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

Anonymous Author(s)

Affiliation

Address

email

Abstract

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

11 1 Introduction

12 The efficacy of deep learning in diagnosing Alzheimer’s Disease (AD) from medical imaging is
13 well-established, with models identifying pathological indicators from MRI scans with remarkable
14 progress [1, 2]. However, model performance is fundamentally dependent on large, diverse datasets, a
15 requirement severely hampered by stringent privacy constraints governing patient health information
16 [3]. Regulations such as HIPAA and GDPR render data centralization for training practically
17 infeasible, creating a critical bottleneck for developing robust clinical AI [4].

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

27 2 Related Works

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

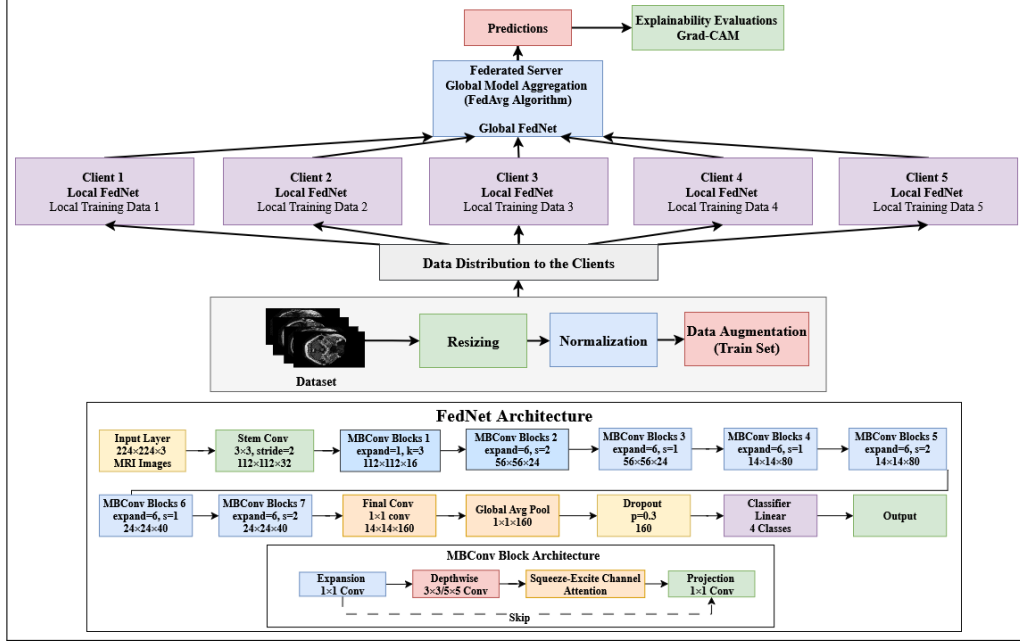


Figure 1: Overview of the AlzFed-XAI methodology.

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, 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 $\mathcal{X} \subset \mathbb{R}^{3 \times 224 \times 224}$. 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 :

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

where the local objective for client k is defined as:

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

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:

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

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 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://anonymous.4open.science/r/AlzFed-XAI-MuslimInML/>

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 visual evidence provides clinical plausibility, suggesting AlzFed-XAI learns neuroanatomically relevant features, thereby enhancing trust and transparency in its predictions.

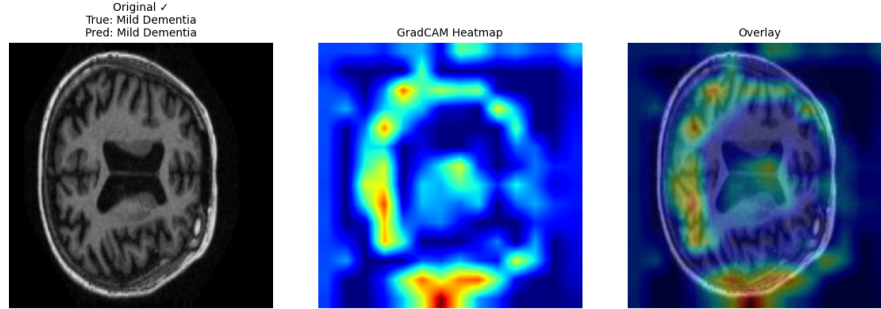


Figure 2: Grad-CAM visualization for a correctly classified ‘Mild Dementia’ patient.

5 Conclusion

In this work, we introduced AlzFed-XAI, a federated learning framework for the accurate and private diagnosis of Alzheimer’s disease. By leveraging a lightweight client-side CNN, our framework achieves exceptional performance, reaching 99.73% accuracy with a negligible drop compared to a centralized baseline, while Grad-CAM visualizations enhance its clinical trustworthiness by confirming it learns neuroanatomically relevant features. We acknowledge our evaluation is currently limited to a simulated federation on a single dataset. Despite this, our results strongly underscore the potential of federated learning to build robust and interpretable AI for critical healthcare challenges. Future work should therefore focus on validating the framework’s scalability and robustness on genuinely multi-institutional, non-IID data to pave the way for wider clinical adoption.

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.

170 A Supplemental Figures and Details

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

174 A.1 Dataset Distribution

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

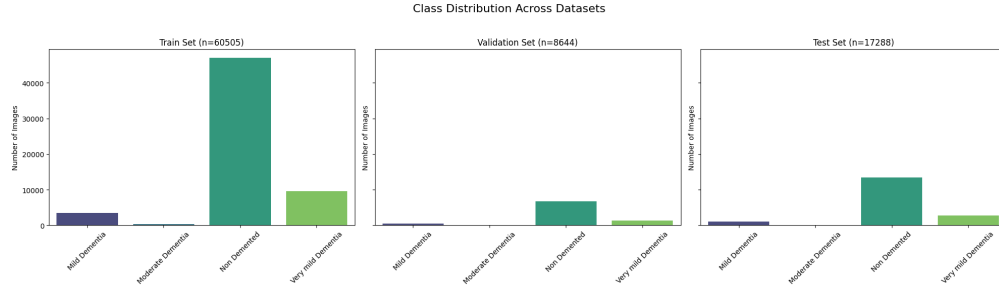


Figure 3: Class distribution of the OASIS-1 dataset. The 'Non Demented' class constitutes the vast majority of samples, creating a significant class imbalance challenge.

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

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

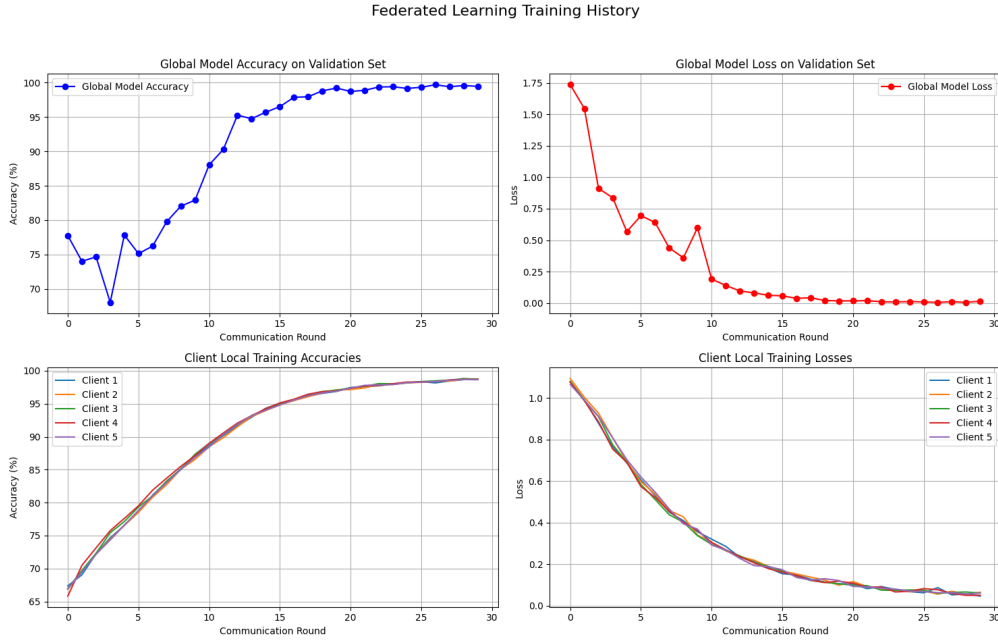


Figure 4: Training dynamics of the AlzFed-XAI framework over 30 communication rounds. Top: The global model shows stable convergence on the validation set. Bottom: Client-side models demonstrate consistent and effective local learning.

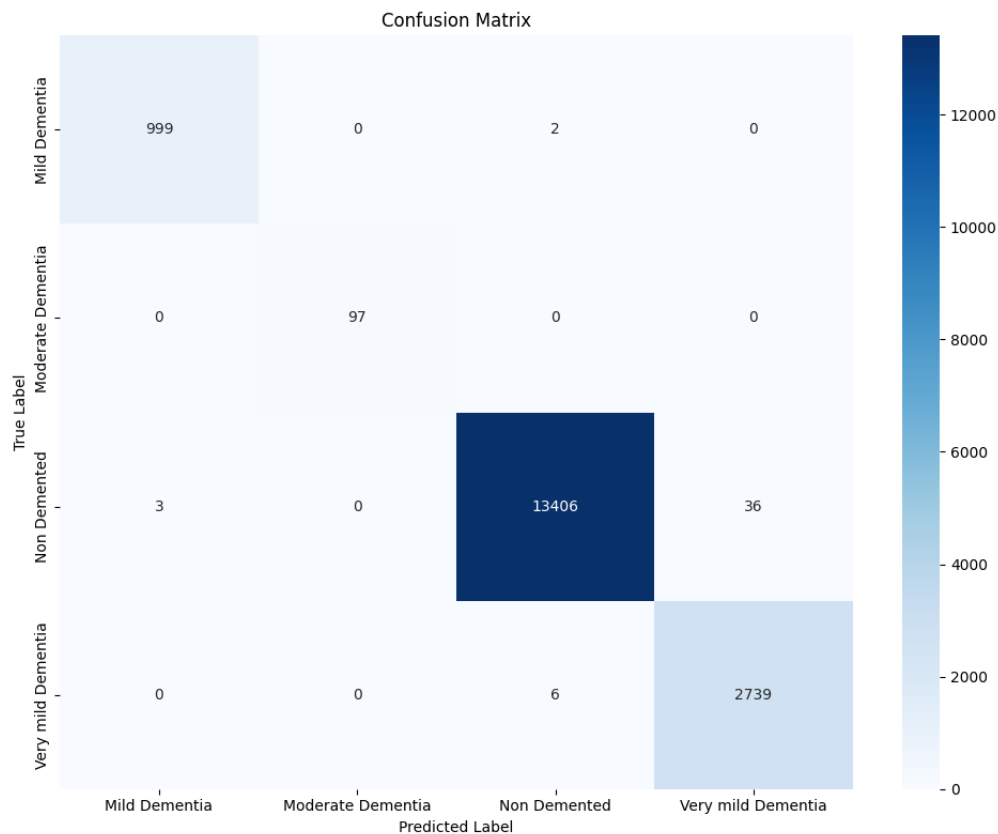


Figure 5: Confusion matrix for the AlzFed-XAI model on the test set. The model shows high accuracy across all classes, including the underrepresented 'Moderate Dementia' class.

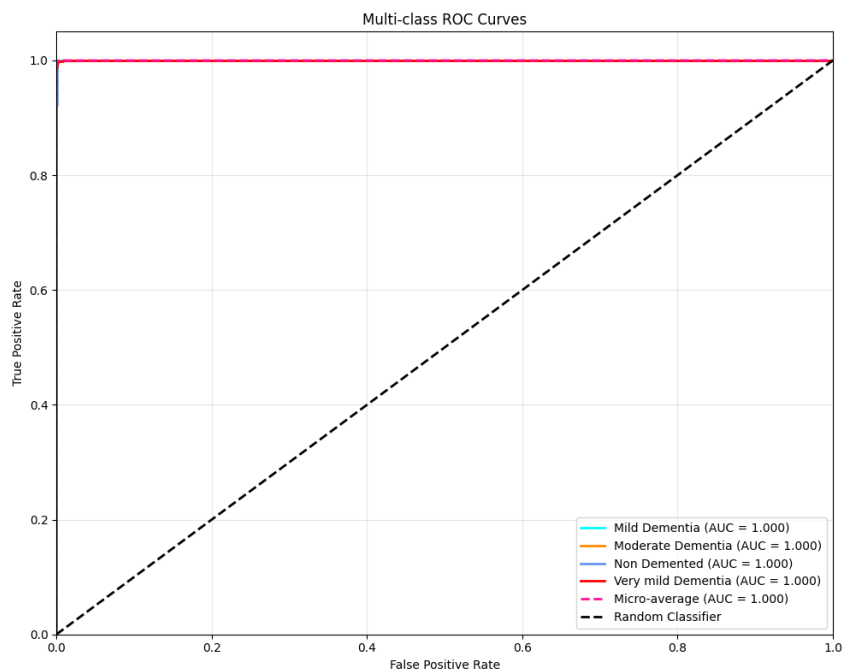


Figure 6: Multi-class ROC curves for the AlzFed-XAI model. The perfect AUC score of 1.000 for all classes indicates excellent discriminative capability.

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?

229 Answer: [No]

230 Justification: The paper reports performance metrics from a single experimental run. Error

231 bars or measures of statistical significance (e.g., mean and standard deviation over multiple

232 runs) are not included.

233 **8. Experiments compute resources**

234 Question: For each experiment, does the paper provide sufficient information on the com-

235 puter resources (type of compute workers, memory, time of execution) needed to reproduce

236 the experiments?

237 Answer: [Yes]

238 Justification: Section 4.1 explicitly states the computational resources used for the experi-

239 ments, including the GPU type (NVIDIA Tesla P100) and VRAM.

240 **9. Code of ethics**

241 Question: Does the research conducted in the paper conform, in every respect, with the

242 NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

243 Answer: [Yes]

244 Justification: The research uses a publicly available, de-identified medical dataset and

245 proposes a methodology (federated learning) designed to enhance data privacy, aligning

246 with the principles of the NeurIPS Code of Ethics.

247 **10. Broader impacts**

248 Question: Does the paper discuss both potential positive societal impacts and negative

249 societal impacts of the work performed?

250 Answer: [Yes]

251 Justification: The paper extensively discusses the positive societal impact of enabling

252 privacy-preserving medical diagnostics. A discussion of potential negative societal impacts,

253 such as model bias or security vulnerabilities, is not included.

254 **11. Safeguards**

255 Question: Does the paper describe safeguards that have been put in place for responsible

256 release of data or models that have a high risk for misuse (e.g., pretrained language models,

257 image generators, or scraped datasets)?

258 Answer: [Yes]

259 Justification: The core methodology of federated learning is itself a safeguard, designed to

260 train models on sensitive medical data without requiring the data to be shared or released.

261 **12. Licenses for existing assets**

262 Question: Are the creators or original owners of assets (e.g., code, data, models), used in

263 the paper, properly credited and are the license and terms of use explicitly mentioned and

264 properly respected?

265 Answer: [Yes]

266 Justification: The paper properly cites the original publication for the OASIS-1 dataset. The

267 specific data license is not explicitly mentioned, but the asset is correctly attributed to its

268 creators.

269 **13. New assets**

270 Question: Are new assets introduced in the paper well documented and is the documentation

271 provided alongside the assets?

272 Answer: [Yes]

273 Justification: The paper introduces a new model architecture, FedNet, and a new framework,

274 AlzFed-XAI. Both are documented with sufficient architectural and procedural detail in

275 Section 3.

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

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

280 Answer: [NA]

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

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

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

289 Answer: [NA]

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

292 **16. Declaration of LLM usage**

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

295 Answer: [NA]

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