AlzFed-XAI: High-Fidelity Interpretable Alzheimer's Diagnosis with Privacy-Preserving Federated Learning

Md. Abdur Rahman[1] **Md. Tofael Ahmed Bhuiyan**[1] **Abdul Kadar Muhammad Masum**[1]
[1] Department of Computer Science and Engineering, Southeast University, Dhaka, Bangladesh
akmmasum@seu.edu.bd

Abstract

Data privacy constraints hinder deep learning in medical imaging by preventing data centralization. We introduce AlzFed-XAI, a federated learning framework for Alzheimer's diagnosis from decentralized MRIs. AlzFed-XAI trains a lightweight CNN (FedNet, 378K parameters) across data silos without exposing raw patient information. On the imbalanced OASIS-1 dataset, our framework achieves 99.73% accuracy and a 0.9970 macro F1-score, demonstrating a negligible performance drop compared to a centralized baseline. To foster clinical trust, Grad-CAM visualizations confirm the model learns neuroanatomically relevant features. Our work presents a robust, privacy-by-design solution, demonstrating a viable pathway for building high-performance, interpretable AI for critical healthcare diagnostics.

1 Introduction

The efficacy of deep learning in diagnosing Alzheimer's Disease (AD) from medical imaging is well-established, with models identifying pathological indicators from MRI scans with remarkable progress [1, 2]. However, model performance is fundamentally dependent on large, diverse datasets, a requirement severely hampered by stringent privacy constraints governing patient health information [3]. Regulations such as HIPAA and GDPR render data centralization for training practically infeasible, creating a critical bottleneck for developing robust clinical AI [4].

We leverage Federated Learning (FL), a decentralized training paradigm that enables multiple parties to build a shared model without exchanging raw data [5]. In this work, we introduce **AlzFed-XAI**, a novel framework for the privacy-preserving diagnosis of AD. AlzFed-XAI orchestrates the training of a custom, lightweight CNN, FedNet, across distributed clients, aggregating only model parameter updates to learn a powerful global model. Furthermore, to address the "black-box" nature of deep learning and foster clinical trust, our framework incorporates Gradient-weighted Class Activation Mapping (Grad-CAM) for model interpretability. We demonstrate that our federated approach achieves performance nearly equivalent to a centralized model, proving that robust diagnostic accuracy need not be sacrificed for patient privacy.

2 Related Works

Federated Learning (FL) has emerged as a critical paradigm for Alzheimer's Disease (AD) diagnostics, enabling multi-institutional collaboration while respecting data privacy. Recent works have focused on enhancing this approach's security and robustness. For instance, frameworks like MetisFL achieve performance comparable to centralized training by leveraging advanced security mechanisms like Fully Homomorphic Encryption (FHE) [6]. Similarly, others have employed Secure Aggregation (SecAgg) to provide strong privacy guarantees against heterogeneous data distributions [7].

Other research aims to address clinical data complexity. Several works propose multi-modal FL systems integrating diverse data types like MRI and blood tests to improve diagnostic accuracy,

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: .

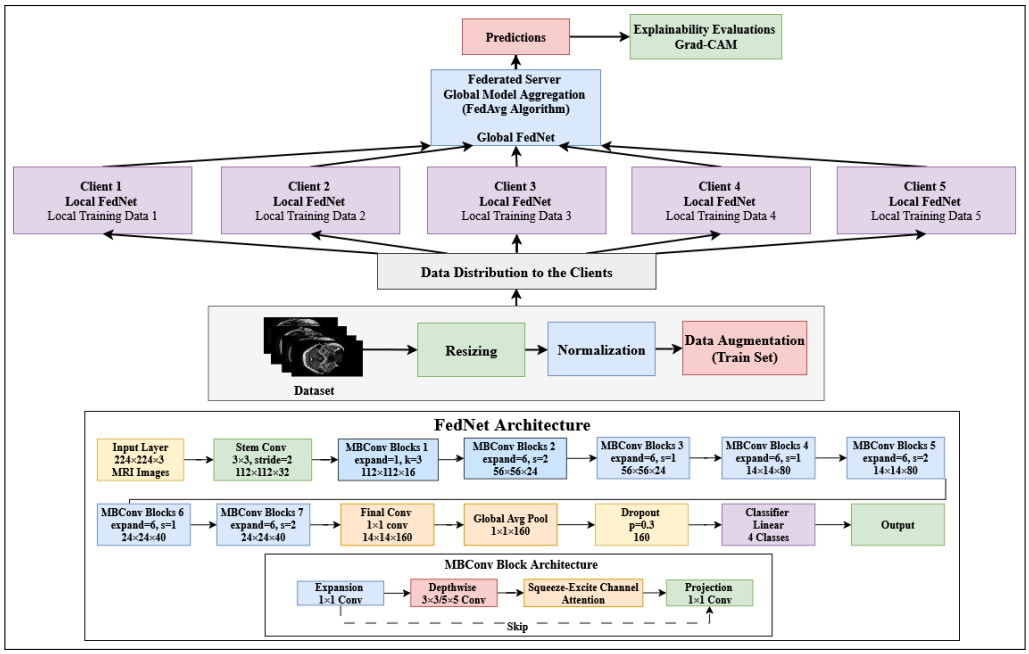

Figure 1: Overview of the AlzFed-XAI methodology.

reporting accuracies up to 99% [8, 9]. Others tackle data imbalance by integrating Generative Adversarial Networks (GANs) within a Split Federated Learning (SFL) architecture [10]. While these approaches advance specific aspects like cryptography or data augmentation, our work distinguishes itself by presenting a holistic framework. **AlzFed-XAI** prioritizes the synergy of three key elements: **(1)** high-fidelity performance with an efficient model, **(2)** inherent privacy via the standard FL protocol, and **(3)** clinical trustworthiness through integrated interpretability.

3 Methodology

We present **AlzFed-XAI**, a federated framework for privacy-preserving AD classification from distributed MRI data (Fig. 1). It employs an efficient client-side model, FedNet, with a decentralized optimization protocol based on Federated Averaging. The global model is learned by aggregating local updates, ensuring raw data never leaves the client environment.

3.1 Dataset and Federated Data Protocol

Our study uses the OASIS-1 MRI dataset [11], which presents a significant class imbalance that complicates classification; the full class distribution is in the Appendix (Figure 3). Let the global dataset be $\mathcal{D}$, with pairs (s, y) of 3D MRI scans and diagnostic labels. We define a transformation T that processes each scan s into a set of 2D axial slices, resized to 224×224 and normalized, yielding our input space $\mathcal{X} \subset \mathbb{R}^{3\times224\times224}$. To simulate a decentralized environment, the global training data is partitioned among $N = 5$ clients into disjoint subsets, $\mathcal{D} = \bigcup_{k=1}^{N} \mathcal{D}_k$, such that $\mathcal{D}_k \cap \mathcal{D}_j = \emptyset$ for $k \neq j$. Each client k has exclusive access to its local partition $\mathcal{D}_k$, forming the basis of our privacy protocol.

3.2 FedNet: Lightweight Client Architecture

For client-side computation, we designed FedNet, a lightweight convolutional neural network. The architecture is built upon the Mobile Inverted Bottleneck Convolution (MBConv) block, a core component of EfficientNet [12], which leverages depthwise separable convolutions and Squeeze-and-Excitation (SE) modules [13] for optimal efficiency. The architecture comprises an initial stem convolution, followed by a sequence of seven MBConv blocks, and a final classification head composed of a 1×1 convolution, global average pooling, dropout, and a linear classifier. This efficient

design results in a compact model with only **378,780** total parameters, making it ideally suited for deployment in resource-constrained federated settings. We represent the model as a parameterized function $f(\cdot;\theta)$, which maps an input $x \in \mathcal{X}$ to a probability distribution over the classes in $\mathcal{Y}$.

3.3 AlzFed-XAI Optimization Protocol

The core of our framework is a federated optimization protocol aimed at minimizing a global objective function $F(\theta)$ without data centralization. The global objective is the weighted average of the local loss functions $L_k(\theta)$ for each client k:

$$\theta^* = \arg\min_{\theta} F(\theta) := \sum_{k=1}^{N} \frac{|\mathcal{D}_k|}{|\mathcal{D}_{\text{train}}|} L_k(\theta) \tag{1}$$

where the local objective for client k is defined as:

$$L_k(\theta) = \frac{1}{|\mathcal{D}_k|} \sum_{(x_i,y_i)\in\mathcal{D}_k} \ell(f(x_i;\theta), y_i) \tag{2}$$

Here, ℓ is the weighted cross-entropy loss function. The training proceeds over a series of communication rounds. In each round t, the following three steps are executed:

1. **Distribution:** The central server broadcasts the current global model parameters θ_g^t to all N clients.
2. **Local Update:** Each client k sets its local model parameters to the global parameters, $\theta_k^t \leftarrow \theta_g^t$. It then performs E local epochs of training using its private data $\mathcal{D}_k$ and the AdamW optimizer [14] to compute its updated parameters, θ_k^{t+1}.
3. **Aggregation:** All clients transmit their updated parameters θ_k^{t+1} to the server. The server then aggregates these to form the new global model by computing their unweighted average:

$$\theta_g^{t+1} \leftarrow \frac{1}{N} \sum_{k=1}^{N} \theta_k^{t+1} \tag{3}$$

This iterative procedure enables collaborative training while strictly preserving data privacy.

4 Experiments

This section outlines the experimental setup, reports quantitative results of **AlzFed-XAI**, and analyzes its interpretability. We benchmark our federated approach against centralized training to evaluate performance.

4.1 Experimental Setup

Experiments were conducted in a Kaggle environment with an NVIDIA Tesla P100 GPU (16 GB VRAM). For AlzFed-XAI, the global model was trained for 30 communication rounds, with 5 clients performing $E = 3$ local epochs per round. The centralized FedNet baseline was trained for 50 epochs. Both paradigms utilized the AdamW optimizer [14] with a learning rate of 1×10^{-3}, a weight decay of 1×10^{-4}, and a weighted cross-entropy loss to address class imbalance. Given the severe class imbalance, we prioritize macro-averaged Precision, Recall, and F1-score. The reported metrics reflect the performance achieved in a single representative training run; future work will incorporate multi-run statistical analysis (mean and standard deviation) to confirm robustness. The full implementation details for our framework are available in our repository.[1]

4.2 Results and Discussion

The quantitative performance of our proposed AlzFed-XAI framework and the centralized FedNet baseline is summarized in Table 1. Our AlzFed-XAI framework achieves an outstanding test accuracy

[1] `https://github.com/borhanitrash/AlzFed-XAI`

Table 1: Performance comparison of FedNet baseline and proposed AlzFed-XAI framework.

Model	Test accuracy (%)	Precision (macro)	Recall (macro)	F1-score (macro)
FedNet	99.9364	0.9980	0.9997	0.9988
AlzFed-XAI	**99.7281**	**0.9959**	**0.9982**	**0.9970**

of 99.7281% and a macro F1-score of 0.9970. This demonstrates the model's exceptional capability in distinguishing between dementia stages within a privacy-preserving environment. The training dynamics (Appendix, Figure 4) illustrate stable global convergence and effective local learning.

To quantify the performance trade-off, we compare AlzFed-XAI to the centralized FedNet model, which achieves a marginally higher accuracy of 99.9364% and F1-score of 0.9988. The performance degradation from federation is minimal ($\approx$0.21% drop in accuracy, $\approx$0.18% in F1-score). This result is highly significant, demonstrating robust, near-centralized performance while providing the critical benefit of data privacy. The confusion matrix and ROC curves (Appendix, Figures 5 and 6) further corroborate the model's discriminative power.

4.3 Model Interpretability

To ensure our model avoids spurious correlations, we employ Gradient-weighted Class Activation Mapping (Grad-CAM) [15] to visualize its decision process. Figure 2 presents a representative visualization for a correctly classified 'Mild Dementia' case. The heatmap highlights activations concentrated within the temporal and parietal lobes, corresponding to regions of visible cortical atrophy, a key neuropathological hallmark of the disease. This consistency between the model's focus and established medical knowledge provides critical clinical plausibility, confirming AlzFed-XAI learns neuroanatomically relevant features, thereby enhancing trust and transparency in its predictions for potential clinical adoption.

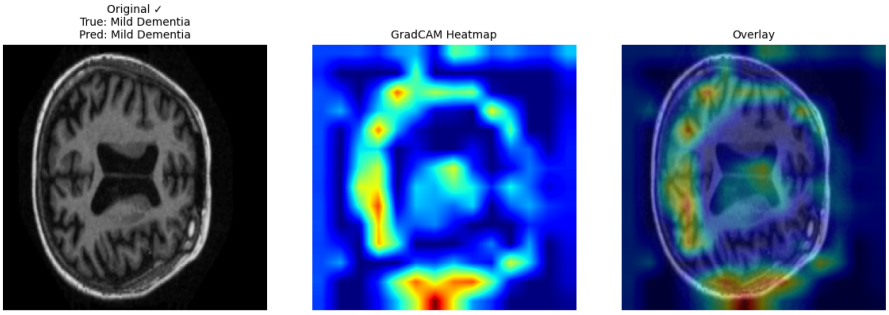

Figure 2: Grad-CAM visualization for a correctly classified 'Mild Dementia' patient.

5 Conclusion

In this work, we introduced AlzFed-XAI, a federated learning framework that provides an accurate and private solution for Alzheimer's disease diagnosis from decentralized MRI data. By leveraging a custom, lightweight CNN (FedNet, 378K parameters), our system achieved exceptional performance, reaching 99.73% accuracy and a 0.9970 macro F1-score with a negligible performance drop compared to a centralized baseline. Furthermore, the integration of Grad-CAM visualizations confirms the model learns neuroanatomically relevant features (such as atrophy in the temporal and parietal lobes), enhancing the transparency and clinical trustworthiness essential for adoption. This work presents a robust, privacy-by-design prototype; however, we acknowledge two primary limitations in the current scope: the reported metrics reflect a single experimental run, necessitating future statistical validation (mean and standard deviation) to confirm robustness; and the evaluation uses a simulated federation on a single, homogeneous dataset. Despite these limitations, our results strongly underscore the foundational potential of federated learning for critical healthcare diagnostics, and we assert that future work must focus on rigorously validating the framework's scalability and stability on genuinely multi-institutional, non-IID data to pave the way for its secure and effective clinical deployment.

References

[1] EL-Geneedy Marwa, Hossam El-Din Moustafa, Fahmi Khalifa, Hatem Khater, and Eman AbdElhalim. An mri-based deep learning approach for accurate detection of alzheimer's disease. *Alexandria Engineering Journal*, 63:211–221, 2023.

[2] Afiya Parveen Begum and Prabha Selvaraj. Multiclass diagnosis of alzheimer's disease analysis using machine learning and deep learning techniques. *International Journal of Image and Graphics*, 24(03):2450031, 2024.

[3] Georgios A Kaissis, Marcus R Makowski, Daniel Rückert, and Rickmer F Braren. Secure, privacy-preserving and federated machine learning in medical imaging. *Nature Machine Intelligence*, 2(6):305–311, 2020.

[4] Nicola Rieke, Jonny Hancox, Wenqi Li, Fausto Milletari, Holger R Roth, Shadi Albarqouni, Spyridon Bakas, Mathieu N Galtier, Bennett A Landman, Klaus Maier-Hein, et al. The future of digital health with federated learning. *NPJ digital medicine*, 3(1):119, 2020.

[5] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR, 2017.

[6] Dimitris Stripelis, Umang Gupta, Hamza Saleem, Nikhil Dhinagar, Tanmay Ghai, Chrysovalantis Anastasiou, Rafael Sánchez, Greg Ver Steeg, Srivatsan Ravi, Muhammad Naveed, et al. A federated learning architecture for secure and private neuroimaging analysis. *Patterns*, 5(8), 2024.

[7] Angela Mitrovska, Pooyan Safari, Kerstin Ritter, Behnam Shariati, and Johannes Karl Fischer. Secure federated learning for alzheimer's disease detection. *Frontiers in aging neuroscience*, 16:1324032, 2024.

[8] Abdullah Lakhan, Mazin Abed Mohammed, Mohd Khanapi Abd Ghani, Karrar Hameed Abdulkareem, Haydar Abdulameer Marhoon, Jan Nedoma, Radek Martinek, and Muhammet Deveci. Fdcnn-as: Federated deep convolutional neural network alzheimer detection schemes for different age groups. *Information Sciences*, 677:120833, 2024.

[9] Nanziba Basnin, Tanjim Mahmud, Raihan Ul Islam, and Karl Andersson. An evolutionary federated learning approach to diagnose alzheimer's disease under uncertainty. *Diagnostics*, 15 (1):80, 2025.

[10] G Narayanee Nimeshika and D Subitha. Enhancing alzheimer's disease classification through split federated learning and gans for imbalanced datasets. *PeerJ Computer Science*, 10:e2459, 2024.

[11] Daniel S Marcus, Tracy H Wang, Jamie Parker, John G Csernansky, John C Morris, and Randy L Buckner. Open access series of imaging studies (oasis): cross-sectional mri data in young, middle aged, nondemented, and demented older adults. *Journal of cognitive neuroscience*, 19 (9):1498–1507, 2007.

[12] Mingxing Tan and Quoc V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. *CoRR*, abs/1905.11946, 2019. URL `http://arxiv.org/abs/1905.11946`.

[13] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.

[14] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[15] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

A Supplemental Figures and Details

This appendix provides additional visualizations and details to support the findings presented in the main paper. This includes the dataset class distribution and a full set of performance graphs for the AlzFed-XAI.

A.1 Dataset Distribution

Figure 3 details the class distribution of the OASIS-1 dataset used in our experiments, highlighting the significant imbalance that poses a challenge for model training and evaluation.

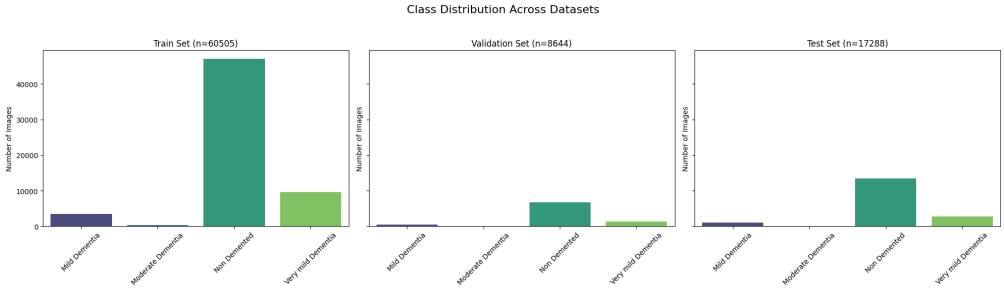

Figure 3: Class distribution of the OASIS-1 dataset. The 'Non Demented' class constitutes the vast majority of samples, creating a significant class imbalance challenge.

A.2 Federated Learning Model Performance (AlzFed-XAI)

This section provides detailed performance visualizations for our proposed AlzFed-XAI framework, as referenced in the main text.

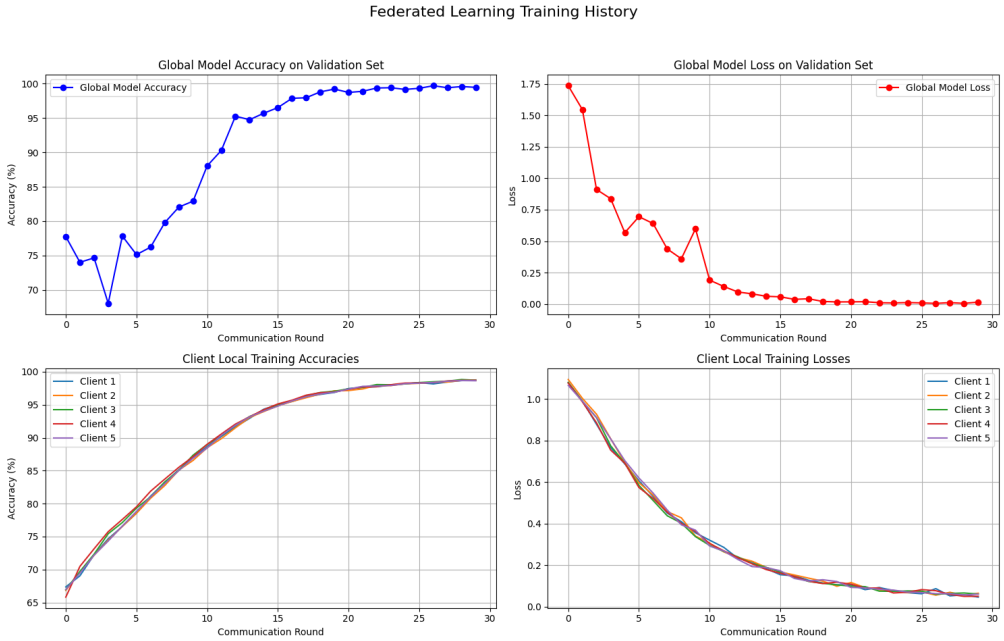

Figure 4: Training dynamics of the AlzFed-XAI framework over 30 communication rounds. Top: The global model shows stable convergence on the validation set. Bottom: Client-side models demonstrate consistent and effective local learning.

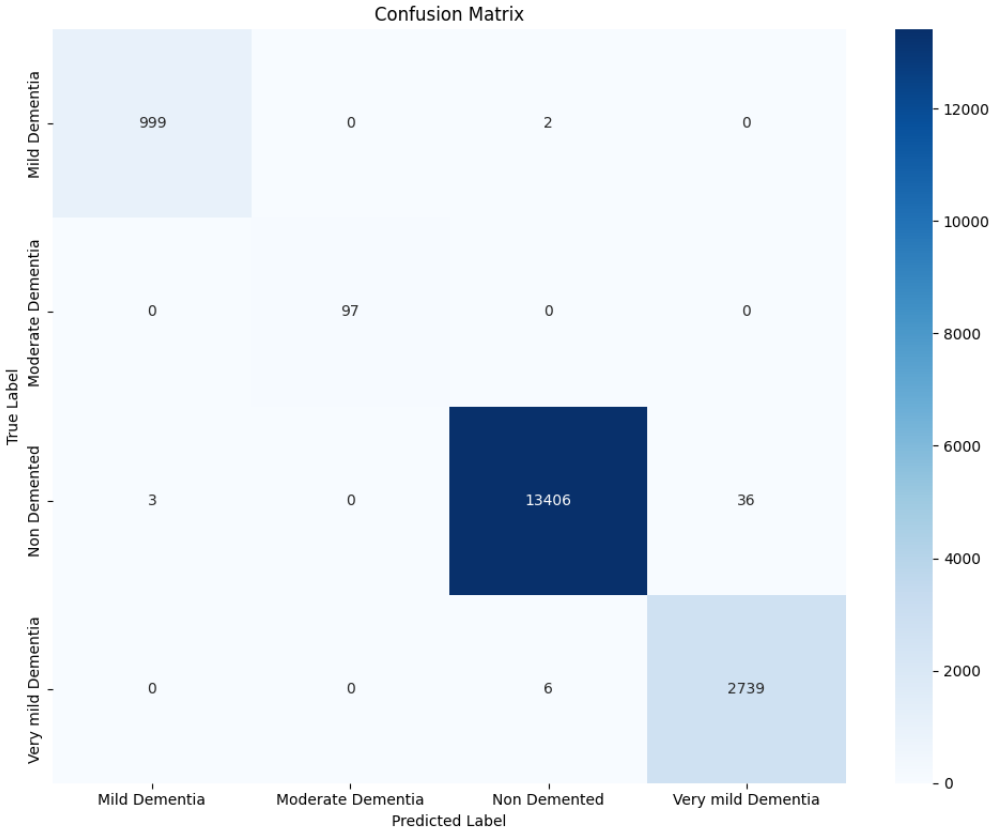

Figure 5: Confusion matrix for the AlzFed-XAI model on the test set. The model shows high accuracy across all classes, including the underrepresented 'Moderate Dementia' class.

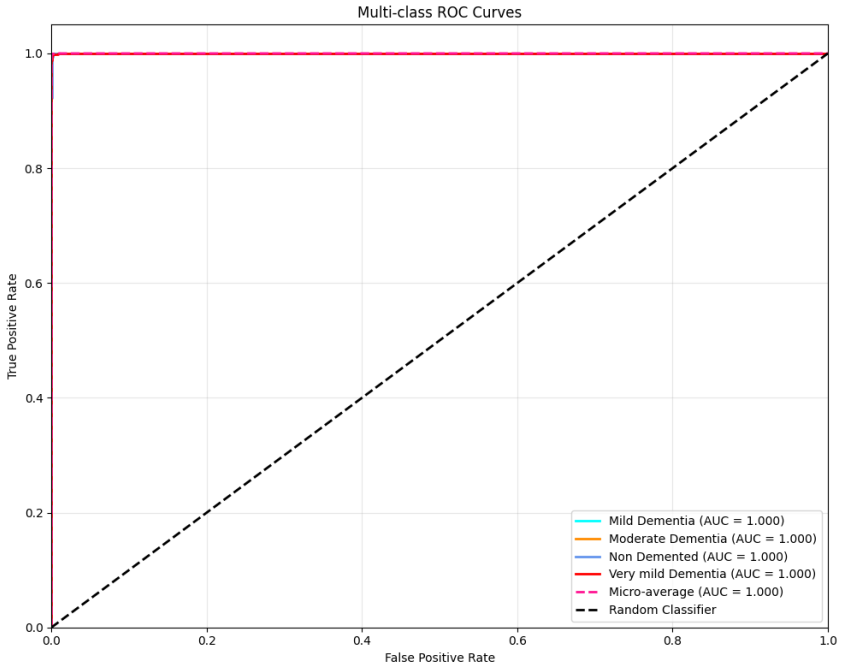

Figure 6: Multi-class ROC curves for the AlzFed-XAI model. The perfect AUC score of 1.000 for all classes indicates excellent discriminative capability.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction claim a high-performance, interpretable, and privacy-preserving federated framework. These claims are directly supported by the experimental results in Section 4, including Table 1 and Figure 2, which validate the performance and interpretability.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The Conclusion (Section 5) explicitly discusses the primary limitations of our work. We acknowledge that our evaluation is based on a simulated federation from a single dataset and suggest that future work should validate the framework's scalability and robustness on genuinely multi-institutional, non-IID data.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: This is an empirical paper focused on the application and evaluation of a federated learning framework. It does not introduce new theoretical results or formal proofs.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper details the model architecture (Section 3.2), dataset (Section 3.1), and key training hyperparameters for both federated and centralized setups (Section 4.1), which are sufficient to reproduce the main experimental results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The OASIS-1 dataset is publicly available. An anonymized version of our code is provided for review, and the final code will be released in a public repository upon publication.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Section 4.1 details the experimental environment (GPU, RAM), training hyperparameters (learning rate, optimizer, epochs, communication rounds), and evaluation metrics, providing a clear basis for understanding the results.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: The paper reports performance metrics from a single experimental run. Error bars or measures of statistical significance (e.g., mean and standard deviation over multiple runs) are not included.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Section 4.1 explicitly states the computational resources used for the experiments, including the GPU type (NVIDIA Tesla P100) and VRAM.

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research uses a publicly available, de-identified medical dataset and proposes a methodology (federated learning) designed to enhance data privacy, aligning with the principles of the NeurIPS Code of Ethics.

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The paper extensively discusses the positive societal impact of enabling privacy-preserving medical diagnostics. A discussion of potential negative societal impacts, such as model bias or security vulnerabilities, is not included.

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [Yes]

    Justification: The core methodology of federated learning is itself a safeguard, designed to train models on sensitive medical data without requiring the data to be shared or released.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The paper properly cites the original publication for the OASIS-1 dataset. The specific data license is not explicitly mentioned, but the asset is correctly attributed to its creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The paper introduces a new model architecture, FedNet, and a new framework, AlzFed-XAI. Both are documented with sufficient architectural and procedural detail in Section 3.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: This research does not involve crowdsourcing or new research with human subjects; it utilizes a pre-existing, publicly available, and de-identified dataset.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: As the study uses a pre-existing and de-identified public dataset, no new Institutional Review Board (IRB) approval was required for this work.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research?

Answer: [NA]

Justification: An LLM was used for assistance in writing, editing, and formatting the manuscript. As per the guidelines, since the LLM did not contribute to the core methodology, experimental design, or results analysis, a formal declaration is not required.