http://dx.doi.org/10.21622/ACE.2025.05.2.1567

A FEDERATED FRAMEWORK FOR SPEECH-BASED EARLY DETECTION OF ALZHEIMER'S DISEASE

Mohamed Mourad Abdellattif[1], Abdelrahman Mohamed Farouk[2], Nada Hamada Ahmed[3], Nadine Ahmed Elquersh[4], Ahmed Hamdy Elshennawy[5] and Noha S. Tawfik[6]

[1:6] Arab Academy for Science, Technology and Maritime Transport, College of Engineering and Technology, Computer Engineering Department, Abou Kir, Alexandria, Egypt

Emails: {Mohamedmourad22002@gmail.com, abdelrahmannmofaroukk@gmail.com, nadahamada1712@gmail.com, nadine.quersh@gmail.com, a.elshennawy282@gmail.com, *Noha.abdelsalam@aast.edu}

Received on, 09 August 2025 - Accepted on, 25 August 2025 - Published on, 11 November 2025

ABSTRACT

The development of artificial intelligence for Alzheimer's disease (AD) diagnostics is often hindered by data privacy regulations that prevent the aggregation of sensitive patient information. Federated Learning (FL) offers a decentralized solution, enabling collaborative model training without sharing raw data. This paper presents a robust FL framework for the early detection of AD using spontaneous speech from the ADReSS dataset. We systematically evaluate the optimal components for a privacy-preserving pipeline by simulating a cross-silo federated environment. Our methodology involves comparing multiple feature extraction techniques, where VGGish audio embeddings proved most effective, and two classification models, with the Multi-Layer Perceptron (MLP) demonstrating superior performance. We further optimized the framework by comparing FedAvg, FedAvgM, and FedProx aggregation strategies, identifying FedAvgM as the most stable and effective. Our results show that the collaborative FL model significantly outperforms models trained on isolated local data. The final optimized framework achieved a state-of-the-art accuracy of 87.50% and 81.25% in a 2-client and 3-client setting, respectively. This study validates the feasibility of using federated learning to build scalable, accurate, and ethical diagnostic tools for Alzheimer's disease.

Keywords: *Machine Learning, Federated Learning, Alzheimer's detection.*

1. INTRODUCTION

Alzheimer's disease (AD) represents a pressing global health crisis, standing as the most prevalent form of dementia and affecting over 55 million individuals worldwide, with projections expecting this number to rise to 139 million by 2050 [1]. The progressive neurodegenerative nature of AD leads to a gradual decline in memory, cognitive function, and behavior, profoundly impacting patients and their families [2, 3]. Early and accurate diagnosis is paramount for effective disease management, allowing for timely therapeutic interventions and care planning. In recent years, the integration of artificial intelligence (AI) has opened new frontiers for non-invasive diagnostics. Among various modalities, the analysis of spontaneous speech has emerged as a particularly promising biomarker, as subtle changes in language and speech can signal underlying cognitive impairment [2, 4]. However, the full potential of AI in medicine is significantly hampered by a fundamental challenge: the data paradox. While robust and generalizable machine learning

http://dx.doi.org/10.21622/ACE.2025.05.2.1567

(ML) models require access to large, diverse, and representative datasets, medical data is inherently sensitive and siloed. Strict data protection regulations, such as the Health

Insurance Portability and Accountability Act (HIPAA) and the General Data Protection Regulation (GDPR), impose severe restrictions on sharing patient information across institutions [5,6]. This creates a landscape of isolated data islands, preventing researchers from aggregating the necessary data to train high-performing models. This limitation not only curtails model accuracy but also introduces biases, as models trained on limited, homogeneous data often fail to generalize to broader patient populations.

This paper addresses this critical problem by proposing a solution rooted in Federated Learning (FL). FL is a decentralized training paradigm that enables collaborative model development without centralizing raw data [5, 6]. In an FL framework, a shared global model is trained iteratively across multiple client institutions. Each client downloads the global model, trains it locally on its private data, and then sends only the anonymized model updates—not the data itself—back to a central server for aggregation. This privacy-by-design approach makes it possible to harness the collective knowledge of multiple institutions while adhering to strict privacy and regulatory standards, thus resolving the conflict between data access and data protection [6, 7].

The primary objective of this study is to design, implement, and evaluate a federated learning framework for the early detection of Alzheimer's disease using spontaneous speech from the publicly available ADReSS dataset. We simulate a real-world cross-silo environment to investigate the feasibility and effectiveness of this approach. Specifically, this paper makes the following contributions: 1) It evaluates various feature extraction techniques and classification models to identify an optimal pipeline for speech-based AD detection within an FL context. 2) It provides a comparative analysis of different federated aggregation algorithms (FedAvg, FedAvgM, and FedProx) to determine their impact on model convergence and performance. 3) It empirically demonstrates that a collaboratively trained federated model significantly outperforms models trained on iso- lated, local datasets, thereby quantifying the benefits of privacy-preserving collaboration in a clinical research setting. By addressing these objectives, this work aims to provide a robust proof-of-concept for the development of scalable, ethical, and accurate AI-driven diagnostic tools for neurodegenerative diseases.

2. LITERATURE REVIEW

The application of artificial intelligence to dementia diagnostics has rapidly evolved, yet progress is often constrained by the difficulty of accessing large, centralized medical datasets due to strict privacy regulations. Federated Learning (FL) has emerged as a key enabling technology, allowing for collaborative model training across institutional silos without sharing raw patient data. While the majority of machine learning research has focused on centralized approaches, the adoption of FL has grown significantly, particularly for neurodegenerative conditions like Alzheimer's disease (AD).

Initial research into FL for AD primarily leveraged neuroimaging data. For instance, several studies have proposed robust FL frameworks for classifying AD stages from MRI scans, demonstrating that decentralized models can achieve high accuracy comparable to centralized methods while preserving data privacy [8–10]. These works established the viability of FL for complex medical imaging tasks and highlighted its potential for building generalizable models from heterogeneous, multi-institutional data.

More recently, the focus has expanded to include other non-invasive biomarkers, with spontaneous speech gaining significant attention. Centralized approaches, such as those developed for the ADReSS-M Challenge, have shown that speech contains rich acoustic and linguistic features capable of distinguishing between cognitively normal individuals and those with dementia [11]. However, these models still face the fundamental limitation of requiring data centralization.

http://dx.doi.org/10.21622/ACE.2025.05.2.1567

To overcome this barrier, a nascent body of research has begun to explore FL for speech-based AD detection. A pioneering example is ADDETECTOR, an IoT-based FL system that collects voice recordings via smart home devices and uses privacy-enhancing techniques like differential privacy and cryptographic aggregation to protect user data [12]. The system successfully extracted acoustic and linguistic features from the ADReSS dataset to achieve real-time classification, providing a strong proof-of-concept for in-home cognitive monitoring. Another relevant multi-modal system, ADMarker, incorporates audio alongside other digital biomarkers in a home-care setting, using a sophisticated FL pipeline to handle data imbalance and limited labels [13]. These studies highlight a clear trend towards developing privacy-preserving, accessible, and scalable tools for remote cognitive assessment. Our work builds directly upon this focused area, aiming to further investigate the optimal model architectures and aggregation strategies specifically for speech-based AD detection within a cross-silo federated environment.

Wenqing Wei et al. [14] proposed a novel federated learning method, FedCPC, to enhance privacy and performance in early-stage Alzheimer's disease detection from speech signals. The proposed approach consists of a two-step federated training process. First, the method performs a federated contrastive pre-training (FedCPC) to learn better representations from raw audio data. This stage utilizes a convolutional encoder and a gated recurrent unit (GRU) to produce context-aware representations. In the second step, the pre-trained FedCPC model acts as a feature extractor, and a classifier is fine-tuned on top using federated learning to detect AD speech. The experiments were conducted on the 2021 NCMMSC AD Recognition Challenge dataset, which includes training, development, and testing splits with labels for Alzheimer's disease, mild cognitive impairment, and healthy control. Another study by Kalabakov et al. [15] investigated a more realistic federated scenario to evaluate cognitive decline detection. Instead of partitioning a single corpus, they simulated three distinct institutions using separate datasets from DementiaBank: the Pitt-ADReSS dataset (containing healthy controls and AD), the Delaware corpus (healthy controls and MCI), and the Lu dataset (healthy controls and AD). This setup enabled multi-class classification task that includes Mild Cognitive Impairment (MCI). Their methodology compared FedAvg against local and centralized training using XGBoost classifiers on a variety of acoustic and deep embedding features. Zhong et al. [16] conducted a systematic investigation into regularized FL for the related and challenging tasks of dysarthric and elderly speech recognition. They explore three distinct regularization strategies designed to stabilize the FL training process: parameter-based, embedding-based, and a novel loss-based regularization. Using the UASpeech and DementiaBank Pitt corpora, they demonstrated that these regularization techniques consistently and significantly improved performance over the baseline FedAvg framework.

3. MATERIALS AND METHODS

DATASET

The dataset used in this work originates from the ADReSS Challenge (Alzheimer's Dementia Recognition through Spontaneous Speech) [17], a curated and publicly available corpus designed to support the automatic detection of Alzheimer's dementia using spontaneous speech. Speech samples were collected from 156 participants during routine clinical evaluations, where each subject was asked to describe the standard "Cookie Theft" picture from the Boston Diagnostic Aphasia Examination. This task elicits naturalistic but structured speech that reflects cognitive and linguistic capabilities relevant to dementia diagnosis. Participants were divided into a training set of 108 and a test set of 48, with the training data including balanced numbers of cognitively healthy controls and individuals with dementia.

Each subject's cognitive status was further quantified using the Mini-Mental State Examination (MMSE), a widely used 30-point questionnaire that evaluates key domains such as orientation, memory, attention, language, and visuospatial abilities. MMSE

http://dx.doi.org/10.21622/ACE.2025.05.2.1567

scores serve as a clinical indicator of cognitive impairment severity and are included as regression targets in the training set.

To prepare the data for analysis, the ADReSS organizers applied a series of audio preprocessing steps. Volume levels were normalized to ensure consistent amplitude across recordings. Long silences and background noise were removed to focus on active speech, and a uniform sampling rate was enforced across all files. The recordings were then segmented into smaller audio chunks using voice activity detection, and all files were anonymized to protect participant privacy. While the dataset was primarily designed to support speech-only analysis, some studies have later aligned the audio with transcripts to enable multimodal approaches combining acoustic and linguistic features.

Two types of audio files were provided: full wave enhanced recordings and normalized audio chunks. The enhanced recordings consist of complete utterances that have been denoised to preserve the natural prosody and speaking patterns of participants. In contrast, the normalized chunks are short-duration segments extracted from these full recordings after applying amplitude normalization and silence trimming. These segments are more suitable for fine-grained acoustic analysis and frame-level feature extraction. In the training set, the cognitively normal group yielded 1358 normalized segments, while the dementia group yielded 1476, for a total of 2834 segments. The test set includes 1243 additional normalized segments, though diagnosis labels are withheld to support blind model evaluation.

Table 1. ADReSS Dataset Summary.

Subset	Group	Participants			Norm. Chunks
		Total	Male	Female	
Training	CC (Cognitively Normal)	52	21	31	1358
Training	CD (Dementia)	54	27	27	1476
Test	Mixed (CC + CD)	48	24	24	1243

This structure allows researchers to explore speech-based cognitive markers at both the holistic (session-level) and granular (utterance-level) scales. The availability of both full-length and segmented audio, combined with transcript and metadata alignment, provides a rich foundation for multimodal and multi-resolution analysis of speech in Alzheimer's disease research.

FEATURE EXTRACTION

To identify the most effective representation of speech for our classification task, we experimented with multiple feature extraction techniques. The process of feature extraction transforms raw audio signals into a structured and informative representation that can be understood and processed by machine learning algorithms. While rich in detail, audio waveforms are highly variable and unstructured, making them unsuitable as direct input for most models. Instead, meaningful characteristics are extracted through signal processing or deep learning methods. In the context of speech analysis for health-related applications, certain acoustic features can serve as indicators of cognitive or neurological conditions. We evaluated three prominent methods, each offering a different philosophy, from interpretable, hand-crafted features to deep, learned embeddings.

eGeMAPS (extended Geneva Minimalistic Acoustic Parameter Set) [18] is a standardized feature set designed for affective computing and clinical voice analysis. It includes 88 low-level descriptors (LLDs) and derived functionals carefully selected for their relevance to emotional expression, cognitive load, and neurological function. The features are extracted from short-time frames and aggregated over longer windows using statistical operations. These fall into several categories: prosodic features (fundamental frequency, loudness), voice quality features (jitter, shimmer,

http://dx.doi.org/10.21622/ACE.2025.05.2.1567

harmonics-to-noise ratio), spectral features (spectral slope, formant frequencies), and temporal features (rate of voiced segments, pause characteristics). The strength of eGeMAPS lies in its interpretability and clinical relevance, as the features are derived from known speech physiology and phonetics, allowing researchers and clinicians to connect model behavior to observable human traits.

ComParE (Computational Paralinguistics Challenge Feature Set) [19] is one of the most comprehensive hand-engineered acoustic feature sets in speech technology. Developed for the INTERSPEECH Computational Paralinguistics Challenge series, it contains over 6,000 features generated from 65 base LLDs, including MFCCs, energy, pitch, jitter, and shimmer, computed over small analysis frames with a 10ms overlap. What distinguishes ComParE is the application of a wide range of statistical functionals—such as mean, standard deviation, percentiles, skewness, kurtosis, and linear regression coefficients—across the entire recording. The result is a high-dimensional feature vector that captures both global and local speech dynamics. While useful for modeling complex paralinguistic behavior, its size often requires dimensionality reduction or strong regularization to avoid overfitting, especially in datasets with limited samples.

VGG-based features [20] is a deep learning-based audio embedding model inspired by the VGG image classification architecture. Trained on the large-scale AudioSet dataset, it is designed to learn general-purpose audio representations. The VGGish pipeline first converts raw audio to a log-mel spectrogram, which is then divided into patches. Each patch is passed through a convolutional neural network, and the output of the final layer is a 128-dimensional embedding vector representing the acoustic characteristics of that segment. While VGGish embeddings are not specifically designed to be interpretable, they are highly effective for transfer learning, as the model learns to recognize abstract patterns in sound including tone, rhythm, and texture. One common strategy is to apply VGGish to all patches of an audio recording and then average the embeddings to produce a single fixed-length vector for downstream classification. VGGish has been successfully used in sound classification, event detection, and increasingly in voice-based health diagnostics.

Together, these feature extraction approaches provide a wide spectrum of tools for modeling human speech. eGeMAPS focuses on clinical interpretability with minimal dimensionality, ComParE offers a rich statistical profile of acoustic behavior, and VGGish leverages deep learning to extract abstract, transferable audio features. The choice among them depends on the task at hand, the availability of data, and whether interpretability or representation power is prioritized.

CLASSIFICATION MODELS

The final stage in the machine learning pipeline is classification, where structured feature representations are mapped to categorical outputs. In the context of speech-based health analysis, this involves determining whether a given audio sample corresponds to a cognitively normal individual or someone exhibiting signs of neurological impairment. Based on the form and complexity of the input features, different model architectures can be employed. We evaluated two commonly used neural network architectures for this task.

Multi-Layer Perceptron (MLP) is a class of feedforward neural network com- posed entirely of fully connected (dense) layers. As one of the most versatile deep learning architectures, it is suitable for problems where input data can be repre- sented as a fixed-length vector [21]. MLPs are particularly well-suited to scenarios where the extracted features are global and non-sequential, such as statistical acous- tic descriptors or the VGGish embeddings used in this study. The architecture of our MLP includes an input layer that accepts the 128-dimensional embedding vector, followed by one or more hidden layers that apply a linear transformation and a nonlinear ReLU activation function to model complex relationships. To mitigate overfitting, dropout layers are used

http://dx.doi.org/10.21622/ACE.2025.05.2.1567

between hidden layers to randomly deactivate a fraction of neurons during training. The network culminates in an output layer with a single neuron and a sigmoid activation function for binary classification. While MLPs are powerful function approximators that are easy to implement, they treat input features as independent dimensions and do not inherently leverage local or sequential structure within the data. This makes them ideal for pre-aggregated global features but less efficient for time-varying inputs like raw spectrograms.

Lightweight Convolutional Neural Network (LW-CNN) is an adapted form of a traditional CNN, specifically designed for efficiency in low-resource environments or with small datasets. CNNs are well-suited for spatial or sequential data, as they learn filters that extract localized patterns through convolution operations. In audio processing, CNNs are commonly applied to 2D time-frequency representations like log-mel spectrograms. An LW-CNN retains the fundamental building blocks of a full CNN but reduces the number of parameters to improve training efficiency and generalization. The architecture consists of convolutional layers applying small, learnable kernels to detect low-level patterns such as formant transitions or local pitch modulations. These are followed by non-linear activation functions like ReLU and pooling layers (e.g., max-pooling) to reduce the spatial resolution of feature maps and provide translational invariance. Regularization techniques such as batch normalization and dropout are also incorporated to improve convergence. After the high-level features are flattened, a dense output layer performs the final classification. Unlike MLPs, LW-CNNs can learn spatial dependencies within the input data, which is particularly useful for spectrogram-like inputs where neighboring frames carry related information [22] .

Both MLPs and LW-CNNs have been widely adopted in speech-related classification tasks, including emotion recognition, speaker verification, and cognitive health assessment. The choice of architecture depends heavily on the structure of the input data and the desired trade-off between interpretability, performance, and computational efficiency. MLPs offer simplicity and are optimal for flat, aggregated features, whereas LW-CNNs are more suitable for capturing temporal or frequency-localized information in structured audio inputs. In this work, we intentionally avoided the use of highly complex or computationally intensive architectures, as the intended real-life deployment scenario involves resource-constrained edge devices where lightweight, efficient models are more practical and feasible.

4. FEDERATED LEARNING SETTING

FL FRAMEWORK

Our experimental design is centered on a distributed machine learning system where a global model is trained collaboratively without centralizing raw data. In this federated learning (FL) approach, a global model is first initiated on a central server and then distributed to a set of clients. Each client trains this model locally on its private dataset and subsequently returns the updated model weights to the server. After all participating clients have completed their local training for a given round, the server employs an aggregation technique to synthesize the collected model weights into an improved global model. This updated model is then redistributed to the clients, marking the completion of one training round. This iterative process allows the model to generalize and learn from diverse datasets that could not be pooled due to privacy regulations or logistical constraints.

The implementation of such a system requires a suitable software framework. FL frameworks are generally designed to support one of two paradigms: Cross-Silo or Cross-Device. Cross-Silo FL is characterized by a small number of stable, institutional clients (such as hospitals) with large datasets and reliable network connections. In contrast, Cross-Device FL involves a massive number of clients (like mobile or IoT devices) with smaller, more heterogeneous datasets and intermittent connectivity. Given that our

http://dx.doi.org/10.21622/ACE.2025.05.2.1567

use case involves collaboration between distinct data-holding entities, a Cross-Silo approach is most appropriate. A comparison of several prominent FL frameworks is provided in Table 2.

Table 2. Comparison of Federated Learning Frameworks.

Feature	FedML	FLWR	OpenFL	FedScale	FATE	IBMFL
Simulation	✓	✓	-	✓	✓	-
Cross Silo	✓	✓	✓	✓	✓	✓
Cross Device	✓	✓	-	✓	-	-
OS Support	Multi-OS	Multi-OS	L/W	L/M	Linux	L/W

Note: L/W: Linux/Windows, L/M: Linux/MacOS, Multi-OS: Linux/MacOS/Windows.

After evaluating the available options, we selected Flower (FLWR) for our implementation. Flower was chosen for its high degree of flexibility and compatibility with major machine learning libraries like TensorFlow and PyTorch, which streamlined the integration of our existing code. Furthermore, Flower is explicitly designed to be scalable and offers a user-friendly API that simplifies the setup of a basic federated system, making it a practical choice for our research.

5. AGGREGATION TECHNIQUES

A critical component of the FL setting is the server-side aggregation of model updates. The choice of aggregation algorithm plays a central role in the convergence, robustness, and efficiency of the training process. We evaluated three widely used techniques in our experiments. The first is **Federated Averaging (FedAvg)** [23], the foundational FL algorithm where the server computes a weighted average of the client models' weights. If w_t represents the model weights from client i at round t, and n_i is the number of local data points, the server update is calculated as:

$$w^{t+1} = \sum_{i=1}^{K} \frac{n_i}{\sum_j n_j} w_i^t$$

While communication-efficient, FedAvg can struggle with heterogeneous (non-IID) data. To address this, we also tested **FedAvg with Momentum (FedAvgM)** [24], which incorporates a server-side momentum term to stabilize convergence. The update rules are:

$$v^{t+1} = \beta v^t + \sum_{i=1}^{K} \frac{n_i}{\sum_j n_j} (w_i^t - w^t)$$

$$w^{t+1} = w^t - \eta v^{t+1}$$

Finally, we explored **Federated Proximal (FedProx)** [25], an algorithm designed to handle statistical heterogeneity by modifying the local training objective. It adds the following proximal term to the local loss function to penalize large deviations from the global model:

$$\frac{\mu}{2} \|w - w^t\|^2$$

Together, these methods provide a comprehensive toolkit for adapting the federated learning process to diverse and unpredictable real-world conditions.

http://dx.doi.org/10.21622/ACE.2025.05.2.1567

6. RESULTS

Our experimental evaluation was designed to systematically identify the optimal components for our federated learning framework and to demonstrate its superiority over local, non-collaborative training. The presentation of results follows a logical progression, starting with the evaluation of data representation strategies and concluding with a comparative analysis with state-of-the-art. The overall system framework is illustrated in Figure 1. This research provides a comparative analysis that systematically evaluates various components of the federated learning pipeline, including feature extraction, classification models, and aggregation techniques. The study explores two to three different configurations for each parameter to determine their impact on model performance. The complete implementation, including server and client code, data partitions, and configuration files, is publicly available[1]

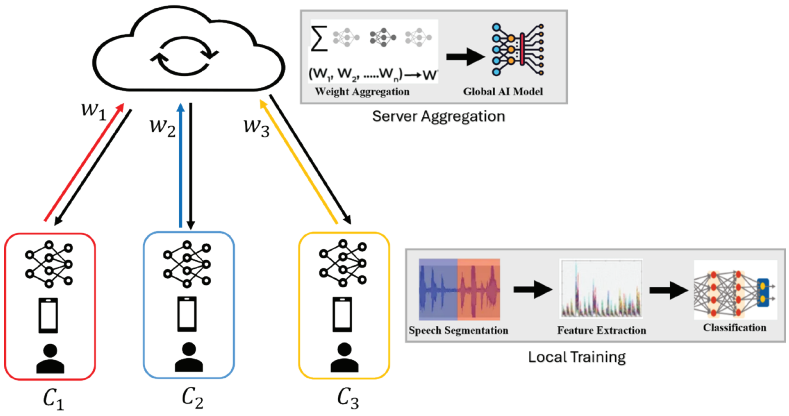

Fig 1. Overview of the proposed federated learning system for Alzheimer's detection from speech.

DATA MODALITY AND FEATURE EXTRACTION

The initial phase of our investigation focused on determining the most effective data representation for the classification task. We first compared the performance of models trained on two different audio modalities: short, normalised audio chunks and the complete, full-wave enhanced audio recordings. This comparison was conducted in a centralized setting, where all data was aggregated on a single node for model training and evaluation. Our findings show that full wave enhanced audio substantially outperformed normalized audio in both accuracy (87.50% vs. 72.11%) and F1-score (87.50% vs. 70.28%). These findings suggest that preserving the full prosodic and contextual structure of speech enhances model discriminability in Alzheimer's classification tasks. Having established the superior modality, we then evaluated three distinct feature extraction techniques to find the most discriminative representation. We compared two handcrafted feature sets, eGeMAPS and ComParE, against VGGish, a deep learning-based embedding. The results, presented in Table 3, show that VGGish features achieved the highest classification accuracy (75.00%) and the best F1-score for the AD class (75%). While eGeMAPS performed competitively, the high-dimensional ComParE set yielded the lowest performance. Consequently, the combination of full-wave enhanced audio and VGGish features was selected as the optimal data representation for all subsequent federated learning experiments.

Table 3. Performance of feature extraction techniques using full wave enhanced audio and centralized classification

Feature	F1-Score (Non-AD)	F1-Score (AD)	Accuracy
eGeMAPS	67%	74%	70.83%
ComParE	49%	68%	60.42%
VGGish	49%	75%	75.00%

http://dx.doi.org/10.21622/ACE.2025.05.2.1567

CLASSIFICATION MODEL

With the optimal data representation determined, we proceeded to select the most effective classification architecture within a federated learning environment. We compared a Multi-Layer Perceptron (MLP) and a Lightweight Convolutional Neural Network (LW- CNN) in a simulated 2-client FL setting. As detailed in Table 4, the MLP model demonstrated superior performance and robustness. It achieved a global accuracy of 77.08% and a weighted F1-score of 76.99%, showing strong generalization despite the performance variation between the individual clients. In contrast, the LW-CNN achieved a lower global accuracy of 66.67%. Despite the LW-CNN's ability to capture temporal features, the MLP proved more robust in this setting, likely due to its compatibility with aggregated embeddings and its lower complexity, which favors edge deployment.

Table 4. Federated Learning Results for MLP and LW-CNN (2 Clients).

Model Metric	Client 1	Client 2	Global
MLP Accuracy	66.67%	87.50%	77.08%
F1 Weighted	-	-	76.99%
LW-CNN Accuracy	66.67%	70.83%	66.67%
F1 Weighted	-	-	65.71%

AGGREGATION TECHNIQUE OPTIMIZATION

To further optimize our federated framework, we evaluated the impact of three different aggregation strategies; FedAvg, FedAvgM, and FedProx—on the performance of the MLP model. The comparison was conducted in both 2-client and 3-client configurations to assess scalability and stability. The results in Table 5 clearly indicate that FedAvgM consistently achieved the highest performance. In the 2-client setup, FedAvgM reached a global accuracy of 87.50%, significantly outperforming FedAvg and FedProx. This advantage was maintained in the 3-client setup, where FedAvgM achieved a global accuracy of 81.25%. While Table 5 focuses on accuracy, we note that the F1 scores generally followed similar trends, with FedAvgM achieving the highest weighted F1 and class-level F1 across both client configurations. These results suggest that FedAvgM, with its momentum-based optimization, facilitates better convergence and generalization, particularly when more clients are involved.

Table 5. Client-wise and Global Accuracy Across Aggregation Methods

Aggregation		Clients	C1 Acc	C2 Acc	C3 Acc	Global Acc
FedAvg	2	Clients	66.67%	87.50%	-	77.08%
	3	Clients	62.50%	81.25%	68.75%	70.83%
FedAvgM	2	Clients	79.17%	95.83%	-	87.50%
	3	Clients	81.25%	87.5%	75.00%	81.25%
FedProx	2	Clients	66.67%	87.50%	-	77.08%
	3	Clients	62.50%	81.25%	68.75%	70.83%

FEDERATED VS. LOCAL PERFORMANCE

Furthermore, to quantify the benefit of our privacy-preserving collaborative approach, we compared the performance of the optimized global federated model (MLP with FedAvgM) against models trained in isolation on each client's local data. While local training also preserves data privacy by avoiding any form of data sharing, it suffers from limited data availability per client, which often leads to suboptimal model performance and reflects

http://dx.doi.org/10.21622/ACE.2025.05.2.1567

realistic deployment scenarios constrained by current data protection regulations.

As illustrated in Table 6, the results show superiority of FL learning over traditional non-collaborative approach. The locally trained models struggled significantly, with accuracies as low as 37.50% and poor F1-scores, highlighting the limitations of training on small, siloed datasets. This improvement stems from the collaborative nature of FL, which allows clients with weaker or less diverse data to benefit from the patterns learned across the entire network. This finding provides clear evidence that federated learning provides a more robust and generalizable solution for medical voice-based AD detection while respecting patient privacy.

Table 6. Comparison of local Vs Federated learning across clients.

Setting	Client 1		Client 2		Client 3	
	Local	FL	Local	FL	Local	FL
2-Client Setting	58.33%	70.83%	58.33%	91.67%	-	-
3-Client Setting	37.50%	81.25%	50.00%	81.25%	56.25%	68.75%

Note: "Local" refers to the accuracy of a model trained exclusively on that client's data. "FL" refers to the client's model accuracy after participating in federated training.

COMPARISON WITH STATE-OF-THE-ART

To contextualize the performance of our proposed framework, we compare our results against a relevant state-of-the-art benchmark for federated AD detection from speech [12]. The study investigating the impact of client scalability reported a global model accuracy of approximately 82% with two clients and 81.8% with three clients, observing a degradation in performance as the number of participants increased [12]. While our results exhibit the same degradation pattern for increasing number of clients, our framework demonstrates highly competitive, and in some cases superior, performance. As presented in Table 7, our model achieved a global accuracy of 87.50% in the 2-client setting, surpassing the benchmark by a significant margin of over 5 percentage points. In the 3-client configuration, our model maintained strong performance with an accuracy of 81.25%, which is closely comparable to the state-of-the-art result.

Table 7. Performance Comparison with Other Models

Model	2-Client Setting	3-Client Setting
ADDetector [12]	82.0%	81.8%
Our Model	87.50%	81.25%

While both our model and the benchmark exhibit a slight decrease in accuracy when scaling from two to three clients—a common characteristic in federated systems due to increased heterogeneity—our framework's substantially higher accuracy in the 2-client scenario highlights its robustness and effectiveness. Furthermore, the inherent simplicity of the MLP architecture ensures minimal computational overhead, making the framework well-suited for deployment on edge devices with limited resources.

7. LIMITATIONS

However, it is important to acknowledge the limitations of this study. The federated experiments were conducted in a simulated environment, which does not capture the full complexities of real-world deployment, such as network latency or asynchronous client updates. While the proposed framework demonstrates the feasibility of federated learning for speech-based Alzheimer's detection, it was implemented in a simulated

http://dx.doi.org/10.21622/ACE.2025.05.2.1567

environment without integrating advanced privacy-preserving or adversarial defense mechanisms. Real-world deployments would require robust safeguards against threats such as model poisoning, model inversion, and membership inference, as well as secure aggregation protocols. Our focus in this study was to establish a performance baseline for federated speech classification, similar to prior dementia-focused FL research on datasets such as OASIS and ADNI, which likewise omitted a full security stack in their initial feasibility evaluations [26–28]. Another key limitation is A key observation is the decrease in global model accuracy from 87.50% in the 2-client setting to 81.25% with three clients. This performance degradation is expected as a direct consequence of increased statistical heterogeneity, a fundamental and well-documented challenge in federated learning. As more clients are added, the data becomes more fragmented and the non-IID (non-identically and independently distributed) nature of the data is amplified. The ADReSS dataset, while a valuable benchmark, is of a finite size; partitioning it across more clients results in smaller, more idiosyncratic local datasets, making it inherently more challenging for a single global model to generalize across all unique data distributions. This effect would likely be less pronounced if the additional clients contributed data of higher quality or a distribution more aligned with the existing data. The 3-client model's accuracy still represents a profound improvement over models trained in isolation (local), confirming the key advantage of the collaborative federated approach. However, this highlights the need for further research into the framework's scalability.

8. CONCLUSION

This study successfully demonstrates the significant practical benefits of a federated learning framework for the early detection of Alzheimer's disease from spontaneous speech, directly addressing the critical challenge of patient privacy in medical AI. Our primary finding is that a collaborative model can be trained effectively across distributed clients without centralizing sensitive raw audio data. By adopting an FL framework, we ensure that all personally identifiable information remains securely on local client devices, adhering to the principles of major data protection regulations. This privacy-by-design approach is a fundamental strength of our work, offering a viable path for building scalable and ethical AI tools that do not force a trade-off between performance and privacy.

The results quantitatively underscore the dramatic advantage of collaborative learning over isolated data analysis. In our 3-client setting, models trained exclusively on local data performed poorly, with an average accuracy of just 47.92%. The collaboratively trained federated global model achieved a robust accuracy of 81.25%. The improvement was even more pronounced in the 2-client scenario, where the federated model reached an accuracy of 87.50% almost 30% increase over the 58.33% average of the locally trained counterparts. This highlights how FL can overcome the limitations of smaller, institution-specific datasets, a crucial factor in medical research where data is often scarce. The systematic evaluation leading to the success of the VGGish embeddings combined with an MLP architecture and the FedAvgM aggregation strategy provides a clear and validated methodology for achieving high performance in this domain. The choice of a relatively simple MLP model, while the choice driven by resource constraints, also proved to be a strength, as its low computational overhead makes it ideal for deployment on edge devices.

In conclusion, this research serves as a strong proof-of-concept for the application of federated learning in speech-based cognitive assessment. Our model achieved a state-of-the-art accuracy of 87.50% in a 2-client configuration, highlighting the tangible benefits of our approach. Future work should focus on several key areas. A significant extension would be to develop a multimodal framework that integrates Natural Language

http://dx.doi.org/10.21622/ACE.2025.05.2.1567

Processing (NLP) features extracted from the speech transcripts, allowing the model to capture linguistic markers of cognitive decline in addition to the acoustic features already analyzed. Validating this enhanced framework in real-world clinical trials with diverse patient populations and exploring the integration of more advanced privacy-enhancing technologies are also critical next steps.

REFERENCES

[1] WHO, "Dementia," 2025. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/dementia

[2] NHS, "Alzheimer's disease," 2025.

[3] "Association A. Dementia vs. Alzheimer's Disease: What Is the Difference?," 2025.

[4] J. T. Becker, F. Boller, O. Lopez, J. Saxton, and K. McGonigle, "The Natural History of Alzheimer's Disease," Arch Neurol, vol. 51, no. 6, p. 585, Jun. 1994, doi: 10.1001/archneur.1994.00540180063015.

[5] "IBM Research. What is Federated Learning?," 2021. [Online]. Available: https://www.ibm.com/think/topics/federated-learning.

[6] "What is federated learning?," 2022. [Online]. Available: https://theodi.org/insights/explainers/what-is-federated-learning/

[7] A. Gooday, "Understanding the Types of Federated Learning - OpenMined," Aug. 2020. [Online]. Available: https://openmined.org/blog/federated-learning-types/

[8] T. Ghosh, M. I. A. Palash, M. A. Yousuf, M. A. Hamid, M. M. Monowar, and M. O. Alassafi, "A Robust Distributed Deep Learning Approach to Detect Alzheimer's Disease from MRI Images," Mathematics, vol. 11, no. 12, 2023, doi: 10.3390/math11122633.

[9] C. Qian, H. Xiong, and J. Li, "FeDeFo: A Personalized Federated Deep Forest Framework for Alzheimer's Disease Diagnosis," Jul. 2023, pp. 572–577. doi: 10.18293/SEKE2023-013.

[10] N. K. Trivedi, S. Jain, and S. Agarwal, "Identifying and Categorizing Alzheimer's Disease with Lightweight Federated Learning Using Identically Distributed Images," in 2024 11th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), IEEE, Mar. 2024, pp. 1–5. doi: 10.1109/ICRITO61523.2024.10522428.

[11] S. Luz, F. Haider, D. Fromm, I. Lazarou, I. Kompatsiaris, and B. MacWhinney, "An Overview of the ADReSS-M Signal Processing Grand Challenge on Multilingual Alzheimer's Dementia Recognition Through Spontaneous Speech," IEEE Open Journal of Signal Processing, vol. 5, pp. 738–749, 2024, doi: 10.1109/OJSP.2024.3378595.

[12] J. Li et al., "A Federated Learning Based Privacy-Preserving Smart Healthcare System," IEEE Trans Industr Inform, vol. 18, no. 3, pp. 2021–2031, Mar. 2022, doi: 10.1109/TII.2021.3098010.

[13] X. Ouyang et al., "ADMarker: A Multi-Modal Federated Learning System for Monitoring Digital Biomarkers of Alzheimer's Disease," in Proceedings of the 30th Annual International Conference on Mobile Computing and Networking, New York, NY, USA: ACM, May 2024, pp. 404–419. doi: 10.1145/3636534.3649370.

[14] W. Wei et al., "FedCPC: An Effective Federated Contrastive Learning Method for Privacy Preserving Early-Stage Alzheimers Speech Detection," in 2023 IEEE Automatic Speech Recognition and Understanding Workshop, ASRU 2023, 2023. doi: 10.1109/ASRU57964.2023.10389690.

[15] S. Kalabakov, M. Gonzalez-Machorro, F. Eyben, B. W. Schuller, and B. Arnrich, "A Comparative Analysis of Federated Learning for Speech-Based Cognitive Decline Detection," in Interspeech 2024, ISCA: ISCA, Sep. 2024, pp. 2455–2459. doi: 10.21437/Interspeech.2024-996.

http://dx.doi.org/10.21622/ACE.2025.05.2.1567

[16] T. Zhong, M. Geng, S. Hu, G. Li, and X. Liu, "Regularized Federated Learning for Privacy-Preserving Dysarthric and Elderly Speech Recognition," Proc. Interspeech , pp. 2103–2107, 2025.

[17] S. Luz, F. Haider, S. de la Fuente Garcia, D. Fromm, and B. MacWhinney, "Editorial: Alzheimer's Dementia Recognition through Spontaneous Speech," Front Comput Sci, vol. 3, Oct. 2021, doi: 10.3389/fcomp.2021.780169.

[18] F. Eyben et al., "The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing," IEEE Trans Affect Comput, vol. 7, no. 2, pp. 190–202, Apr. 2016, doi: 10.1109/TAFFC.2015.2457417.

[19] B. Schuller et al., "The INTERSPEECH 2013 computational paralinguistics challenge: social signals, conflict, emotion, autism," in Interspeech 2013, ISCA: ISCA, Aug. 2013, pp. 148–152. doi: 10.21437/Interspeech.2013-56.

[20] S. Hershey et al., "CNN architectures for large-scale audio classification," in 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, Mar. 2017, pp. 131–135. doi: 10.1109/ICASSP.2017.7952132.

[21] I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning. MIT Press, 2016.

[22] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," Proceedings of the IEEE, vol. 86, no. 11, pp. 2278–2324, 1998, doi: 10.1109/5.726791.

[23] H. Brendan McMahan, E. Moore, D. Ramage, S. Hampson, and B. Agüera y Arcas, "Communication-efficient learning of deep networks from decentralized data," in Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, AISTATS 2017, 2017.

[24] Tzu Ming Hsu, Hang Qi, and Matthew Brown, "Measuring the Effects of Non-Identical Data Distribution for Federated Visual Classification," arXiv preprint , 2019.

[25] T. Li, A. K. Sahu, A. Talwalkar, and V. Smith, "Federated Learning: Challenges, Methods, and Future Directions," IEEE Signal Process Mag, vol. 37, no. 3, 2020, doi: 10.1109/MSP.2020.2975749.

[26] T. Ghosh, M. I. A. Palash, M. A. Yousuf, Md. A. Hamid, M. M. Monowar, and M. O. Alassafi, "A Robust Distributed Deep Learning Approach to Detect Alzheimer's Disease from MRI Images," Mathematics, vol. 11, no. 12, p. 2633, Jun. 2023, doi: 10.3390/math11122633.

[27] N. K. Trivedi, S. Jain, and S. Agarwal, "Identifying and Categorizing Alzheimer's Disease with Lightweight Federated Learning Using Identically Distributed Images," in 2024 11th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), IEEE, Mar. 2024, pp. 1–5. doi: 10.1109/ICRITO61523.2024.10522428.

[28] J. Li et al., "A Federated Learning Based Privacy-Preserving Smart Healthcare System," IEEE Trans Industr Inform, vol. 18, no. 3, pp. 2021–2031, Mar. 2022, doi: 10.1109/TII.2021.3098010.