

Federated Explainable AI-Based Alzheimer's Disease Prediction With Multimodal Data

Saniya Aslam Patel, Prof. H. B. Jadhav, Prof. D. K. Gund, Prof. P. T. Dhotre

Dept. of Computer Engineering

Adsul's Technical Campus, Ahilyanagar, India

Abstract: *Alzheimer's disease (AD) is a progressive neurodegenerative disorder whose early diagnosis benefits from combining several kinds of patient data-brain imaging, cognitive tests, clinical records, and genetic markers. However, these data are sensitive and distributed across hospitals, which limits the central pooling that conventional deep learning requires, and clinicians are reluctant to trust black-box predictions. This paper presents a conceptual framework that unifies three ideas: multimodal fusion to improve accuracy, federated learning (FL) to train a shared model across institutions without sharing raw data, and explainable AI (XAI) to make predictions interpretable. We describe a layered architecture, outline a training and explanation workflow, and discuss an illustrative evaluation plan on public datasets such as ADNI and OASIS. We emphasise that the reported analysis is a design-level study rather than a completed experiment, and we identify open challenges in data heterogeneity, communication cost, privacy-utility trade-offs, explanation faithfulness, and clinical validation.*

Keywords: Alzheimer's disease, federated learning, explainable AI, multimodal data, deep learning, privacy preservation, healthcare.

I. INTRODUCTION

Alzheimer's disease is the most common cause of dementia and progresses gradually, often beginning with mild cognitive impairment (MCI) before advancing to severe decline. Timely prediction can support earlier intervention and care planning, but no single data source captures the full clinical picture. Structural magnetic resonance imaging (MRI) shows atrophy, positron emission tomography (PET) reflects metabolic change, cognitive tests measure functional decline, and clinical and genetic records add context. Combining these modalities-multimodal learning-generally improves prediction over any single source [1], [2].

Two obstacles complicate the practical use of such models. First, patient data are sensitive and legally protected, so hospitals usually cannot pool their records into one central dataset. Federated learning (FL) addresses this by training a shared model across institutions while raw data remain local; only model updates are exchanged and aggregated [4], [5]. Second, medical staff need to understand why a model produced a prediction. Explainable AI (XAI) methods such as SHAP [7], LIME [8], and gradient-based saliency [9] expose the features and regions that influenced a decision, supporting trust and review.

This paper proposes a conceptual framework that unifies multimodal fusion, federated training, and explanation for AD prediction. Our contributions are: (i) a layered architecture that keeps raw data on each client while producing a shared, interpretable model; (ii) a training and explanation workflow that integrates FL aggregation with XAI; and (iii) a discussion of an illustrative evaluation plan and the open challenges that must be resolved before clinical deployment. This is a design-level study; we do not claim completed experiments or measured accuracy.



II. RELATED WORK

Explainable multimodal AD detection has been studied by El-Sappagh et al. [1], who proposed a multilayer model that fuses several data types and attaches interpretation so that predictions can be examined rather than treated as a black box. Jahan et al. [2] presented an explainable AI framework for AD prediction and management using multimodal data, showing that transparent reasoning can support both diagnosis and ongoing care.

Federated learning was introduced by McMahan et al. [4] through the Federated Averaging (FedAvg) algorithm, which trains deep networks from decentralized data by averaging locally computed updates. Rieke et al. [5] surveyed the role of FL in digital health and highlighted its potential for privacy-preserving medical modelling, while Yang et al. [6] formalised federated machine learning concepts, including horizontal and vertical settings. A vertical federated approach to AD detection on multimodal data [3] is directly relevant, since different hospitals often hold different modalities for overlapping patients.

On the explanation side, SHAP [7] provides a unified, game-theoretic measure of feature contribution, LIME [8] fits local surrogate models to explain individual predictions, and Grad-CAM [9] produces visual saliency maps for convolutional networks. Public datasets such as ADNI [10] and OASIS [11] provide the imaging and clinical data commonly used to study AD. To our knowledge, few student-level frameworks bring federation, multimodality, and explanation together in one coherent design, which is the gap this paper addresses.

III. BACKGROUND AND PRELIMINARIES

A. Multimodal Data for AD

Each modality captures a complementary signal: MRI and PET reveal structural and metabolic change, cognitive scores quantify functional decline, and clinical and genetic records (for example the APOE marker) add risk context. Fusion can occur early (concatenating features), late (combining per-modality decisions), or at an intermediate level, and the right choice depends on data availability and alignment [1], [2].

B. Federated Learning

In FL, a central server coordinates several clients. Each client trains on its local data and sends model updates to the server, which aggregates them-typically by weighted averaging as in FedAvg [4]-into a global model that is returned to the clients. The cycle repeats for several rounds. In vertical FL, clients hold different features for the same patients, which suits multimodal medical data [3], [6].

C. Explainable AI

Feature-attribution methods estimate how much each input contributed to a prediction. SHAP [7] assigns each feature a contribution based on cooperative game theory, LIME [8] trains an interpretable surrogate around a single prediction, and Grad-CAM [9] highlights influential image regions. In AD prediction, such explanations help a clinician judge whether a model relies on medically plausible evidence.

IV. PROPOSED FRAMEWORK

A. Layered Architecture

The proposed system has five layers. The data layer holds each institution's multimodal records locally. The client layer preprocesses data and trains per-modality encoders and a fusion head. The aggregation layer runs on a coordinating server and combines client updates using FedAvg with optional secure aggregation. The explanation layer applies SHAP, LIME, or Grad-CAM to the global model's predictions. Finally, the application layer presents risk scores, disease stage, and the supporting explanation to clinicians.

B. Training and Explanation Workflow

Training proceeds in rounds. (1) The server initialises a global model and sends it to all participating hospitals. (2) Each hospital trains locally on its multimodal data for a few epochs. (3) Clients send only model updates-never raw data-to the server. (4) The server aggregates the updates into an improved global model. (5) Steps 2-4 repeat until convergence.



After training, the explanation layer generates feature attributions and saliency maps for each prediction so that outputs are interpretable and auditable.

C. Privacy and Security Measures

Because exchanged updates can still leak information, the framework allows secure aggregation and differential privacy

[12] to be layered on top of FedAvg. These measures reduce leakage risk at some cost to accuracy, so the trade-off must be tuned per deployment. Keeping raw data on site also helps institutions comply with privacy regulation.

V. ILLUSTRATIVE EVALUATION PLAN

This section describes how the framework could be evaluated; it does not report completed experiments or measured results. A realistic study would use public multimodal datasets such as ADNI [10] and OASIS [11], partitioned across simulated hospital clients to emulate a federated, non-identically distributed setting.

A. Datasets and Setup

ADNI provides MRI, PET, cognitive, and biomarker data, while OASIS provides cross-sectional and longitudinal MRI. Data would be split so that each client holds a subset of patients (horizontal FL) or a subset of modalities (vertical FL), and a held-out set would be reserved for testing. Three classes are typical: cognitively normal, MCI, and AD.

B. Baselines and Metrics

Reasonable baselines include a single-modality model, a centrally trained multimodal model (upper reference), and single-site models (lower reference). Suitable metrics are accuracy, sensitivity, specificity, F1-score, and area under the ROC curve, reported with confidence intervals. Communication cost per round and the effect of privacy noise on accuracy should also be reported.

C. Expected Behaviour

Based on the reviewed literature [1]-[6], multimodal fusion is expected to outperform single-modality models, and federated training is expected to approach-though not necessarily match-centrally trained performance while preserving privacy. Non-identical data across clients and added privacy noise are expected to reduce accuracy, which motivates the open problems discussed next. These are expectations drawn from prior work, not measured outcomes of this study.

VI. DISCUSSION AND OPEN CHALLENGES

Several challenges remain. Data heterogeneity-different scanners, protocols, and populations-can degrade a federated model and complicate fusion. Communication cost grows with the number of rounds and clients, motivating update compression and selective participation. Privacy techniques such as differential privacy [12] protect data but reduce accuracy, so the balance must be chosen carefully. Explanations may be unstable or may not faithfully reflect the model's reasoning, so their quality must itself be evaluated. Finally, fairness across demographic groups and prospective clinical validation are essential before any such system is used in care.

VII. CONCLUSION

We presented a conceptual framework that unifies multimodal fusion, federated learning, and explainable AI for Alzheimer's disease prediction. The design keeps sensitive data on each institution, leverages complementary modalities to improve prediction, and produces interpretable outputs for clinicians. We outlined a layered architecture, a training-and-explanation workflow, and an illustrative evaluation plan on public datasets, while being explicit that no experiments were conducted in this study. Future work includes implementing the framework, handling missing modalities, strengthening privacy guarantees, and validating faithfulness and fairness under realistic clinical conditions.



REFERENCES

- [1] S. El-Sappagh, J. M. Alonso, S. M. R. Islam, A. M. Sultan, and K. S. Kwak, "A multilayer multimodal detection and prediction model based on explainable artificial intelligence for Alzheimer's disease," *Sci. Rep.*, vol. 11, art. 2660, 2021.
- [2] N. Jahan et al., "Explainable AI-based Alzheimer's prediction and management using multimodal data," *PLOS ONE*, vol. 18, no. 11, e0294253, 2023.
- [3] "Vertical Federated Alzheimer's detection on multimodal data," *arXiv:2312.10237*, 2023.
- [4] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Proc. AISTATS, PMLR*, vol. 54, 2017, pp. 1273-1282.
- [5] N. Rieke et al., "The future of digital health with federated learning," *npj Digit. Med.*, vol. 3, art. 119, 2020.
- [6] Q. Yang, Y. Liu, T. Chen, and Y. Tong, "Federated machine learning: concept and applications," *ACM Trans. Intell. Syst. Technol.*, vol. 10, no. 2, art. 12, 2019.
- [7] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proc. NeurIPS*, 2017, pp. 4765-4774.
- [8] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should I trust you?': explaining the predictions of any classifier," in *Proc. ACM SIGKDD*, 2016, pp. 1135-1144.
- [9] R. R. Selvaraju et al., "Grad-CAM: visual explanations from deep networks via gradient-based localization," in *Proc. IEEE ICCV*, 2017, pp. 618-626.
- [10] C. R. Jack Jr. et al., "The Alzheimer's Disease Neuroimaging Initiative (ADNI): MRI methods," *J. Magn. Reson. Imaging*, vol. 27, no. 4, pp. 685-691, 2008.
- [11] D. S. Marcus et al., "Open Access Series of Imaging Studies (OASIS): cross-sectional MRI data in young, middle aged, nondemented, and demented older adults," *J. Cogn. Neurosci.*, vol. 19, no. 9, pp. 1498-1507, 2007.
- [12] C. Dwork and A. Roth, "The algorithmic foundations of differential privacy," *Found. Trends Theor. Comput. Sci.*, vol. 9, no. 3-4, pp. 211-407, 2014.

