# An Explainable Hybrid Multimodal Model for Alzheimer's Disease Detection

**Md Siyamul Islam**
Information and Computer Science Department
King Fahd University of Petroleum and Minerals, Dhahran 31261, Saudi Arabia
g202313550@kfupm.edu.sa

**Noushath Shaffi**
Department of Computer Science
Sultan Qaboos University, P.O. Box 50, P.C. 123, Al-khod, Sultanate of Oman
n.shaffi@squ.edu.om

**Md Mahfuzur Rahman**
Information and Computer Science Department
King Fahd University of Petroleum and Minerals, Dhahran 31261, Saudi Arabia
mdmahfuzur.rahman@kfupm.edu.sa

**Mufti Mahmud**
Information and Computer Science Department
SDAIA-KFUPM Joint Research Center for AI
Interdisciplinary Research Center for Bio Systems and Machines
King Fahd University of Petroleum and Minerals, Dhahran 31261, Saudi Arabia
mufti.mahmud@kfupm.edu.sa, muftimahmud@gmail.com

## Abstract

Alzheimer's disease (AD) is a progressive neurodegenerative disorder and a major global health concern. Early and accurate prediction of AD stages, particularly during the Early and Late Mild Cognitive Impairment (EMCI, LMCI), is crucial for timely intervention. While deep learning (DL) models have shown promise, most prior work relies on single-data modality, leading to limited diagnostic accuracy. This work presents a novel multimodal DL model that integrates neuroimage and tabular clinical data to improve AD detection. Trained and tested on the OASIS dataset, the proposed model combines the extracted embeddings from the image data through a dense network with selected clinical features, identified via SHAP-based feature attribution and cumulative contribution thresholding. This integration enables a four-way classification across Normal Cognition (NC), EMCI, LMCI, and AD that surpasses the state-of-the-art performance with a precision of 96.02%, a recall of 95.84%, and an F1 score of 95.92%, alongside an overall accuracy of 95.84%.

## 1 Introduction

Alzheimer's disease (AD) is an irreversible neurological disorder that leads to cognitive decline. Over 55 million people worldwide live with AD, and the symptoms include memory loss, behavioural instability, vision issues, and reduced mobility. AD is the seventh leading cause of

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

death globally (3), and its diagnosis relies on chronic symptom observation, neuroimaging and clinical tests (16). The AD continuum includes Mild Cognitive Impairment (MCI), with fewer symptoms (13) and is categorized into Early (EMCI) and Late (LMCI) for mild and severe deficits, respectively (7), and their early detection is crucial for intervention.

Artificial Intelligence (AI) methods show promise in AD prediction. Early studies relied on single-modality data (i.e., Magnetic Resonance Imaging (MRI), Positron Emission Tomography (PET), Computed Tomography (CT), or clinical records) and traditional Machine Learning (ML) models. Due to their limited performance, the paradigm shifted to more advanced techniques such as deep learning (DL), ensemble learning and vision transformers (13). However, in reality, physicians use multimodal data (imaging, clinical tests, demographics) for any clinical diagnosis, highlighting the need for multimodal decision-support systems (5). Implementing this is challenging due to missing data, heterogeneity, and class imbalance. Notably, few studies integrate MRI with cross-sectional clinical data, despite its diagnostic value.

To address this issue, this work proposes a framework that emulates the clinical AD diagnostic workflow, integrating neuroimaging, cognitive assessments, and patient history - through a consensus SHAP and CCT approach by adopting the OASIS (9) dataset. with two main contributions: (1) a novel multimodal framework for AD prediction, and (2) an automated clinical feature selection approach. The rest of the paper, Section 2 reviews related works, while Sections 3 and 4.1 detail the methodology and the results, and Section 5 concludes.

## 2 Related Works

AD prediction traditionally relies on cognitive assessments, biomarkers, and imaging modalities like CT, MRI, and PET (14). Tools such as MoCA (10) and MMSE (2) are widely used, while biomarker testing faces challenges of sensitivity, specificity, and invasiveness (4).

Classical ML methods (SVM, RF, LR, KNN) have shown strong performance using MRI-derived features. Acharya et al. (1) achieved 94.54% accuracy with KNN and Shearlet Transform. Shaffi et al. (14) reported that RF and SVM can rival or outperform DL models on MRI data, questioning DL superiority. More recent approaches leverage Vision Transformers (VTs) with ensemble strategies for MRI feature extraction (13), though single-modality limits diagnostic accuracy.

With advances in deep learning, multimodal models have emerged. Liu et al. combined CNN-based imaging with non-imaging data; Qiu et al. fused MRI likelihood maps with clinical features (11); Liu et al. proposed cross-attention architectures (8); and Rahim et al. used 3D CNN with RNNs (12). Venugopalan et al. integrated imaging, clinical, and genetic data, outperforming unimodal models but still constrained by missing data (15). Notably, no study has explicitly combined MRI with clinical history, a gap this work addresses through feature selection and multimodal integration.

## 3 Methodology

Figure 2 in the appendix represents the proposed framework of the multimodal model. This framework utilizes the OASIS dataset which provides nine clinical features (age, education, SES, MMSE, CDR, eTIV, nWBV, ASF, delay). The importance of these features was estimated by training six ML models (GB, XGBoost, CatBoost, LightGBM, SVM, LR) and finding the SHAP values subsequently (appendix figure 3). In this study, an ensemble-based consensus strategy was designed by averaging SHAP values for each feature across the applied model to mitigate the bias and instability of feature importance estimates from any single algorithm. By integrating perspectives from both linear (e.g., Logistic Regression, SVM) and nonlinear (e.g., XGBoost, CatBoost) learners, this method favors features that are consistently important across heterogeneous models, thereby improving robustness. Additionally, we conducted a statistical analysis (details in the section 4.1) to determine the threshold for key feature selection, and only these key features were fed into the multimodal model. Let $\phi_m(f)$ be the mean absolute SHAP value for feature $f$ under model $m$ from a set of heterogeneous learners $\mathcal{M}$. We define the consensus attribution as $\bar{\phi}(f) = \frac{1}{|\mathcal{M}|}\sum_{m\in\mathcal{M}}\phi_m(f)$. Sort features by $\bar{\phi}$ and pick the smallest prefix $S$ such that $\frac{\sum_{f\in S}\bar{\phi}(f)}{\sum_f \bar{\phi}(f)} \geq \tau$

with $\tau = 0.95$. This model-agnostic, distribution-aware selector reduces single-model bias and avoids normality assumptions that make $Z$-scores brittle under right skew.

In parallel, we utilized preprocessed 2D MRI slices (reorientation, skull stripping, enhancement) (14), which were collected from raw 3D MRI scanning, and then fed into an image network. EfficientNet was utilised as a feature extractor from 2D images (6). Its original classification layer was replaced with a projection head that outputs a 128-dimensional embedding, obtained through a linear layer followed by ReLU activation and dropout regularization. The early fusion technique is applied on extracted image features that undergo dense, dropout, and batch-normalization. In parallel, the metadata is processed as structured inputs using a fully connected network with a 64-dimensional representation, also regularised with dropout. The outputs from both branches are concatenated into a 192-dimensional fused representation, which is passed through a combined head consisting of two fully connected layers (128 and 4 units). The final Softmax layer is used for a four-way classification, i.e. NC, EMCI, LMCI, and AD.

Let the input image be denoted as $I \in \mathbb{R}^{H\times W\times C}$, where $H$, $W$, and $C$ represent the image height, width, and number of channels. The classification head is removed, and the image $I$ is passed through a pre-trained EfficientNet model, $f_{\text{img}}$. The resulting output $z_{\text{img}}$ is a feature embedding computed as $z_{\text{img}} = f_{\text{img}}(I) = \text{GlobalAvgPool}(\phi(I)) \in \mathbb{R}^n$, where $\phi(I) \in \mathbb{R}^{h\times w\times d}$ is the output of the final convolutional block, and global average pooling reduces this to a vector of size $n = d$. For EfficientNet, $n = 1024$.

Alongside the image features, suppose we have $l$ selected features from the tabular data, determined within a 90% confidence interval, represented as $F = [f_1, f_2, \ldots, f_l]$. Additionally, each feature $f_i \in F$ has $m$ metadata feature vectors represented by $m_i = [\theta_1, \theta_2, \ldots, \theta_d]^\top \in \mathbb{R}^d$, where each $m_i$ is a selected clinical or demographic feature.

Subsequently, the metadata feature vectors are passed through a two-layer feedforward neural network: $f_{\text{meta}}(m_i) = \sigma(W_2(\sigma(W_1 m_i + b_1)) + b_2) \in \mathbb{R}^k$, where $W_1 \in \mathbb{R}^{h\times d}$ and $b_1 \in \mathbb{R}^h$ are the first layer parameters, such as weights and bias, respectively. Similarly, $W_2 \in \mathbb{R}^{k\times h}$ and $b_2 \in \mathbb{R}^k$ represent weights and bias of the second layer. $\sigma$ denotes a nonlinear activation function (e.g., ReLU), and $k$ is the output dimension of the metadata embedding (e.g., 128). The metadata embeddings are concatenated as $z_{\text{meta}} = [f_{\text{meta}}(m_1) \,\|\, f_{\text{meta}}(m_2) \,\|\, \ldots \,\|\, f_{\text{meta}}(m_l)] \in \mathbb{R}^k$. Followingly, the image and metadata embeddings are concatenated to form a joint representation along the feature dimension. The combined vector $z_{\text{concat}} = [z_{\text{img}} \,\|\, z_{\text{meta}}] \in \mathbb{R}^{n+k}$ fuses both modalities into a single latent space.

The concatenated vector is passed through a classification head composed of fully connected layers: $\hat{y} = \text{Softmax}(W_4(\sigma(W_3 z_{\text{concat}} + b_3)) + b_4)$, where $W_3 \in \mathbb{R}^{p\times(n+k)}$, $b_3 \in \mathbb{R}^p$, $W_4 \in \mathbb{R}^{c\times p}$, and $b_4 \in \mathbb{R}^c$, with $c$ being the number of output classes. The model is trained using the categorical cross-entropy loss function: $L = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{c} y_{\text{true},i,j}\log(\hat{y}_{i,j})$, where $\hat{y}_{i,j}$ and $y_{\text{true},i,j}$ are the model prediction and true label, respectively, and $N$ is the number of training samples. To enhance model robustness, dropout is employed to reduce overfitting, early stopping is used to halt training when validation loss stagnates, and model checkpointing ensures the best-performing model is saved based on validation performance.

## 4 Results and Discussion

### 4.1 Experiments and Results

The widely used OASIS dataset was utilized in this research for conducting experiments. Table 1 summarizes the demographic and clinical statistics of the OASIS-1 cross-sectional MRI dataset. The dataset consists of four diagnostic groups: NC (CDR = 0.0), EMCI (CDR = 0.5), LMCI (CDR = 1.0), and AD (CDR = 2.0). For each group, the table reports the number of subjects, the male-to-female distribution, the mean age with standard deviation, and the number of available MRI scans (3–4 per subject). Overall, the dataset includes 436 MRI scans from 168 male and 268 female participants, with a mean age of $51.4 \pm 25.3$ years.

Preprocessed 2D MRI slices incorporating clinical features were fed into the proposed multi-modal model. Therefore, a new tabular dataset was developed that linked clinical characteristics

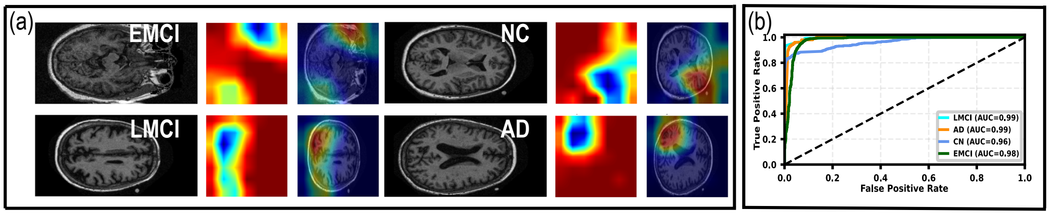


Figure 1: (a) Grad-CAM visualizes the brain regions and (b) ROC curve for the multimodal model's Alzheimer's disease prediction.

to 2D MRI slices for each patient. The most significant clinical features were identified by training six ML models on tabular clinical data and estimating feature importance using SHAP values (see Figure 3a–f in the appendix). Global importance scores were then averaged across models to obtain the final ranking (Figure 3g in the appendix). Two statistical methods were applied based on SHAP values: Z-score thresholding and cumulative contribution threshold (CCT). The Z-score method ($z > 1$) identified only *MMSE* as significantly important. In contrast, CCT retained features explaining 95% of total SHAP contribution, selecting *MMSE*, *nWBV*, *Age*, *eTIV*, and *SES*. CCT outperformed Z-score as it accounts for the right-skewed SHAP distribution without assuming normality.(see Figure 3 and Table 2).

Table 1: Clinical features statistics from the OASIS dataset

| Feature | Mean ± SD | Min–Max | Median |
|---|---|---|---|
| Age | 51.36 ± 25.27 | 18.0–96.0 | 54.0 |
| Educ | 3.18 ± 1.31 | 1.0–5.0 | 3.0 |
| SES | 2.49 ± 1.12 | 1.0–5.0 | 2.0 |
| MMSE | 27.06 ± 3.7 | 14.0–30.0 | 29.0 |
| CDR | 0.29 ± 0.38 | 0.0–2.0 | 0.0 |
| eTIV | 1481.9 ±158.7 | 1123.0–1992.0 | 1475.5 |
| nWBV | 0.79 ± 0.06 | 0.644–0.893 | 0.809 |
| ASF | 1.20 ± 0.13 | 0.881–1.563 | 1.19 |
| Delay | 20.55 ± 23.86 | 1.0–89.0 | 11.0 |

Table 2: Avg. SHAP values and CCT

| Feature | Avg. SHAP | CCT |
|---|---|---|
| MMSE | 0.3692 | 0.3692 |
| nWBV | 0.2665 | 0.6358 |
| Age | 0.2122 | 0.8479 |
| eTIV | 0.0524 | 0.9004 |
| Educ | 0.0341 | 0.9345 |
| SES | 0.0327 | 0.9672 |
| ASF | 0.0318 | 0.9990 |
| Delay | 0.0010 | 1.0000 |

To prevent patient data leakage, all 2D slices from a subject were kept in a single split. A stratified subject-wise hold-out test set (20%) was created. On the remaining 80% of subjects, a 5-fold `StratifiedGroupKFold` was used for hyperparameter tuning and early stopping. A pre-trained EfficientNet extracted image features, while a fully connected branch processed metadata; both were fused for the classification. Training used Adam ($10^{-3}$), cross-entropy loss, early stopping, and ran for up to 100 epochs on a P100 GPU. The Adam optimizer with a learning rate of 0.001 was applied. Due to the early stopping criteria, the model stopped training at the 30th epoch.

Tables 3 and 4 compare the proposed multimodal model with ML and DL baselines using Accuracy, Precision, Recall, Macro-F1 score, and AUC. Table 3 reports the performance of tabular baselines, CB, XGB, LGBM, GB, LR, and SVM trained only on clinical features. Among these, CB achieved the best baseline performance with an accuracy of 0.9091 and an AUC of 0.9529. On the other hand, 4 compares the performance of several convolutional neural networks (CNNs) against the proposed multimodal approach, including EfficientNet-B0, ResNet50, DenseNet121, MobileNet-V2, VGG16, and ConvNeXt-Tiny trained with only MRI 2D images. For both of the cases, the proposed model significantly outperformed all baselines, achieving the highest accuracy (0.9584), precision (0.9602), recall (0.9584), F1-score (0.9592), and AUC (0.9864). Figure 1 (a) shows the Grad-CAM visualization, highlighting MRI regions most influential in model predictions, with warmer colors indicating higher importance. The model achieved near-perfect AUCs (0.99 for LMCI and AD, 0.98 for EMCI, and 0.96 for CN), demonstrating strong generalization across disease stages.

| Model | Accuracy | Precision | Recall | F1-Score | AUC |
|---|---|---|---|---|---|
| CB | 0.904 ± 0.015 | 0.768 ± 0.022 | 0.843 ± 0.025 | 0.805 ± 0.020 | 0.950 ± 0.012 |
| XB | 0.882 ± 0.018 | 0.720 ± 0.025 | 0.792 ± 0.024 | 0.754 ± 0.021 | 0.946 ± 0.013 |
| LGBM | 0.879 ± 0.017 | 0.725 ± 0.023 | 0.788 ± 0.026 | 0.752 ± 0.019 | 0.941 ± 0.014 |
| GB | 0.869 ± 0.020 | 0.732 ± 0.028 | 0.692 ± 0.029 | 0.710 ± 0.025 | 0.934 ± 0.015 |
| LR | 0.848 ± 0.023 | 0.648 ± 0.030 | 0.742 ± 0.033 | 0.689 ± 0.028 | 0.930 ± 0.018 |
| SVM | 0.845 ± 0.025 | 0.660 ± 0.027 | 0.698 ± 0.031 | 0.677 ± 0.026 | 0.922 ± 0.017 |
| Proposed | **0.961 ± 0.017** | **0.958 ± 0.016** | **0.962 ± 0.015** | **0.960 ± 0.014** | **0.987 ± 0.012** |

Legend– Acc: Accuracy, Prc: Precision, Rec: Recall, CB: CatBoost, XB: XGBoost, LGBM: LightGBM, GB: Gradient Boosting, LR: Logistic Regression, SVM: Support Vector Machine.

Table 3: Performance comparison of models (mean ± std over 5 runs) for clinical data.

| Model | Accuracy | Precision | Recall | F1-Score | AUC |
|---|---|---|---|---|---|
| EN | 0.853 ± 0.018 | 0.875 ± 0.020 | 0.867 ± 0.019 | 0.865 ± 0.017 | 0.967 ± 0.010 |
| RN50 | 0.839 ± 0.020 | 0.854 ± 0.021 | 0.860 ± 0.022 | 0.857 ± 0.019 | 0.964 ± 0.011 |
| DN121 | 0.888 ± 0.016 | 0.906 ± 0.018 | 0.897 ± 0.019 | 0.902 ± 0.017 | 0.974 ± 0.009 |
| MNV2 | 0.852 ± 0.019 | 0.880 ± 0.020 | 0.865 ± 0.021 | 0.872 ± 0.018 | 0.963 ± 0.010 |
| VGG16 | 0.883 ± 0.017 | 0.900 ± 0.019 | 0.896 ± 0.020 | 0.898 ± 0.018 | 0.972 ± 0.009 |
| TCN | 0.891 ± 0.015 | 0.909 ± 0.017 | 0.899 ± 0.018 | 0.905 ± 0.016 | 0.973 ± 0.008 |
| Proposed | **0.961 ± 0.017** | **0.958 ± 0.016** | **0.962 ± 0.015** | **0.960 ± 0.014** | **0.987 ± 0.012** |

Legend– EN: EfficientNet-B0, RN50: ResNet50, DN121: DenseNet121, MNV2: MobileNet-V2, TCN: ConvNeXt-Tiny.

Table 4: Comparison of deep learning models (mean ± std over 5 runs) for MRI image data.

### 4.2 Limitations and Future Works

To conduct the experiments, a dataset linking MRI images with clinical data for each patient is required; therefore, we constructed such a dataset to train our model. Future work will focus on building additional datasets with similar criteria from open-source multimodal resources (e.g., ADNI) and benchmarking our model against other comparable multimodal approaches and diverse datasets.

## 5 Conclusion

In conclusion, we introduced a multimodal deep learning framework that integrates MRI imaging with clinical data to enhance prediction of Alzheimer's disease (AD) stages. Leveraging automated clinical feature selection and OASIS MRI data, the model overcomes limitations of single-modality approaches. It achieved strong performance (Accuracy:95.84%, Precision: 96.02%, Recall: 95.84%, F1 score: 95.92%), effectively distinguishing normal cognition, EMCI, LMCI, and AD despite class imbalance. This work provides a robust diagnostic tool aligned with clinical practice that supports timely intervention strategies. The code base of the work is available at this GitHub repository: `https://github.com/brai-acslab/multimodal-explainable-ad-detection`.

## References

[1] U. Rajendra Acharya et al. Automated Detection of Alzheimer's Disease Using Brain MRI Images- A Study with Various Feature Extraction Techniques. *Journal of Medical Systems*, 43(9):302, 2019.

[2] Marshal F. Folstein et al. "Mini-mental state": A practical method for grading the cognitive state of patients for the clinician. *J. Psychiatr. Res.*, 12:189–198, 1975.

[3] Serge Gauthier. World Alzheimer Report 2022: Life after diagnosis: Navigating treatment, care and support. 2022. URL: `https://www.alzint.org/resource/world-alzheimer-report-2022/`.

[4] Harald Hampel et al. Blood-based biomarkers for Alzheimer disease: mapping the road to the clinic. *Nature Reviews. Neurology*, 14(11):639–652, 2018.

[5] Sobhana Jahan, Kazi Abu Taher, M Shamim Kaiser, Mufti Mahmud, Md Sazzadur Rahman, ASM Sanwar Hosen, and In-Ho Ra. Explainable ai-based alzheimer's prediction and management using multimodal data. *Plos one*, 18(11):e0294253, 2023.

[6] Sobhana Jahan, Md. Rawnak Saif Adib, Mufti Mahmud, and M. Shamim Kaiser. Comparison Between Explainable AI Algorithms for Alzheimer's Disease Prediction Using EfficientNet Models. In *Brain Informatics*, volume 13974, pages 357–368. Springer Nature Switzerland, Cham, 2023.

[7] Alessandro Leparulo et al. Dampened Slow Oscillation Connectivity Anticipates Amyloid Deposition in the PS2APP Mouse Model of Alzheimer's Disease. *Cells*, 9(1):54, 2019.

[8] Linfeng Liu et al. Cascaded Multi-Modal Mixing Transformers for Alzheimer's Disease Classification with Incomplete Data. *NeuroImage*, 277:120267, 2023.

[9] 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.

[10] Ziad S. Nasreddine et al. The Montreal Cognitive Assessment, MoCA: A Brief Screening Tool For Mild Cognitive Impairment. *J. Am. Geriatr. Soc.*, 53(4):695–699, 2005.

[11] Shangran Qiu et al. Development and validation of an interpretable deep learning framework for Alzheimer's disease classification. *Brain: A Journal of Neurology*, 143(6):1920–1933, 2020.

[12] Nasir Rahim et al. Prediction of Alzheimer's progression based on multimodal Deep-Learning-based fusion and visual Explainability of time-series data. *Inf. Fusion*, 92(C):363–388, 2023.

[13] Noushath Shaffi et al. Ensemble of vision transformer architectures for efficient Alzheimer's Disease classification. *Brain Informatics*, 11(1):25, 2024.

[14] Noushath Shaffi et al. Performance Evaluation of Deep, Shallow and Ensemble Machine Learning Methods for the Automated Classification of Alzheimer's Disease. *Int. J. Neural Syst.*, 34(07):2450029, 2024.

[15] Janani Venugopalan et al. Multimodal deep learning models for early detection of Alzheimer's disease stage. *Scientific Reports*, 11(1):3254, 2021.

[16] Viswan Vimbi et al. Interpreting artificial intelligence models: a systematic review on the application of LIME and SHAP in Alzheimer's disease detection. *Brain Informatics*, 11(1):10, 2024.

## A Technical Appendices and Supplementary Material

The figure 2 illustrates the framework of proposed multimodal model. Firstly, the clinical feature were collected. Then six different ML models, such as: Gradient Boosting, XGBoost, CatBoost, LightGBM, Support Vector Machine, and Logistic Regression, were trained with the clinical feature. Followingly, the SHAP feature importance algorithm was applied to find out the features that contribute most to the model prediction shown in figure 3. Subsequently, the SHAP feature importance scores of each feature across different models were averaged to avoid the bias of a single algorithm. SHAP values presented in figure 3 (a)-(g) ranged from 0.00 to 0.35, with an interquartile range of 0.05 to 0.21, indicating that most features contributed moderately to the model's predictions. A smaller subset reached values up to 0.35, reflecting stronger localized influence. In the context of Alzheimer's disease classification, this suggests that while many features have balanced effects, a few provide disproportionately higher predictive value, consistent with key biomarkers. Next, we performed a statistical analysis using the z-score thresholding and the cumulative contribution threshold (CCT) to define an automated thresholding to select the key features where CCT performed the best thresholding (Details in section 4.1). With the defined confidence interval by CCT the key features are selected and are ready to be fed into the model.

In addition, the 2D MRI images were sliced from 3D MRI scan and then pre-processed (14). Subsequently, the selected key features and 2D MRI images were fed to the multimodal model in parallel and the model performs an early fusion strategy to perform the four-way Alzheimer disease classification.

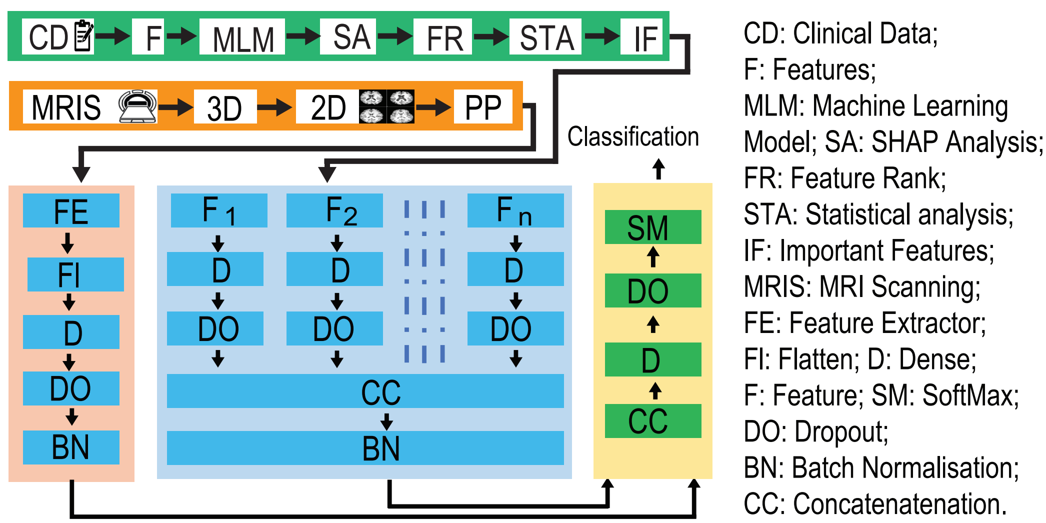


Figure 2: Proposed framework of multimodal model where two parallel pipelines of tabular data and MRI scans are preprocessed and fed to the multimodal model.

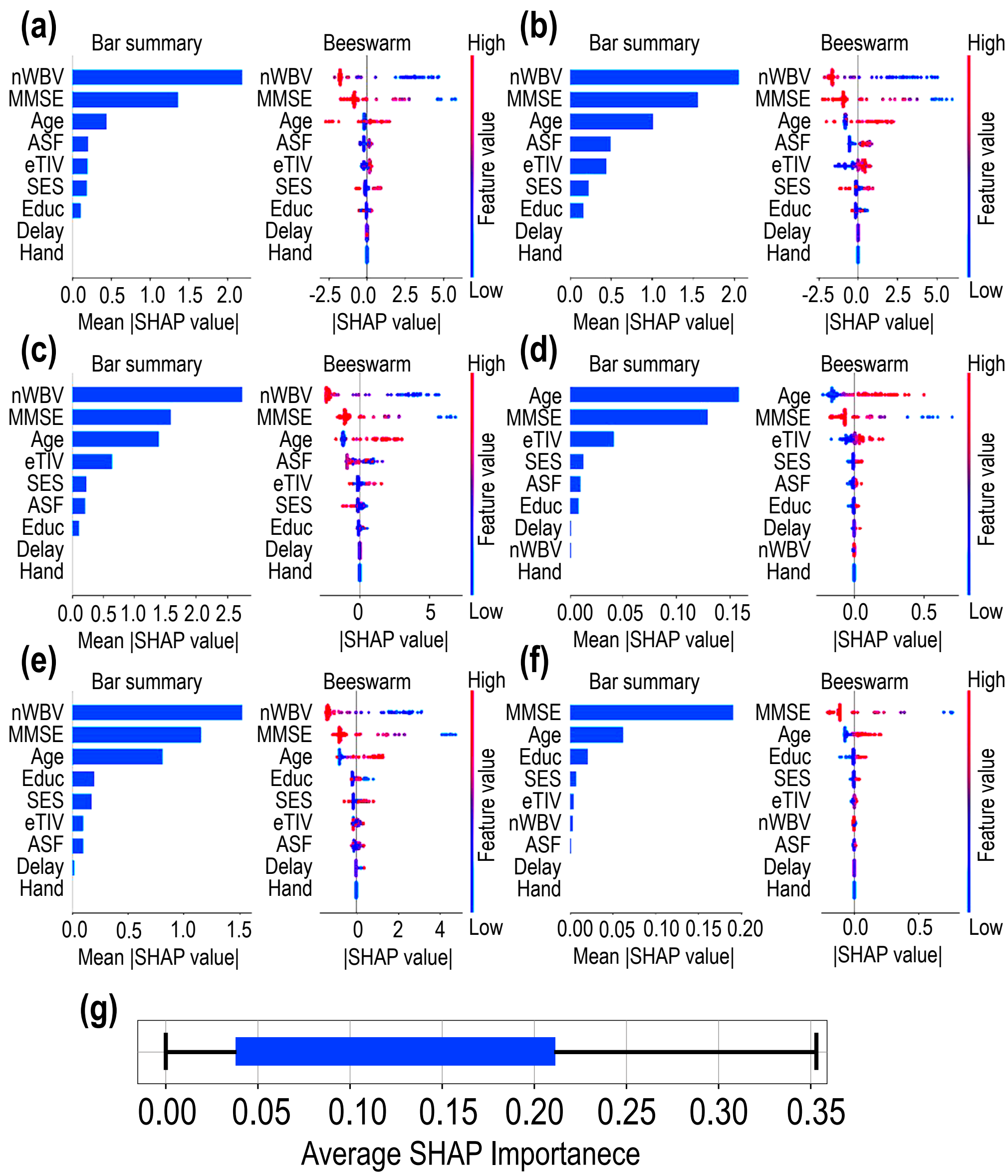


Figure 3: SHAP feature importance across different models: (a) Gradient Boosting, (b) XGBoost, (c) LightGBM, (d) Logistic Regression, (e) CatBoost, and (f) SVM. (g) shows the boxplot of global SHAP values of different features. The Bar-Summary and Beeswarm represent the mean SHAP values on the X-axis, and the feature ranks are on the Y-axis.

## NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .

- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

**The checklist answers are an integral part of your paper submission.** They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

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 clearly state the contributions—a multimodal deep learning model integrating MRI and clinical features for AD prediction, with automated SHAP-based feature selection. The superiority of the multimodal model over the single modality is directly supported by results (accuracy 95.84%, precision 96.02%, etc.) shown in Section 4.1.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

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

   Answer: [Yes]

   Justification: Section 4.2 explicitly acknowledges limitations, noting the reliance on constructing a dataset linking MRI and clinical data, and the need for validation on larger, more diverse multimodal datasets such as ADNI.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.

- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

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: The paper does not present theoretical results or formal proofs; it is an empirical deep learning study.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

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: Section 3 (Methodology) and Appendix 5 describe preprocessing, SHAP-based feature selection, model architecture, and training setup (optimizer, early stopping, dropout, GPU type). These details enable reproduction.

   Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the 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 paper provides a reference to the data used in the experiment. Additionally, a GitHub link is provided to access the code base for the reproduction of the results.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.

- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

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

   Answer: [Yes]

   Justification: The paper specifies experiments ran on a P100 GPU for up to 100 epochs, with early stopping at epoch 30

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

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: Results report performance metrics (accuracy, precision, recall, F1, AUC) but do not include error bars or statistical significance tests (Tables 3 and 4).

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

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: The paper specifies experiments ran on a P100 GPU for up to 100 epochs, with early stopping at epoch 30.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

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 complies with the NeurIPS Code of Ethics. It uses de-identified, publicly available OASIS MRI/clinical data, avoiding privacy risks.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

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

    Answer: [Yes]

    Justification: Positive societal impact: earlier AD diagnosis can aid intervention and treatment. Negative risk is addressed as reliance on limited datasets.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

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: [NA]

Justification: The work does not release high-risk generative models or large-scale datasets requiring safeguards.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

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 OASIS dataset is cited properly (9).

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's 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 code base and data are well documented throughout the manuscript and the appendix, along with a link to the GitHub repository.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.

- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

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 hasn't been done within this research.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

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: Publicly accessible dataset was used; therefore, the IRB was needed for this work.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

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? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: LLMs were not used in this work.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.