Original Research Article

DIGITAL HEALTH
Volume 10: 1–15
© The Author(s) 2024
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/20552076241295577
journals.sagepub.com/home/dhj

Towards privacy-preserving Alzheimer's disease classification: Federated learning on T1-weighted magnetic resonance imaging data

Md Abdus Sahid, Md Palash Uddin, Hasi Saha and Md Rashedul Islam

Abstract

Objective: This study aims to address the challenge of privacy-preserving Alzheimer's disease classification using federated learning across various data distributions, focusing on real-world applicability. The goal is to improve the efficiency of classification by minimizing communication rounds between clients and the central server.

Methods: The proposed approach leverages two key strategies: increasing parallelism by utilizing more clients in each communication round and increasing computation per client during the intervals between rounds. To reflect real-world scenarios, data is divided into three distributions: identical and independently distributed, non-identical and independently distributed equal, and non-identical and independently distributed unequal. The impact of extreme quantity distribution skew is also examined. A convolutional neural network is used to evaluate the performance across these setups.

Results: The empirical study demonstrates that the proposed federated learning approach achieves a maximum accuracy of 84.75%, a precision of 86%, a recall of 85%, and an F1-score of 84%. Increasing the number of local epochs improves classification performance and reduces communication needs. The experiments show that federated learning is effective in handling heterogeneous datasets when all clients participate in each round of training. However, the results also indicate that extreme quantity distribution skew negatively impacts classification performance.

Conclusions: The study confirms that federated learning is a viable solution for Alzheimer's disease classification while preserving data privacy. Increasing local computation and client participation enhances classification performance, though extreme distribution imbalances present a challenge. Further investigation is needed to address these limitations in real-world scenarios.

Keywords

Alzheimer's disease, deep learning, federated learning, magnetic resonance imaging

Submission date: 25 March 2024; Acceptance date: 8 October 2024

Introduction

Neurodegenerative disorder mild cognitive impairment (MCI) is characterized as the prodromal stage of cognitive decline that falls between normal aging and the onset of dementia.[1–4] It is estimated that ∼15%–20% of individuals aged 65 and older experience MCI.[3] Alarmingly, MCI carries a substantial risk of progression to Alzheimer's disease (AD) or other forms of dementia, with a ∼54% chance of such progression.[5,6] The prevalence of dementia is a global concern, and AD accounts for a significant portion of dementia cases.[3,7,8] Statistics reveal that around 60%–70% or 60%–80% of all dementia cases are attributed to AD.[3,8,9]

Department of Computer Science and Engineering, Hajee Mohammad Danesh Science and Technology University, Dinajpur, Bangladesh

Corresponding author:
Md Palash Uddin, Department of Computer Science and Engineering, Hajee Mohammad Danesh Science and Technology University, Dr M. A. Wazed Building, Dinajpur, Rangpur 5200, Bangladesh.
Email: palash_cse@hstu.ac.bd

AD is a severe and progressive neurological condition that manifests through a range of cognitive impairments, including gradual memory loss, reasoning difficulties, behavioral changes, communication issues, motor dysfunction, and impaired daily activities, among other cognitive deficits.[10–13] The pathology of AD is characterized by abnormal proteins that disrupt brain cells and neurons, leading to the breakdown of signal transmitters crucial for memory retention.[14] AD extends beyond the realm of clinical symptoms; it has profound social, economic, and global implications.[15] The burden on caregivers, often family members, is immense, encompassing emotional, financial, and practical challenges.[16]

According to the World Alzheimer Report 2021, Alzheimer's Disease International (ADI) has reported that up to 75% of people with dementia worldwide remain undiagnosed, and this figure may be as high as 90% in some low- and middle-income countries, where stigma and lack of awareness about dementia pose significant barriers to diagnosis. The number of people living with dementia has surpassed 55 million globally, and this number continues to grow, with projections indicating it could reach 78 million by 2030.[14] The socioeconomic impact of dementia is also substantial, with the worldwide cost expected to have exceeded US$1.3 trillion in 2019 and predicted to rise to over US$2.8 trillion by 2030 due to the increasing number of individuals living with dementia and the associated caregiving expenses.[17] Consequently, it is imperative to focus on preventive measures to combat this debilitating disease for the sake of both healthcare and the economy.

While the precise causes of AD are still a subject of ongoing research, several risk factors have been identified. These risk factors include genetics, family history, head injuries, down syndrome, heart disease, diabetes, stress, stroke, high blood pressure, high cholesterol, and, notably, age.[18,19] There are promising anti-amyloid treatments (e.g. aducanumab, lecanemab, etc.) that are clinically indicated in patients with earlier stages of AD, so earlier and accurate detection of AD is important when considering the potential use of these medications.

To detect AD, conventional centralized machine learning models require data to be transmitted to a central server, posing significant privacy vulnerabilities. In contrast, federated learning (FL) offers a methodology for model training that eliminates the need to centralize data, thereby safeguarding the privacy of individuals or entities providing their data. This privacy-preserving approach aligns with the imperative need to protect the confidentiality of such information in an era where data security and privacy concerns are paramount. As such, in this study, we have employed FL to classify AD.

FL is a machine learning paradigm that involves training a model using data from multiple clients, such as mobile phones, tablets, and hospitals, without the need to directly share sensitive training data. This collaborative approach is often facilitated by a central server that orchestrates the learning process.[20–22] There are two common settings for FL, which are determined by the network's size and characteristics: cross-device FL and cross-silo FL.[23] Cross-device FL pertains to scenarios where numerous clients participate, each having limited data, bandwidth, and availability, such as mobile devices, laptops, and tablets. In contrast, cross-silo FL involves a smaller number of clients, but each possesses more substantial resources, including institutions like banks, schools, and hospitals. Notably, in cross-silo FL, each client is required to actively engage in the entire training process, which is feasible due to the limited number of clients, typically ranging from two to 100. Figure 1 provides a visual representation and working mechanism of the FL. This privacy-preserving mechanism has significant potential applications in areas like privacy-preserving disease detection and classification. The abovementioned works of FL (explained in more detail in the literature review section) did not conduct experiments based on either different real-world scenarios or with the aim of reducing communication costs while classifying AD. Therefore, the novelty of this research is summarized below:

- Conducted an empirical and rigorous study focused on the detection and classification of AD utilizing privacy-preserving FL.
- Generated synthetic dataset representative of real-world scenarios from the original dataset and conducted experiments to validate the approach's efficacy.
- Implemented experiments exploring strategies of increasing parallelism and computation per client to reduce communication costs while classifying AD.

The rest of the article is organized as follows. The "Literature review" section delves into related literature, highlighting potential areas for further research. In the "Methods and materials" section, we describe our method for privacy-preserving AD classification using the FL scheme and provide the materials used in the experimental results analysis. The "Results and discussions" section furnishes the experimental outcomes and discussions, and finally, "Conclusion and future work" section presents the conclusion and outlines future research directions.

Literature review

This section provides several recent studies that were conducted by applying centralized learning (CL) and FL. To safeguard data privacy and address data heterogeneity issues, the researchers by Lei et al.[24] introduced a framework for multi-site federated domain adaptation based on the transformer model. They utilized the transformer to uncover relationships among features from multi-template regions of interest and to harness the complementary information from these templates.

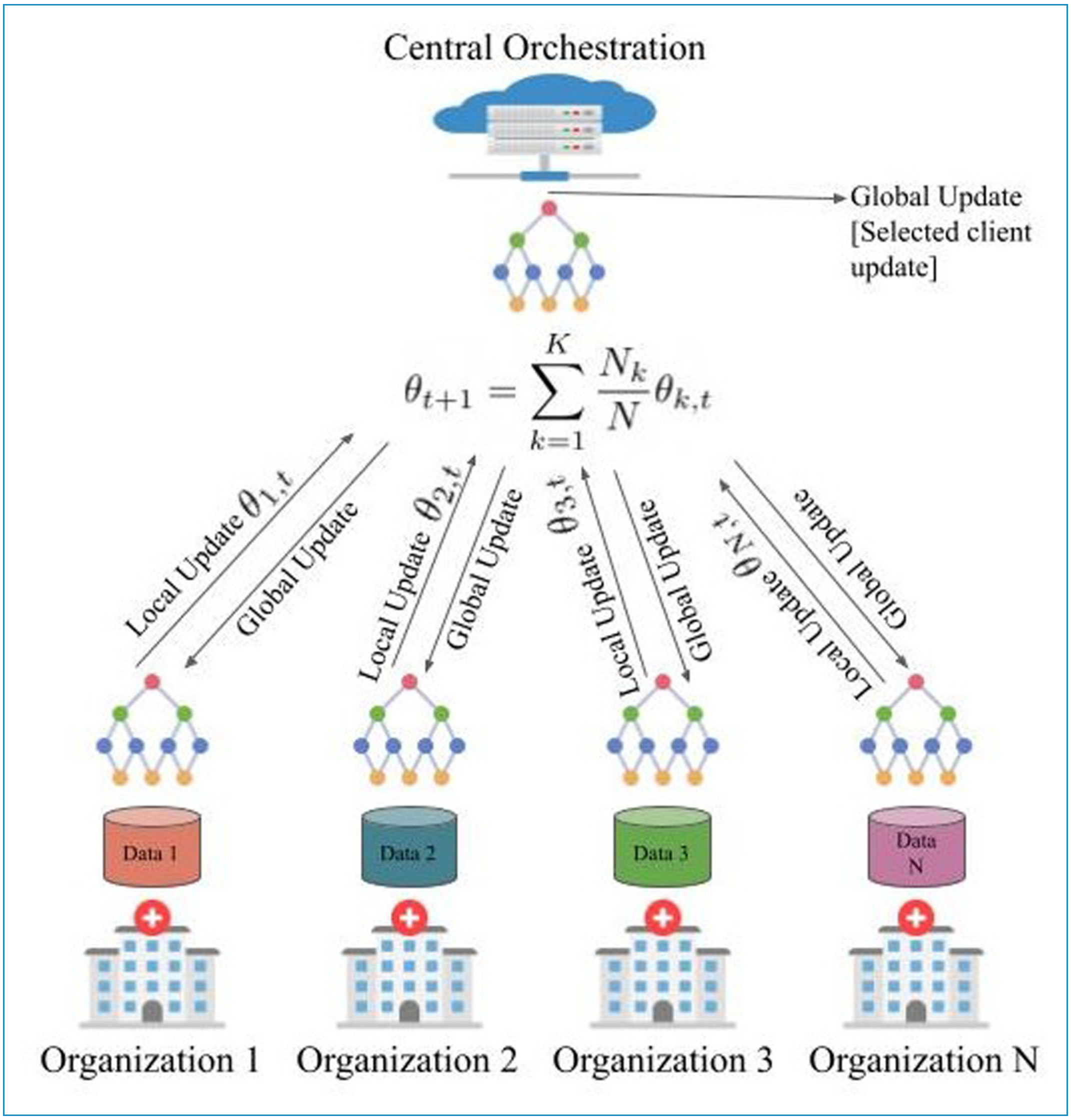
Central Orchestration
Global Update [Selected client update]
$\theta_{t+1} = \sum_{k=1}^{K} \frac{N_k}{N} \theta_{k,t}$
Local Update $\theta_{1,t}$
Global Update
Local Update $\theta_{2,t}$
Global Update
Local Update $\theta_{3,t}$
Global Update
Local Update $\theta_{N,t}$
Global Update
Data 1
Data 2
Data 3
Data N
Organization 1
Organization 2
Organization 3
Organization N

Figure 1. The working mechanism of federated learning (FL) framework. At each communication round t, C clients are selected and trained on their local model by their private dataset and also send their local model parameters. The central orchestration aggregates all the updates sent by selected client C through the FedAvg or other algorithms and trains the global model. After that, the central orchestration sends the updated global model to the C clients for retraining purposes. This process continues until reaching training convergence.

Their study encompassed three distinct datasets: the Alzheimer's Disease Neuroimaging Initiative (ADNI), the Australian Imaging, Biomarker and Lifestyle Flagship Study of Aging (AIBL), and AI4AD data. They conducted both two-way classifications, distinguishing between AD versus healthy control (HC), MCI versus HC, and AD versus MCI, as well as a three-way classification task, separating AD, MCI, and HC. The results showed accuracy rates of 88.75%, 69.51%, and 69.88% for the two-way classification tasks involving AD versus HC, MCI versus HC, and AD versus MCI, respectively.

In their work,[25] the authors introduced an evolutionary deep convolutional neural network (EDCNN) designed for the identification of AD within a privacy-protected FL framework. Their primary emphasis was on convex optimization, to enhance the computational efficiency and accuracy of AD detection. The study involved the utilization of multimodal datasets, including MRI, EEG, and blood test data, each serving as a distinct node within the federated settings. The researchers by Huang et al.[26] dedicated their efforts to developing a privacy-preserving AD classification framework called federated conditional mutual learning (FedCM), designed for client-aware mutual learning. They validated their proposed framework using three different AD datasets: ADNI, Open Access Series of Imaging Studies (OASIS), and AIBL. They conducted two and three-class classifications on the labeled data, differentiating AD, MCI, and HC subjects. In the context of AD versus HC, they achieved remarkable results, attaining a maximum accuracy of 91.9%, a recall of 100%, and a specificity of 91.1% by applying three-dimensional-convolutional neural network (3D-CNN) to the OASIS dataset.

In another secure FL-based approach presented by the researcher,[27] the focus was on the neuroimaging modality. The authors implemented a fully homomorphic encryption mechanism to ensure secure communication and aggregation. They tested their method with the AD datasets (ADNI, OASIS, and AIBL) and the 3D-CNN model. They trained the neural model using three (ADNI phases), four (ADNI phases + OASIS), and five (ADNI phases + OASIS + AIBL) learners or clients. The best performance was achieved with a five-learner setup, resulting in an accuracy of 86%, precision of 80.98%, recall of 81.32%, and an F1-score of 81.14% when considering the ADNI, OASIS, and AIBL datasets. Using the same five learners, the authors achieved an accuracy of 86.12%, a precision of 79.77%, a recall of 82.87%, and an F1-score of 81.22% for centralized settings.

Movahed and Rezaeian[28] introduced a machine learning framework for diagnosing mild AD, focusing on extracting spectral, functional connectivity, and nonlinear features from EEG signals. They utilized the sequential backward feature selection (SBFS) algorithm to choose the most suitable feature subset. Multiple classifiers, including support vector machines (SVM) with linear and RBF kernels, logistic regression (LR), k-nearest neighbor (KNN), decision tree (DT), gentleBoost, naive Bayes (NB), and RushBoost (RB), were examined. Among the classifiers, SVM with 10-fold cross-validation (CV) exhibited the best performance.

Siuly et al.[3] devised a framework for distinguishing mild AD from HC. They employed the stationary wavelet transformation (SWT) method to eliminate low-frequency (including baseline drift) and high-frequency (including power line interference) noise. They analyzed data in non-overlapping 2-second sliding windows. The study introduced the piecewise aggregate approximation (PAA) technique. To evaluate their approach, they used extreme learning machines (ELM), SVM, and KNN with 10-fold CV. Plant et al.[29] proposed a classification framework for distinguishing AD from HC using frequency and time–frequency features extracted from EEG data. They conducted experiments under resting-state eyes open (EO) and eyes closed (EC) conditions, segmenting the EEG signal into 4-second windows. After preprocessing the data, they applied the KNN model with 10-fold CV.

Sarraf et al.[30] presented a method for AD and HC detection using CNN. They utilized MRI and fMRI as their experimental modalities, achieving accuracy rates of 99.9% and 98.84% for the fMRI and MRI pipelines, respectively. Khatun et al.[4] employed single-channel EEG data from Fpz (near the forehead) to classify individuals with MCI from those with normal cognitive functioning. They analyzed the data in 25 ms windows with a 50% overlap across the entire signal. For feature selection, they used the random forest (RF). As classifiers, they utilized SVM with a radial basis function (RBF) kernel and LR with leave-one-out CV.

For early AD detection, an empirical analysis was performed by McBride et al.[5] The researchers explored spectral and complexity features as EEG-based biomarkers to differentiate between normal older individuals, those with MCI, and AD subjects. They considered 24 features and calculated their average values for each channel and 12 brain regions, including left and right regions, as well as a global region representing the average of all regions. They devised a three-way classifier based on a two-way classifier (HC vs. MCI, HC vs. AD, and MCI vs. AD) using a pairwise coupling approach. They applied a quadratic kernel with an SVM classifier for classification under EO, EC, and counting task (CT) conditions.

Aghajani et al.[31] proposed a method for mild AD detection using EEG signals. They aimed to distinguish between healthy individuals and those with mild AD by mapping EEG signals to their corresponding distributed sources via standardized low-resolution brain electromagnetic tomography based on a realistic head model. They proposed using the relative logarithmic transformed power spectrum density of estimated sources as a feature. Singular value decomposition was employed to reduce the number of features and enhance separability.

Plant et al.[29] introduced a sulcal feature-based approach for classifying AD and HC. They computed various features of the sulcal medial surface, including depth, length, mean curvature, Gaussian curvature, and surface area. These features were used in conjunction with an SVM for classification. When tested using 10-fold CV, the model achieved an accuracy of 87.9%, sensitivity of 90.0%, specificity of 86.7%, and an area under the receiver operating characteristic curve (AUC) of 89%.

A comparison of the potentially related works is tabulated in Table 1. Several factors underlie the motivation for this planned research. While conventional CL has been extensively explored for disease detection and classification, limited attention has been given to employing FL. Those few researchers who have ventured into healthcare applications of FL have often omitted empirical analysis.

Methods and materials

Method overview

Our primary goal is to optimize the efficiency of AD classification through advancements in the FL framework by minimizing the number of communication rounds required

Table 1. Comparison of the previously conducted relevant literature for the detection and classification of AD.

Reference	Approach	Data	Model	Performance metrics	Strengths	Shortcomings	Year
Lei et al.[24]	FL	MRI	Transformer	Accuracy rates of 88.75%, 69.51%, 69.88% for AD versus HC, MCI versus HC, and AD versus MCI, respectively	Utilizes advanced transformer model to integrate multi-site data effectively	Requires large datasets to train effectively, complex model architecture	2023
Lakhan et al.[25]	FL	MRI, EEG, blood test	EDCNN	Focus on convex optimization for AD detection, commendable performance	Enhances computational efficiency; incorporates multimodal data	May overlook unique dataset characteristics due to optimization focus	2023
Stripelis et al.[27]	FL	MRI	3D-CNN	Accuracy of 86%, precision 80.98%, recall 81.32%, and F1-score 0.8114.	High privacy preservation; robust against data distribution variability	Performance can be dependent on the number and quality of FL participants	2022
Huang et al.[26]	FL	MRI	3D-CNN	Maximum accuracy 91.9%, recall 100%, specificity 91.1%	Excellent classification accuracy; perfect recall rate	Could benefit from further validation across more diverse datasets	2021
Movahed and Rezaeian[28]	CL	EEG	SVM, LR, KNN, DT, gentleBoost, NB, RB	SVM with 10-fold CV showed best performance	Diverse classifier testing; robust feature selection via SBFS	Single modality (EEG) may limit diagnostic applicability	2022
Siuly et al.[3]	CL	EEG	ELM, SVM, KNN	ELM maximum accuracy 98.78%, precision 99.69%, recall 98.32%, F1-score 98.95%	High accuracy; robust against noise with SWT method	Limited to EEG data; may not generalize well to other modalities	2020
Durongbhan et al.[7]	CL	EEG	KNN	In EO state, accuracy 83.32%, recall 72.57%, specificity 87.52%	Performs well under different sensory conditions (EO and EC)	Performance varies significantly between conditions	2019
Khatun et al.[4]	CL	EEG	SVM, LR	Accuracy 87.90%, recall 84.90%, specificity 95%	Effective in distinguishing MCI from normal cognitive functioning	High dependency on EEG placement and quality	2019

(continued)

Table 1. Continued.

Reference	Approach	Data	Model	Performance metrics	Strengths	Shortcomings	Year
Sarraf et al.[30]	CL	MRI, fMRI	CNN	Accuracy 99.9% for fMRI, 98.84% for MRI	High accuracy; utilizes advanced imaging modalities	fMRI may not be accessible in all clinical settings	2016
Plant et al.[29]	CL	MRI	SVM	Accuracy 87.9%, sensitivity 90.0%, specificity 86.7%, AUC 89%	Uses detailed sulcal features for high diagnostic specificity	Complexity of feature extraction may hinder practical application	2016
McBride et al.[5]	CL	EEG	SVM	Accuracy 84.4% for EO, 96.9% for CT, and 71.9% for EC	Utilizes complex EEG features for detailed analysis	Results vary significantly across tasks; complex setup	2014
Aghajani et al.[31]	CL	EEG	SVM	Accuracy 84.40%, recall 75%, specificity 93.70%	Novel use of sLORETA for feature extraction from EEG	May require high computational resources for feature analysis	2013

AD: Alzheimers disease; FL: federated learning; CL: centralized learning; MRI: magnetic resonance imaging; fMRI: functional magnetic resonance imaging; EEG: electroencephalography; EDCNN: evolutionary deep convolutional neural network; SVM: support vector machine; KNN: k-nearest neighbor; 3D-CNN: three-dimensional convolutional neural network; ELM: extreme learning machine; LR: logistic regression; DT: decision tree; NB: Naive Bayes; RB: RushBoost; HC: healthy control; MCI: mild cognitive impairment; EO: eyes open; EC: eyes closed; CT: counting task; AUC: area under the receiver operating characteristic curve; CV: cross-validation.

during model training. This reduction in communication demands additional computational input, a resource-intensive proposition that we address through two principal strategies: (1) increased parallelism, which entails engaging a larger cohort of clients to perform tasks concurrently during each communication round, leveraging their collective computational power to potentially accelerate the learning process while maintaining or reducing the frequency of required communications; and (2) increased computation per client, where each client is tasked with performing more complex computational tasks within each training cycle, such as advanced processing tasks that contribute to the model's learning phase, thereby enriching the training process within the same communication interval. To ensure the relevance and applicability of our FL enhancements, we divide the original dataset into three distinct data distributions: identical and independently distributed (IID), non-IID equal, and non-IID unequal, each presenting unique challenges and scenarios that closely mimic the variety of real-world conditions under which FL systems must operate. This setup allows us to thoroughly test the resilience and adaptability of our proposed methodologies in different data environments, providing comprehensive insights into their effectiveness and potential areas for further refinement.

Firstly, the collected image data are preprocessed and then used to create IID, non-IID equal, and non-IID unequal synthetic datasets. These datasets are subsequently used for further experiments. The overview of the proposed methodology is depicted in Figure 2.

Dataset

Our experimental Alzheimer's dataset contains 6400 T1-weighted MRI images from four classes, each 128 × 128 pixels.[32] The statistical description of this dataset is outlined in Table 2. Note that the collected image data are originally sourced from the following recognized sources: ADNI,[33] alzheimers.net,[34] MRI and Alzheimer's,[35] Alzheimer's Disease and Healthy Aging Data,[36] and the European Prevention of Alzheimer's Dementia.[37]

Data preprocessing

This research employed fundamental image processing techniques to prepare the data dynamics for training. The preprocessing procedures are outlined below:

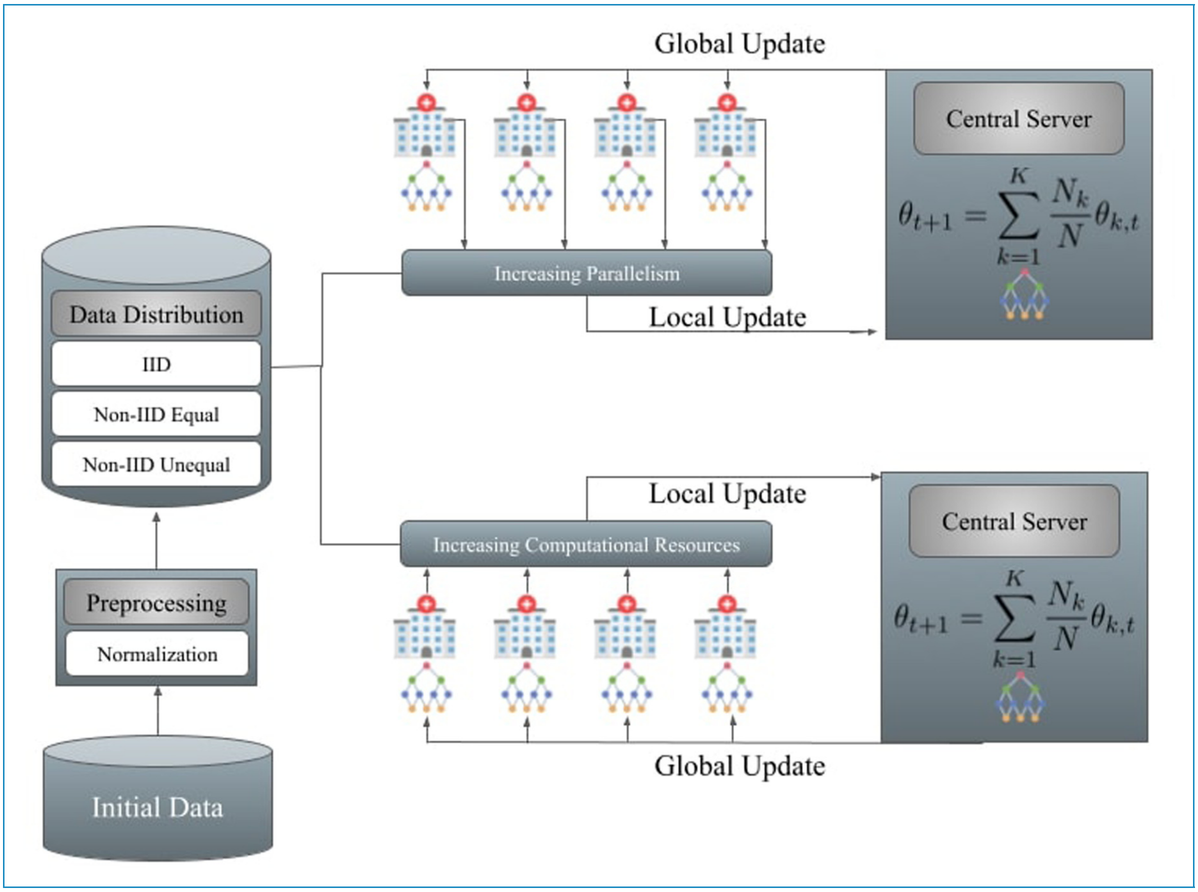

Figure 2. An overview diagram of the proposed framework for privacy-preserving AD classification using the FL scheme. AD: Alzheimers disease; FL: federated learning.

Table 2. Statistical description of AD dataset.

Type	Subject	Train	Test	Channel
Mild demented	896	672	224	3
Moderate demented	64	48	16	3
Very mild demented	2240	1680	560	3
HC	3200	2400	800	3

AD: Alzheimer's disease; HC: eyes closed.

Normalization. Image normalization was carried out to scale every pixel in the image to a range between 0 and 1. This involves transforming the pixel intensity range of the image into a standardized scale, facilitating model learning, and enhancing training effectiveness.[38] Variations in lighting conditions, contrast levels, and color distributions among images can all introduce biases, which normalization helps mitigate. The steps involved in image normalization are as follows. The mean pixel value is subtracted from each pixel in the image, thereby eliminating any inherent data dynamics bias, resulting in pixel values centered at zero. Subsequently, the image is divided by the standard deviation of the pixel values, ensuring that the pixel values have a unit standard deviation. This helps to equalize the scale of different features within the image. Mathematically, it can be defined as follows:

$$Normalized\_pixel = \frac{pixel - min\_pixel}{max\_pixel - min\_pixel} \tag{1}$$

Labeling. The dataset encompasses 416 subjects aged 18 to 96, with 3–4 T1-weighted MRI scans per subject. All subjects are right-handed and of both genders. Among those over 60, 100 have been clinically diagnosed with very mild to moderate AD. Additionally, a reliability subset includes 20 non-demented subjects imaged within 90 days of their initial session. On the other hand, the longitudinal data comprises 150 subjects aged 60–96, scanned at least twice over a year, totaling 373 imaging sessions. Each subject's scans are obtained in single sessions. Of

Table 3. Impact of the client fraction C on the classification of AD with $E = 10$. We use $K = 10$ client for our experiment.

			C					
Data distribution	B	Performance metric	0.0 (%)	0.1 (%)	0.2 (%)	0.5 (%)	0.8 (%)	1.0 (%)
IID	10	Accuracy	84.06	84	80	82.88	80.31	83.25
		Precision	85	87	83	84	81	85
		Recall	84	84	80	83	80	83
		F1-score	82	81	76	81	78	81
	32	Accuracy	81.44	80.06	81.88	81.31	81.12	80.69
		Precision	84	84	85	85	84	84
		Recall	81	80	82	81	81	81
		F1-score	78	77	79	78	79	77
Non-IID equal	10	Accuracy	68	67.38	70.25	72.44	69.87	74.62
		Precision	75	70	78	77	81	80
		Recall	68	67	70	72	70	75
		F1-score	54	62	59	72	70	68
	32	Accuracy	65	68	66.69	69.94	74.19	74.69
		Precision	40	59	49	76	80	80
		Recall	65	68	67	70	74	75
		F1-score	48	59	52	61	71	70
Non-IID unequal	10	Accuracy	80.69	76.06	82.31	81.25	83	82.88
		Precision	82	80	85	81	85	82
		Recall	81	76	82	81	83	83
		F1-score	78	70	82	80	81	82
	32	Accuracy	81.25	70.19	80.6	83.5	82.06	84.12
		Precision	83	83	83	83	82	85
		Recall	81	70	80	84	82	84
		F1-score	81	63	78	83	80	82

Note: $C = 0.0$, $C = 0.1$, and $C = 1.0$ corresponds to one, 10%, and 100% client participation per round, respectively; AD: Alzheimers disease; IID: identical and independently distributed.

these subjects, 72 remain consistently non-demented, while 64 are consistently demented, including 51 with mild to moderate Alzheimer's. Another 14 subjects transitioned from non-demented to demented over time.

Artificial partitioning (synthetic data creation) of centralized dataset. To create a synthetic dataset, we take a labeled centralized dataset and employ some scheme described below to pathologically partition the dataset among a set of K clients. These K clients contain local datasets $D_1, D_2, \ldots, D_i, \ldots, D_K$, respectively. As such, the global dataset is $D\Delta = D_1 \cup D_2 \cup \cdots \cup D_i \cup \cdots \cup D_K$. Assume that $D_i \cap D_j = \varnothing$ for $i \neq j$. We define the number of samples in node i as $|D_i|$, where $|\cdot|$ denotes the size of the set. In contrast, this approach does result in a heterogeneous dataset, our adopted scheme to build a synthetic non-IID dataset is outlined below:

1. Label distribution skew (prior probability shift). Consider a scenario with K representing the number of clients, each characterized by the $P_k(x, y)$ data distribution. This distribution can be reformulated as $P_k(x|y)P_k(y)$. Within this context, we can distinguish two distinct scenarios, both of which involve non-identical conditions. The first scenario is referred to as label distribution skew, wherein the label distributions $\{P_k(y)\}_{k=1}^{K}$ exhibit variation among different clients, while the conditional generating distributions $\{P_k(x|y)\}_{k=1}^{K}$ are assumed to remain consistent. This situation may arise when specific types of data are inadequately represented within the local context.[38]
2. Data quantity disparity (unbalancedness or quantity skew). Furthermore, variations in the volume of data held by different clients can result in unequal levels of uncertainty in locally updated models and heterogeneity in the frequency of local updates. In real-world applications, the quantity of data may vary significantly among clients, with large institutions, such as hospitals, typically possessing considerably more medical records than smaller clinics. Notably, the distribution of data quantities frequently demonstrates a pattern where substantial datasets are primarily concentrated in a few specific locations, while a vast number of locations have smaller dataset sizes distributed across them.[38]

In a non-IID equal distribution, the data quantity is equal, whereas in a non-IID unequal data distribution, the quantity varies. In our study, we'll conduct an experiment based on IID, non-IID equal, and non-IID unequal distribution of our adopted Alzheimer's dataset. Formally, for the IID settings let us standardize the stochastic optimization problem,

$$\min_{x\in\mathbb{R}^m} F(x) := x \sim D\mathbb{E}[l(w; x)] \tag{2}$$

In non-IID settings, each of $k \in [K]$ clients has a local data distribution D_k and a local objective function,

$$f_k(x) := x \sim D_k\mathbb{E}[l(w; x)] \tag{3}$$

where we recall that $l(w; x)$ is the empirical loss of a model w at non-identical data x. We typically wish to minimize

$$F(x) = \frac{1}{K}\sum_{k=1}^{K} f_k(x) \tag{4}$$

for our experimental Alzheimer's dataset.

Federated setups

Cross-silo. A discrete entity or organization that manages and controls its unique dataset is commonly referred to as a data silo. When multiple data silos or distinct organizations collaborate to collectively train a unified global model, this variant of FL is recognized as cross-silo FL. Cross-silo FL represents a scenario in which there is a restricted count of participating clients, encompassing entities like banks, schools, and hospitals, each of which possesses more abundant resources. It's noteworthy that these same data silos can be utilized in both the training phase and the subsequent inference stage. Specifically, in the context of cross-silo FL, the number of clients involved typically ranges from 2 to 100, as noted by Kairouz et al.[38]

Client sampling. In cross-silo FL experiments, each client is required to engage in the full training process, since there are only a few clients (about 2–100), so the client sampling rate is *100%* or $C_{sampled} = \{D_1, D_2, \ldots, D_i, \ldots, D_K\}$. As well as, in this study, we also scale up our analysis to partial participation, that is, $C_{sampled} \subseteq \{D_1, D_2, \ldots, D_i, \ldots, D_K\}$ (such as 10%, 20%, 60%, and 100%) of clients in each round of training.

Federated algorithm. In our experimental setup, we employ the Federated averaging (FedAvg) algorithms to aggregate updates originating from each client, where each client $k \in [K]$. The description of this algorithm is as follows:

The FedAvg algorithm is recognized as the most straightforward aggregator method, as mentioned in the reference.[23] To express this mathematically,

$$\theta_{t+1} = \sum_{k=1}^{K} \frac{N_k}{N}\theta_{k,t} \tag{5}$$

In the equation above, θ_{t+1} signifies the updated global model at iteration $(t+1)$. Here, K represents the total count of participating clients, N_k corresponds to the number of samples contributed by client k, and N encompasses the overall sample count across all clients. Lastly, $\theta_{k,t}$ denotes the update originating from a local model of client k during the communication round t.

Optimizers. For simplicity of hyperparameter tuning and experimental controls, we use minibatch stochastic gradient descent (SGD) for client-local training for all experiments.

Hyperparameters. In our experiment, we set clients to train for $E \in \{1, 5, 20\}$ local epochs in every round. Local epochs are a popular technique to reduce communication costs. The local batch size across all clients is fixed with $B \in \{10, 50, 32\}$. For each IID and non-IID data distribution, we set $T = 50$ communication rounds. We set the learning rate $\eta = 0.01$. The total number of clients $K = 10$ and client selection $C \in \{0.0, 0.1, 0.2, 0.5, 0.8, 1.0\}$. Thus, C controls the *global* batch size, with $C = 1$ corresponding to full-batch (non-stochastic) gradient descent.

Model structure. For image classification feed-forward deep networks, and in particular convolutional networks, are well-known to provide state-of-the-art results.[39,40] Our experiments include a non-convex LeNet5 CNN model.

Evaluation protocol. To assess the efficacy of CNN-based FL settings for the classification of AD, this study employed various essential performance metrics. These metrics primarily rely on the widely used tool known as the confusion matrix. The confusion matrix is illustrated in Table 4. It serves as a performance assessment tool that summarizes the performance of the applied classification model by quantifying true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN).

- TP: This represents the number of cases correctly predicted as positive.
- TN: This signifies the number of cases correctly predicted as negative.
- FP: It accounts for the instances where negative cases were incorrectly predicted as positive.
- FN: This denotes the count of positive cases erroneously predicted as negative.

Table 4. Confusion matrix.

Total sample	Predicted	
	Negative	Positive
Actual negative	TN	FP
Actual positive	FN	TP

TP: true positives; TN: true negatives;FP: false positives; FN: false negatives.

Accuracy. Accuracy is defined as the ratio of correctly classified data instances to all data instances.[41] Mathematically,

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{6}$$

Precision. Precision evaluates the accuracy of the minority class.[42] Mathematically,

$$Precision = \frac{TP}{TP + FP} \tag{7}$$

Recall. Recall quantifies how many of the actual positive cases were correctly identified among all positive instances.[42,43] Mathematically,

$$Recall = \frac{TP}{TP + FN} \tag{8}$$

F1-score. The F1-score is the harmonic mean of Precision and Recall.[41,44] Mathematically:

$$F1 - score = \frac{2 * Precision * Recall}{Precision + Recall} \tag{9}$$

Note that the employed accuracy, precision, recall, and F1-score as our primary evaluation metrics are all calculated using the values from the confusion matrix. Additionally, metrics such as precision and recall are directly related to the ROC curve, as they reflect the TP and FP rates, which are fundamental in calculating the AUC. While we did not explicitly plot ROC curves or report *p*-values, the chosen metrics are representative of the same underlying evaluation framework.

Results

We first conducted our experiment based on increasing parallelism and then increasing computation per client approach. The findings are presented below:

Increasing parallelism

In the context of our experiment aimed at increasing parallelism, we kept the local epoch, E, constant at 10 and varied the local batch size $B \in \{10, 32\}$. The client fraction C, representing the proportion of multi-client parallelism, was manipulated across the values $C \in \{0.0, 0.1, 0.2, 0.5, 0.8, 1.0\}$. The impact of these changes on AD classification using LeNet5 CNN model is summarized in Table 3. Notably, when using the smaller local batch size $B = 10$, we observed significant improvements in AD classification performance under various client parallelism conditions, except for cases,

Table 5. Empirical results for the classification of AD according to increasing computation per client fashion.

			$C = 1.0$		
E	*B*	Performance Metric	IID (%)	Non-IID equal (%)	Non-IID unequal (%)
1	32	Accuracy	72.5	72.63	80.69
		Precision	79	73	82
		Recall	72	73	81
		F1-score	63	67	78
5	32	Accuracy	82.06	70.38	83.75
		Precision	85	81	83
		Recall	82	70	84
		F1-score	79	65	82
1	50	Accuracy	73.5	71.12	79.12
		Precision	79	73	79
		Recall	73	71	79
		F1-score	65	67	77
20	32	Accuracy	81.25	58.5	83.38
		Precision	83	72	83
		Recall	81	58	83
		F1-score	78	51	82
1	10	Accuracy	81.06	74.5	84
		Precision	86	77	85
		Recall	82	74	84
		F1-score	79	69	83
5	50	Accuracy	82.67	69.12	84.75
		Precision	85	72	86
		Recall	83	69	85
		F1-score	81	58	84
20	50	Accuracy	81.56	77.81	82.81
		Precision	83	78	82
		Recall	82	78	83
		F1-score	79	74	81
5	10	Accuracy	82.38	76.62	83.88
		Precision	85	77	83
		Recall	82	77	84
		F1-score	80	74	83

AD: Alzheimers disease; IID: identical and independently distributed.

where $C \in \{0.2, 0.8\}$ in the IID scenario and $C = 0.8$ in the non-IID equal scenario, as well as $C \in \{0.0, 0.5, 1.0\}$ in all three data distributions. Based on these findings, we determined that, for most of our subsequent experiments, a client fraction of $C = 1.0$ offers optimal performance for our heterogeneous real-world Alzheimer's data, resulting in a favorable convergence rate.

Increasing computation per client

In this approach of experiment, we fix $C = 1.0$, and add more computation per client on each round, either decreasing B, increasing E, or both. The quantitative results are provided in Table 5 and test set accuracy versus communication rounds are depicted in Figure 3. From the empirical results tabulated in Table 5, it is clear that instead of a single local epoch multiple local epochs increase the performance of AD detection using LeNet5 CNN model in federated settings which also reduces the communication costs between central server and clients. Moreover, in comparison with the existing literature, this study achieved comparatively similar performance, as outlined in Table 6.

Discussions

Data confidentiality in every field is crucial and highly demanded in the present world. With that in mind, this study focuses on the classification of AD using the emerging FL approach. The adopted FL approaches for privacy-preserving AD classification show comparatively satisfactory performance while keeping data private. The following explains this in great detail.

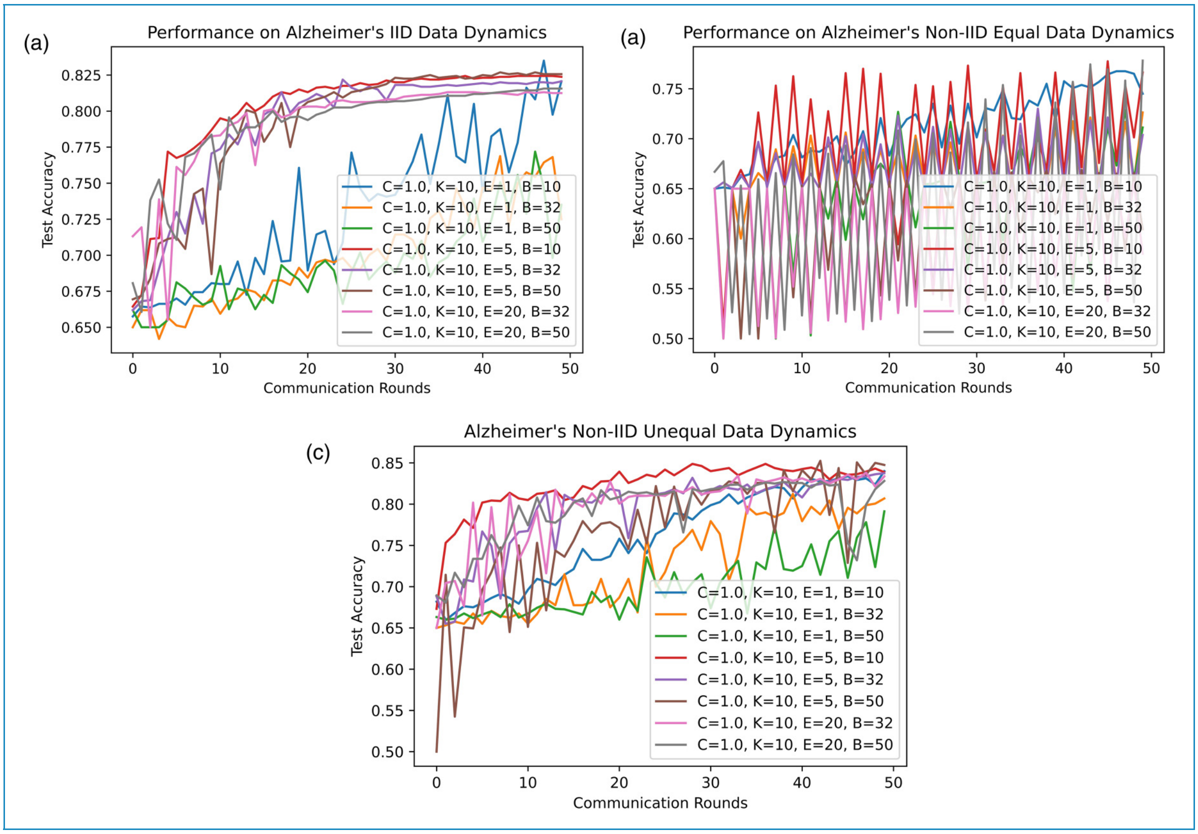

Figure 3. Depiction of convergence for the increasing computation per client experiment. (a) Test accuracy versus communication on the non-identical and independently distributed (IID) data distribution. (b) Test accuracy versus communication on the non-IID equal data distribution. (c) Test accuracy versus communication on the non-IID unequal data distribution.

Increasing parallelism

In terms of this approach, the results are presented in Table 3. By selecting $C = 1.0$ and $B = 10$, we achieved an accuracy of 83.25%, precision of 85%, recall of 83%, and an F1-score of 81% for the IID data distribution of AD. However, the minimum accuracy achieved was 80% when using 20% client participation. With $B = 32$ and IID data, almost all client participation ratios yielded similar performance. In the case of the non-IID equal data distribution, we reached a maximum accuracy of 74.69%, precision of 80%, recall of 75%, and an F1-score of 70% when $C = 1.0$ and $B = 32$ were employed. Conversely, the minimum accuracy reached was 65%, with precision at 40%, recall at 65%, and F1-score at 48%. On the other hand, with $B = 10$, we achieved a maximum accuracy of 74.62%, precision of 80%, recall of 75%, and an F1-score of 68% with 100% client participation. However, with 10% client participation, the minimum performance achieved was an accuracy of 67.38%, precision of 70%, recall of 67%, and an F1-score of 62%. For the non-IID unequal data distribution, selecting $C = 1.0$ and $B = 32$ led to a maximum accuracy of 84.12%, precision of 85%, recall of 84%, and an F1-score of 82%. With the same value of B and client participation ratios of 1%, 20%, 50%, and 80%, we obtained roughly the same performance. However, with $B = 10$, we achieved a maximum accuracy of 83%, precision of 85%, recall of 83%, and F1-score of 81%. It is worth noting that our empirical multi-client parallelism experiment highlighted that in the dataset with an extreme data quantity skew, FL performs better when all clients $C = 1.0$, participate in each round. Moreover, it also revealed that extreme data quantity skew significantly hampers the performance of FL settings.

Increasing computation per client

In our experiment, the result obtained by increasing computation per client is presented in Table 5. Involving the detection of AD using an IID data distribution, we attained optimal performance with 82.67% accuracy, 85% precision, 83% recall, and 81% F1-score when we set $E = 5$ and $B = 50$. This outcome closely resembles the results obtained under the settings of $E = 5$ and $B = 10$. However, with the values of $E = 1$ and $B = 32$, we obtained a comparatively minimum accuracy of 72.5%, precision of 79%, recall of 72%, and F1-score of 63%.

Table 6. Comparison with previously conducted research using FL.

Reference	Data	Model	Result
Lei et al.[24]	MRI	Transformer	Achieved accuracy rates of 88.75%, 69.51%, and 69.88% for the two-way classification tasks differentiating AD versus HC, MCI versus HC, and AD versus MCI, respectively
Huang et al.[26]	MRI	CNN	Attained maximum accuracy of 91.9%, a recall of 100%, and a specificity of 91.1%
Stripelis et al.[27]	MRI	CNN	Demonstrated a maximum performance with accuracies of 86%, precision of 80.98%, recall of 81.32%, and an F1-score of 81.14% using ADNI, OASIS, and AIBL dataset for five learners
Proposed	MRI	CNN	Achieved a maximum accuracy of 84.75%, precision of 86%, recall of 85%, and an impressive F1-score of 84%

FL: federated learning; AD: Alzheimers disease; CL: centralized learning; MRI: magnetic resonance imaging; CNN: convolutional neural network; HC: healthy control; MCI: mild cognitive impairment; ADNI: Alzheimers Disease Neuroimaging Initiative; OASIS: Open Access Series of Imaging Studies; AIBL: Australian Imaging, Biomarker and Lifestyle Flagship Study of Aging.

For the scenario of non-IID equal data distribution, we achieved an accuracy of 77.81%, precision of 78%, recall of 78%, and an F1-score of 74% by configuring $E = 20$ and $B = 50$. Conversely, we achieved a comparatively minimum accuracy of 58.5%, precision of 72%, recall of 58%, and F1-score of 51% with $E = 20$ and $B = 32$.

In the case of a non-IID unequal data distribution, we obtained exceptional results by selecting $E = 5$ and $B = 50$ yielding a maximum accuracy of 84.75%, precision of 86%, recall of 85%, and an impressive F1-score of 84% in the classification of AD. However, we obtained a minimum accuracy of 79.12%, precision of 79%, recall of 79%, F1-score of 77 with $E = 1$ and $B = 50$.

This study, while providing valuable insights, does possess several limitations that present opportunities for further investigation. The experimental results showcased herein are derived from a singular CNN model. Consequently, these findings could exhibit variability when replicated with alternative models or under different FL configurations. This suggests a potential area for future research to explore the robustness of our results across a broader range of models and settings. Additionally, the noted decline in classification performance attributed to quantity distribution skew is primarily based on empirical observations. A more comprehensive understanding could be achieved through theoretical analysis, which would provide a deeper insight into the underlying mechanisms affecting performance in varied data distribution scenarios. This dual approach of combining theoretical explorations with empirical validations could significantly enhance the generalizability and reliability of future studies.

In our study, privacy preservation is achieved through the fundamental structure of FL itself, where the raw data remains decentralized on the clients' devices. The only information shared between the clients and the central server is the model weight updates (gradients), rather than the actual data, which ensures that sensitive information (such as the MRI scans) is not transmitted. This communication of weights during each training round inherently addresses privacy concerns by preventing data leakage. Moreover, while we have not implemented additional privacy-enhancing techniques like differential privacy or secure aggregation in this current work, the FL process alone already mitigates a large portion of privacy risks by keeping the data localized. The communication of only the model weights is a key mechanism to ensure privacy, as it abstracts the raw data from being accessible or shared across entities. However, we agree that adding such techniques could further strengthen privacy guarantees in future work.

Conclusion and future work

In this article, we have conducted an empirical and rigorous analysis of AD detection while prioritizing privacy preservation. Our study delves into strategies for increasing parallelism and computation per client to mitigate communication costs, as evidenced by experiments on AD classification. We have evaluated these methodologies across different data distributions: IID, non-IID equal, and non-IID unequal. Across all distributions, our selected approach has demonstrated satisfactory performance in AD detection. Note that the choice of the non-convex LeNet5 model in this article is driven by its established theoretical advantages and computational efficiency, which we believe suited our initial study aims. To support the robustness of our approach, our evaluation is conducted across diverse datasets, which demonstrated consistent performance. However, we recognize the need for further validation using a variety of models to fully ascertain the generalizability and robustness of our findings. Plans are underway to include additional CNN-based models in our future work to address these critical aspects comprehensively. In our current study, we focused on a four-class classification to demonstrate the capabilities of our proposed methodology in a more complex scenario. We acknowledge the importance of binary and three-class classifications as they may be fundamental in many clinical applications. While these were not included in this article, we are considering these simpler classification tasks for future work to provide a comprehensive evaluation of our methodology across different classification scenarios. This approach will allow us to further validate the versatility and applicability of our proposed methods. Looking ahead, our future research endeavors will focus on exploring personalized FL configurations aimed at developing models tailored to individual clients.

Furthermore, we aim to investigate the incorporation of differential privacy techniques to bolster the security of communication between the central server and each client. These directions promise to enhance both the effectiveness and privacy assurances of FL-based AD detection systems. In addition, to enhance the sophistication of AD detection, it would be beneficial to incorporate multi-modal data, such as blood-based data, PET scans, EEG, and other relevant sources. Applying test cases in active learning and validating the results through biological or medical data would be an important step in further assessing the scalability and robustness of the model in the future.

Acknowledgements: We thank the Computer Science and Engineering Department of Hajee Mohammad Danesh Science and Technology University, Bangladesh for supporting the study process.

Availability of data and materials: The dataset utilized in this study is available at Kaggle, https://www.kaggle.com/datasets/sachinkumar413/alzheimer-mri-dataset.

Contributorship: Md Abdus Sahid: conceptualization, methodology, software, validation, data curation, data analysis, visualization, and manuscript drafting. Md Palash Uddin: conceptualization, methodology, resources, investigation, manuscript review, manuscript finalization, and supervision. Hasi Saha: investigation, manuscript review, and supervision. Md Rashedul Islam: co-supervision.

Declaration of conflicting interests: The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Ethical approval: This research is not subject to ethical approval since the research did not have participants (humans or animals).

Funding: The author(s) received no financial support for the research, authorship, and/or publication of this article.

Guarantor: MPU.

Informed consent: In accordance with the ICMJE guidelines, we confirm that patient consent was not applicable to this study. This study did not involve human subjects, and all data were obtained from publicly available sources.

ORCID iD: Md Palash Uddin https://orcid.org/0000-0002-4429-6590

References

1. Perez-Heydrich C, Walker C, Pile M, et al. Comparison of digital recruitment strategies for Alzheimer's disease patients. *Digital Health* 2024; 10: 20552076241229164.
2. Petersen RC, Smith GE, Waring SC, et al. Mild cognitive impairment: clinical characterization and outcome. *Arch Neurol* 1999; 56: 303–308.
3. Siuly S, Alçin ÖF, Kabir E, et al. A new framework for automatic detection of patients with mild cognitive impairment using resting-state EEG signals. *IEEE Trans Neural Syst Rehabil Eng* 2020; 28: 1966–1976.
4. Khatun S, Morshed BI and Bidelman GM. A single-channel EEG-based approach to detect mild cognitive impairment via speech-evoked brain responses. *IEEE Trans Neural Syst Rehabil Eng* 2019; 27: 1063–1070.
5. McBride JC, Zhao X, Munro NB, et al. Spectral and complexity analysis of scalp EEG characteristics for mild cognitive impairment and early Alzheimer's disease. *Comput Methods Programs Biomed* 2014; 114: 153–163.
6. Tsai CL, Pai MC, Ukropec J, et al. The role of physical fitness in the neurocognitive performance of task switching in older persons with mild cognitive impairment. *J Alzheimer's Dis* 2016; 53: 143–159.
7. Durongbhan P, Zhao Y, Chen L, et al. A dementia classification framework using frequency and time-frequency features based on EEG signals. *IEEE Trans Neural Syst Rehabil Eng* 2019; 27: 826–835.
8. Alzheimer's Association. 2015 Alzheimer's disease facts and figures. *Alzheimer's Dement* 2015; 11: 332–384.
9. Yang S, Bornot JM, Wong-Lin K, et al. M/EEG-based biomarkers to predict the MCI and Alzheimer's disease: a review from the ML perspective. *IEEE Trans Biomed Eng* 2019; 66: 2924–2935.
10. Butler J, Watermeyer TJ, Matterson E, et al. The development and validation of a digital biomarker for remote assessment of Alzheimer's diseases risk. *Digital Health* 2024; 10: 20552076241228416.
11. Braunwald E, Fauci AS, Kasper DL, et al. *Harrison's principles of internal medicine 15th*. NY: McGraw-Hill Book Company, 2001.
12. Gallego-Jutglà E, Solé-Casals J, Vialatte FB, et al. A hybrid feature selection approach for the early diagnosis of Alzheimer's disease. *J Neural Eng* 2015; 12: 016018.

13. Ghorbanian P, Devilbiss DM, Verma A, et al. Identification of resting and active state EEG features of Alzheimer's disease using discrete wavelet transform. *Ann Biomed Eng* 2013; 41: 1243–1257.
14. Ju R, Hu C and Li Q. Early diagnosis of Alzheimer's disease based on resting-state brain networks and deep learning. *IEEE/ACM Trans Comput Biol Bioinform* 2017; 16: 244–257.
15. Li K, Ma X, Chen T, et al. A new early warning method for mild cognitive impairment due to Alzheimer's disease based on dynamic evaluation of the "spatial executive process". *Digital Health* 2023; 9: 20552076231194938.
16. Feeney Mahoney D, Coon DW and Lozano C. Latino/Hispanic Alzheimer's caregivers experiencing dementia-related dressing issues: corroboration of the preservation of self model and reactions to a "smart dresser" computer-based dressing aid. *Digital Health* 2016; 2: 2055207616677129.
17. Zhang D, Wang Y, Zhou L, et al. Multimodal classification of Alzheimer's disease and mild cognitive impairment. *Neuroimage* 2011; 55: 856–867.
18. Ahmed OB, Benois-Pineau J, Allard M, et al. Recognition of Alzheimer's disease and mild cognitive impairment with multimodal image-derived biomarkers and multiple kernel learning. *Neurocomputing* 2017; 220: 98–110.
19. Mesrob L, Sarazin M, Hahn-Barma V, et al. DTI and structural MRI classification in Alzheimer's disease. *Adv Mol Img* 2012; 2: 12.
20. Uddin MP, Xiang Y, Lu X, et al. Federated learning via disentangled information bottleneck. *IEEE Trans Serv Comput* 2022; 32: 1874–1889.
21. Uddin MP, Xiang Y, Yearwood J, et al. Robust federated averaging via outlier pruning. *IEEE Sig Process Lett* 2021; 29: 409–413.
22. Uddin MP, Xiang Y, Lu X, et al. Mutual information driven federated learning. *IEEE Trans Parallel Distrib Syst* 2020; 32: 1526–1538.
23. McMahan B, Moore E, Ramage D, et al. Communication-efficient learning of deep networks from decentralized data. In: *Artificial intelligence and statistics*, 10 Apr 2017, pp.1273–1282. PMLR.
24. Lei B, Zhu Y, Liang E, et al. Federated domain adaptation via transformer for multi-site Alzheimer's disease diagnosis. *IEEE Trans Med Imaging* 2023; 42: 3651–3664.
25. Lakhan A, Grønli TM, Muhammad G, et al. EDCNNS: federated learning enabled evolutionary deep convolutional neural network for Alzheimer disease detection. *Appl Soft Comput* 2023; 147: 110804.
26. Huang YL, Yang HC and Lee CC. Federated learning via conditional mutual learning for Alzheimer's disease classification on T1W MRI. In: *2021 43rd annual international conference of the IEEE engineering in medicine & biology society (EMBC)*, 1 Nov 2021, pp.2427–2432. IEEE.
27. Stripelis D, Gupta U, Saleem H, et al. Secure federated learning for neuroimaging. arXiv preprint arXiv:2205.05249, 11 May 2022.
28. Movahed RA and Rezaeian M. Automatic diagnosis of mild cognitive impairment based on spectral, functional connectivity, and nonlinear EEG-based features. *Comput Math Methods Med* 2022; 2022: 1–17.
29. Plant C, Teipel SJ, Oswald A, et al. Automated detection of brain atrophy patterns based on MRI for the prediction of Alzheimer's disease. *Neuroimage* 2010; 50: 162–174.
30. Sarraf S, DeSouza DD, Anderson J, et al. DeepAD: Alzheimer's disease classification via deep convolutional neural networks using MRI and fMRI. *BioRxiv* 2016; 2016: 1–32.
31. Aghajani H, Zahedi E, Jalili M, et al. Diagnosis of early Alzheimer's disease based on EEG source localization and a standardized realistic head model. *IEEE J Bio Health Info* 2013; 17: 1039–1045.
32. Alzheimer MRI Preprocessed Dataset. https://www.kaggle.com/datasets/sachinkumar413/alzheimer-mri-dataset/data, 2 Jan 2023.
33. No Author. Alzheimer's Disease Neuroimaging Initiative. https://adni.loni.usc.edu/, 2 Jan 2023.
34. No Author. Alzheimers.net. https://www.alzheimers.net/, 2 Jan 2023.
35. No Author. MRI and Alzheimers. https://www.kaggle.com/datasets/jboysen/mri-and-alzheimers, 2 Jan 2023.
36. No Author. Alzheimer's Disease and Healthy Aging Data. https://catalog.data.gov/dataset/alzheimers-disease-and-healthy-aging-data, 2 Jan 2023.
37. No Author. European Prevention of Alzheimer's Dementia. https://cordis.europa.eu/article/id/429468-the-final-epad-dataset-is-now-available-on-the-alzheimer-s-disease-workbench, 2 Jan 2023.
38. Kairouz P, McMahan HB, Avent B, et al. Advances and open problems in federated learning. *Found Trends Mach Learn* 2021; 14: 1–210.
39. Krizhevsky A, Sutskever I and Hinton GE. Imagenet classification with deep convolutional neural networks. *Adv Neural Inf Process Syst* 2012; 25: 1–9.
40. LeCun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition. *Proc IEEE* 1998; 86: 2278–2324.
41. Powers DM. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. arXiv preprint arXiv:2010.16061, 11 Oct 2020.
42. Buckland M and Gey F. The relationship between recall and precision. *J Am Soc Inf Sci* 1994; 45: 12–19.
43. Sahid MA, Hasan M, Akter N, et al. Effect of imbalance data handling techniques to improve the accuracy of heart disease prediction using machine learning and deep learning. In: *2022 IEEE region 10 symposium (TENSYMP)*, 1 Jul 2022, pp.1–6. IEEE.
44. Hasan M, Sahid MA, Uddin MP, et al. Performance discrepancy mitigation in heart disease prediction for multisensory inter-datasets. *PeerJ Comput Sci* 2024; 10: e1917.