Federated Learning for Brain Tumor Classification from MRI: A Comparison of MLP and ConvNeXt Approaches under IID and non-IID Data Scenarios

Giustino Claudio Miglionico
University of Pisa
giustino.miglionico@phd.unipi.it

Michela Fazzolari
CNR - IIT
m.fazzolari@iit.cnr.it

Pietro Ducange
University of Pisa
pietro.ducange@unipi.it

Francesco Marcelloni
University of Pisa
francesco.marcelloni@unipi.it

Fabrizio Ruffini
University of Pisa
fabrizio.ruffini@unipi.it

Abstract

Federated Learning (FL) effectively addresses privacy concerns in medical imaging by enabling collaborative model training without sharing sensitive patient data. This paper compares two neural network approaches for Brain Tumor Classification (BTC) from Magnetic Resonance Imaging (MRI) in a FL setting. Both models operate on ROIs (Regions of Interest) extracted from brain MRI scans. The first is a lightweight Multi-Layer Perceptron (MLP) that classifies ROIs based on radiomic features extracted from them. The second is a Deep Learning (DL) approach based on the ConvNeXt architecture, which performs classification directly on the ROI images. Two experimental scenarios are considered: a balanced (IID) and an unbalanced (non-IID) distribution of data among federated clients. Results show that the radiomics-based MLP achieves performance comparable to the more complex ConvNeXt model, while requiring significantly lower computational resources. Moreover, FL consistently outperforms isolated local training, particularly under non-IID conditions, emphasizing its potential for clinical deployment.

Keywords: Federated Learning, Brain Tumor Classification, Magnetic Resonance Imaging, Radiomics, Artificial Neural Networks

1. Introduction

The rapid advancement of Artificial Intelligence (AI) and Machine Learning (ML) in medical imaging has substantially improved clinical decision support systems, particularly in the critical task of BTC. Accurate and timely diagnosis of brain tumors is essential, given their high mortality and morbidity rates worldwide. MRI is the gold standard imaging modality for assessing such tumors due to its superior contrast and detailed anatomical representation, providing rich quantitative information for robust tumor characterization (Ranjbarzadeh et al. (2023)).

In recent years, DL architectures, especially Convolutional Neural Networks (CNNs), have achieved remarkable performance in BTC, often surpassing traditional ML methods by directly extracting informative patterns from imaging data (Younis et al. (2023)). However, these image-based approaches require substantial computational resources and raise concerns about data privacy and centralization when integrating data from multiple healthcare institutions.

FL, a decentralized training paradigm, addresses these challenges by enabling institutions to collaboratively train models without sharing sensitive patient data. Despite the growing interest in healthcare applications (Amin et al. (2025)), comparative studies between traditional neural networks and modern DL architectures in federated contexts remain limited.

This paper presents a comparative study of two neural network approaches for BTC in an FL setting, evaluated in terms of predictive performance and model complexity. The first approach employs a lightweight MLP trained on radiomic features extracted from ROIs, while the second leverages a modern DL model based on the *ConvNeXt* architecture, fine-tuned directly on MRI ROIs.

Two scenarios are considered to reflect realistic clinical data distributions: a balanced (IID) and an unbalanced (non-IID) setting. The outcomes provide insights into trade-offs between performance, computational cost, and data distribution challenges, offering guidance for deploying AI models in federated

URI: https://hdl.handle.net/10125/111630
978-0-9981331-9-5

clinical environments.

The remainder of the paper is organized as follows: Section 2 reviews the state of the art on brain tumor classification in federated settings. Section 3 describes the proposed models. In Section 4, the envisioned federated BTC environment is presented. Section 5 details the dataset, experimental scenarios, and training process. Section 6 presents the results, and Section 7 discusses conclusions and future directions.

2. State of the art

The adoption of DL, particularly CNNs, has substantially advanced automated brain tumor classification. Architectures such as VGG16, ResNet, DenseNet, and EfficientNet have achieved strong performance in medical imaging by leveraging powerful feature extraction and hierarchical representations, clearly surpassing traditional ML methods (Albalawi et al. (2024) and Kalpana et al. (2023)).

More recently, *ConvNeXt* (Liu et al. (2022)) has emerged as an evolution of CNN architectures, integrating innovations inspired by vision transformers. It has demonstrated excellent performance in medical imaging, including BTC, combining high accuracy with improved computational efficiency (Mehmood and Bajwa (2024)). Thanks to this balance, *ConvNeXt* is well-suited for FL, where resource constraints and communication efficiency are critical. For instance, Viet et al. (2023) successfully integrated *ConvNeXt* into an FL framework for brain tumor classification, achieving high performance under both IID and non-IID conditions.

FL itself has attracted growing attention for enabling collaborative training without sharing sensitive data, addressing privacy concerns while supporting generalizable diagnostic models (Guan et al. (2024)). Nonetheless, FL faces important challenges, especially the presence of non-Independent and Identically Distributed (IID) data across institutions, which can hinder convergence and generalization (Narula et al. (2024)).

In parallel, MLP models applied to radiomic features have shown promising results in BTC, capturing complex nonlinear relationships among diagnostic variables (Kale et al. (2024) and Xia et al. (2025)). For example, (Xia et al. (2025)) combined radiomic feature extraction with MLP and other classical ML models, achieving high accuracy in distinguishing tumor from non-tumor scans.

Despite these advances, direct comparisons between radiomics-based neural networks (e.g., MLPs) and modern image-based architectures (e.g., *ConvNeXt*) in FL remain scarce. Our work addresses this gap by evaluating the performance, robustness, and computational trade-offs of these two approaches under both balanced (IID) and realistically unbalanced (non-IID) scenarios, providing insights for their practical deployment in clinical settings.

3. Preliminaries

This section outlines the methodological background of our study. We first describe the standard pipeline for brain BTC from MRI images, then introduce the radiomic features used to encode medical data into a structured format. We then present the two models adopted as baselines: a lightweight MLP classifier and the *ConvNeXt* DL architecture.

3.1. Standard pipeline for BTC from MRI

A widely used approach in Computer-Aided Diagnosis (CAD) systems for BTC using MRI images, as described in (Muhammad et al. (2020)), involves several steps, namely (i) image acquisition, (ii) image pre-processing, (iii) image segmentation, (iv) feature extraction, (v) feature selection and dimensionality reduction, (vi) image classification. In the following, these steps are described in detail.

First, the images are acquired and collected. Then, the images are pre-processed to improve their quality.

Next, the segmentation step is performed, either manually or automatically. In the first case, an expert radiologist identifies and delineates the ROIs; in the second case, algorithms are applied to automatically detect and segment suspicious areas without human supervision (Yu et al. (2024)).

Once the ROIs are segmented, the feature extraction phase begins. In this phase, quantitative features are computed from the segmented areas to characterize the tumor tissue. These features can follow two main approaches. In the first step, interpretable features, commonly known as radiomic features, are computed using algorithms that quantify quantitative aspects of the image. These features provide transparency and can be directly associated with biological or anatomical properties. In the second approach, non-interpretable features are extracted automatically by the convolutional layers of DL models.

An optional next step involves feature selection and dimensionality reduction to reduce overfitting, enhance efficiency, and highlight the most relevant features, thereby improving model accuracy and interpretability.

In the classification phase, supervised or unsupervised models can be used to categorize segmented lesions, distinguishing between benign and

malignant tumors or between different tumor types and severity levels.

As a result of these steps, a CAD system, based on MRI image analysis, can provide a preliminary diagnosis of the tumor, visually presenting the results to support clinical interpretation.

3.2. Radiomic features

Before medical images can be processed by an MLP, they need to be converted into a structured numerical format. In this study, radiomics is employed to extract quantitative features from the images, enabling their use in the learning model.

Radiomic features describe an image as a feature vector that captures aspects of the visual information, including pixel intensities, spatial relationships between pixels, morphology of regions of interest, and textural properties. This abstract representation of the image is particularly useful in the medical field, where subtle local variations in texture or intensity distribution can be correlated with different pathological conditions (Bera et al. (2022)).

In this paper, radiomic features are extracted following the pipeline proposed by (Carré et al. (2020)), which ensures a consistent and reproducible process. In particular, we used PyRadiomics[1], an open-source Python library widely adopted by the scientific community for its flexibility and compliance with the Imaging Biomarker Standardization Initiative (IBSI) guidelines (Zwanenburg et al. (2020)). By adhering to these standards, we ensured that the extracted features are comparable across studies and reproducible in diverse clinical settings.

Images were first Z-score normalized to reduce intensity variability across samples, then discretized to facilitate textural matrix computation. Extracted features include first-order statistics and higher-order descriptors from GLCM, GLRLM, and GLSZM, capturing voxel intensity distributions, sequence lengths, and homogeneous area sizes to characterize image texture.

Typically, more than 110 radiomic features can be extracted from a medical image, covering a wide range of morphological, statistical, and textural descriptors. In this work, we selected a specific subset of features, giving priority to geometric features, which describe the morphology of two-dimensional ROIs, and to first-order statistical features, which capture the distribution of pixel intensities within the analyzed regions.

In addition to radiomic features, two spatial features were extracted based on the typical central location of pituitary tumors. Specifically, the average horizontal and vertical coordinates of the pixels within each ROI were computed to compactly represent tumor position.

The features used in the experimental analysis of this study were selected through a thorough evaluation that considered not only their discriminative power but also their interpretability and readability by human experts. A detailed overview of the selected features is provided in Table 1.

Table 1. Description of the selected radiomic features.

Type	ID	Description
Geometric	F_0	The elongation of the ROI
	F_1	The major axis length of the ROI
	F_2	The minor axis length of the ROI
	F_3	Perimeter: the total boundary length of the lesion in the ROI
	F_4	The maximum distance between any two points within the ROI
	F_5	Degree to which the ROI approximates a circle
Pixel Intensity	F_6	The average gray level intensity value across all pixels in the ROI
	F_7	The median gray level intensity value of the pixels in the ROI
	F_8	The minimum gray level intensity value of the pixels in the ROI
	F_9	The maximum gray level intensity value of the pixels in the ROI
	F_{10}	The difference between F_9 and F_8
	F_{11}	A measure of the homogeneity level of the intensity of pixels
Spatial	F_{12}	X-coordinate of the center of mass of the ROI
	F_{13}	Y-coordinate of the center of mass of the ROI
	F_{14}	Anatomical plane in which the ROI is located

3.3. The MLP classifier

As an initial approach to the BTC task, a lightweight neural network was adopted, designed to operate on structured radiomic features extracted from ROIs in MRI images. The model is based on an MLP architecture.

As introduced in Section 3.2, the feature set used in this study includes a curated subset of radiomic and spatial features, selected for their interpretability and relevance for BTC.

The adopted MLP model comprises an input layer with 18 neurons, corresponding to the dimensionality of the encoded feature vector, followed by two fully connected hidden layers, with 32 and 16 neurons, respectively. Each hidden layer is activated via a Rectified Linear Unit (ReLU) function. The final classification layer returns a probability distribution over tumor classes. In this work, as discussed in Section 5.1, we consider three classes, namely meningioma, glioma, and pituitary tumor. Due to its low complexity and small number of parameters, this architecture is particularly suitable for implementation in distributed environments or on devices with limited computational resources. Implementation details and training procedures are described in Section 5.

[1] https://pyradiomics.readthedocs.io/en/latest

3.4. The DL base model: *ConvNeXt*

This study adopts *ConvNeXt* as the DL baseline, a state-of-the-art convolutional architecture for BTC introduced in (Liu et al. (2022)). The model was chosen for its strong empirical performance and widespread use in recent literature. As with the previously described MLP, *ConvNeXt* processes ROIs extracted from MRI scans, aiming to classify each sample into one of three tumor categories.

Architecturally, *ConvNeXt* can be seen as a modern evolution of ResNet (He et al. (2016)). It organizes the network into four hierarchical stages with a computational load distributed in a ratio of 1:1:3:1. The initial convolutional stem of *ResNet* is replaced with a 4×4 convolution with a stride size of 4, improving downsampling efficiency. Each stage uses depthwise separable convolutions and inverted bottleneck structures with an expansion factor of 4, reminiscent of the feedforward blocks typical of Transformer architectures. To emulate the large receptive fields of attention mechanisms, *ConvNeXt* uses large depthwise convolutions of 7×7. It also replaces batch normalization with layer normalization and swaps ReLU activations with GELU, which introduces smoother nonlinearities. The block design is also simplified, reducing the number of activation and normalization layers per block. Downsampling between stages is handled by dedicated layers, each preceded by a LayerNorm to ensure training stability.

In our study, it serves as a reference architecture. Details about the training protocols, model adaptation, and computational resources used are presented in Section 5.

4. The envisioned federated BTC environment

This section presents the envisioned federated BTC environment. Let's consider a network of autonomous healthcare centers, each equipped with MRI machines and involved in the diagnosis of brain tumors. Within each center, an AI-based diagnostic support application is deployed, designed to perform BTC starting from already segmented images, in which the ROIs have been previously identified (manually by radiologists or automatically by algorithms). Each application includes dedicated modules for federated learning of BTC models and for using them for local inference to classify brain tumors.

The overall architecture of the system is illustrated in Figure 1.

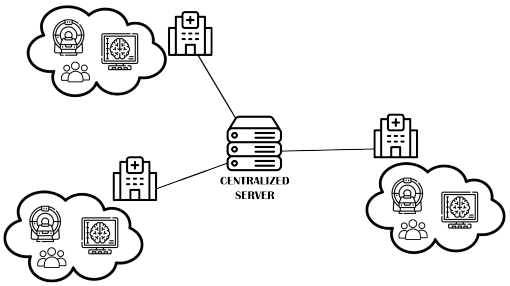

Figure 1. Distributed system architecture for federated brain tumor classification.

Each healthcare center operates autonomously and is managed by different organizations, which prevents the sharing of sensitive patient data. To address this constraint, a FL approach is adopted, enabling each facility to contribute to the construction of a global and shared classification model without exchanging raw data. In this framework, each center trains a *local model* on its own dataset and transmits only the *updated weights* to a central server. The server aggregates these updates to generate a new *federated model*, which is then redistributed to all participating centers as the starting point for the next training round. This iterative process leads to continuous improvement of the global model. The final federated model is subsequently deployed locally at each healthcare center and used to perform brain tumor classification, enabling decentralized diagnosis while preserving data privacy.

5. Experimental setup

This section describes the dataset used, the experimental scenarios designed to simulate realistic clinical conditions, and the training protocols adopted for each model under different learning schemes.

As already introduced in Section 1, the main objective of this work is to compare two neural network models for BTC in an FL context, as illustrated in Figure 2.

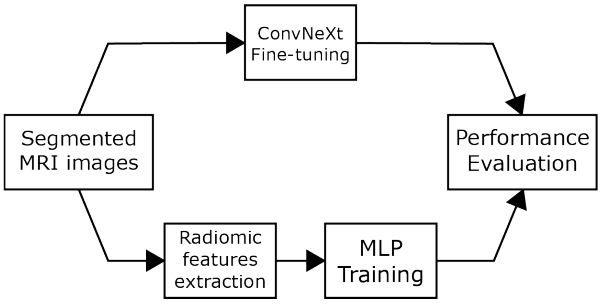

Figure 2. Experimental pipeline used in this study.

The first approach employs an MLP that processes

structured numerical features extracted from ROIs. The second approach adopts a DL model based on the *ConvNeXt* architecture, which directly operates on ROI image data. These two approaches are compared in terms of classification performance and computational efficiency, considering two different scenarios.

To evaluate model performance under varying data-sharing constraints, experiments were conducted across three paradigms: Centralized Learning (CL), FL, and Local Learning (LL) (see Section 5.3 for details).

5.1. Dataset description

For the experimental analysis, the *Brain Tumor Public Dataset* was used. This dataset was initially introduced in (Cheng et al. (2015)) and recently adopted in a number of papers, such as (Miglionico et al. (2024)).

The dataset comprises T1-weighted and contrast-enhanced MRI images, collected from a cohort of 233 patients across two Chinese hospitals between 2005 and 2010. The corpus includes a total of 3,064 axial slices, each with a resolution of 512×512 pixels. The images have a thickness ranging from 1 to 6 mm, with a constant interslice distance of 6 mm. Tumor regions were manually segmented by three experienced radiologists and subsequently annotated into three diagnostic categories: *meningioma* (708 images), *glioma* (1,426 images), and *pituitary tumor* (939 images).

5.2. Experimental scenarios

To simulate the behavior of a network of healthcare centers, the dataset described above was used to construct two realistic experimental scenarios. In each scenario, five healthcare facilities were considered, each equipped with a system on which the BTC application was deployed. Each healthcare facility acts as a client in the FL framework.

The first scenario simulates an FL environment with IID data distribution, where each local node contains data representative of the global problem. Table 2 shows, for each client, the number of training data instances per class and the corresponding number of patients. Each client has a training set of approximately 538 to 558 instances. Most clients have access to all classes, although the distribution is not perfectly uniform.

The second scenario, described in Table 3, introduces a more complex and non-IID data distribution, more closely reflecting the clinical reality. It is observed that at least three healthcare facilities do not have data for one of the three diagnostic classes considered. This heterogeneity reflects common situations in medical practice, where the distribution of pathologies can vary from one facility to another, influenced by geographical factors, specialization of the centers, and size of the catchment area. This variability increases the complexity of training the federated model.

Table 2. Class distribution across clients in Scenario 1 (# instances (# patients)).

Class	Client 1	2	3	4	5
Meningioma	82 (9)	94 (12)	139 (17)	175 (19)	131 (15)
Glioma	346 (19)	229 (17)	245 (15)	254 (13)	252 (15)
Pituitary tumor	113 (7)	217 (13)	171 (11)	109 (10)	175 (11)
Total	541 (35)	540 (42)	555 (43)	538 (42)	558 (41)

Table 3. Class distribution across clients in Scenario 2 (# instances (# patients)).

Class	Client 1	2	3	4	5
Meningioma	160 (19)	0 (0)	146 (17)	142 (18)	173 (18)
Glioma	324 (19)	369 (24)	0 (0)	282 (17)	351 (19)
Pituitary tumor	0 (0)	169 (12)	190 (12)	223 (15)	203 (13)
Total	484 (38)	538 (36)	336 (29)	647 (50)	727 (50)

To evaluate the generalization capability of the federated models in both scenarios, a balanced test set was created by selecting 10 patients per tumor class from the overall dataset, prior to the construction of the two experimental scenarios. This approach represents a more conservative evaluation compared to testing on the original, often unbalanced, distribution, as it requires the model to perform equally across all classes. In a clinical context, where each pathology holds equal diagnostic relevance, this strategy provides a more rigorous and unbiased estimate of model performance.

The detailed distribution of the classes in the test set is reported in Table 4.

Table 4. Class distribution in the test set (# instances (# patients)).

Class	Test Set
Meningioma	87 (10)
Glioma	100 (10)
Pituitary tumor	145 (10)
Total	332 (30)

5.3. Model training

To evaluate the performance of the classification models, three training paradigms were considered: CL, FL, and LL.

In the CL setup, all available training data were aggregated into a single dataset to train a global model. While this configuration does not comply with privacy constraints, it serves as an upper bound for performance comparison.

In the FL scenario, data remained decentralized, with each client retaining its local dataset. A global model was obtained through iterative aggregation of local updates via the Federated Averaging (FedAVG) algorithm. This approach enables collaborative training while preserving data confidentiality by avoiding direct data exchange.

Conversely, the LL scheme involved clients independently training models on their local data, with no communication or aggregation. Although this setup ensures maximum data privacy, it often leads to suboptimal generalization due to limited data diversity and volume.

In all experiments, early stopping was employed to prevent overfitting. For the CL setting, models were trained for a maximum of 1,000 epochs, with training halted if no improvement was observed on the validation set for 50 consecutive epochs. In the FL setting, training was conducted over 1,000 rounds, with each client performing 5 local epochs per round before sharing model updates. For LL, each client trained its model independently for up to 1,000 epochs, also using early stopping with a patience of 50 epochs based on local validation performance.

Both the MLP and *ConvNeXt* models were federated using the FedAVG algorithm (McMahan et al. (2017)). During each federated round, clients train locally on private data and transmit model updates to a central server, where weights are averaged to update the *global model*. FedAVG was selected for its simplicity, widespread adoption, and its prior application in similar BTC tasks using the *ConvNeXt* architecture (Viet et al. (2023)).

The MLP model was trained from scratch across all three paradigms. To address class imbalance, a weighted cross-entropy loss function was employed, with weights derived from the class distribution. Model selection was performed via grid search over hyperparameters including learning rate, number of hidden layers, and units per layer, using validation accuracy as the criterion. The optimal configuration used a learning rate of 0.001.

For the *ConvNeXt* model, the pre-trained ConvNeXt-Tiny backbone from the PyTorch library[2] was adopted, initialized with ImageNet weights. The final convolutional block (Block 3) was fine-tuned, while the remainder of the network was kept frozen. The classification head was replaced with a fully connected layer comprising three output neurons, corresponding to the target classes. Training was performed using the Adam optimizer, with a learning rate of 0.001 and a batch size of 32. To further mitigate the class imbalance problem, the *Focal Loss* loss function with a focus parameter of 2 was used. The class weighting factor was defined as a vector of class-specific weights, calculated as the inverse of the class frequency in each client's local dataset. In particular, the weight assigned to each class was determined by dividing the total number of samples by the number of samples belonging to that class. These weights were then normalized to ensure numerical stability and recalculated at each training round.

All experiments were conducted on a workstation equipped with an NVIDIA RTX 3070 GPU, 64 GB RAM, and a 12th-generation Intel Core i7 processor.

[2]https://pytorch.org/vision/stable/models/generated/torchvision.models.convnext_tiny.html

6. Experimental results

In this section, the results obtained from the experiments described in the previous section are presented and analyzed. The analysis begins with an evaluation of model performance based on standard classification metrics, comparing the three learning paradigms across both experimental scenarios. Particular attention is devoted to the impact of data distribution and training paradigm on model generalization, especially in the presence of non-IID data. A second analysis was conducted to evaluate the complexity of the two models. Specifically, the evaluation considered the number of trainable parameters, disk usage, training time, and inference speed to assess their computational and operational feasibility across different deployment scenarios.

6.1. Accuracy analysis

The results obtained for each scenario, considering the three different learning paradigms, are shown in Table 5 and Table 6.

For each class, in the FL and CL settings, the values of precision, recall, and F1-score metrics are reported, calculated on the common test set. In the case of the LL scheme, the table presents the mean and standard deviation of the same metrics, computed on the test set by evaluating the predictions made by each local model.

Although the FL approach does not reach the same performance levels as CL, it maintains a strong ability to distinguish between the classes. The relatively narrow gap between the two methods suggests that federated models are capable of achieving effective

Table 5. Comparison of MLP Performance in Scenario 1 and Scenario 2, considering CL, FL and LL.

Scenario	Experiment Type	Meningioma			Glioma			Pituitary Tumor		
		Precision	Recall	F1-score	Precision	Recall	F1-score	Precision	Recall	F1-score
1	Centralized	0.80	0.79	0.79	0.92	0.82	0.86	0.90	0.97	0.94
	Federated	0.74	0.68	0.71	0.90	0.76	0.82	0.84	0.97	0.90
	Local AVG	0.68 ± 0.09	0.73 ± 0.05	0.70 ± 0.03	0.86 ± 0.06	0.69 ± 0.08	0.75 ± 0.04	0.82 ± 0.04	0.87 ± 0.06	0.84 ± 0.02
2	Centralized	0.78	0.87	0.82	0.93	0.80	0.86	0.93	0.96	0.95
	Federated	0.80	0.82	0.81	0.89	0.83	0.86	0.91	0.93	0.92
	Local AVG	0.46 ± 0.26	0.64 ± 0.33	0.52 ± 0.27	0.61 ± 0.33	0.60 ± 0.30	0.59 ± 0.30	0.61 ± 0.32	0.76 ± 0.38	0.68 ± 0.34

Table 6. Comparison of ConvNeXt Performance in Scenario 1 and Scenario 2, considering CL, FL and LL.

Scenario	Experiment Type	Meningioma			Glioma			Pituitary Tumor		
		Precision	Recall	F1-score	Precision	Recall	F1-score	Precision	Recall	F1-score
1	Centralized	0.85	0.85	0.85	0.84	0.86	0.85	0.90	0.88	0.89
	Federated	0.82	0.83	0.83	0.79	0.83	0.80	0.87	0.84	0.85
	Local AVG	0.77 ± 0.08	0.76 ± 0.05	0.76 ± 0.03	0.74 ± 0.06	0.65 ± 0.10	0.68 ± 0.04	0.75 ± 0.04	0.79 ± 0.08	0.76 ± 0.04
2	Centralized	0.85	0.85	0.85	0.84	0.86	0.85	0.90	0.88	0.89
	Federated	0.87	0.80	0.83	0.75	0.78	0.76	0.80	0.82	0.81
	Local AVG	0.51 ± 0.27	0.66 ± 0.33	0.56 ± 0.28	0.61 ± 0.33	0.54 ± 0.30	0.55 ± 0.27	0.54 ± 0.29	0.72 ± 0.37	0.61 ± 0.30

generalization, even in the absence of centralized data access.

The LL models consistently show the lowest overall performance across both scenarios. These limitations become especially evident in Scenario 2, where the data distribution across clients is highly non-IID. The lack of representative and diverse data at the local level significantly hinders the models' ability to generalize, resulting in a sharper decline in performance, particularly for underrepresented classes. This outcome highlights the challenges of relying solely on isolated local training when data heterogeneity and scarcity are present.

Figures 3 and 4 provide a comparative overview of the F1 scores obtained for each class across the three learning paradigms. These visualizations support the observations discussed above, clearly showing the superior performance of CL, the minor degradation observed with FL, and the significantly lower scores obtained by LL models. The performance gap is especially pronounced in Scenario 2, where the highly non-IID data distribution exacerbates the limitations of isolated training, particularly for underrepresented diagnostic categories.

To analyze in more detail the predictive behavior of individual clients on the test set, Figures 5 and 6 present the Empirical Cumulative Distribution Function (ECDF) (Ducange et al. (2024)) of F1-score differences between the FL and LL approaches for the MLP model in Scenario 1 and Scenario 2, respectively. Similarly, Figures 7 and 8 show the same analysis for the *ConvNeXt* model.

For each client, the F1 score difference is calculated by subtracting the score obtained by the local model from that of the corresponding federated model, using

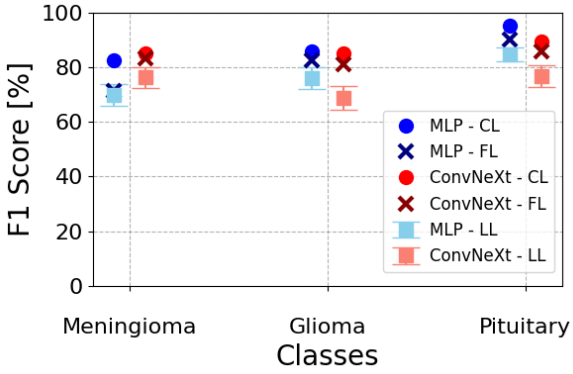

Figure 3. Class-wise F1-score in Scenario 1 for MLP and ConvNeXt using CL, FL, and LL.

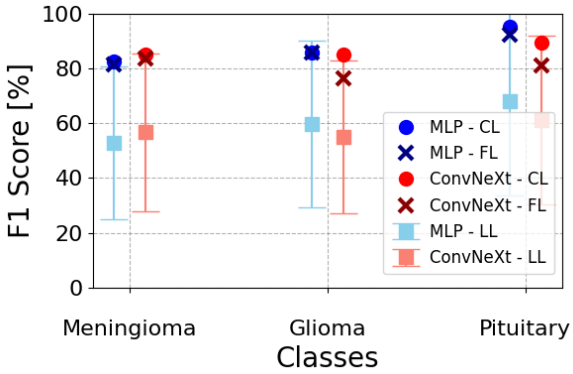

Figure 4. Class-wise F1-score in Scenario 2 for MLP and ConvNeXt using CL, FL, and LL.

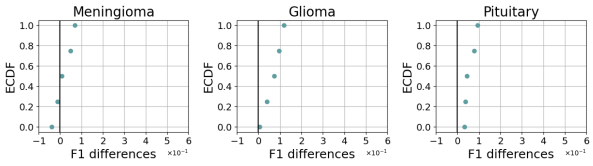

Figure 5. Empirical Cumulative Distribution Function of F1-score differences between FL and LL in Scenario 1 using the MLP model.

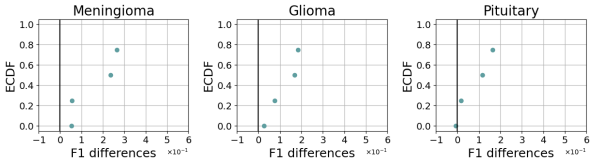

Figure 6. Empirical Cumulative Distribution Function of F1-score differences between FL and LL in Scenario 2 using the MLP model.

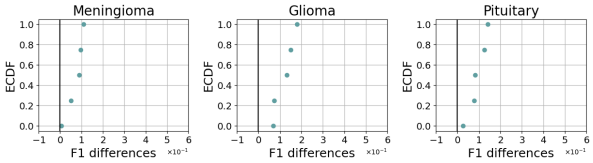

Figure 7. Empirical Cumulative Distribution Function of F1-score differences between FL and LL in Scenario 1 using the ConvNeXt model.

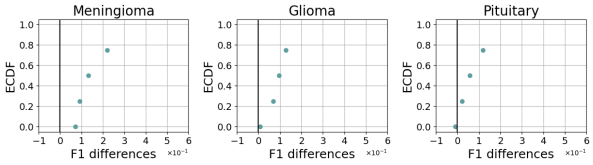

Figure 8. Empirical Cumulative Distribution Function of F1-score differences between FL and LL in Scenario 2 using the ConvNeXt model.

the common test set for evaluation. The plot can be interpreted as follows: if a point lies in the positive half-plane, it indicates that the federated model consistently outperforms its local counterpart on all clients. In both scenarios, most F1-score differences fall within the range [0, 0.3], confirming a systematic advantage of the federated approach. In particular, Scenario 2 shows a larger dispersion in the differences, reflecting the increased variability introduced by the non-IID data distribution. This variability highlights the increasing challenges faced by local models when exposed to unbalanced and biased data, further reinforcing the benefits of collaborative training in such contexts.

6.2. Complexity analysis

To evaluate the computational and operational feasibility of the proposed architectures in the different scenarios, we conducted a complexity analysis focusing on four main dimensions: number of parameters, disk usage, training time, and inference speed. The results are shown in Table 7.

The lightweight MLP model consists of only 1,155 trainable parameters, distributed across three fully connected layers. The total disk size of the saved model is approximately 8 KB.

In contrast, the *ConvNeXt*-based model contains a total of 27.8 million parameters. However, to keep the training complexity low, a partial tuning was applied: only the final block (Block 3) was updated during training, while the rest of the network remained fixed. This results in approximately 1 million trainable parameters on the full model. The saved model occupies approximately 109 MB of disk space.

The MLP model showed excellent training efficiency. In the CL setting, the optimal model was reached after 237 epochs, with a total training time of 24.1 seconds. In the FL setting, training took 483 rounds in Scenario 1 and 312 rounds in Scenario 2, corresponding to 88.3 and 45.2 seconds, respectively, with an average of less than 1 second per round (with 5 local epochs per round). These times refer only to computation and exclude communication overhead, which can vary depending on network conditions and system architecture.

The *ConvNeXt* model required the longest training time, as expected, due to its higher architectural complexity and image-based input. In CL, the best performance was achieved after only 5 epochs, with a total training time of 310.8 seconds. In FL, the model converged after only one round in Scenario 1 and three rounds in Scenario 2, with local training times of 94.4 and 128.8 seconds, respectively.

In terms of inference speed, the MLP model operates on compact radiomics vectors and achieves sub-millisecond latency even on the CPU, making it ideal for implementation on lightweight devices. However, in practical scenarios, radiomics features must first be extracted from the input image, which typically takes a few milliseconds on the CPU, depending on the image resolution and the number of computed features.

The *ConvNeXt* model, which directly processes 224×224 pixel ROIs, requires longer inference times but remains efficient on GPU hardware, typically staying below 30 milliseconds per image.

Table 7. Computational and operational comparison of the proposed models.

Model	Total Parameters	Trainable Parameters	Disk Size	Training Time (CL)	Training Time (FL)	Inference Time
MLP	1,155	1,155	8 KB	24.1 s (237 epochs)	88.3 s (Sc.1) / 45.2 s (Sc.2)	<1 ms (CPU) + radiomics extraction
ConvNeXt	27.8M	∼1M (Block 3)	109 MB	310.8 s (5 epochs)	94.4 s (Sc.1) / 128.8 s (Sc.2)	<30 ms (GPU)

7. Conclusion

This study presents a comparative analysis of two architectures for brain tumor classification within a federated learning framework: a lightweight multilayer perceptron (MLP) that processes structured features extracted from ROIs, and a *ConvNeXt* model that operates directly on MRI-derived ROI images. The results indicate that both models are capable of achieving high performance in federated training settings under both IID and non-IID scenarios. Despite its significantly lower complexity and computational demands, the MLP model produced outcomes comparable to those of the more complex *ConvNeXt* architecture. Under non-IID data distributions, which better reflect real-world clinical heterogeneity, FL consistently outperformed LL approaches, confirming the advantages of collaborative training.

Some aspects of the study deserve further consideration. The analysis was conducted on a widely adopted public dataset, which ensures reproducibility and comparability with previous work, but naturally does not capture the full variability of clinical practice. Furthermore, the dataset was collected from two hospitals over a restricted time frame (2005–2010), raising concerns about its representativeness with respect to current imaging protocols. Finally, the study focuses exclusively on the classification stage of the diagnostic pipeline, assuming that ROI segmentation has already been performed (either manually or through automatic methods). As such, the results do not account for potential variability introduced in the detection phase, but rather target the diagnostic decision-making step once the tumor region has been identified.

Building on these limitations, several directions emerge for future research. First, extending the analysis to larger and more diverse multi-institutional datasets would allow for validating the robustness of the proposed approaches under heterogeneous acquisition protocols and more recent imaging standards. Second, an important step will be to move beyond the assumption of pre-segmented ROIs and to integrate detection and classification into a unified federated pipeline, thereby capturing the full diagnostic workflow. Third, the adoption of advanced FL strategies, such as personalized federated learning and federated transfer learning, could help mitigate inter-institutional heterogeneity and enhance model adaptability. Finally, the integration of Explainable Artificial Intelligence (XAI) techniques will be crucial: for the MLP, post-hoc methods may clarify the contribution of radiomic features, while for *ConvNeXt*, saliency-based explanations could improve clinical interpretability. Combining imaging with complementary modalities (e.g., clinical, histopathological, or genomic data) represents another promising avenue to further strengthen clinical applicability.

Acknowledgements

This work has been partly funded by the PNRR - M4C2 - Investimento 1.3, Partenariato Esteso PE00000013 - *FAIR - Future Artificial Intelligence Research* - Spoke 1 *Human-centered AI* under the NextGeneration EU program, and from the PNRR *Tuscany Health Ecosystem* (THE) (*Ecosistemi dell'Innovazione*) - Spoke 6 - *Precision Medicine & Personalized Healthcare* (CUP I53C22000780001) under the NextGeneration EU program. Additionally, this research has been supported by the Italian Ministry of University and Research (MUR) in the framework of the *FoReLab* and *CrossLab* projects (*Departments of Excellence*).

References

Albalawi, E., TR, 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.

Amin, M. S., Ahmad, S., & Loh, W.-K. (2025). Federated learning for healthcare 5.0: A comprehensive survey, taxonomy, challenges, and solutions. *Soft Computing*, 1–28.

Bera, K., Braman, N., Gupta, A., Velcheti, V., & Madabhushi, A. (2022). Predicting cancer outcomes with radiomics and artificial intelligence in radiology. *Nature reviews Clinical oncology*, *19*(2), 132–146.

Carré, A., Klausner, G., Edjlali, M., Lerousseau, M., Briend-Diop, J., Sun, R., Ammari, S., Reuzé, S., Alvarez Andres, E., Estienne, T., et al. (2020). Standardization of brain MR images across machines and protocols: Bridging

the gap for MRI-based radiomics. *Scientific reports*, *10*(1), 12340.

Cheng, J., Huang, W., Cao, S., Yang, R., Yang, W., Yun, Z., Wang, Z., & Feng, Q. (2015). Enhanced performance of brain tumor classification via tumor region augmentation and partition. *PloS one*, *10*(10), e0140381.

Ducange, P., Marcelloni, F., Renda, A., & Ruffini, F. (2024). Federated learning of xai models in healthcare: A case study on parkinson's disease. *Cognitive Computation*, *16*(6), 3051–3076.

Guan, H., Yap, P.-T., Bozoki, A., & Liu, M. (2024). Federated learning for medical image analysis: A survey. *Pattern Recognition*, 110424.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Kale, P. V., Gadicha, A. B., & Dalvi, G. (2024). Detection and classification of brain tumor using machine learning. *2024 Third International Conference on Smart Technologies and Systems for Next Generation Computing (ICSTSN)*, 1–6.

Kalpana, V., Chowdary, J. S., Sravya, T. N. L., Reddy, A. A., Pravallika, P., & Gnanasri, V. (2023). Implemented global model for brain tumor detection using federated learning. *2023 7th International Conference on Electronics, Materials Engineering & Nano-Technology (IEMENTech)*, 1–5.

Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., & Xie, S. (2022). A convnet for the 2020s. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 11976–11986.

McMahan, B., Moore, E., Ramage, D., Hampson, S., & y Arcas, B. A. (2017). Communication-efficient learning of deep networks from decentralized data. *Artificial intelligence and statistics*, 1273–1282.

Mehmood, Y., & Bajwa, U. I. (2024). Brain tumor grade classification using the convnext architecture. *Digital Health*, *10*, 20552076241284920.

Miglionico, G. C., Ducange, P., Marcelloni, F., & Pedrycz, W. Deep learning and multi-objective evolutionary fuzzy classifiers: A comparative analysis for brain tumor classification in mri images. In: In *Proceedings of the 1st international conference on explainable ai for neural and symbolic methods - explains*. INSTICC. SciTePress, 2024, 108–115.

Muhammad, K., Khan, S., Del Ser, J., & De Albuquerque, V. H. C. (2020). Deep learning for multigrade brain tumor classification in smart healthcare systems: A prospective survey. *IEEE Transactions on Neural Networks and Learning Systems*, *32*(2), 507–522.

Narula, M., Meena, J., & Vishwakarma, D. K. (2024). A comprehensive review on federated learning for data-sensitive application: Open issues & challenges. *Engineering Applications of Artificial Intelligence*, *133*, 108128.

Ranjbarzadeh, R., Caputo, A., Tirkolaee, E. B., Ghoushchi, S. J., & Bendechache, M. (2023). Brain tumor segmentation of MRI images: A comprehensive review on the application of artificial intelligence tools. *Computers in biology and medicine*, *152*, 106405.

Viet, K. L. D., Le Ha, K., Quoc, T. N., & Hoang, V. T. (2023). Mri brain tumor classification based on federated deep learning. *2023 Zooming Innovation in Consumer Technologies Conference (ZINC)*, 131–135.

Xia, X., Wu, W., Tan, Q., & Gou, Q. (2025). Interpretable machine learning models for differentiating glioblastoma from solitary brain metastasis using radiomics. *Academic Radiology*.

Younis, A., Li, Q., Khalid, M., Clemence, B., & Adamu, M. J. (2023). Deep learning techniques for the classification of brain tumor: A comprehensive survey. *IEEE Access*, *11*, 113050–113063.

Yu, Z., Li, X., Li, J., Chen, W., Tang, Z., & Geng, D. (2024). Hsa-net with a novel cad pipeline boosts both clinical brain tumor mr image classification and segmentation. *Computers in Biology and Medicine*, *170*, 108039.

Zwanenburg, A., Vallières, M., Abdalah, M. A., Aerts, H. J., Andrearczyk, V., Apte, A., Ashrafinia, S., Bakas, S., Beukinga, R. J., Boellaard, R., et al. (2020). The image biomarker standardization initiative: Standardized quantitative radiomics for high-throughput image-based phenotyping. *Radiology*, *295*(2), 328–338.