared with deeper architectures such as DenseNet121 and EfficientNet, the ResNet-18 architecture has fewer parameters and less computational complexity for distributed clients, which means less training overhead and communication cost. Although lightweight architectures such as MobileNetV2 enable efficient deployment, but ResNet-18 has shown stable and competitive performance in medical image analysis, especially in the MRI-based tumor classification task. Moreover, convolutional neural networks (CNNs) like ResNet-18 are efficient in extracting local spatial features, which are important for tumor detection in MRI images. More recent architectures, such as Vision Transformers (ViTs), on the other hand, typically require larger datasets and substantially more computational resources to achieve optimal performance, which may be less practical in privacy-preserving federated settings. Therefore, we chose ResNet-18 as an efficient and robust baseline model to meet the requirements of distributed medical learning systems.

3.5 Model Interpretability and Visual Analysis (Grad-CAM)

To make the proposed FL model more interpretable, Grad-CAM was used to visualize the regions that affect the model's predictions. It produces heatmaps based on the gradients of the target class with respect to the last convolutional layer. In this work, Grad-CAM was applied to the trained ResNet18 global model, and the heatmaps were overlaid on MRI images to show the most important regions for classification. This helps improve model transparency and shows that the model focuses on tumor-related areas, which increases confidence in its use for medical applications.
The Grad-CAM heatmap is created by weighting the feature maps from the last convolutional layer:

$$L^{c}_{Grad-CAM}\ ReLU\ (\sum_{k} \alpha^{c}_{k}\ A^{k}) \qquad (2)$$

Where A^{k} represents the feature map of channel k., α^{c}_{k} denotes the weight derived from gradients of class c, and ReLU ensures that only positive contributions are evaluated.

4. Results and Discussion

This section gives a detailed evaluation of the suggested federated deep learning system for MRI-based brain tumor classification. This paper focuses on the model's qualitative behavior as well as its quantitative performance across various distributions of data (IID and Non-IID). The problems are provided by statistical heterogeneity and drift domain, as a result of multi-source datasets that combine, which are highlighted in detail in this research.
In order to enhance the stability of convergence and generalization, the performance of the mode is evaluated during communication rounds. The effects of several crucial design decisions, including transfer learning, adaptive fine-tuning, and the addition of a proximal term (FedProx), are also investigated. This discussion's objectives are to describe performance differences, evaluate the observed results, and propose potential areas for the suggested system. A comparison with a centralized (single machine) baseline was conducted, to evaluate the effectiveness of the suggested FL framework. The centralized model achieved an accuracy of 90.62% and an F1-score of 0.8966, which demonstrate well performance due to complete access to the entire dataset during training. Despite this advantage, centralized training requires aggregating private medical data in a single location, which is often impractical due to regulatory limitations and privacy constraints. In contrast, by distributing data across clients, the proposed FL framework achieves comparable performance (92% under IID and 91.5% under Non-IID conditions) while preserving data privacy. These results suggest that FL offers similar performance while ensuring data confidentiality, providing a practical alternative to centralized training, and enabling collaboration across multiple medical centers.

4.1 Global Performance

Table 4 compared the global model performance under IID and Non-IID distributions of data, at the best-performing communication round (round 8). By using the same round in both cases, to ensure fair and uniform evaluation. The results show that the IID model performs slightly better than the Non-IID model, achieving an accuracy of 92%, as opposed to 91.5%. For this improvement, the consistent distribution of data among clients, which enables more reliable and efficient model aggregation, is responsible. On the other hand, the precision value is achieved in the Non-IID scenario (0.985), compared to the IID condition of (0.997), which has a greater precision value. This behavior results in fewer false positive predictions. The confusion matrix verified this behavior, by showing that the value of false positives increases from one in the IID setting to five in Non-IID conditions. In contrast, both scenarios have the same recall values (0.842), which demonstrates that the ability of the model to detect tumors cases remains consistent regardless of the data distribution.
This finding indicates that the F1-score is same in both scenarios, with a slight enhancement observed in the IID conditions. Overall, the results suggest that the proposed FL operates efficiently in both IID and non-IID scenarios. Although, the IID setting creates a more stable environment for training. While in the Non-IID case, the performance is reduced, because of the impact of heterogeneous data, which can affect the consistency of local model updates and decrease the effectiveness of global aggregation.

In addition to accuracy, precision, recall, and F1-score, sensitivity and specificity were computed to provide a more clinically relevant assessment of diagnostic performance. The results indicate high specificity under both IID and Non-IID settings, while maintaining stable sensitivity for tumor detection.

Table 4: *COMPARISON OF RESNET18 MODEL.*

Metrics	Best Round	Accuracy	Loss	Precision	Recall	F1-Score	TN	FP	FN	TP	Sensitivity	Specificity
IID	8	0.92	0.4216	0.997	0.842	0.913	399	1	63	337	0.842	0.998
Non-IID	8	0.915	0.4204	0.985	0.842	0.908	395	5	63	337	0.842	0.988

To further assess the discriminative capability of the proposed framework, Receiver Operating Characteristic (ROC) curve analysis and Area Under the Curve (AUC) scores were computed for both IID and Non-IID settings, as illustrated in Figure 2. The ROC curves demonstrate strong classification capability in both scenarios, with the IID configuration achieving a slightly higher AUC value compared with the Non-IID setting. This behavior is consistent with the global performance metrics reported in Table 4 and indicates that the proposed federated model maintains robust tumor versus non-tumor discrimination despite heterogeneous client data distributions.

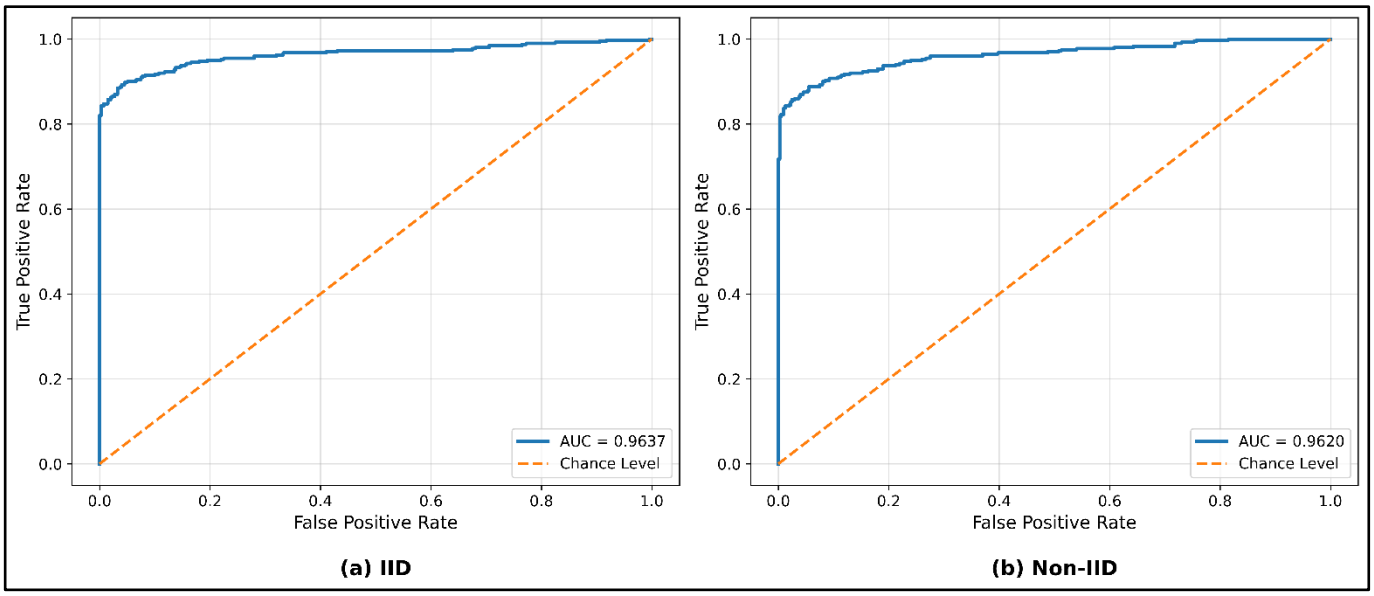

Figure 2: *ROC curves and AUC analysis of the proposed ResNet18 federated model under different data distribution settings:*

(a) IID setting and (b) Non-IID setting

4.2 Confusion Matrix Analysis

The success of the suggested system is evaluated by the confusion matrices, which are utilized to illustrate the model's reliability and performance of classification. The best-performing communication round (round 8) under both IID and Non-IID settings were displayed by Figures (3) and (4). A complete overview of the suggested FL's categorization behavior was provided by these matrices, which shows the model's class-wise performance.

In addition, a confusion matrix can be beneficial because it can quantify how each class can be classified. In the IID scenario, as illustrated in Figure (3), the model performs very effectively, with a high number of true negatives (TN = 399), and true positives (TP = 337), and a low number of false positives (FP = 1). This suggests that the model operates quite well at distinguishing between cases with tumor and those without.

On the other hand, a comparatively significant number of false negatives (FN = 63) was observed, which indicated that some tumor cases were incorrectly categorized as non-tumor. Where precision is preferred over recall, which suggests that the model frequently uses a conservative prediction strategy. Despite reducing false positives, this tendency may lead to missed tumor detections, which a major problem in clinical applications.

Under the Non-IID setting, the observe a small decrease in performance. As shown in Figure 3, the number of false positives rises (FP = 5), while the true negatives fall slightly (TN = 395), which indicates less consistency in identifying non-tumor samples. The data heterogeneity among clients, which may result conflicting updates throughout the aggregation process, that cause of this behavior.

Notably, the number of false negatives (FN = 63) is same in both settings, demonstrating that the model maintains a stable sensitivity to tumor detection, regardless of the data distribution. This observation shows the robustness of proposed framework, especially with the use of proximal regularization (FedProx), which helps mitigate the negative effects of Non-IID data.

Furthermore, when looking at the confusion matrices at subsequent training rounds (e.g., round 40), it shows that the model maintains consistent classification behavior with little variation. This indicates that the model converges at an early stage and preserves its performance over subsequent communication rounds.

Finally, the confusion matrix analysis shows that the suggested FL model performs reliably in identifying non-tumor cases, while reliably recognizing tumor classification performance, with very slight degradation under Non-IID conditions.

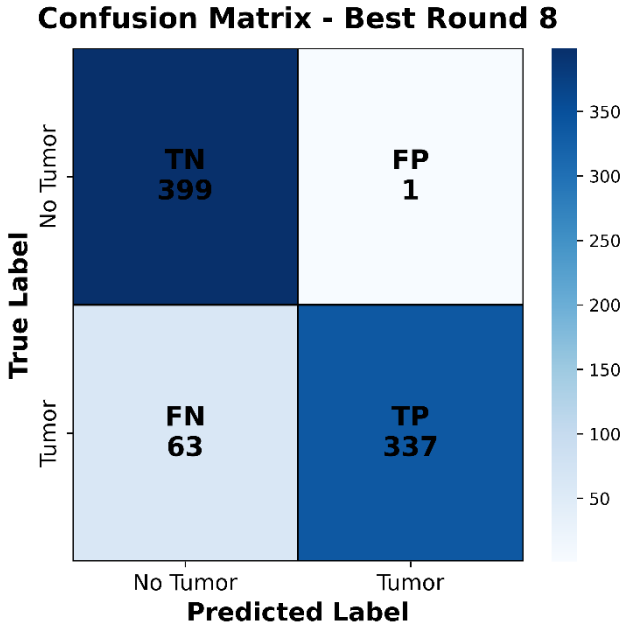

Figure 3: *Confusion matrix for ResNet18 under IID settings for the best Round*

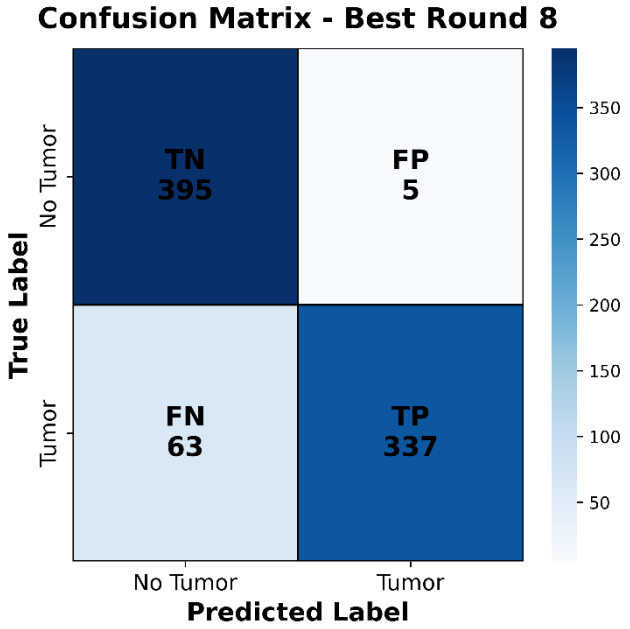

Figure 4: *Confusion matrix for ResNet18 under Non- IID settings for the best Round*

4.3 Convergence Behavior and Training Stability

Figures (5) and (6) show the evaluation of the convergence and training stability for the proposed system under both IID and Non-IID data distributions. This helps in analyzing the evolution of global metrics, including accuracy, precision, recall, and F1-score, together with the loss values across communication rounds, which is better understands the model's training behavior.

Figure 5 illustrates the IID setting; the model converges consistently and quickly. This shows a clear increase in global accuracy in the early rounds, increasing at round 2 from about 75% to over 90%. The performance stabilizes, after this stage, it achieves accuracy remaining around 91%, and minimal fluctuations. A similar pattern is shown by the F1-score, which indicates balanced classification performance. A value near 1.0, the precision quickly approaches, which indicates an extremely low false positive rate. On the other hand, recall has stabilized around 0.82, which indicates a consistent, but slightly conservative detection of cases of tumors.

This behavior is also supported by the loss curve, in the initial rounds, which displays a sharp decrease, followed by small fluctuations, and indicates steady training. The results show that the model learns important features early in training, and then reaches a stable stage. The slight variations following convergence reflect that the aggregation process is stable and is not heavily affected by noise.

In contrast, Figure 6 illustrates the non-IID setting, where the model convergence is slower and less stable, because of variations in the distribution of data under clients. When compared to the IID setting, the global metrics fluctuate more throughout the early and middle communication rounds, but overall performance is still competitive. This behavior is mainly

caused by variations between client domains and unbalanced data distributions, which lead to inconsistent local updates. Despite these limitations, after around 8–10 communication rounds, the model can stabilize, with relatively consistent performance.

The FedProx is also relevant in this process, as it limits local updates and contributes to enhancing stability during aggregation. Consequently, there is a small performance difference between IID and Non-IID settings.

The loss curves in the Non-IID scenario are also a little more varying than in IID environments, as it is harder to optimize in heterogeneous data conditions. However, the indicators of instability or divergence are observed, which means that the suggested framework remains robust.

Generally, the model is stable and reliable with Non -IID conditions and converges more rapidly and smoothly when IID conditions are considered. These results showed the capability of the proposed framework to deal with real-world FL scenarios, where data distributions are often heterogeneous between clients.

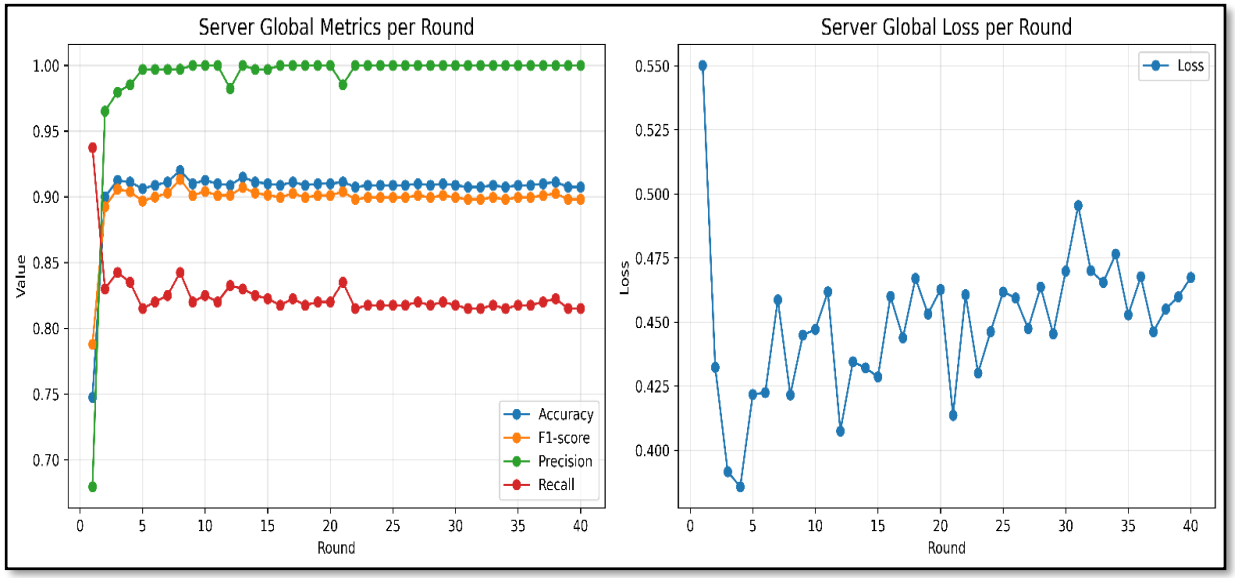

Figure 5: *Training convergence behavior of the ResNet18 under IID settings*

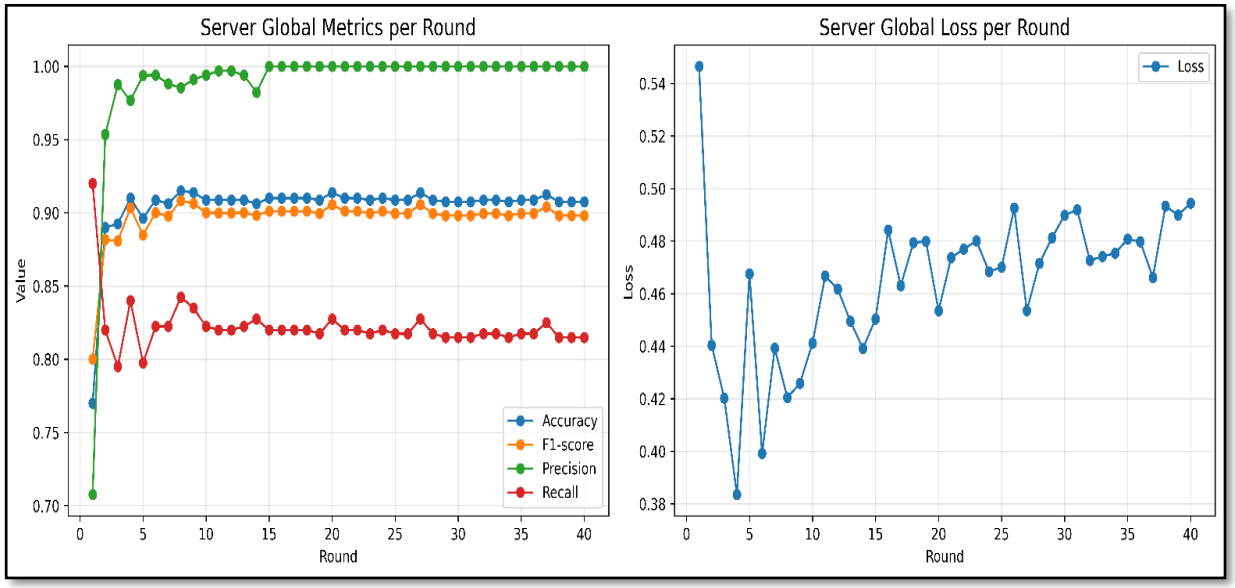

Figure 6: *Training convergence behavior of the ResNet18 under non-IID settings*

4.4 Client-Level Performance Analysis

As shown in Figures 7 and 8, this section examines the performance of individual clients within the FL framework. The analysis based on validation accuracy, loss, and F1-score between communication rounds to better understand how each client learns when using IID and Non-IID data distributions.

The clients in the IID setting (Figure 7) demonstrate consistent and well-aligned learning behavior. Most clients achieve a rapid improvement in validation accuracy, then a steady convergence at high performance levels. This is also true of the F1-score, which is equalized over clients. The loss curves also decrease fast in the early rounds, and then, it reaches stability, similar to efficient optimization and minimal variation between clients. This consistency points out to the fact that reducing differences between local models and ensuring fair learning may be achieved by consistent data distribution.

On the other hand, due to disparities in data distribution, the Non-IID setting (Figure 8) introduces noticeable variation in client-level performance. Although the convergence behavior is not as predictable across clients, accuracy and F1-score generally improve over time. Some clients converge rapidly, while others show fluctuations or need more rounds to stabilize. This difference is more observed in the loss curves, where some clients show oscillation across multiple rounds, or have higher loss values.

The skewness and imbalance in local datasets can explain these differences due to the exposure of some clients to skewed and small samples. As a result, inconsistent updates during aggregation may result from local models learning representations that are less in line with the global objective.

Nonetheless, the global model can successfully transfer knowledge among the participants, as all clients gradually achieve the same performance levels in later rounds. The convergence of values of the F1-score level through time suggests that the aggregation process helps reduce local biases and more generalized features.

Moreover, the FedProx is significant in stabilizing client updates by restricting large deviations of the global model. This helps reduce the impact of client drift and improves convergence stability, especially in the Non-IID scenario.

Overall, the results indicate that the suggested framework offers stable and consistent learning when the conditions are IID, whereas it is robust when the conditions are Non-IID. This indicates that the proposed approach can handle real-world FL scenarios where the data is distributed, and there is heterogeneity across clients.

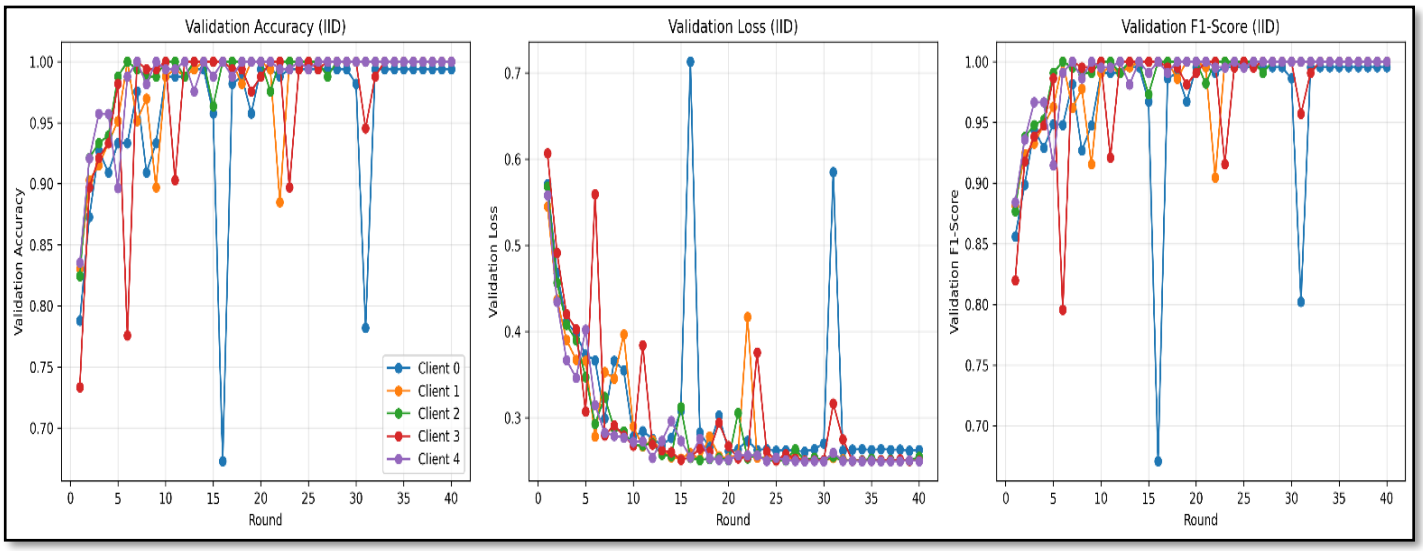

Figure 7: *Clients Performance for ResNet18 under IID settings*

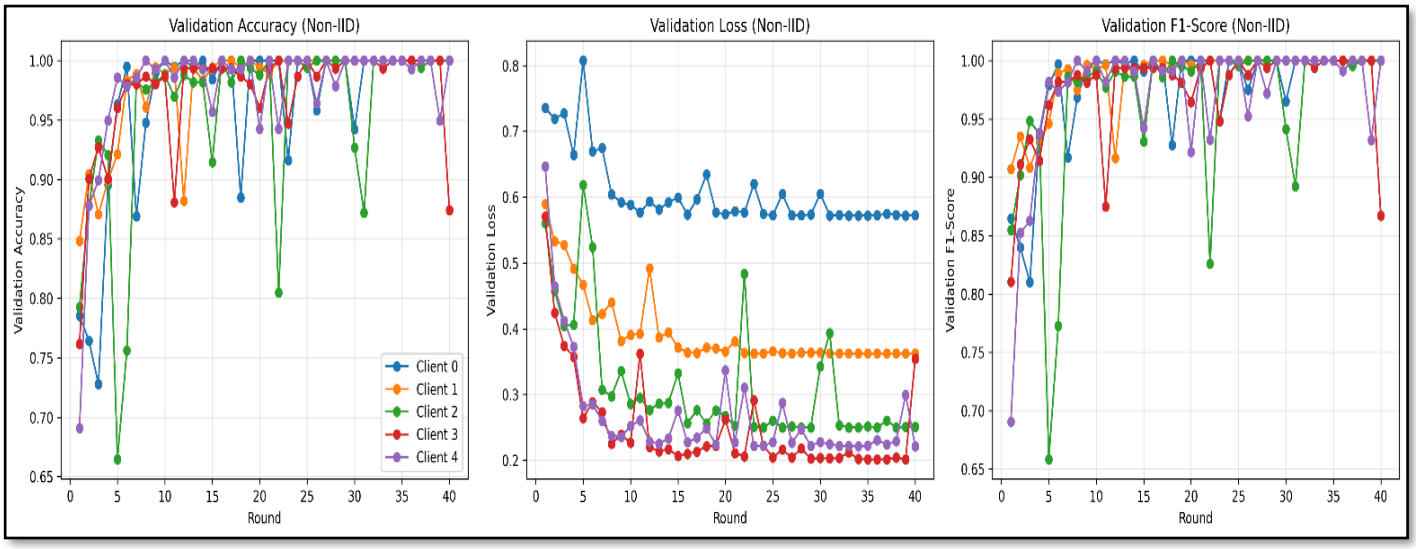

Figure 8: *Clients Performance for ResNet18 under Non-IID settings*

4.5 Model Interpretability and Visual Analysis (Grad-CAM)

The Gradient-weighted Class Activation Mapping (Grad-CAM) approach was implemented to visualize the regions of the image that influence what the model predicts, and thus, improves the interpretability of the proposed FL framework. This qualitative analysis is a complement to the quantitative results, in which the ResNet18 global model attained an accuracy of about 91–92%, with the F1-score of over 0.90 under IID conditions and slightly lower, more stable performance under Non-IID conditions.

The Grad-CAM maps also confirm these results, as indicated in Figures 9 and 10. In no-tumor conditions, the model predicts the class with high confidence (often above 98%), and the heatmaps of the conditions indicate weak and diffuse activations throughout the brain.

With the absence of localized abnormalities, this pattern is still consistent and suggests that the model does not rely on irrelevant focal regions when identifying normal images. In the case of tumors, the model also generates predictions with high confidence (typically above 90%), which is also consistent with the recall value in the confusion matrix.

The Grad-CAM maps discover greater and more concentrated activations in relevant regions of the MRI scans, which reveal that the model is learning features that are associated with the presence of tumor. Even though these activations still provide useful localization cues, they are not strictly confined to exact tumor boundaries, which is expected, so the model performs classification rather than segmentation.

When comparing IID and Non-IID settings, the same interpretability is only due to the differences in data distribution. The Activation maps continue to depend on meaningful anatomical regions, although the Non-IID setup adds a degree of variation in the performance metrics. This indicates that the FL process does not change the model generalization to heterogeneous or distributed datasets.

Notably, the visual explanations also show that the model is sensitive to the internal structures of the brain, rather than variations related to the acquisition or background artifacts. Overall, both the quantitative metrics (accuracy ≈ 91%, F1-score ≈ 0.90) and the Grad-CAM visualization illustrate that the model is robust and reliable. These results demonstrate that it can be used in clinical medical image applications, so the predications base on clinically relevant features.

It should be noted that the proposed framework addresses image classification rather than tumor segmentation. Therefore, quantitative localization metrics such as IoU, overlap score, or localization accuracy were not computed because pixel-level tumor annotations were not available in the utilized datasets. Instead, Grad-CAM was employed as a qualitative explainability tool to visualize model attention regions and assess whether predictions rely on clinically meaningful anatomical structures.

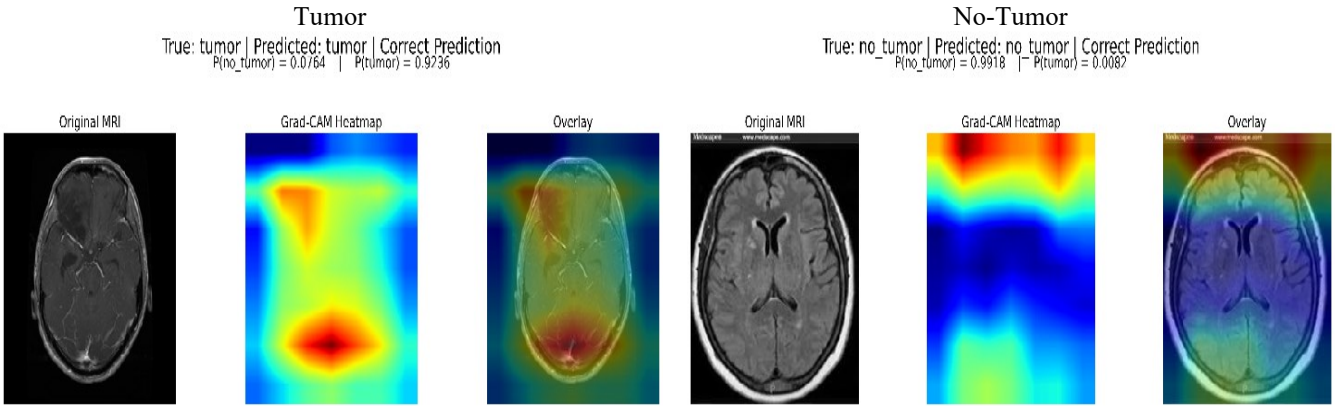

Figure 9: *Grad-CAM for best model (ResNet18) under IID settings*

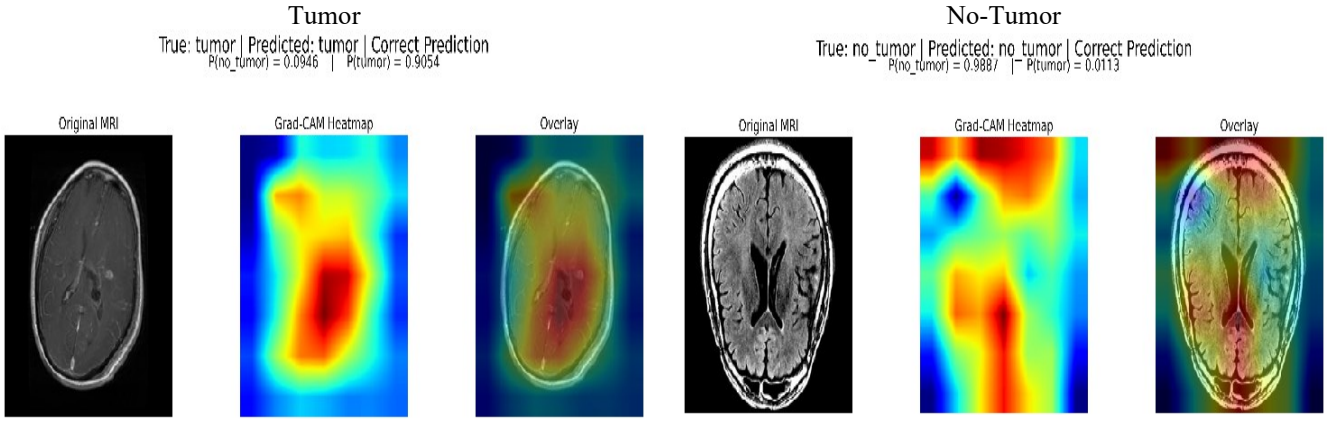

Figure 10: *Grad-CAM for best model (ResNet18) under non-IID settings*

4.6 Discussion and Key Findings

To assess the performance of the proposed system in the case of classifying brain tumors, the following section includes a discussion of the experimental results, based on both quantitative metrics and qualitative observations. The results show that the ResNet18-based model achieves strong performance.

For IID conditions, the model reaches an accuracy of about 91–92% and an F1-score close to 0.90. These results indicate that the model can be trained to learn features that can distinguish between tumor and no-tumor MRI images. Under Non-IID conditions, only a slight reduction in performance was observed despite heterogeneous client distributions. This behavior suggests that the proposed framework is capable of maintaining stable global learning even when local client data are imbalanced and statistically inconsistent. One important contributing factor is the use of FedProx regularization, which constrains excessive deviation of local updates from the global model. By reducing client drift, FedProx helps improve aggregation stability and limits performance degradation under heterogeneous training conditions. As a result, the proposed framework preserves competitive accuracy and stable convergence in realistic distributed medical environments.

The convergence analysis further demonstrates that the model is an efficient learner over communication rounds. In the IID case, the training process is smooth, with only a few differences in accuracy and loss, which denotes consistent learning behavior. Due to the imbalanced and skewed distribution of client data, the model in the Non-IID case exhibits greater variation in the initial rounds. However, the stability is gradually achieved during later communication rounds, which means that the process optimization can reduce the drift of the client and gradually local updates align toward a shared global goal.

At the client level, there is a clear difference between IID and Non-IID. In the IID case. Clients have very similar learning patterns with consistent performance. On the other hand, Non-IID clients exhibit different convergence speeds and more

noticeable fluctuations. However, all clients tend to reach comparable performance levels over time, indicating that the aggregation would facilitate effective knowledge sharing among clients. This implies that the suggested approach can deal with heterogeneous data without significantly affecting generalization.

The confusion matrix analysis will give further insights into the behavior of the model. The results show that the tumor class has a higher number of false negatives compared to the no-tumor class, which has very few false positives. This implies that the model is conservative in its tumor prediction, which could be because of the complexity of the patterns of the tumors or data imbalance. However, the F1-score is strong, which also suggests that a good balance between recall and precision.

In addition, the visualizations of Grad-CAM support the decision-making in the model. In tumor cases, the heatmaps show concentrated activations in relevant regions. While the activations are more dispersed in no-tumor cases. This indicates that the model is based on meaningful anatomical features in making predictions.

Another important observation relates to the use of heterogeneous multi-source datasets. The proposed framework integrates MRI images originating from different repositories and image formats (e.g., JPG and PNG), which introduces variability in imaging characteristics and data distributions. Although such heterogeneity increases learning complexity, the experimental results indicate that the federated model can successfully learn generalized tumor representations across sources. This finding highlights the potential applicability of the proposed framework in realistic clinical scenarios, where medical institutions often rely on heterogeneous imaging protocols and distributed datasets.

Despite variations in data distribution, and the behavior of model, the model exhibits consistent behavior in both IID and Non-IID settings, which makes it learn stable and clinically relevant features. Although, preserve data secret and dealing with distributed datasets, the results indicate that FL can achieve good classification accuracy. The model' ability makes suitable for realistic medical applications to generalize across different data sources (e.g., PNG and JPG formats).

Overall, this system shows a balance between accuracy, stability, and interpretability, which shows that federated deep learning is a useful and scalable approach for privacy preserving medical images analysis. Despite the promising results, several limitations should be considered. Initially, only high-grade and low-grade tumor cases from the Mendeley dataset were used in the data selection process, lesion-like cases were excluded, which may introduce selection bias. Second, some spatial information may be lost when converting 3D MRI volumes into 2D slices. Additionally, the preprocessing steps, especially the conversion of NIfTI-to-PNG, adds extra computational cost. Lastly, in real-world deployment, such preprocessing steps are expected to be performed locally, which may result in additional computational requirements at each client (e.g., hospitals).

While ResNet-18 performs strong performance in both IID and Non-IID settings, it is important to consider other architectures such as Vision Transformers (ViTs). ViT-based models have shown promising results in capturing global contextual relationships in images. However, they typically need higher computational resources, and large-scale datasets, which may not be suitable for FL environments with communication constraints and distributed clients. In contrast, ResNet-18 provides a more stable and effective federated training solution because of its lightweight architecture and ability to extract local spatial features effectively. Therefore, while Vision Transformers are a promising direction, integrating them into FL approaches is remain challenging and will require more research.

In federated learning settings, system performance is affected by factors, including data heterogeneity, communication cost, and model convergence. First, data heterogeneity is a critical role, so clients may have different data distributions (IID vs. Non-IID), which can lead to difference in updating in local model and affect global aggregation. This was observed in the Non-IID scenario, where higher variability and slightly lower performance were observed. Second, communication cost is an important factor in FL systems, as model parameters must be sharing between clients and the central server across multiple communication rounds. The use of a lightweight architecture, like ResNet-18, helps lower communication overhead in comparison to more complex models, although this study does not explicitly minimize communication efficiency. Finally, model convergence behavior is an essential indicator for training stability. Despite initial fluctuations, the results show that under IID conditions the proposed model converges efficiently, and under Non-IID settings remains stable. This demonstrates the robustness of the proposed framework in dealing with realistic distributed scenarios.

4.7 Comparison with Current Approaches

Table 5 presents a comparison between the proposed model and previous studies in the field of FL for brain tumor classification. The comparison includes different aspects like model architecture, number of clients, training rounds, data distribution settings, and reported performance metrics.

In many previous studies, very high accuracy scores were reported, which are sometimes above 97% and, in some cases, close to 99%. However, the majority of these results were achieved under simplified conditions, such as IID data distributions, centralized training, or the use of a single dataset. In contrast, this system is tested in more realistic conditions by combining the FL with heterogeneous (Non-IID) data from different sources.

The integration of multi-source datasets, which involve MRI images from different sources (Kaggle Brain Tumor and Mendeley Data), is a significant addition to this research. This setup increases the diversity in data distribution, acquisition conditions, and image characteristics. Despite this complexity, the model performs competitively, demonstrating high generalization under diverse sources of data, with an accuracy of roughly 92% under IID settings, and 91.5% in Non-IID cases.

In distributed medical settings, the proposed system provides a more realistic evaluation compare to previous studies that based on single datasets and controlled environments. Additionally, the integration of Grad-CAM enhances model interpretability by visualizing model decisions, which improves transparency and confidence in the results, when many other approaches focus just on performance. Another important aspect, the proposed framework allows for data privacy, enables training collaboratively instead of patient data sharing, unlike a centralized system. This is particularly important in medical applications, while data confidentiality is a major concern.

In general, while the achieved accuracy is slightly lower than that of some highly optimized centralized models, the suggested system exceeds them in terms of robustness, scalability, and practical application. Its ability to maintain steady performance under Non-IID conditions and across multiple datasets highlights its relevance for real-world deployment. These results show that the proposed system provides a good balance between performance and usability, making it an acceptable alternative for the analysis of privacy-preserving in medical images.

The reviewed studies differ considerably in terms of datasets, number of participating clients, communication rounds, prediction tasks, evaluation protocols, and data distribution settings (IID vs. Non-IID). Consequently, reported performance metrics may not be directly comparable across studies. Therefore, the comparison in this work considers not only accuracy values but also experimental realism, data heterogeneity, privacy preservation, interpretability mechanisms, model architecture, and federated training conditions. This broader perspective enables a more balanced and realistic assessment of the proposed framework relative to existing approaches.

Table 5: *COMPARISON WITH EXISTING METHODOLOGIES.*

Reference	Dataset	No of clients	No of Rounds	Accuracy	Loss	Model	Data Distributed	Grad-CAM
[5]	Brain Tumor MRI Dataset	10	N/A	98	N/A	VGG16	IID	no
[14]	Brain Tumor MRI Dataset	10	50	94.24	0.18	GoogLeNet	IID	yes
[9]	Brain Tumor MRI Dataset	5 5	10	99,76 99,71	0.0107 0.0107	MobileNetV2	IID Non-IID	no
[10]	Brain Tumor MRI Dataset	4	10	98	N/A	ResNe-18	IID	yes
[11]	Brain Tumor MRI Dataset	5	10	97.19	N/A	CNN MODEL	Non-IID	no
[12]	Brain Tumor MRI Dataset	4	16	98.4	N/A	VGG-16	IID	no
[15]	Brain Tumor MRI Dataset	5	50	92 90 88	N/A N/A N/A	ResNe-18	Centralized CNN Non-IID Non-IID	yes
Proposed research	Brain Tumor MRI Dataset + Brain lesion MRI and co-related MRS Spectroscopy Dataset	5	40 40	92 91.5	0.42 0.42	ResNet18	IID Non-IID	yes

5. Conclusion

In this paper, the use of an FL system based on a deep learning approach is presented as a robust and privacy-preserving which is to classifying brain tumor for MRI data. To collaborate across multiple clients on FL instead of sharing sensitive data, the ResNet18 model was combined into a decentralized training paradigm. This research integrated heterogeneous multi-source data from two sources: one from Kaggle Brain Tumor (in JPG format) and another from Mendeley Data (in NIfTI converted into PNG formats), and evaluated the model under both IID and Non-IID contexts. This context is offering

a more realistic representation of clinical environments than many earlier studies that use centralized training or single-source datasets.

The experimental results demonstrated that the model achieved steady and reliable performance, with an F1-score of about 0.90 and accuracy close to 92% in IID cases and 91.5% in Non-IID conditions. The client-level analysis shows that the system can deal with differences in the distribution of data, while the convergence results show that the training process remains stable. The balanced classification performance was reflected by the confusion matrix, especially in recognizing non-tumor cases.

Furthermore, Grad-CAM visuals verify that the model depends on relevant brain regions, which increases confidence and improves prediction interpretability. A major contribution to this work, is the integration of FL with multi-source medical datasets, which adds substantial heterogeneity and improves the capacity of model for generalization. This model retains competitive performance despite the task's increased complexity, proving its robustness and suitability for privacy-sensitive medical applications in practical implementation.

For future work, several directions can be explored. In addition to tumor segmentation, the proposed framework can be extended to multi-class brain tumor classification tasks, which would further evaluate its scalability and generalization capabilities in more complex clinical scenarios. Generalization may also be enhanced by incorporating larger multi-institutional datasets. Furthermore, integrating advanced privacy-preserving techniques such as differential privacy, along with more comprehensive explainability methods and clinical validation, and repeated experimental evaluation with additional statistical validation measures (e.g., standard deviation and confidence interval analysis), could further strengthen the robustness and reproducibility of the system. Finally, exploring adaptive and personalized federated learning strategies may help address client heterogeneity.

Conflicts of Interest

The authors declare no conflicts of interest.

REFERENCES

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

[2] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[3] N. Rieke *et al.*, "The future of digital health with federated learning," *NPJ Digit. Med.*, vol. 3, no. 1, Dec. 2020, doi: 10.1038/s41746-020-00323-1.

[4] W. Li *et al.*, "Privacy-preserving federated brain tumour segmentation," in *International workshop on machine learning in medical imaging*, Springer, 2019, pp. 133–141.

[5] E. Albalawi *et al.*, "Integrated approach of federated learning with transfer learning for classification and diagnosis of brain tumor," *BMC Med. Imaging*, vol. 24, no. 1, Dec. 2024, doi: 10.1186/s12880-024-01261-0.

[6] M. Islam, M. T. Reza, M. Kaosar, and M. Z. Parvez, "Effectiveness of federated learning and CNN ensemble architectures for identifying brain tumors using MRI images," *Neural Process. Lett.*, vol. 55, no. 4, pp. 3779–3809, 2023.

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

[8] J. B. Awotunde, C. O. Abikoye, B. Brahma, E. T. Oladipupo, and A. Bandyopadhyay, "Federated learning augmented with convolutional neural networks for brain cancer classification," *Procedia Comput. Sci.*, vol. 258, pp. 2617–2626, 2025.

[9] S. Sharma *et al.*, "A privacy-preserved horizontal federated learning for malignant glioma tumour detection using distributed data-silos," *PLoS One*, vol. 20, no. 2 February, Feb. 2025, doi: 10.1371/journal.pone.0316543.

[10] S. Anoosha and B. Seetharamulu, "Federated Learning-Based ResNet-18 Model for Brain Tumor Classification in MRI Scans," *Ingénierie des systèmes d information*, vol. 30, no. 8, pp. 2021–2031, Aug. 2025, doi: 10.18280/isi.300808.

[11] N. Sivakumar *et al.*, "A Hybrid Brain Tumor Classification Using FL With FedAvg and FedProx for Privacy and Robustness Across Heterogeneous Data Sources," *IEEE Access*, vol. 13, pp. 57705–57719, 2025, doi: 10.1109/ACCESS.2025.3549440.

[12] G. Appasami and N. Savarimuthu, "Federated learning for secure medical MRI brain tumor image classification," *European Physical Journal: Special Topics*, Oct. 2025, doi: 10.1140/epjs/s11734-025-01516-z.

[13] M. D. Z. Muntaqim and T. A. Smrity, "Federated learning framework for brain tumor detection using MRI images in non-IID data distributions," *Journal of Imaging Informatics in Medicine*, vol. 38, no. 6, pp. 3909–3927, 2025.

[14] Q. U. A. Mastoi *et al.*, "Explainable AI in medical imaging: an interpretable and collaborative federated learning model for brain tumor classification," *Front. Oncol.*, vol. 15, 2025, doi: 10.3389/fonc.2025.1535478.

[15] S. Gupta, M. Gupta, R. Kumar, and A. Abraham, "A Federated Learning and Explainable AI Framework for Privacy-Preserving Brain Tumor Diagnosis Using Multi-Institutional MRI Data," *IEEE Access*, 2026.

[16] R. J. Gohari, L. Aliahmadipour, and E. Valipour, "FedBrain-Distill: Communication-Efficient Federated Brain Tumor Classification Using Ensemble Knowledge Distillation on Non-IID Data," International Conference on Computer and Knowledge Engineering (ICCKE), Nov. 2024, pp. 49–54. [Online]. Available: http://arxiv.org/abs/2409.05359

[17] A. Al-Saleh, G. G. Tejani, S. Mishra, S. K. Sharma, and S. J. Mousavirad, "A federated learning-based privacy-preserving image processing framework for brain tumor detection from CT scans," *Sci. Rep.*, vol. 15, no. 1, p. 23578, 2025.

[18] M. A. A. Jabbar, A. S. Jaddoa, and U. S. Mahmoud, "A Systematic Review of Federated Learning: Emerging Techniques, Challenges, and Research Directions," *Iraqi Journal for Computers and Informatics*, vol. 51, no. 2, pp. 86–97, 2025.