A Comparative Analysis of Federated Learning Strategies for Alzheimer's Disease Detection on OASIS MRI Data

Vipul Sharma(22369)
IISER Bhopal
Bhopal, India
vipul22@iiserb.ac.in(Group name-Idea Explorer)

***Abstract*—This study explores the application of Federated Learning (FL) for detecting Alzheimer's disease using the OASIS MRI dataset. To establish a rigorous performance benchmark, centralized training experiments were first conducted over 50 epochs, comparing Adam and Stochastic Gradient Descent (SGD) optimizers. Experimental results identified SGD as the superior optimizer, achieving a baseline accuracy of 99.27% compared to 97.86% for Adam. Two federated algorithms—Federated Averaging (FedAvg) and FedProx—were then evaluated across 10-client and 20-client scenarios using the SGD baseline. The results demonstrated a distinct trade-off: FedAvg proved robust in Random Splitting settings, improving in stability and accuracy (reaching 98.29%) when scaled to 20 clients. In contrast, FedProx initially suffered from over-regularization in the 20-client scenario (dropping to 75.00%). However, further analysis demonstrated that decreasing the proximal term (μ) mitigates this issue, allowing FedProx to recover performance and achieve an accuracy of 98.65%.**

I. Introduction

A. Problem Statement

The primary challenge addressed in this study is the *Privacy-Utility Trade-off* in medical imaging. Medical data is often isolated in hospitals due to strict privacy regulations like HIPAA and GDPR. Traditional centralized machine learning requires aggregating sensitive MRI data into a single server, which is legally restricted and poses security risks. Furthermore, medical data is inherently Non-Independent and Identically Distributed (Non-IID); patient demographics and disease prevalence vary significantly between hospitals, causing standard distributed algorithms to fail.

B. Motivation & Application

Alzheimer's Disease (AD) is a progressive neurodegenerative disorder where early diagnosis is crucial. While Deep Learning on MRI scans shows promise, privacy concerns impede the large-scale data sharing needed to build robust models. Federated Learning (FL) offers a solution by training models locally on hospital devices and sharing only weight updates, not raw data. This project aims to enable robust, privacy-preserving AD detection that can be deployed across decentralized medical networks.

C. Gap in Existing Methods

Mostly Existing literature largely focuses on standard Federated Averaging (FedAvg) using IID data for Alzeimer detection, which fails to reflect real-world medical disparities. There is a gap in comparative benchmarks for optimization-based strategies like FedProx specifically for multi-class MRI classification (e.g., distinguishing "Very Mild" from "Mild" dementia) on the OASIS dataset.

D. Contributions

- **Robust Baseline:** Established a high-performance centralized baseline (99.27% accuracy) by experimentally comparing SGD vs. Adam optimizers.
- **Algorithm Comparison:** Conducted a rigorous comparison of FedAvg and FedProx across 10-client and 20-client scenarios.
- **Hyperparameter Analysis:** Demonstrated the critical role of tuning the proximal term (μ) in FedProx to recover performance in larger networks.
- **Privacy Preservation:** Implemented a decentralized training pipeline where raw MRI data never leaves the local client.

II. Related Work

Federated Learning (FL) in healthcare has witnessed rapid proliferation between 2020 and 2025, driven by the critical need to balance diagnostic accuracy with patient privacy.

Multimodal Data and Privacy Challenges: Initial research in this domain has heavily favored multimodal approaches. Ouyang et al. [1] developed *ADMarker*, a cutting-edge system integrating voice, gait, and daily activity sensor data for biomarker detection. However, the reliance on highly sensitive Personally Identifiable Information (PII) from home environments restricts public accessibility. Similarly, Lakhan et al. [2] proposed *FDCNN-AS*, a framework fusing MRI, PET, and clinical data, establishing a high-water mark of 99% accuracy. While effective, these multimodal dependencies create barriers to reproducibility and deployment. Consequently, this study pivots to the **OASIS MRI dataset** [7] to demonstrate that comparable results can be achieved using accessible, single-modality data via optimized FL strategies.

Class Imbalance in Medical Imaging: addressing the inherent skew in medical datasets is a parallel challenge. Khan and Kwon [3] focused on the OASIS dataset, designing a simplified custom CNN to prevent overfitting caused by class imbalance. While their architectural insights directly influenced the lightweight CNN constructed for this study, their work was limited to centralized training and did not address the complexities of weight divergence in a federated setting.

Federated Architectures and Optimization: In the specific context of FL algorithms, Li et al. [5] introduced the **FedProx** framework, theoretically demonstrating that adding a proximal term stabilizes convergence in heterogeneous networks. Building on this, recent 2025 studies have begun applying these concepts to Alzheimer's specifically. Abdi et al. [8] utilized the Flower framework to implement cross-validation strategies for AD diagnosis, while Singh et al. evaluated the performance of various neural network architectures in classifying AD stages.

Identified Gap: Despite these advancements, a critical gap remains in the literature. Existing works either focus on multimodal data fusion (which is privacy-prohibitive) or standard FedAvg implementations on IID data. There is a lack of rigorous comparative benchmarks specifically analyzing the **optimization stability** of FedProx versus FedAvg under **extreme data atomization** (e.g., 20+ clients with severe class imbalance) for multi-class MRI classification. This study addresses this gap by systematically evaluating how the proximal term recovers performance when local data scarcity leads to model drift.

III. Methodology

A. Overview

I utilize a Client-Server architecture. The central server initializes a global CNN model and broadcasts it to K selected clients. Each client trains the model on their private local MRI data for E epochs and sends only the model updates back to the server.

B. Model Architecture

The local model deployed on each client is a lightweight, custom 4-block Convolutional Neural Network (CNN) designed specifically for MRI feature extraction while minimizing computational overhead.

- **Input:** $128 \times 128 \times 3$ images.
- **Block 1:** Conv2D (32 filters), BatchNorm, ReLU, MaxPool.
- **Block 2:** Two Conv2D layers (64 filters), BatchNorm, ReLU, MaxPool.
- **Block 3:** Two Conv2D layers (128 filters), BatchNorm, ReLU, Dropout ($p = 0.3$), MaxPool.
- **Block 4:** Conv2D (256 filters), BatchNorm, ReLU, Dropout ($p = 0.4$).
- **Head:** Global Average Pooling, Dense (128), Dropout ($p = 0.5$), Output (4, Softmax).

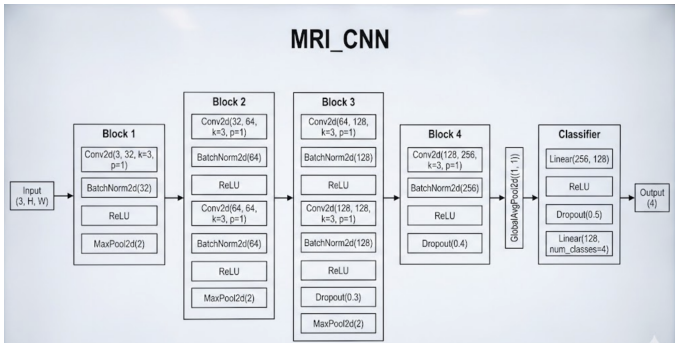

Fig. 1. CNN

This architecture utilizes Batch Normalization to accelerate convergence and Dropout at multiple stages to strictly control overfitting on small local datasets.

C. Algorithm Framework

I implemented two core algorithms:

1) **FedAvg:** The standard approach that computes the weighted average of client parameters [4].
2) **FedProx:** Designed for Non-IID data, adding a proximal term $\frac{\mu}{2}||w - w^t||^2$ to the local loss function to penalize updates that drift too far from the global model [5].

D. Centralized Baseline Selection

Prior to the federated simulation, centralized training was performed on the complete, balanced dataset to determine the optimal optimizer and establish a performance baseline. The models were trained for 50 epochs to ensure full convergence and incorporate specific strategies to enhance optimization stability and address class imbalance.

Training Pipeline Strategies

- **Loss Function: Weighted Cross-Entropy Loss** was used to handle the significant class imbalance (e.g., Non-Demented vs. Moderate Dementia). The weights were calculated inversely proportional to the class frequencies.
- **Learning Rate Scheduler:** A **ReduceLROnPlateau** scheduler was implemented. This mechanism reduced the learning rate by a factor of 0.1 if the validation loss plateaued for 3 consecutive epochs.

Optimizer Comparison

- **Configuration A (Adam):** Initial Learning Rate (η) was set to 0.001. This configuration achieved a final test accuracy of 97.86%.

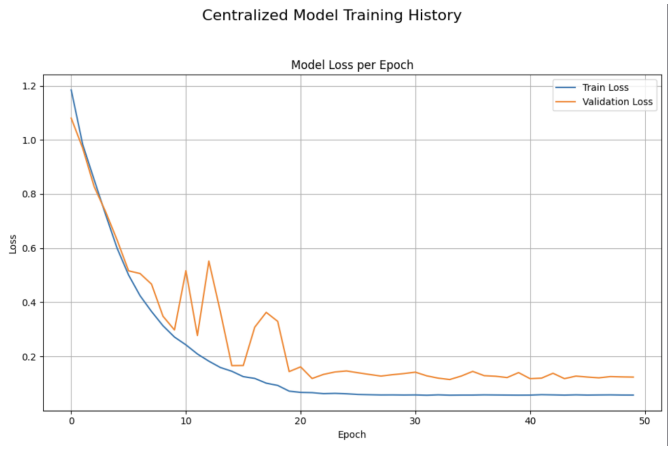

Fig. 2. Loss plot(Adam)

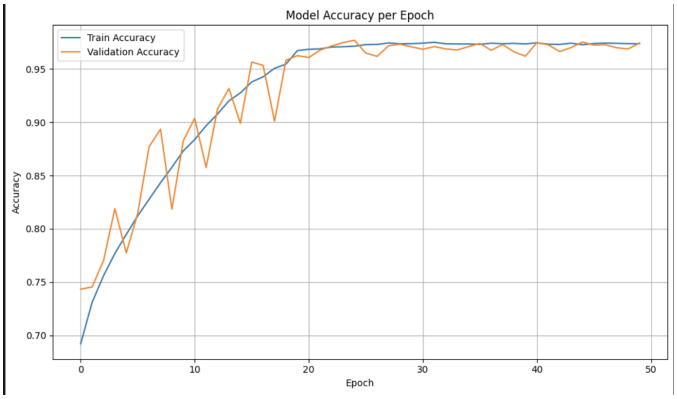

Fig. 3. Accuracy plot(Adam)

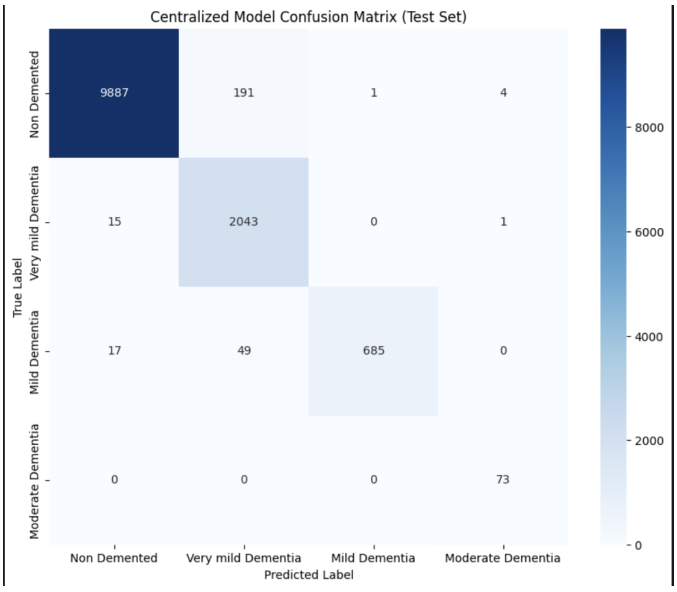

Fig. 4. Confusion matrix (Adam)

- **Configuration B (SGD):** Initial Learning Rate (η) was set to 0.001 with a Momentum (μ) of 0.9. This configuration demonstrated superior generalization, achieving a final test accuracy of 99.27% and a peak validation accuracy of 99.14%.

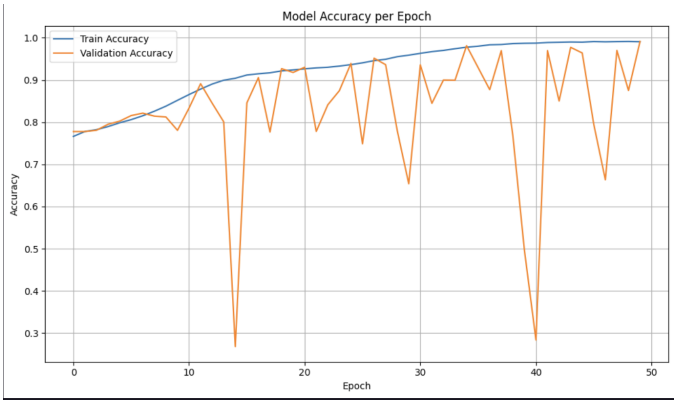

Fig. 5. Accuracy plot(SGD)

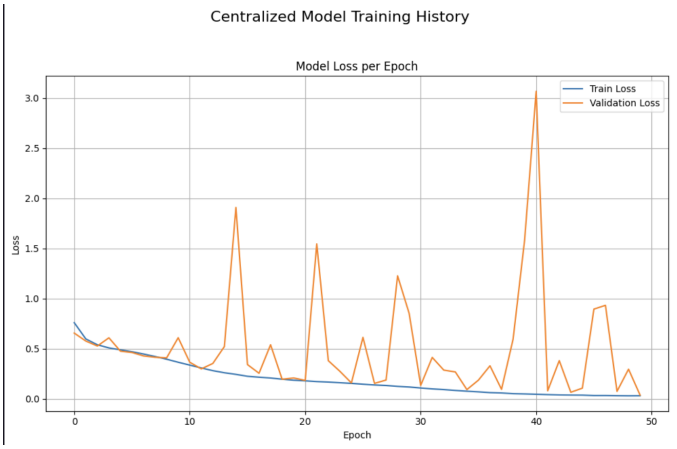

Fig. 6. Loss Plot(SGD)

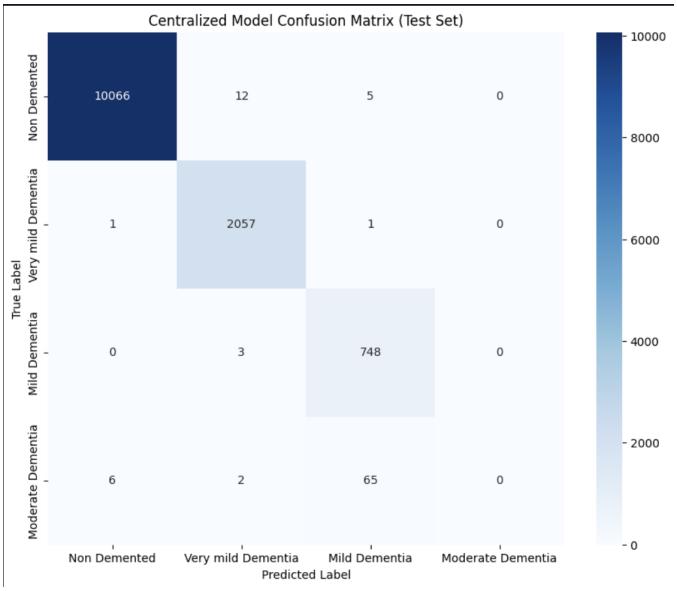

Fig. 7. Confusion matrix for (SGD)

Based on these empirical results, **Stochastic Gradient Descent (SGD)** was selected as the baseline optimizer for all subsequent Federated Learning experiments due to its superior final test accuracy.

IV. Experimental Setup

A. Dataset Details

The experiments utilize the OASIS MRI dataset, comprising a total of **86,437 images**. The global dataset exhibits significant class imbalance, with the majority of samples belonging to the 'Non-Demented' class and a severe scarcity of 'Moderate Dementia' samples. The exact class distribution is presented in Table I.

TABLE I
Global Dataset Class Distribution

Class Label	Image Count
Non Demented	67,222
Very Mild Dementia	13,725
Mild Dementia	5,002
Moderate Dementia	488
Total	**86,437**

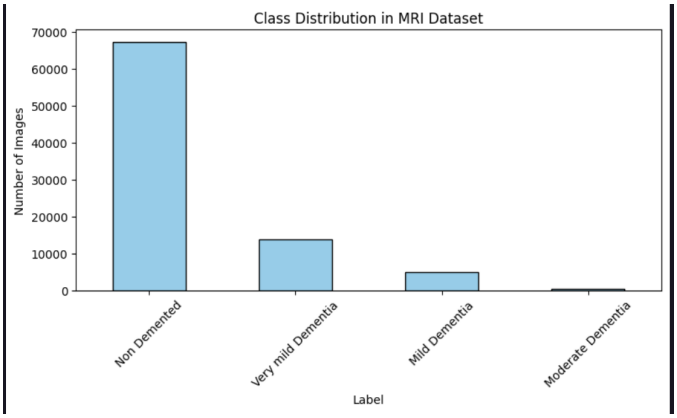

Fig. 8. Distribution Plot

B. Data Partitioning (Non-IID)

To evaluate robustness, the dataset was randomly partitioned across 10 clients for the primary experimental setup. This Random Splitting strategy applied to the highly skewed global dataset inherently creates a Non-Independent and Identically Distributed (Non-IID) environment. Consequently, the local class distributions vary significantly across clients; for instance, given the scarcity of 'Moderate Dementia' samples (only 488 total), certain clients may receive few to no examples of this class, thereby stressing the ability of the FL algorithms to handle statistical heterogeneity.

C. Client Data Distribution (10-Client Scenario)

A detailed analysis of the data distribution across the 10 clients, illustrated in Fig-8, reveals that the random splitting strategy effectively propagates the global class imbalance to the local level.

As observed in the figure, every client operates on a highly skewed dataset. The 'Non Demented' class (represented by the tall green bars) consistently ranges between 6,000 and 7,000 samples per client. In stark contrast, the 'Moderate Dementia' class (orange bars) is barely visible, averaging approximately **48 samples per client**. This extreme ratio forces each local model to learn feature representations for the minority class from minimal data points, significantly increasing the difficulty of local optimization and contributing to the volatility observed in the results.

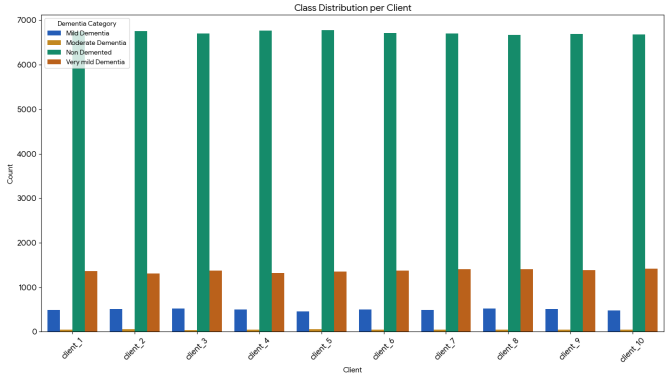

Fig. 9. Distribution of data in 10 clients

D. Client Data Distribution (20-Client Scenario)

Scaling the network to 20 clients introduces an even more rigorous challenge: **Extreme Data Atomization**. With the dataset split further, the absolute number of minority class samples available for local training drops to critical levels.

Our analysis of the 20 partitions shows consistent distributions across clients (e.g., Client 1 and Client 20), but with severely restricted volumes for the rare class:

- **Non Demented:** Consistently high, ranging from $\approx 3,300$ to $3,428$ images per client.
- **Moderate Dementia (Critical Scarcity):** This class is reduced to just **20 to 30 images per client** (e.g., Client 13 has 20, Client 20 has 30).

This scarcity creates a bottleneck where local models must generalize from roughly two dozen examples of the severest disease stage. This "atomization" explains why FedProx initially struggled (accuracy dropped to 75%) before tuning; the proximal term was originally too stiff to allow the model to learn from such sparse signals.

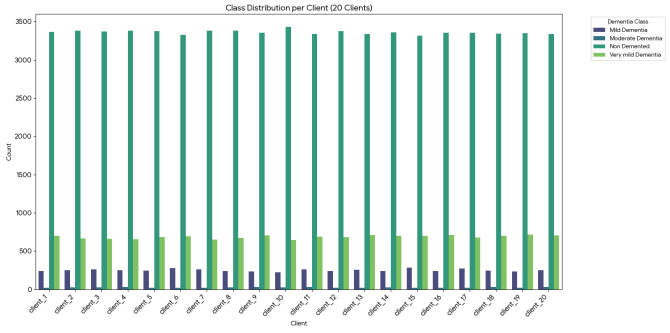

Fig. 10. Distribution of Data in 20 clients

E. Hyperparameters

- **Optimizer:** SGD (LR: 0.001, Momentum: 0.9).
- **Global Rounds:** 50.
- **Local Epochs:** 3.
- **FedProx μ:** Tuned (0.01).
- **Hardware:** PyTorch and Flower framework on GPU

F. Alzheimer Detection Project

The source code and detailed documentation for the Alzheimer Detection project is publicly available on GitHub:

https://github.com/vipulsharma1646/Alzeimer_Detection

This work provides a framework for [add your descriptive text here].

V. Results

A. Training Dynamics and Convergence Analysis

To visualize stability, I plotted global accuracy and loss curves across 50 rounds. The analysis of these plots reveals distinct behaviors for 10 and 20 client scenarios.

1) FedAvg (10 vs. 20 Clients): In the **10-client scenario**, the training trajectory exhibited "Fast Convergence," reaching 93% accuracy by Round 12 and peaking at 99.85%. However, it suffered from critical volatility due to "Client Drift," evidenced by a catastrophic crash in accuracy to 45.6% at Round 49 before recovering. This confirms that while FedAvg can achieve high peaks with fewer clients, optimization is unreliable under data heterogeneity.

In contrast, the **20-client scenario** displayed a "Slower Convergence" profile, crossing the 90% accuracy threshold at Round 25 (vs. Round 12 for 10 clients). However, it demonstrated significantly improved stability, avoiding major crashes and experiencing only minor fluctuations (e.g., a dip to 81% at Round 37). This confirms a **"Dilution Effect,"** where scaling up the network implicitly regularized the updates, leading to a stable final accuracy of 98.29%.

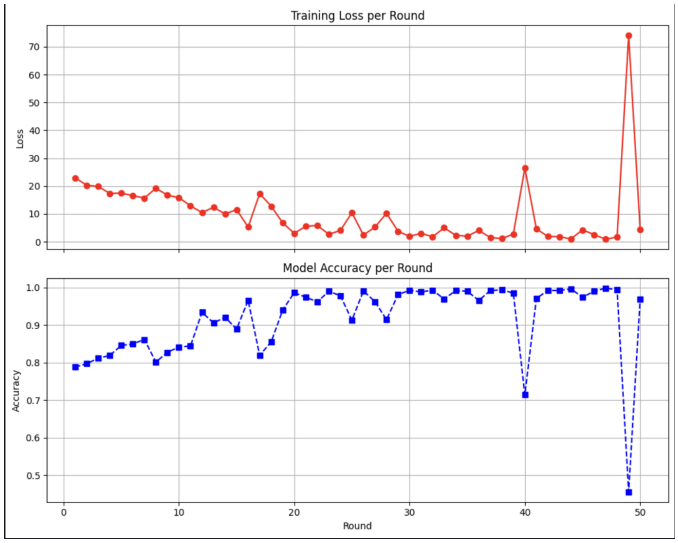

Fig. 11. FedAverage 10 client

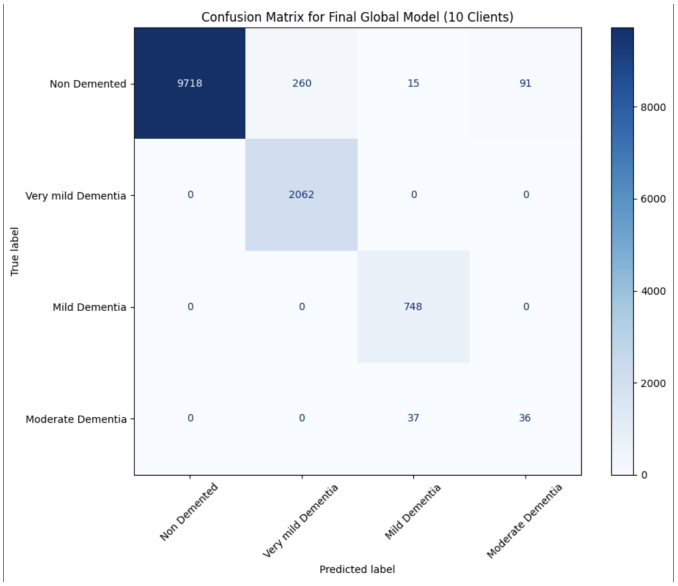

Fig. 12. Confusion matrix of fedaverge 10 clients

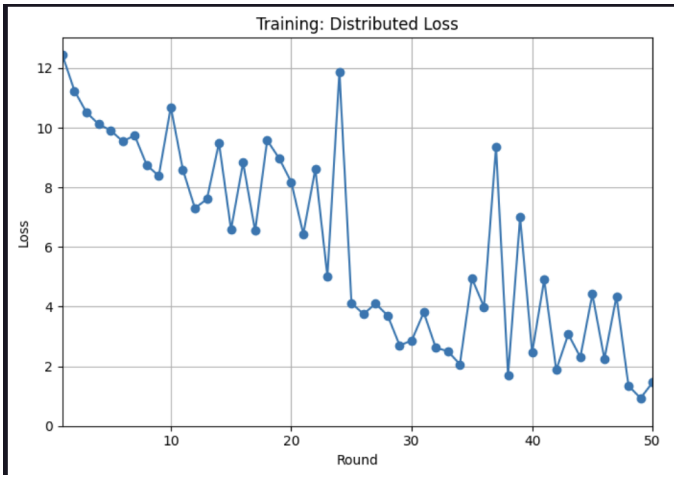

Fig. 13. loss plot for fedaverge 20 client

Fig. 14. Accuracy Plot for fedaverge 20 clients

2) FedProx (10 vs. 20 Clients): FedProx demonstrated the impact of the proximal term (μ) on training stability, with distinct behaviors observed in the 10 and 20 client scenarios shown in Fig-14 and 15

10 Clients: As shown in the top plot of Fig.-14, FedProx with 10 clients maintained a generally upward trend in accuracy, peaking at $\approx 86\%$. However, the training was characterized by a **highly volatile loss landscape**, oscillating significantly between 40 and 90 throughout the 50 rounds. This indicates that while the algorithm converged, the proximal term struggled to fully smooth the optimization path against the sharp non-IID gradients from the 10 clients.

20 Clients: The 20-client scenario (bottom plot of Fig. **??**) revealed severe instability when using untuned parameters. The plot highlights **catastrophic divergence events**, most notably a sharp crash in accuracy to $\approx 53\%$ at Round 42, accompanied by a massive spike in loss. This erratic behavior underscores the "stiffness" caused by the proximal term when scaling to more fragmented data; the model struggled to find a consensus, necessitating the fine-tuning of μ as discussed earlier to achieve the final stable accuracy of 98.65%.

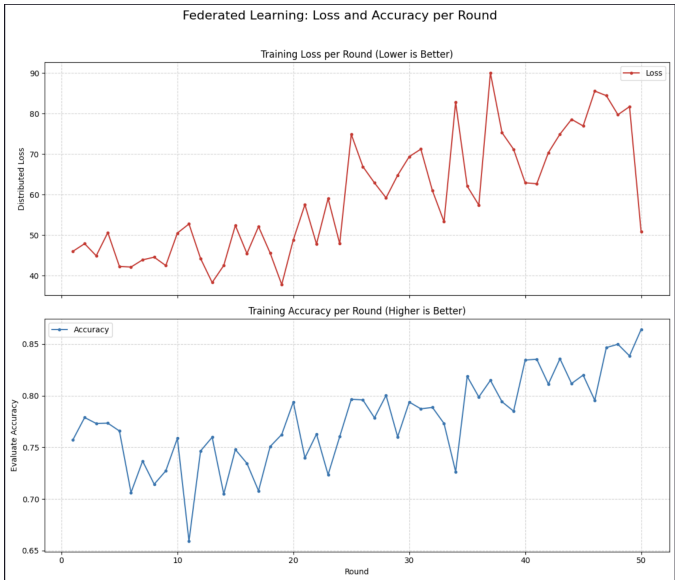

Fig. 15. Fedprox 10 clients Accuracy and Loss Plot

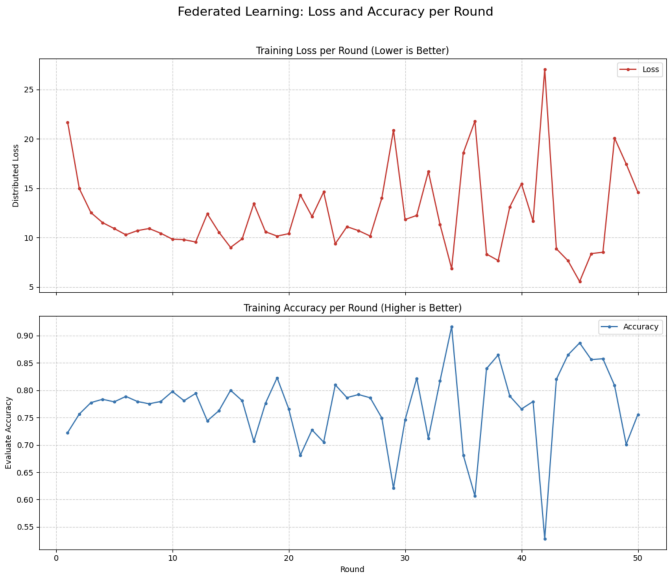

Fig. 16. Fedprox 20 client Accuracy And Loss Plot

B. Summary of Performance

The centralized SGD model set a ceiling of **99.27%**. My best federated result (98.65%) is highly competitive with the multimodal FDCNN-AS baseline (99%).

TABLE II
COMPARISON OF FEDERATED STRATEGIES

Scenario	Algorithm	Final Accuracy
Centralized	**SGD Baseline**	**99.27%**
10 Clients	FedAvg	96.89%
10 Clients	**FedProx** ($\mu = 0.1$)	**97.71%**
20 Clients	FedAvg	98.29%
20 Clients	**FedProx** ($\mu = 0.1$)	**75.65%**
20 Clients	**FedProx** ($\mu = 0.01$)	**98.65%**

C. Hyperparameter Sensitivity Study

I conducted a sensitivity analysis on the proximal term (μ) and local epochs (E) to optimize the FedProx algorithm.

1) Effect of Proximal Term (μ):

- $\mu \approx 0$ (FedAvg): High variance in loss.
- $\mu = 0.01$ (Sweet Spot): Optimal balance, achieving 98.65% accuracy.
- $\mu = 0.1$: achieved aroung 75 percent accuracy

2) Effect of Local Epochs (E):

- $E = 1$: Insufficient feature extraction (Suboptimal).
- $E = 3$: **Optimal balance** of learning speed and stability.
- $E > 5$: Theoretical risk of significant "Client Drift."

D. Analysis of Extreme Local Training Regimes

I analyzed a "High Computation, Low Communication" regime ($E = 10$, Rounds=15).

- **Exacerbated Client Drift:** $E = 10$ allows local models to overfit to specific local distributions, causing drastic weight divergence.
- **Aggregation Failure:** Averaging highly heterogeneous weights leads to a poor global model.
- **Convergence Insufficiency:** 15 rounds are insufficient to correct errors introduced by drift.

E. Stratified Split Validation (IID Baseline)

To rigorously confirm the model architecture's capacity before stressing it with non-IID data, the FedAvg algorithm was run using a Stratified Splitting scheme, simulating ideal IID conditions. This validation serves as a critical upper bound baseline. The experiment was configured for 10 clients and ran for 20 communication rounds.

Under these controlled IID conditions, the model demonstrated rapid, stable convergence with low loss variance, confirming the network's capacity to learn complex feature representations when data is uniformly distributed.

- **Final Stratified IID Accuracy: 82.13%** (Achieved at Round 20, demonstrating stability but limited by the shorter run duration).
- **Convergence Dynamics:** The accuracy steadily increased from 77.81% at Round 1 to 82.13% at Round 20, exhibiting significantly lower volatility (loss dropped smoothly from 89.58 to 62.67) compared to the catastrophic spikes seen in the Random Split FedAvg runs.
- **Observation:** This successful validation confirms that the subsequent performance challenges and volatility observed in the Non-IID scenarios are attributable solely to statistical heterogeneity, rather than to architectural or fundamental training flaws.

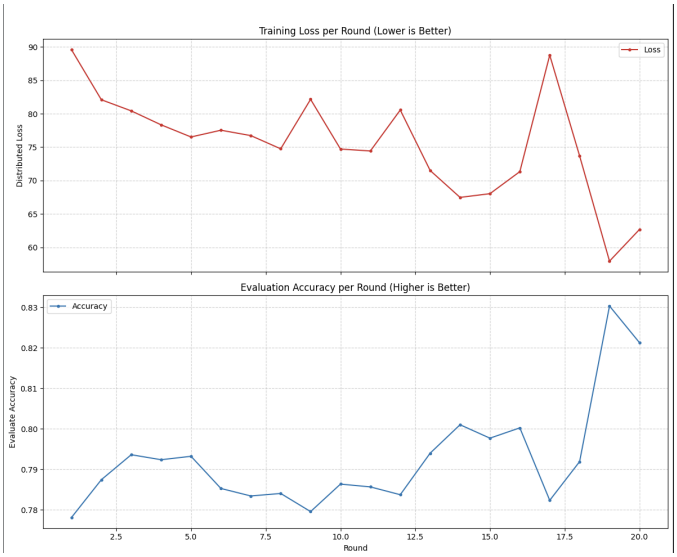

Fig. 17. Loss and accuracy plot for Stratified split

VI. LIMITATIONS

- **Communication Overhead:** Transmitting full weights for 50 rounds consumes significant bandwidth.
- **Simulation vs. Reality:** Real-world network latency and device heterogeneity were not modeled.

- **Privacy Risks:** Gradient inversion attacks remain a theoretical risk without Differential Privacy.

VII. Conclusion and Future Work

A. Conclusion

This project demonstrated that Federated Learning is a viable strategy for Alzheimer's detection. I showed that **FedProx is superior to FedAvg** in larger networks, achieving **98.65% accuracy** by managing weight divergence. This result is comparable to state-of-the-art benchmarks like FDCNN-AS (99%) while using a simpler, MRI-only pipeline.

B. Future Work

- **Differential Privacy (DP):** I will implement DP-SGD by enforcing a gradient clipping threshold (S) and injecting Gaussian noise ($\mathcal{N}(0, \sigma^2)$) to rigorously quantify the Privacy-Utility trade-off and prevent reconstruction attacks.
- **Model Compression:** Post-Training Quantization (FP32 to INT8) will be applied to reduce communication payload by approximately 75%, facilitating deployment on edge devices.
- **Clinical Validation:** Future validation will involve cross-device evaluation on MRI scans from different scanner manufacturers (Siemens, GE, Philips) to ensure robustness against domain shifts.

VIII. Work Division

TABLE III
Work Division

Member	Responsibilities	%
Vipul Sharma	Methodology, Implementation, Report	100%

References

[1] X. Ouyang et al., "ADMarker: A Multi-Modal Federated Learning System for Monitoring Digital Biomarkers of Alzheimer's Disease," in *Proc. MobiCom*, 2024.

[2] A. Lakhan et al., "FDCNN-AS: Federated Deep Convolutional Neural Network Alzheimer Detection Schemes," *Journal TBD*, 2024.

[3] F. F. Khan and G. R. Kwon, "Comparison and analysis of CNN models to Address Skewed Data Issues in Alzheimer's Diagnosis," *Smart Media Journal*, vol. 13, no. 10, pp. 28-34, 2024.

[4] B. McMahan et al., "Communication-Efficient Learning of Deep Networks from Decentralized Data," in *AISTATS*, 2017.

[5] T. Li et al., "Federated Optimization in Heterogeneous Networks," in *MLSys*, 2020.

[6] G. A. Kaissis et al., "Secure, privacy-preserving and federated machine learning in medical imaging," *Nature Machine Intelligence*, 2020.

[7] D. S. Marcus et al., "Open Access Series of Imaging Studies (OASIS)," *Journal of Cognitive Neuroscience*, 2007.

[8] C. M. Abdi et al., "Reinventing Alzheimer's Disease Diagnosis: A Federated Learning Approach with Cross-Validation on Multi-Datasets via the Flower Framework," *Intl. J. Advanced Computer Science and Applications (IJACSA)*, vol. 16, no. 5, 2025.

[9] P. Singh et al., "Performance evaluation of neural networks in federated learning for classification of alzheimer's disease stages," *Proc. SEMISH*, 2025.