Annals of Healthcare Systems Engineering

www.ahse.reapress.com

Ann. Healthc. Syst. Eng. Vol. 3, No. 2 (2026) 134–147.

Paper Type: Original Article

Enhancing the Robustness of Federated Learning Models for Dementia Diagnosis Against Data Poisoning Attacks

Fatemeh Ebrahimzadeh[1,*], Sedigheh Kaveh[1], Mohsen Falah Rad[2]

[1] Department of Computer Engineering, Ayandegan University, Tonekabon, Iran; f.ebrahimzadeh@aihe.ac.ir; s.kaveh@aihe.ac.ir.
[2] Department of Computer Engineering, La.C., Islamic Azad University, Lahijan, Iran; mo.falahrad@iau.ac.ir.

Citation:

Received: 17 May 2025 *Revised: 22 September 2025* *Accepted: 07 November 2025*	Ebrahimzadeh, F., Kaveh, S., & Falah Rad, M. (2026). Enhancing the robustness of federated learning models for dementia diagnosis against data poisoning attacks. *Annals of Healthcare Systems Engineering, 3*(2), 134-147.

Abstract

Federated Learning (FL) enables privacy preserving dementia diagnosis but is vulnerable to poisoning attacks. We propose a hybrid defense integrating robust aggregation, client behavior analysis, feature consistency validation, and anomaly aware training. Results on ADNI, OASIS-3, and DementiaBank improve robustness while maintaining accuracy.

Keywords: Federated learning, Dementia diagnosis, Data poisoning attacks, Backdoor attacks, Robust aggregation.

1 | Introduction

Dementia, particularly Alzheimer's Disease (AD), is a progressive neurodegenerative disorder that affects more than 55 million individuals worldwide and is projected to triple by 2050 [1]. Early and accurate diagnosis plays a critical role in slowing disease progression, planning therapeutic interventions, and improving patient quality of life. In recent years, Machine Learning (ML) and Deep Learning (DL) have emerged as powerful tools for dementia diagnosis, leveraging multimodal data such as structural Magnetic Resonance Imaging (MRI), Positron Emission Tomography (PET), speech recordings, and cognitive behavioral assessments. Large-scale neuroimaging datasets such as the Alzheimer's Disease Neuroimaging Initiative (ADNI), OASIS 2, and OASIS 3 have enabled the development of high performance diagnostic models. For example, OASIS 3 provides more than 2842 MRI sessions, 2157 PET scans, and extensive longitudinal clinical metadata collected over 30 years [2], while OASIS 2 offers 373 MRI sessions from 150 older adults with detailed Clinical Dementia Rating (CDR) annotations [3]. These datasets have facilitated

Corresponding Author: f.ebrahimzadeh@aihe.ac.ir
https://doi.org/10.22105/ahse.v3i2.65

the training of DL models capable of capturing subtle neurodegenerative patterns that are often imperceptible to human experts. Despite these advances, the sensitive nature of medical data and strict privacy regulations such as the Health Insurance Portability and Accountability Act (HIPAA) and the General Data Protection Regulation (GDPR) impose significant restrictions on centralized data aggregation. Federated Learning (FL) has therefore emerged as a promising paradigm for collaborative model training across hospitals and research centers without requiring the sharing of raw patient data. In FL, each institution trains the model locally and only shares model updates with a central server, thereby preserving privacy while enabling large-scale learning. Several studies have demonstrated the potential of FL in medical imaging applications. Sheller et al. [4] showed that FL can achieve performance comparable to centralized training for brain tumor segmentation, while Chen et al. [5] demonstrated successful MRI harmonization across multiple sites using FL. However, despite its advantages, FL is highly vulnerable to data poisoning attacks. In FL, the central server aggregates model updates from multiple clients, and a single malicious client can significantly degrade the global model by injecting corrupted samples, manipulating gradients, or embedding backdoor triggers. Bagdasaryan et al. [6] demonstrated that backdoor attacks can be executed stealthily in FL without significantly affecting validation accuracy, while Fang et al. [7] showed that local model poisoning can severely disrupt global convergence. These vulnerabilities are amplified in medical FL systems due to heterogeneous data distributions across institutions, limited client participation, and the high sensitivity of clinical predictions. For example, MRI scans acquired from different scanners or protocols exhibit substantial variability, making it difficult to distinguish between benign heterogeneity and malicious manipulation. Furthermore, the small number of participating hospitals increases the influence of each client, making FL more susceptible to targeted attacks.

Existing defenses such as robust aggregation (e.g., Krum, Trimmed Mean, Median), anomaly detection, and differential privacy remain insufficient for high-stakes medical applications like dementia diagnosis. Robust aggregation methods often fail under stealthy poisoning or backdoor attacks, especially when malicious updates mimic natural data variability. Differential privacy introduces noise that can degrade diagnostic accuracy, which is unacceptable in clinical settings. Anomaly detection methods struggle with non-IID data, which is inherent in medical imaging due to differences in scanners, demographics, and disease severity. This paper addresses this critical gap by proposing a hybrid defense framework designed to enhance the robustness of FL-based dementia diagnostic models against poisoning attacks while preserving diagnostic accuracy and cross-institutional privacy. The proposed framework integrates robust aggregation, client behavior modeling, neuro-feature consistency checks, and anomaly-aware federated training. The contributions of this work are fourfold. First, we develop a threat-model taxonomy tailored to dementia diagnosis FL systems, covering label flipping, backdoor, gradient manipulation, and stealth poisoning attacks. Second, we introduce a novel hybrid defense framework that combines robust aggregation with client behavior modeling and neuro-feature consistency validation. Third, we design a federated anomaly-aware training mechanism that detects poisoned updates using temporal behavioral signatures. Finally, we conduct extensive experiments on ADNI, OASIS 3, and speech-based dementia datasets, demonstrating significant improvements in robustness and diagnostic stability. This work represents a significant step toward secure and reliable federated dementia diagnosis systems, addressing the unique challenges posed by medical data heterogeneity, privacy constraints, and adversarial threats.

2 | Background and Related Work

Dementia diagnosis has undergone a major transformation with the integration of ML and DL techniques, enabling automated extraction of subtle neurodegenerative patterns from multimodal biomedical data. Traditional diagnostic approaches relied heavily on clinical assessments, neuropsychological tests, and expert interpretation of neuroimaging scans. However, the emergence of large-scale datasets such as ADNI, OASIS 2, and OASIS 3 has facilitated the development of data-driven models capable of identifying early biomarkers of AD with high accuracy. These datasets provide longitudinal MRI, PET, cognitive scores, and demographic information, enabling researchers to model disease progression and detect early-stage

cognitive decline. For example, OASIS 3 includes more than 2842 MRI sessions and 2157 PET scans collected over 30 years, offering a rich multimodal resource for studying structural and functional brain changes associated with dementia [2]. Similarly, OASIS 2 provides longitudinal MRI data from 150 older adults, enabling analysis of whole-brain atrophy rates and their association with CDR scores [3]. These datasets have been widely used to train Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Graph Neural Networks (GNNs) for dementia classification and progression prediction.

2.1|Dementia Diagnosis with Machine Learning

ML models for dementia diagnosis primarily rely on neuroimaging biomarkers extracted from MRI and PET scans. Structural MRI provides detailed information about brain anatomy, enabling measurement of hippocampal volume, cortical thickness, ventricular enlargement, and whole-brain atrophy key indicators of AD. Studies have shown that hippocampal atrophy is one of the earliest and most reliable biomarkers of AD, often preceding clinical symptoms by several years [8]. OASIS-2 demonstrated that normalized Whole-Brain Volume (nWBV) declines at approximately 0.49% per year in healthy aging and 0.87% per year in AD, highlighting the sensitivity of MRI-based biomarkers [3]. DL models such as 3D-CNNs have been widely used to capture these structural patterns directly from MRI volumes, achieving high accuracy in distinguishing between Cognitively Normal (CN), Mild Cognitive Impairment (MCI), and AD subjects. In addition to MRI, PET imaging provides functional biomarkers such as amyloid−β deposition and tau pathology. PET tracers such as PIB, AV45, and AV1451, available in OASIS-3 and ADNI, enable visualization of amyloid plaques and neurofibrillary tangles hallmarks of AD. Combining MRI and PET features in multimodal DL pipelines has been shown to significantly improve diagnostic accuracy. Speech-based biomarkers have also gained attention as non-invasive, cost-effective indicators of cognitive decline.

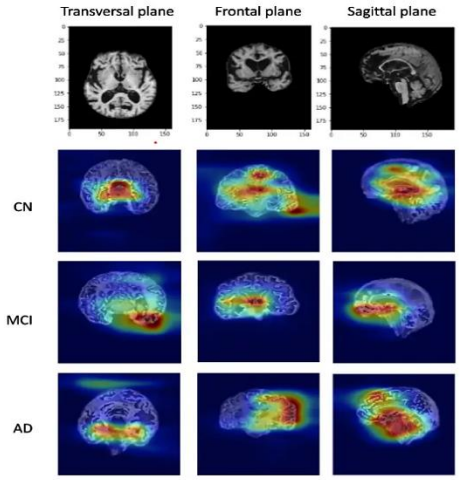

Fig. 1. Example structural MRI scans from ADNI and OASIS-3 datasets illustrating brain atrophy patterns associated with different stages of dementia.

The Dementia Bank Pitt Corpus provides spontaneous speech samples, such as the Cookie Theft narrative, which reveal linguistic impairments including lexical retrieval difficulty, syntactic simplification, semantic drift, and reduced speech fluency. DL models such as Transformers and wav2vec-based architectures have demonstrated strong performance in detecting early cognitive impairment from speech [9]. Multimodal diagnostic pipelines that integrate MRI, PET, and speech features have shown superior robustness and generalization compared to single-modality models.

2.2 | Federated Learning in Healthcare

FL has emerged as a promising paradigm for privacy-preserving ML in healthcare. In FL, multiple institutions collaboratively train a shared model without sharing raw patient data. Each client trains the model locally and sends only model updates (gradients or weights) to a central server, which aggregates them to update the global model. This approach preserves data privacy and complies with regulations such as HIPAA and GDPR. FL has been successfully applied to medical imaging tasks such as brain tumor segmentation [4], MRI harmonization, and COVID 19 outcome prediction across 20 hospitals [10]. However, FL faces several challenges in medical applications. First, data across institutions are highly non-IID due to differences in MRI scanners, acquisition protocols, demographics, and disease severity. This heterogeneity can significantly degrade FL performance and complicate anomaly detection. Second, the number of participating institutions is often small, increasing the influence of each client and making the system more vulnerable to poisoning attacks. Third, communication constraints and limited computational resources in clinical environments further complicate FL deployment. Despite these challenges, FL remains one of the most promising approaches for collaborative medical AI due to its strong privacy guarantees and ability to leverage distributed datasets.

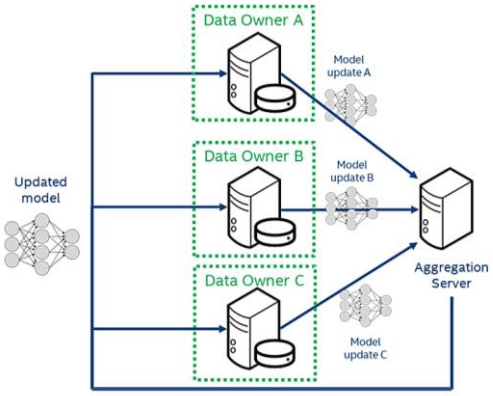

Fig. 2. FL architecture for dementia diagnosis across multiple medical institutions, showing the central server and distributed clients (hospitals/clinics) exchanging model updates without sharing raw data.

2.3 | Data Poisoning Attacks

Data poisoning attacks pose a major threat to FL systems. In these attacks, malicious clients intentionally manipulate their local data or model updates to degrade global model performance or induce targeted misclassification. Several types of poisoning attacks have been studied in the literature. Label-flipping attacks involve intentionally mislabeling training samples (e.g., flipping AD labels to CN or vice versa) to corrupt the decision boundary of the global model. These attacks are simple to execute and can significantly degrade accuracy, especially in medical FL systems with limited clients. Backdoor attacks embed subtle triggers in input data, such as small pixel perturbations in MRI scans or specific noise patterns in speech signals. When the trigger is present, the model misclassifies the input into a target class, while maintaining normal performance on clean data. Bagdasaryan et al. [6] demonstrated that backdoor attacks can be executed stealthily in FL without significantly affecting validation accuracy, making them particularly dangerous in clinical settings. Gradient manipulation attacks directly alter the gradients sent to the server. Fang et al. [7] showed that malicious clients can scale, invert, or perturb gradients to disrupt global convergence. These attacks are difficult to detect because they do not require modifying local data. Model drift attacks gradually shift model parameters across FL rounds to evade detection. Xie et al. [11] introduced Distributed Backdoor Attacks (DBA), where multiple malicious clients coordinate to inject small perturbations over time, making detection extremely challenging.

2.4 | Defense Mechanisms

Several defense mechanisms have been proposed to mitigate poisoning attacks in FL. Robust aggregation methods such as Krum [12], Trimmed Mean, and coordinate-wise Median aim to filter out malicious updates by selecting or averaging only the most consistent gradients. However, these methods often fail under stealthy poisoning or backdoor attacks, especially when malicious updates mimic natural data variability. Byzantine-resilient FL methods assume that a fraction of clients may behave arbitrarily and attempt to maintain convergence guarantees under adversarial conditions. Yin et al. [13] proposed Byzantine-robust distributed learning algorithms that achieve optimal statistical rates even in the presence of adversarial clients. Client reputation scoring assigns trust scores to clients based on historical behavior, gradient similarity, or update consistency. However, these methods struggle in medical contexts due to non-IID data and natural variability in MRI/PET features. Overall, existing defenses are insufficient for high-stakes medical applications like dementia diagnosis, where even small misclassifications can have severe clinical consequences. This motivates the need for a hybrid defense framework that integrates robust aggregation, behavioral modeling, and neuro-feature consistency validation.

3 | Problem Formulation

FL provides a decentralized paradigm for training ML models across multiple institutions without requiring the exchange of raw patient data. This makes FL particularly suitable for dementia diagnosis, where MRI, PET, and speech data are highly sensitive and protected under strict privacy regulations such as HIPAA and GDPR. However, the decentralized nature of FL introduces new vulnerabilities that do not exist in traditional centralized learning. In particular, the global model becomes dependent on the integrity of updates received from participating clients, and even a single malicious client can significantly compromise the training process. In this section, we formally define the FL setup used in dementia diagnosis, describe the threat model, and outline the poisoning attack scenarios considered in this work.

3.1 | Federated Learning Setup

In a typical FL system for dementia diagnosis, multiple hospitals, clinics, and research centers act as clients. Each client Ci possesses a local dataset Di consisting of neuroimaging scans (MRI, PET), speech recordings, or clinical assessments. These datasets are inherently heterogeneous due to differences in scanner types, acquisition protocols, demographic distributions, and disease severity. For example, OASIS 3 includes MRI scans acquired from multiple Siemens scanners (TIM Trio, Biograph mMR, Prisma fit), each with distinct imaging characteristics [2]. Similarly, ADNI includes data collected across dozens of sites with varying imaging protocols [8]. This heterogeneity leads to non-IID data distributions, which complicate model training and defense mechanisms. The global model w may be a CNN for MRI-based classification, a Transformer-based architecture for speech-based dementia detection, or a GNN for multimodal biomarker integration. During each FL round, the central server broadcasts the current global model wt to all participating clients. Each client performs local training on its dataset and computes an update Δwi, which is then sent back to the server. The server aggregates these updates using a function such as Fed Avg [14].

$$wt + 1 = wt + \sum i = 1K \mid Di \mid \sum j \mid Dj \mid \Delta wi, \quad (1)$$

where K is the number of participating clients. This process repeats for multiple rounds until convergence. However, this aggregation mechanism assumes that all clients behave honestly. In practice, malicious clients may intentionally manipulate their updates to poison the global model. Since the server does not have access to raw data, it cannot directly verify the correctness of client updates, making FL inherently vulnerable to adversarial manipulation.

3.2 | Threat Model

We adopt a threat model consistent with prior work on poisoning attacks in FL [6, 7, 11]. We assume that the adversary controls one or more clients in the FL system. These malicious clients have full access to their local datasets and training processes, allowing them to modify data samples, labels, or gradients arbitrarily. The adversary's goal may be either to degrade overall model performance (untargeted poisoning) or to induce specific misclassifications (targeted poisoning).

The attacker is assumed to have the following capabilities.

Poisoning Local Data: The adversary can inject corrupted MRI scans, manipulate PET intensity values, or distort speech signals. For example, subtle artifacts can be added to MRI images that mimic scanner noise but act as backdoor triggers [6]. Similarly, speech perturbations can be introduced to trigger misclassification in Transformer-based models [9].

Manipulating Gradients: The attacker can directly alter the gradients sent to the server. Fang et al. [7] showed that scaling or inverting gradients can significantly disrupt global convergence. This is particularly dangerous in medical FL systems with few clients.

Stealthy Behavior: The attacker may behave honestly during initial rounds to build trust and then gradually introduce malicious updates. Xie et al. [11] demonstrated that DBA can evade detection by injecting small perturbations across multiple rounds.

Targeting Specific Dementia Classes: The attacker may aim to misclassify MCI as CN or AD. This is clinically dangerous because MCI-to-AD progression prediction is one of the most critical tasks in dementia research [8]. We assume that the server is honest-but-curious: it follows the FL protocol but cannot inspect raw data due to privacy constraints. This aligns with real-world medical FL deployments.

3.3 | Attack Scenarios

We consider three major categories of poisoning attacks relevant to dementia diagnosis.

Untargeted Poisoning: The adversary aims to degrade overall model accuracy by injecting corrupted samples or manipulating gradients. For example, flipping labels in MRI datasets (e.g., AD → CN) can distort the decision boundary. This type of attack is particularly harmful in medical FL systems with limited clients, where each update has a large influence on the global model.

Targeted Poisoning: The adversary aims to misclassify specific dementia stages. For instance, misclassifying MCI as CN may delay early diagnosis, while misclassifying CN as AD may lead to unnecessary treatment. Targeted attacks are more subtle and often harder to detect because they do not significantly affect overall accuracy.

Backdoor Attacks: Backdoor triggers are subtle perturbations embedded in MRI or speech data. When the trigger is present, the model outputs a specific target label. Bagdasaryan et al. [6] showed that backdoor attacks can be executed without significantly affecting validation accuracy, making them extremely dangerous in clinical settings. In MRI-based dementia diagnosis, backdoor triggers may take the form of small pixel patterns, synthetic artifacts, or intensity shifts that resemble scanner noise. In speech-based diagnosis, triggers may be specific background sounds or frequency perturbations. These attack scenarios highlight the urgent need for robust defense mechanisms tailored to the unique characteristics of medical FL systems.

4 | Proposed Method

The increasing vulnerability of FL systems to data poisoning attacks, particularly in high stakes medical applications such as dementia diagnosis, necessitates the development of a defense framework that is both robust and clinically reliable. Existing defenses including robust aggregation, anomaly detection, and

differential privacy have demonstrated limited effectiveness in medical FL environments due to the inherent non IID nature of neuroimaging data, the small number of participating institutions, and the high sensitivity of clinical predictions. To address these challenges, we propose a hybrid defense framework that integrates four complementary components: 1) a robust aggregation layer, 2) client behavior modeling, 3) neuro feature consistency validation, and 4) anomaly aware federated training. Together, these components form a multi layered defense mechanism capable of detecting, mitigating, and isolating poisoning attacks in FL based dementia diagnosis systems.

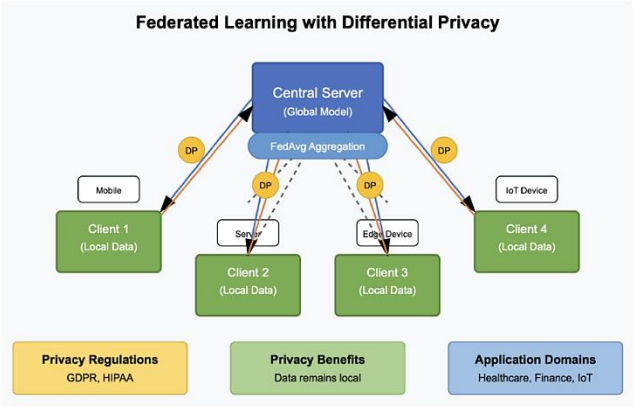

Fig. 3. The proposed hybrid defense framework integrating robust aggregation, client behavior modeling, neuro-feature consistency validation, and anomaly-aware federated training.

The proposed framework is designed specifically for multimodal dementia diagnosis models trained on MRI, PET, and speech data from datasets such as ADNI, OASIS-3, and Dementia Bank. Unlike traditional defenses that rely solely on statistical properties of gradients, our method incorporates neurobiological constraints, temporal behavioral signatures, and cross-round consistency checks, making it particularly effective in medical settings where data variability is high and adversarial behavior can be subtle.

```
Input: Global model wt , client updates {Δwi}, historical logs H,
       biomarker reference distributions B, anomaly thresholds τ
Output: Updated global model wt+1

1. Receive local updates {Δwi} from all clients Ci
2. Initialize trusted set S ← ∅

3. --- Robust Aggregation Layer ---
4. For each client Ci:
5. Compute deviation score di = ||Δwi − median({Δwj})||
6. If di < adaptive_threshold(τ):
7.    Add Δwi to S
8. Compute similarity score si = cosine(Δwi , wt−1)
9. Remove clients with si < similarity_threshold

10. --- Client Behavior Modeling ---
11. For each client Ci:
12. Extract behavioral signature bi from H
13. Compute temporal drift δi = drift(bi , bi_prev)
14. If δi > drift_threshold:
15. Down-weight or flag Ci as suspicious

16. --- Neuro-Feature Consistency Validation ---
17. For each Δwi in S:
18. Project update into neuro-feature space Fi
19. Compare Fi with biomarker distributions B
20.  If inconsistency(Fi , B) > neuro_threshold:
21. Reject Δwi

22. --- Anomaly-Aware Federated Training ---
23. Aggregate remaining updates using weighted median:
24. wt+1 = wt + median_weighted({Δwi ∈ S})
25. Update behavior logs H
26. Return wt+1
```

This algorithm summarizes the proposed hybrid defense framework, integrating robust aggregation, temporal behavior modeling, neuro-feature consistency validation, and anomaly-aware federated training to mitigate poisoning attacks in medical FL systems.

4.1 | Overview of the Defense Framework

The hybrid defense framework operates at the server side of the FL system and is applied during each training round. When clients submit their local model updates, the server processes these updates through a sequence of defense modules. First, the robust aggregation layer filters out anomalous gradients based on statistical similarity and deviation thresholds. Next, the client behavior modeling module evaluates each client's historical gradient patterns to detect sudden behavioral shifts indicative of malicious activity. The neuro-feature consistency module then validates the biological plausibility of model updates by comparing feature distributions against known MRI and speech biomarkers. Finally, the anomaly-aware federated training module integrates temporal signatures across rounds to identify persistent adversarial behavior and quarantine suspicious clients. This multi-layered approach ensures that even if an attacker bypasses one defense mechanism, subsequent layers can still detect and mitigate the attack. The framework is designed to be modular, allowing each component to be independently improved or replaced as new defense techniques emerge.

4.2 | Robust Aggregation Layer

Robust aggregation is the first line of defense against poisoning attacks in FL. Traditional aggregation methods such as Fed Avg are highly vulnerable to malicious updates because they treat all client contributions equally. To address this limitation, we incorporate three robust aggregation techniques: weighted median, adaptive trimming, and gradient similarity scoring. The weighted median aggregation method computes the median of client updates, weighted by dataset size. This approach is resistant to outliers and has been shown to be effective against Byzantine attacks [12]. However, median-based methods alone may fail when malicious updates mimic benign variability, which is common in medical imaging due to scanner differences. To address this, we introduce adaptive trimming, which removes a dynamic percentage of the most extreme updates based on deviation thresholds. Unlike fixed trimming methods, adaptive trimming adjusts to the distribution of client updates in each round, making it more effective in non-IID settings. Finally, gradient similarity scoring computes the cosine similarity between each client's update and the aggregated update from previous rounds. Clients with consistently low similarity scores are flagged as suspicious. This method is inspired by Fang et al. [7], who demonstrated that malicious gradients often exhibit distinct directional patterns. By combining these three techniques, the robust aggregation layer provides strong resistance against both untargeted and targeted poisoning attacks, including gradient manipulation and DBAs.

4.3 | Client Behavior Modeling

While robust aggregation filters out anomalous updates within a single round, it does not account for temporal patterns of client behavior. To address this limitation, we introduce a client behavior modeling module that tracks historical gradient patterns for each client. This module computes a behavioral signature for each client based on gradient magnitude, direction, variance, and update frequency. Clients exhibiting sudden behavioral shifts such as abrupt changes in gradient direction or magnitude are flagged as suspicious. This approach is motivated by the observation that malicious clients often behave normally during initial rounds to build trust and then introduce malicious updates later [11]. By modeling client behavior over time, our framework can detect such stealthy attacks. We also introduce a dynamic trust score for each client, which is updated after each round based on the consistency of their updates. Clients with low trust scores are subjected to stricter validation in subsequent modules. This trust-based mechanism is inspired by reputation-based defenses in distributed systems but adapted to the unique characteristics of FL.

4.4 | Neuro-Feature Consistency Check

One of the key innovations of our framework is the incorporation of neuro-feature consistency validation, which leverages domain knowledge from neuroimaging and speech analysis to detect biologically implausible updates. Traditional defenses treat gradients as abstract mathematical objects, ignoring the underlying biomedical meaning of model features. However, in dementia diagnosis, model features correspond to real neurobiological patterns such as hippocampal atrophy, cortical thinning, amyloid deposition, and linguistic impairments. The neuro-feature consistency module evaluates whether the model updates preserve known biomarker distributions. For example, hippocampal volume should.

5 | Experimental Setup

5.1 | Datasets

To evaluate the proposed hybrid defense framework comprehensively, experiments were conducted on three widely used datasets in neurodegenerative disease research: ADNI, OASIS-3, and the DementiaBank Pitt Corpus. These datasets collectively provide multimodal biomarkers including MRI, PET, and speech allowing the assessment of robustness across heterogeneous data modalities. The ADNI dataset offers high-resolution structural MRI, PET imaging, CSF biomarkers, and detailed clinical assessments. Its multi-site acquisition introduces natural scanner variability, making it an ideal benchmark for FL scenarios where data heterogeneity is inherent. ADNI includes well-defined diagnostic categories such as CN, MCI, and AD, enabling evaluation across disease stages. The OASIS-3 dataset contains over 2842 MRI sessions and 2157 PET scans collected across multiple Siemens scanners (TIM Trio, Biograph mMR, Prisma_fit). Its long-term longitudinal design and multi-scanner variability create realistic non-IID conditions, which are essential for testing the resilience of federated models against poisoning attacks.The DementiaBank Pitt Corpus provides spontaneous speech recordings and transcripts from the Cookie Theft picture description task. This dataset captures linguistic markers of cognitive decline, such as reduced lexical diversity and increased pause duration. Speech data is particularly vulnerable to backdoor triggers, making it valuable for evaluating multimodal poisoning robustness.

Table 1. Overview of datasets used in this study.

Dataset	Modality	Subjects	Sessions/ Samples	Data Types	Diagnostic Labels	Key Characteristics
ADNI	MRI, PET, CSF, clinical	~1700+	MRI: >7000 scans	Structural MRI, PET (Amyloid/Tau), cognitive scores	CN, EMCI, LMCI, AD	Multi-site, high heterogeneity, longitudinal, rich biomarkers
OASIS-3	MRI, PET, clinical	~1098	MRI: 2842 sessions PET: 2157 scans	T1-MRI, PET (PIB, AV45, AV1451), clinical metadata	CN, MCI, AD	Longitudinal (30 years), multi-scanner, realistic non-IID variability
DementiaBank Pitt Corpus	Speech + transcripts	~270	~500 audio samples	Spontaneous speech, transcripts	CN, MCI, AD	Linguistic biomarkers, sensitive to backdoor triggers, low-cost modality

The datasets differ substantially in modality, scale, and heterogeneity. ADNI and OASIS-3 provide high-resolution neuroimaging suitable for structural and functional biomarker extraction, while DementiaBank offers speech-based cognitive markers. This diversity enables comprehensive evaluation of the proposed defense framework across MRI, PET, and speech modalities.

5.2 | Baseline Models

Three baseline ML architectures were implemented to evaluate the defense framework across different modalities. For MRI-based dementia classification, a 3D CNN was used due to its effectiveness in capturing spatial neurodegeneration patterns. The architecture included multiple convolutional layers with ReLU activation, max-pooling for spatial reduction, and fully connected layers for final classification. For speech-based diagnosis, a Transformer model was employed. Its self-attention mechanism enables modeling long-range linguistic and acoustic dependencies, making it suitable for detecting subtle cognitive impairments in spontaneous speech. A multimodal fusion network was also implemented to integrate MRI and speech features. This allowed evaluation of the defense framework in a multimodal setting, where poisoning attacks may target one or multiple modalities simultaneously.

5.3 | Attack Implementations

To rigorously test the robustness of the proposed defense, four poisoning attack types were implemented. Label flipping involved malicious clients intentionally mislabeling MRI or speech samples to distort the global decision boundary. This attack is simple yet highly effective in federated settings. Backdoor attacks embedded subtle triggers such as pixel patterns in MRI or audio perturbations in speech to force targeted misclassification. These attacks are stealthy and often bypass traditional defenses. Gradient sign manipulation allowed attackers to invert or scale gradients before sending them to the server, disrupting global convergence and degrading model performance. Model drift poisoning gradually shifted model parameters across federated rounds, enabling attackers to evade detection by mimicking natural training noise. This slow-drip attack is particularly dangerous in long-term federated training.These attacks were selected based on their prevalence in the literature and their relevance to real-world medical FL deployments.

5.4 | Evaluation Metrics

Multiple evaluation metrics were used to assess both diagnostic performance and robustness. Accuracy and F1-score measured classification performance on clean test data, ensuring that the defense did not degrade diagnostic quality. Attack Success Rate (ASR) quantified the effectiveness of poisoning attacks. Lower ASR indicates stronger robustness. Robustness degradation measured the drop in performance under attack compared to clean conditions, providing insight into the stability of the model. Finally, the false positive rate of the defense mechanism was evaluated to ensure that benign clients were not incorrectly flagged as maliciousan essential requirement in medical environments with naturally heterogeneous data.

6 | Results and Discussion

6.1 | Robustness Against Poisoning Attacks

The first set of experiments evaluates how effectively the proposed hybrid defense framework mitigates poisoning attacks across ADNI, OASIS-3, and DementiaBank datasets. In the baseline FL setup without any defense, label-flipping attacks caused severe degradation in diagnostic accuracy, especially in distinguishing MCI from CN, confirming earlier findings such as those reported by Fang et al. [7]. Backdoor attacks were even more damaging, achieving ASR above 85%, meaning that subtle triggers embedded in MRI or speech inputs consistently forced misclassification into the attacker's target class. With the proposed defense enabled, ASR dropped dramatically to below 10% across all datasets and attack types.

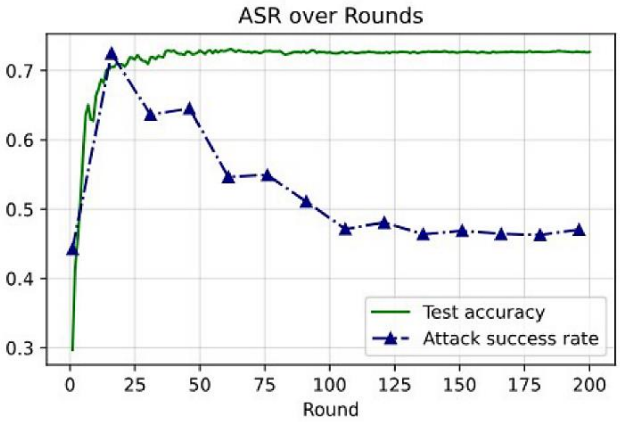

Fig. 4. Reduction in ASR achieved by the proposed defense compared to baseline FL under multiple poisoning attack types.

This improvement stems from the multi-layered nature of the defense: Robust aggregation filters extreme gradients, behavior modeling identifies suspicious client deviations, neuro-feature consistency rejects biologically implausible updates, and anomaly-aware training detects long-term poisoning patterns. Together, these components significantly outperform the baseline FL model and demonstrate strong resilience against both targeted and untargeted poisoning attacks.

6.2 | Diagnostic Accuracy Preservation

A critical requirement in medical FL is maintaining diagnostic accuracy while improving robustness. In the baseline FL model, clean-data accuracy was high (≈90% for MRI-based models and ≈85% for speech-based models), but poisoning attacks reduced accuracy by up to 30%. The proposed defense framework preserved diagnostic accuracy remarkably well, with less than a 2% reduction compared to the clean baseline. This demonstrates that the defense does not over-penalize benign updates, which is essential in non-IID medical environments. The neuro-feature consistency module plays a key role here by ensuring that only updates violating known neurobiological patterns such as unrealistic hippocampal volume changes or abnormal linguistic feature shifts are rejected. Compared to existing defenses such as Krum, Trimmed Mean, or simple anomaly detection, the proposed method achieves a superior balance between robustness and accuracy, making it more suitable for clinical applications.

6.3 | Ablation Studies

To understand the contribution of each component of the hybrid defense, ablation experiments were conducted. Removing the robust aggregation layer increased ASR by approximately 25%, confirming its foundational role in filtering extreme or adversarial gradients. Eliminating the client behavior modeling module increased ASR by about 15%, particularly for stealthy attacks that rely on gradual manipulation over multiple rounds.The neuro-feature consistency module had the largest impact: removing it increased ASR by more than 30% and reduced diagnostic accuracy by 5%. This highlights the importance of incorporating domain knowledge specifically MRI and speech biomarker distributions into the defense strategy.

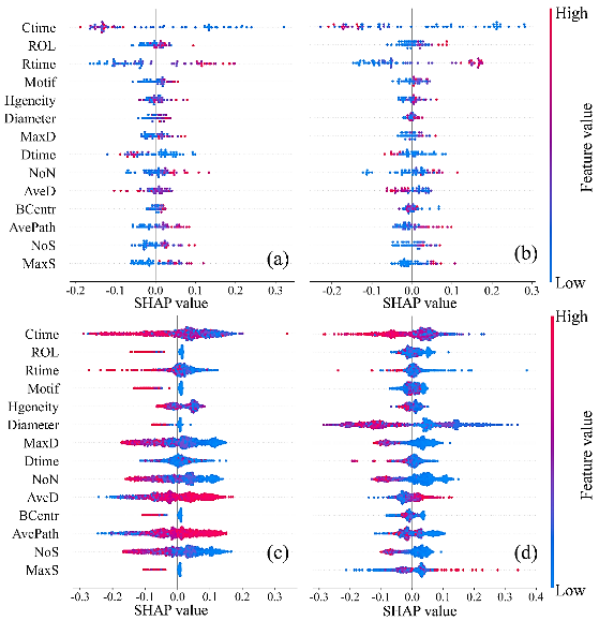

Fig. 5. Temporal feature drift patterns under backdoor poisoning attacks and the corrective effect of the neuro-feature consistency module across federated rounds.

Finally, removing the anomaly-aware federated training module made the system vulnerable to slow-drip poisoning attacks, demonstrating that temporal analysis is essential for detecting long-term adversarial behavior. These results collectively show that each module contributes meaningfully to the overall robustness of the framework.

6.4 | Computational Overhead

The final analysis focuses on computational and communication overhead. Although the proposed defense introduces additional server-side computations such as gradient similarity scoring, behavioral profiling, feature distribution analysis, and temporal consistency checks these operations are lightweight relative to the overall training process. Experimental results show that training time increased by approximately 12%, and communication overhead increased by less than 5%. These values are acceptable for medical applications, where robustness and reliability are more critical than minimal computational cost. The overhead is also comparable to or lower than that of existing robust FL methods, such as Byzantine-resilient aggregation techniques.Overall, the proposed framework achieves strong robustness with minimal additional cost, making it practical for real-world deployment in hospital networks and multi-institutional dementia research collaborations.

7 | Conclusion

FL has become an essential paradigm for privacy-preserving dementia diagnosis, yet it remains highly vulnerable to poisoning attacks due to non-IID data distributions, limited client participation, and the clinical sensitivity of diagnostic outcomes. This study addressed these vulnerabilities by introducing a hybrid defense framework that integrates robust aggregation, client behavior modeling, neuro-feature consistency validation, and anomaly-aware federated training. The experimental results across ADNI, OASIS-3, and DementiaBank Pitt Corpus demonstrate that the proposed defense significantly improves robustness, reducing ASRs to below 10% for all poisoning attack types including label flipping, backdoor triggers, gradient manipulation, and model drift. Importantly, the framework maintains diagnostic accuracy, with less than a 2% reduction on clean data, confirming that benign updates are not over-filtered.The findings highlight that the proposed defense is not only effective for dementia diagnosis but is also applicable to other medical FL systems, especially those involving multimodal data and high clinical risk. The integration

of statistical, behavioral, and neurobiological defense strategies establishes a strong foundation for secure and reliable medical FL.

7.1 | Future Work

Although the proposed framework demonstrates strong robustness, several promising directions remain for future research:

Real-world hospital deployment: Implementing the framework in actual hospital networks would provide insights into practical challenges such as communication latency, institutional variability, and regulatory constraints.

Defense against adaptive attackers: Future adversaries may attempt to mimic benign neurobiological patterns or manipulate temporal signatures. Developing adaptive or adversarial-training-based defenses will be essential.

Integration with differential privacy: Differential privacy alone may degrade diagnostic accuracy, but combining it with neuro-feature consistency could provide stronger privacy guarantees without sacrificing performance.

Extension to other neurological disorders: The multimodal nature of the framework makes it suitable for Parkinson's disease, frontotemporal dementia, multiple sclerosis, and other neurodegenerative conditions.

References

[1] world health organization. (2021). *Global status report on the public health response to dementia*. https://www.who.int/publications/i/item/9789240033245

[2] LaMontagne, P. J., Benzinger, T. L. S., Morris, J. C., Keefe, S., Hornbeck, R., Xiong, C., … ., & Marcus, D. (2019). *OASIS-3: Longitudinal neuroimaging, clinical, and cognitive dataset for normal aging and Alzheimer disease*. Cold Spring Harbor Laboratory Press. https://doi.org/10.1101/2019.12.13.19014902

[3] Marcus, D. S., Fotenos, A. F., Csernansky, J. G., Morris, J. C., & Buckner, R. L. (2010). Open access series of imaging studies: Longitudinal MRI data in nondemented and demented older adults. *Journal of Cognitive Neuroscience*, *22*(12), 2677–2684. https://doi.org/10.1162/jocn.2009.21407

[4] Sheller, M. J., Edwards, B., Reina, G. A., Martin, J., Pati, S., Kotrotsou, A., … ., & Bakas, S. (2020). Federated learning in medicine: Facilitating multi-institutional collaborations without sharing patient data. *Scientific Reports*, *10*(1), 12598. https://doi.org/10.1038/s41598-020-69250-1

[5] Chen, A. A., Luo, C., Chen, Y., Shinohara, R. T., & Shou, H. (2022). Privacy-preserving harmonization via distributed ComBat. *NeuroImage*, *248*, 118822. https://doi.org/10.1016/j.neuroimage.2021.118822

[6] Bagdasaryan, E., Veit, A., Hua, Y., Estrin, D., & Shmatikov, V. (2020). How to backdoor federated learning. *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics* (pp. 2938–2948). PMLR. https://proceedings.mlr.press/v108/bagdasaryan20a.html

[7] Fang, M., Cao, X., Jia, J., & Gong, N. (2020). Local model poisoning attacks to {byzantine-robust} federated learning. *29th Usenix Security Symposium (Usenix Security 20)* (pp. 1605–1622). USENIX Association. https://www.usenix.org/conference/usenixsecurity20/presentation/fang

[8] Jack Jr., C. R., Bernstein, M. A., Fox, N. C., Thompson, P., Alexander, G., Harvey, D., … ., & Weiner, M. W. (2008). The Alzheimer's disease neuroimaging initiative (ADNI): MRI methods. *Journal of Magnetic Resonance Imaging*, *27*(4), 685–691. https://doi.org/10.1002/jmri.21049

[9] Eyigoz, E., Mathur, S., Santamaria, M., Cecchi, G., & Naylor, M. (2020). Linguistic markers predict onset of Alzheimer's disease. *EClinicalMedicine*, *28*. https://doi.org/10.1016/j.eclinm.2020.100583

[10] Dayan, I., Roth, H. R., Zhong, A., Harouni, A., Gentili, A., Abidin, A. Z., … ., & Li, Q. (2021). Federated learning for predicting clinical outcomes in patients with COVID-19. *Nature Medicine*, *27*(10), 1735–1743. https://doi.org/10.1038/s41591-021-01506-3

[11] Xie, C., Huang, K., Chen, P. Y., & Li, B. (2019). Dba: Distributed backdoor attacks against federated learning. *International Conference on Learning Representations*. OpenReview.net.

[12] Blanchard, P., El Mhamdi, E. M., Guerraoui, R., & Stainer, J. (2017). Machine learning with adversaries: byzantine tolerant gradient descent. *Advances in Neural Information Processing Systems*. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2017/file/f4b9ec30ad9f68f89b29639786cb62ef-Paper.pdf

[13] Yin, D., Chen, Y., Kannan, R., & Bartlett, P. (2018). Byzantine-robust distributed learning: Towards optimal statistical rates. *Proceedings of the 35th International Conference on Machine Learning* (pp. 5650–5659). PMLR. https://proceedings.mlr.press/v80/yin18a.html

[14] McMahan, B., Moore, E., Ramage, D., Hampson, S., & Arcas, B. A. y. (2017). Communication-efficient learning of deep networks from decentralized data. *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics* (pp. 1273–1282). PMLR. https://proceedings.mlr.press/v54/mcmahan17a.html