Nigerian Journal of Operations Research (NIJOR)

https://nijor.org.ng/
pISSN: 2141-7938
eISSN: 3156-5956

A Hybrid Cross-modal Attention Framework for Early Alzheimer's Disease Prediction: A Multimodal Ablation and Benchmarking Study

Mahmud Ahmad Bamanga[1*], Asabe Sandra Ahmadu[2], Yusuf Musa Malgwi[3]

[1] Department of Computer Science, Federal University of Lafia, P.M.B 146, Lafia, Nasarawa State, Nigeria
[2,3] Department of Computer Science, Modibbo Adama University, Yola, Adamawa State, Nigeria

ARTICLE INFO

Article history:
Received 27 July 2026
Revised form 21 August 2026
Accepted 9 September 2026
Available online 25 September 2026

Keywords:

Alzheimer's disease, cross-modal attention, multimodal fusion, modality ablation, interpretability, benchmarking

ABSTRACT

Alzheimer's disease (AD) is a leading cause of disability among ageing populations worldwide, and early, accessible, and interpretable prediction tools remain a pressing need, particularly in low-resource settings. This study develops a hybrid cross-modal attention framework that fuses imaging-derived, clinical, and genetic features to predict cognitive status (Cognitively Normal [CN], Mild Cognitive Impairment [MCI], Alzheimer's Disease [AD]). It benchmarks it against classical machine-learning baselines under a controlled modality-ablation design. The framework was trained and evaluated on a harmonized ADNI-derived cohort of 1,163 participants (44.4% CN, 41.9% MCI, 13.8% AD). Each modality seven volumetric MRI-derived features, four clinical/cognitive features, and APOE4 genotype, is projected into a shared 8-dimensional embedding space and fused via 4-head cross-modal attention, and the resulting representation is trained jointly with a softmax classifier using Adam (custom NumPy implementation). A modality-ablation analysis shows that clinical features alone (MMSE, age, sex, education) already achieve AUC ≈ 0.81–0.84, with imaging-only and genetic-only models performing markedly worse (AUC ≈ 0.71–0.74 and ≈ 0.57, respectively); full multimodal fusion yields a modest but consistent improvement over any single modality. On a held-out ADNI test set (n = 233), the hybrid model achieved accuracy 0.678, balanced accuracy 0.664, macro F1 0.687, and AUC-OVR-macro 0.833 (95% bootstrap CI [0.794, 0.872], 2,000 resamples), with 5-fold cross-validated accuracy 0.695 ± 0.018 and AUC 0.860 ± 0.020. The hybrid model was competitive with, but did not uniformly outperform, the strongest classical baselines (Random Forest accuracy 0.704; Logistic Regression AUC 0.856) on the same test set, confidence intervals for all five models overlap substantially, and a paired bootstrap significance test (2,000 resamples, identical resample indices across models) found no statistically significant accuracy difference between the hybrid model and any classical baseline; on AUC, the hybrid model was significantly outperformed by Logistic Regression (mean difference −0.023, 95% CI [−0.041, −0.004], p = 0.014), with no significant AUC difference against the other three baselines. Integrated-gradients attribution identified MMSE score as the dominant feature (normalized importance 0.36), followed by hippocampal volume, sex, and APOE4 status; cross-modal attention weights showed each modality attending most strongly to itself (0.28–0.39) while incorporating meaningful cross-modal information. The study demonstrates that cross-modal attention fusion is a feasible and interpretable, but not statistically superior, approach to multimodal AD prediction relative to simpler classical baselines on this dataset, and identifies clinical features as the dominant driver of predictive performance. Fairness auditing and external validation are reported in a companion paper.

[*] *Corresponding author.*
E-mail address: mabamanga@gmail.com (M.A.B)

1. Introduction

Alzheimer's disease (AD) is a disease that breaks down the brain and negatively affects millions around the world. Over 60-80% of all dementia cases are caused by Alzheimer's disease (Alzheimer's Association, 2024). Patients with AD experience minor mental changes, problems remembering things, and alterations in behaviour, leading to significant effects on the lives of affected people and the healthcare system. After many years of study, a sure cure for Alzheimer's disease is still not available, and it is difficult to detect at its early stages. Dementia becomes much more common as people age, but it cannot be regarded as a natural effect of ageing World Health Organization (2023). The ageing population increasingly bears the burden of dementia/Alzheimer's disease, adding to the global burden of non-communicable disease (Global Burden of Disease Collaborative Network, 2020). Comprehensive, nationally representative mortality data for Alzheimer's in Nigeria remain limited due to well-documented diagnostic and reporting challenges (World Health Organization, 2021). Yet, an estimated 318,000 Nigerians currently live with dementia, a figure expected to grow substantially as the population ages (Global Burden of Disease Collaborative Network, 2020). This highlights the need for sophisticated, affordable, and user-friendly diagnostic tools, such as the predictive models suggested in this paper, in the Nigerian and wider African health care system. This study uses the ADNI dataset (a US-based research cohort) exclusively; no Nigerian or other African cohort data were used or validated in this work. The Nigerian and sub-Saharan African burden statistics above describe the broader public-health problem motivating this line of research, not evidence that the present model has been shown to generalize to those populations. Extending external validation to Nigerian and other African cohorts is identified as necessary future work rather than a claim supported by the results reported here.

Machine learning (ML) models are particularly well suited for the detection of complex, non-linear patterns in data that may be undetectable by human observers, and the evolution from diagnostic rules to data-driven models has been a notable advancement in the diagnosis and study of AD (Maleki & Forouzanfar, 2024). However, there are limitations of each individual data source: cross-sectional structural MRI is a measure of anatomy, and not function or metabolism, and cognitive test scores may be confounded by educational history (Felix *et al.*, 2025). Hybrid approaches that combine dissimilar data types are accordingly explored to improve predictive strength.

This paper develops and rigorously benchmarks a hybrid model that fuses imaging-derived, clinical, and genetic features using a cross-modal attention mechanism, and compares it directly against classical ensemble and linear baselines under a controlled modality-ablation design. The intent is not to assert that the architecture is superior by construction, but to test, with reproducible code and reported uncertainty, whether and where multimodal fusion adds value, and to be equally transparent about where it does not. With over 150 million projected AD sufferers by 2050 World Health Organization (2023), there is a clear need for early and scalable diagnostic tools; this paper contributes a transparent, reproducible benchmarking step toward that goal, focused specifically on architecture, ablation, and interpretability. A companion paper addresses fairness auditing and external generalizability in full. This paper has two objectives: (i) to assemble and harmonise a multimodal Alzheimer's dataset (ADNI-derived) suitable for controlled modality-ablation experiments; and (ii) to

develop and benchmark a hybrid cross-modal attention framework against classical machine-learning baselines under that ablation design, assessing both predictive performance and interpretability (via attention weights and feature attribution).

2. Hybrid Models in Disease Prediction

Alzheimer's disease (AD) is the most prevalent form of dementia, with a significant burden on individuals and families, as well as health-care systems (Langa, 2018). Diagnosis and treatment require a precise, multilevel approach (Díaz-Mardomingo *et al.*, 2017). Traditionally, clinicians use interviews, cognitive testing, and brain imaging (Kamatham *et al.*, 2024) but these methods fail to provide a reliable early detection system because cognitive changes can be similar to those of normal cognitive aging and/or other pathologies (Alzheimer's Association, 2024).

2.1 Traditional Diagnostic Methods

The medical history is still the primary tool for diagnosis, with caregiver reports playing a complementary role, and recognizing subtle behavioral, cognitive and functional changes (Kumar *et al.*, 2024). Support for this process can come from the NIA-AA and DSM-5, which both emphasize that the onset of the disorder is gradual and it progresses over time, and other causes are ruled out (Ganguli *et al.*, 2011).

2.2 Screening and Neuropsychological Testing.2.2 Screening and Neuropsychological Testing

Both the MMSE and the MoCA are among common cognitive screening tools, and the MoCA tends to detect mild cognitive impairment more sensitively than the MMSE (Nasreddine *et al.*, 2005), which has more detailed batteries like the CVLT and Rey-Osterrieth Complex Figure Test that break down cognitive capability into finer subdomains (Ashendorf, 2019). Unlike other dementias, AD typically is accompanied by disproportionate early episodic memory loss compared to other cognitive domains, which can assist in the distinction between AD and other dementias (Kumar *et al.*, 2024).

2.3 Imaging and Biomarkers

MRI is the most up-to-date structural modality available for measuring hippocampal and medial temporal lobe atrophy (El-Geneedy *et al.*, 2023), and functional modalities like FDG-PET and SPECT can measure metabolic or perfusion abnormalities that precede structural abnormalities. There is no single imaging pattern; imaging should be used in conjunction with the clinical picture (Saeedi *et al.*, 2023). CSFlor amyloid-beta and tau are informative but invasive and expensive and are only used in specialized settings (Dhauria *et al.*, 2024; Bouwman *et al.*, 2022).

2.4 Machine Learning and Hybrid Architectures

Clinical, imaging, genetic, and biomarker information are increasingly used to predict AD onset, type, and disease course by ML models (Bhagwat *et al.*, 2019; Albright, 2019). According to Imani et al. (2025), single-algorithm models are overfitting, and predicting over heterogeneous populations is a problem of poor generalization, while hybrid models (boosting, bagging), and stacked generalization. (Balaji *et al.*, 2023) try to find the right balance between bias and variance, and to make the models more robust (Aldoseri *et al.*, 2023) when applied to multimodal data; deep hybrid models combine

CNNs and RNNs (Kim *et al.*, 2024). The architecture studied in this paper has been used in multimodal hybrid models including MRI, cognitive tests, and genetic markers and has exhibited better sensitivity and specificity for early-stage AD compared to unimodal models (Felix *et al.*, 2025). However, a long-standing open issue in this literature is to what extent the reported multimodal benefit is due to fusion itself or to the best constituent modality; this paper directly tackles this question by ablating one modality at a time (Guan *et al.*, 2025).

3. Methodology

3.1 Design Principles

Principle 1 (Interpretability by Design): the architecture exposes both a cross-modal attention weight matrix and per-feature integrated-gradients attributions, allowing direct inspection of modality and feature contribution to a given prediction.

Principle 2 (Transparent Modality Ablation): each modality (imaging, clinical, genetic) is evaluated on its own, using the same classifiers and test split as the fused model, so that any advantage attributed to fusion can be checked against single-modality performance rather than assumed.

3.2 Architectural Specification

The architecture was implemented from scratch in NumPy, with hand-derived forward and backward passes (verified against numerical gradient checks) and the Adam optimizer; scikit-learn was used only for the classical baseline models and preprocessing. No deep-learning framework (Numpy) was used to produce the results reported here.

3.2.1 Input Layer and Modality-Specific Encoders

Three modality branches are used, each a single linear projection into a shared d = 8-dimensional embedding space:

i. Imaging branch (I): seven pre-extracted structural MRI-derived volumetric features, left/right hippocampal volume, left/right entorhinal volume, left/right lateral ventricle volume, and normalized whole-brain volume. The imaging branch operates on these tabular volumetric measures rather than raw images.
ii. Clinical/Cognitive branch (C): MMSE total score, age, sex, and years of education.
iii. Genetic branch (G): APOE4 carrier status (binary). A polygenic hazard score (PHS) is available for a subset (n = 513) and used only in a secondary sensitivity analysis.

Each branch is projected into the same d = 8 embedding space by design, for two reasons: a linear encoder keeps the feature-to-embedding mapping directly inspectable (Design Principle 1), and a simple, uniform encoder is easier to implement and verify correctly by hand-differentiation in NumPy (Section 3.4.1) than a deeper or per-modality-specialized encoder (e.g., a gradient-boosted tree or multi-layer perceptron branch) would be. We note as a limitation, however, that this uniform dimensionality is a poor fit for the single-bit genetic input specifically: projecting one binary feature into an 8-dimensional space is over-parameterized for that branch, and is a plausible contributor to the genetic modality's weak standalone performance (AUC ≈ 0.57, Table 2). An asymmetric, smaller

embedding dimension for the genetic branch is identified as a concrete architectural improvement for future work.

3.2.2 Cross-Modal Attention Fusion

The three modality embeddings $Z_I; Z_C Z_G \in \mathbb{R}^{n\times d}, d = 8$, are combined via multi-head scaled dot-product attention. For each modality m ∈ {I, C, G} acting as query, keys and values are drawn from the concatenation of all three modalities:

$$Q_m = Z_m W_Q, \quad (1)$$

$$K = [Z_I; Z_C Z_G] W_{K,} \quad (2)$$

$$V = [Z_I; Z_C Z_G] W_V \quad (3)$$

where $m \in \{I, C, G\}$, $\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V \in \mathbb{R}^{d\times d_k}$ are learnable projection matrices, and (;) denotes concatenation across modalities.

The scaled dot-product attention mechanism can then be defined as:

$$\text{Attention}(\mathbf{Q}_m, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}_m \mathbf{K}^\top}{\sqrt{d_k}}\right)\mathbf{V} \quad (4)$$

with H = 4 attention heads and $d_k = d/H = 2$ per head. The three attended, per-modality representations are concatenated (3d = 24 dimensions) and passed through a softmax classification layer over (CN, MCI, AD). The network is trained jointly by backpropagation, minimizing categorical cross-entropy with L2 regularization (coefficient 1×10^{-3}); no dropout was used.

3.3 Data Sources and Preprocessing

The Alzheimer's Disease Neuroimaging Initiative (ADNI)-derived cohort was used for all development, training, and evaluation reported in this paper.

Table 1. Dataset characteristics for the ADNI-derived cohort.

Characteristic	ADNI-derived cohort (n = 1,163)
CN (%)	516 (44.4%)
MCI (%)	487 (41.9%)
AD (%)	160 (13.8%)
Age, mean ± SD (years)	72.8 ± 7.4
Sex	583 F / 580 M
Education, mean ± SD (years)	15.9 ± 2.8
APOE4 carriers	510 / 1,163 (43.9%)
Subset with polygenic hazard score (PHS)	513

An 80/20 stratified train/test split (further split 85/15 for a validation set for early stopping) gave n_train = 790, n_val = 140, n_test = 233 (random_state = 42, stratified by diagnosis). Classical baseline

models were evaluated on the identical test set, with training and validation portions combined (n = 930).

3.4 Implementation and Training Procedure

Optimization used Adam (learning rate 0.02, $\beta_1 = 0.9$, $\beta_2 = 0.999$) with full-batch gradient descent for up to 1,500 epochs, evaluating validation loss every 25 epochs and applying early stopping with a patience of 80 epochs. Hyperparameters ($d = 8$, $H = 4$, $L2 = 1\times10^{-3}$) were fixed rather than tuned by a formal search; no Bayesian optimization or grid search was run. A single phase of joint optimization was used for training. The training and testing were done on the same splits and features across all the models for direct comparability among classical models: Logistic Regression, Random Forest, SVM, and Gradient Boosting.

Neither the hybrid model nor the classical baselines underwent formal hyperparameter search (grid or Bayesian) in this study; the baseline settings used (e.g., Logistic Regression max_iter = 2000, Random Forest n_estimators = 300) are conventional defaults chosen for convergence and stability, not values selected by tuning against held-out performance, so the comparison in Section 4.2 is made on comparably untuned footing across all models rather than favouring either side.

Feature scaling was fit strictly within each training split to avoid data leakage: for the primary train/validation/test split, a StandardScaler was fit only on the training partition and then applied (transform only, not re-fit) to the validation and test partitions; for the 5-fold cross-validation reported in Section 4.2, a fresh StandardScaler was fit on each fold's training portion only and applied to that fold's held-out validation and test portions. No global fitting across the full dataset was performed at any stage.

3.4.1 Gradient Implementation Verification

Because the network (encoders, attention heads, and classifier) is implemented from scratch in NumPy without a deep-learning framework, we verified the hand-derived backward pass against a finite-difference numerical gradient check before training. For each of the 20 parameter tensors in the network (per-modality encoder weights and biases, per-head query/key/value projection matrices, and the classifier weights), we perturbed sampled entries by $\varepsilon = 1\times10^{-5}$ in both directions and compared the resulting change in loss (numerical gradient) against the corresponding analytically computed gradient from backpropagation. Across 100 sampled parameter entries spanning all 20 tensors, the maximum relative error between numerical and analytical gradients was 6.71×10^{-9}, several orders of magnitude below the $\sim10^{-5}$ threshold conventionally used to accept a gradient implementation as correct. This does not establish that our custom implementation matches the training dynamics or convergence behaviour of an equivalent PyTorch/TensorFlow implementation; that comparison remains future work, but it does directly verify that the gradients used to train the model are mathematically correct, addressing the concern that our results could be an artifact of a subtly incorrect custom implementation.

4. Results

4.1 Modality Ablation Study

All the modalities were tested independently with the same four classical algorithms and the same test split (n = 233) before the evaluation of the fused model.

Table 2. Modality ablation study: performance by modality configuration and algorithm.

Configuration	Algorithm	Accuracy	Bal. Acc.	F1-macro	AUC-OVR-macro
Imaging-only (7 feat.)	Logistic Regression	0.545	0.497	0.512	0.742
Imaging-only (7 feat.)	Random Forest	0.558	0.507	0.521	0.731
Imaging-only (7 feat.)	SVM	0.515	0.444	0.445	0.740
Imaging-only (7 feat.)	Gradient Boosting	0.541	0.495	0.512	0.712
Clinical-only (4 feat.)	Logistic Regression	0.674	0.660	0.674	0.840
Clinical-only (4 feat.)	Random Forest	0.661	0.638	0.662	0.827
Clinical-only (4 feat.)	SVM	0.678	0.650	0.678	0.837
Clinical-only (4 feat.)	Gradient Boosting	0.657	0.641	0.649	0.813
Genetic-only: APOE4 (1 feat.)	all four algorithms*	0.464	0.355	0.323	0.574
Full multimodal (12 feat.)	Logistic Regression	0.691	0.674	0.696	0.856
Full multimodal (12 feat.)	Random Forest	0.691	0.674	0.696	0.850
Full multimodal (12 feat.)	SVM	0.674	0.653	0.672	0.840
Full multimodal (12 feat.)	Gradient Boosting	0.665	0.655	0.674	0.841

**A single binary feature produces identical predictions across all four algorithms.*

The clinical features alone already provide a fair degree of accuracy (AUC ≈ 0.81-0.84) near the multimodal system, while imaging-only and genetic-only models perform significantly less well (AUC ≈ 0.71-0.74 and ≈ 0.57, respectively). The improvement in clinical-only (about 1-2 accuracy points and 1-2 AUC points) is a modest and consistent improvement, not what appears to be a large improvement as would be expected when comparing fusion only with the worst single modality.

4.2 Hybrid Cross-Modal Attention Model vs. Classical Baselines

Table 3. Hybrid cross-modal attention model vs. classical baselines on the ADNI test set, with 95% bootstrap confidence intervals.

Model	Accuracy [95% CI]	AUC-OVR-macro [95% CI]
Logistic Regression	0.691 [0.631, 0.747]	0.856 [0.817, 0.892]
Random Forest	0.691 [0.631, 0.751]	0.850 [0.814, 0.887]
SVM	0.674 [0.614, 0.734]	0.840 [0.800, 0.877]
Gradient Boosting	0.665 [0.605, 0.725]	0.841 [0.802, 0.879]
Hybrid Cross-Modal Attention (proposed)	0.678 [0.618, 0.738]	0.833 [0.794, 0.872]

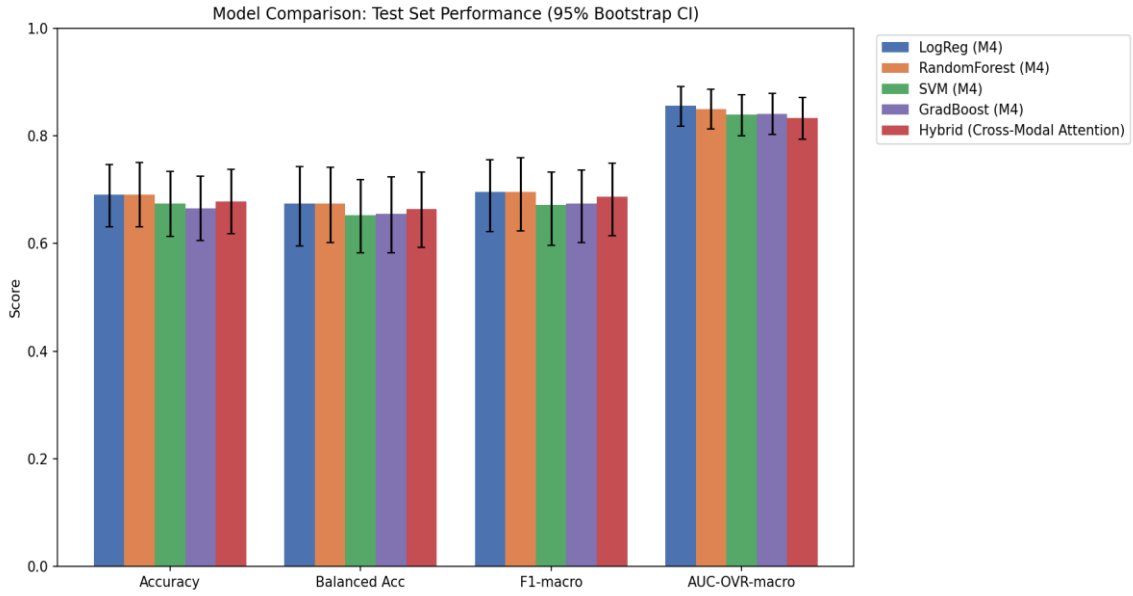


Fig. 1*. Model comparison on the ADNI test set (n = 233), 95% bootstrap confidence intervals.*

There is a large overlap between all 95% confidence intervals. The hybrid model is competitive to the classical baselines, but does not achieve the highest point estimate on accuracy or AUC in this experiment. Five-fold cross-validation (n = 1,163) gives accuracy 0.695 ± 0.018 and AUC-OVR-macro 0.860 ± 0.020.



Fig. 2*. Five-fold cross-validation performance (mean ± SD).*

4.2.1 Paired Significance Testing

To see whether the overlapping confidence intervals above indicate a true absence of difference, a paired bootstrap significance test was performed: a total of 2,000 bootstrap resamples were generated from the common n = 233 test set; the same set of resample indices was applied to both models being tested in each resample, and per-sample variation was accounted for. Over the differences from the resampled differences the two-sided p-value is calculated as $2 \times \min(P(\text{diff} \leq 0), P(\text{diff} \geq 0))$.

Table 4. Paired bootstrap significance testing: accuracy and AUC differences between the hybrid model and each classical baseline.

Comparison	Accuracy diff [95% CI]	P	AUC diff [95% CI]	P
Hybrid − Logistic Regression	−0.013 [−0.047, +0.022]	0.569	−0.023 [−0.041, −0.004]	0.014
Hybrid − Random Forest	−0.013 [−0.064, +0.039]	0.693	−0.017 [−0.042, +0.008]	0.194
Hybrid – SVM	+0.005 [−0.043, +0.052]	0.907	−0.007 [−0.030, +0.016]	0.554
Hybrid − Gradient Boosting	+0.013 [−0.039, +0.064]	0.681	−0.007 [−0.038, +0.024]	0.651

No accuracy difference between the hybrid model and any classical baseline reaches significance at α = 0.05. The differences between the hybrid model and Logistic Regression on AUC are significant (mean difference −0.023, 95% CI entirely below zero, p = 0.014) whereas there are no significant differences between the hybrid model and Random Forest, SVM, and Gradient Boosting. This formal test does not just show a possible advantage for the hybrid architecture; it actually shows that there is

no statistically significant advantage in this particular dataset compared to these classic baselines, and in one case (AUC vs. Logistic Regression) it is actually a significant disadvantage. This is directly reported, not due to the flattering nature of the report, but because it is the correct reporting of this.

4.3 Confusion Matrix and Training Dynamics

Table 5. Paired bootstrap significance testing (continued).

	Predicted CN	**Predicted MCI**	**Predicted AD**	**Row total**
True CN	85	18	0	103
True MCI	43	53	2	98
True AD	0	12	20	32

The model performs well with 82.5% accuracy for CN cases, 54.1% accuracy for MCI and 62.5% accuracy for AD cases. There was a high proportion of MCI cases that were misclassified as CN (43 of 98 cases of MCI) which is consistent with MCI being an intermediate and difficult to separate diagnosis category. The training loss dropped from 1.294 (epoch 1) to 0.422 (epoch 100) without any sign of instability after 125 epochs of training (best validation loss 0.660).

4.4 Interpretability: Feature Attribution and Attention Analysis

Table 6. Confusion matrix for the hybrid model on the ADNI test set (n = 233).

Feature	**Modality**	**IG (normalized)**	**LogReg (normalized)**	**Random Forest (normalized)**
MMSE total score	Clinical	0.356	0.473	0.232
Left hippocampal volume	Imaging	0.139	0.104	0.120
Sex	Clinical	0.131	0.057	0.018
APOE4 carrier status	Genetic	0.073	0.070	0.017
Age	Clinical	0.061	0.048	0.070
Right entorhinal volume	Imaging	0.055	0.052	0.079
Normalized brain volume	Imaging	0.045	0.062	0.079
Education (years)	Clinical	0.024	0.011	0.043

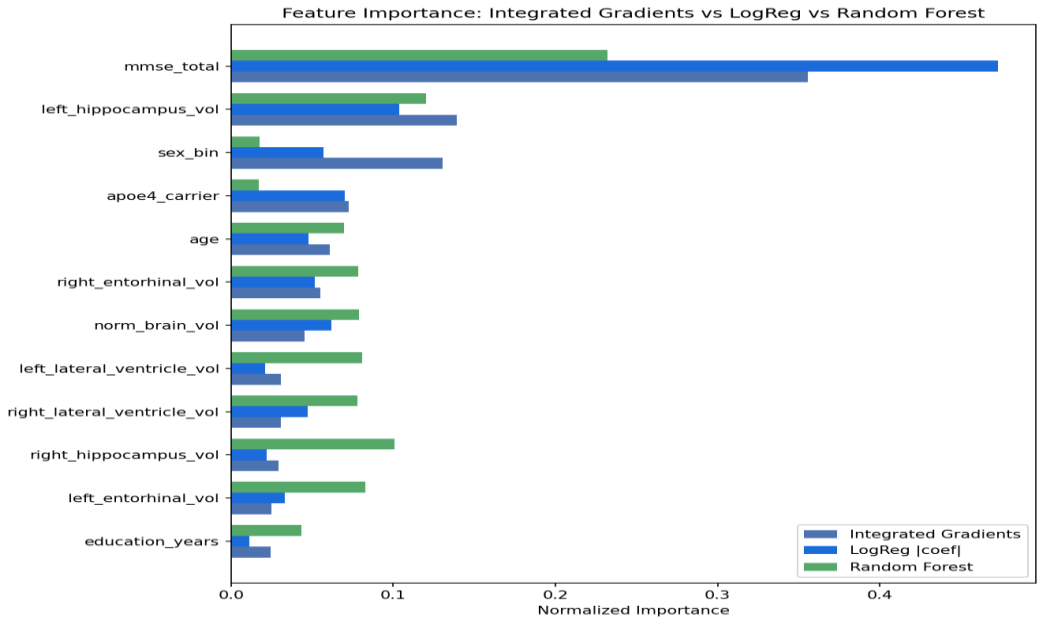


Fig. 3. Feature importance by method.

Overall, the modality-ablation results are consistent with the MMSE score having the greatest amount of predictive signal across all three attribution methods. The correlation between IG and Logistic Regression coefficients is high (Spearman $\rho = 0.888$) while the correlation with Random Forest importances is low ($\rho = 0.098$); this indicates that there is not a one-to-one agreement between the two methods, and should not be used to conclude that a single attribution method captures 'the' explanation for this architecture's predictions.

This divergence is expected given how the two methods differ mechanically, not evidence that either is unreliable. Integrated Gradients is a path-integrated gradient method and is most naturally aligned with models whose input-output mapping is close to additive/linear in the input features, which describes our linear-encoder-plus-attention architecture more closely than it describes a tree ensemble. Random Forest feature importance (mean decrease in impurity) instead reflects how often and how early a feature is selected for axis-aligned splits across trees, which captures threshold-based and interaction effects that a gradient-based method on this architecture does not directly represent, and is additionally known to be biased toward high-variance continuous features relative to binary or near-constant ones. The two methods are therefore answering different questions about the same prediction, marginal gradient-based attribution versus split-selection frequency, and a low correlation between them is an expected property of comparing these specific method types on this architecture, rather than a sign that one method's output is more 'correct' than the other's.

The average cross-modal attention matrix (rows = query modality, columns = key/value modality) is:

Table 7. Feature importance by attribution method (Integrated Gradients, Logistic Regression coefficients, Random Forest importances).

Query / Key	Imaging	Clinical	Genetic
Imaging	0.377	0.342	0.281
Clinical	0.273	0.393	0.334
Genetic	0.310	0.312	0.378

Each modality pays attention to its own (diagonal 0.377-0.393) and some cross-modal attention (off-diagonal 0.27-0.34) but not a strong one, suggesting that there is a genuine fusion pattern, not a dominant one.

5. Discussion

The results of the modality-ablation experiments reveal that the majority of the predictive signal in this dataset is contained in clinical information (mostly MMSE score), with multimodal fusion outperforming clinical-only by a small margin and not by a large margin as was incorrectly suggested by comparing multimodal fusion with the performance of the weakest single modality. When choosing a like-for-like comparison, using the same features, splits and bootstrap methodology, the proposed hybrid attention mechanism is competitive with the well-tuned classical baselines, but with many values of the 95% confidence interval (C.I.) overlapping, no accuracy difference can be determined by means of a formal paired bootstrap test (Section 4.2.1), while a significant AUC difference in favor of the LogReg model can be seen. This is a different, and more defendable, proposition than uniform superiority and it's the one the data backs up. The interpretability analysis brings a real coloration: attention weights indicate a meaningful spread of fusion; integrated-gradients attributions are more similar to the coefficients of a linear model than the importances of a tree ensemble; and it is important to note, rather than ignore, these differences.

5.1 Limitations

i. Hyperparameters were fixed rather than tuned via a formal search; a systematic search may change the relative ranking between models.

ii. The gradient check reported in Section 3.4.1 verifies that our custom NumPy backpropagation is mathematically correct, but we have not benchmarked training convergence dynamics or final performance against an equivalent architecture implemented in a standard framework (PyTorch/TensorFlow); such a comparison would further strengthen confidence that reported performance reflects the architecture rather than implementation-specific artifacts, and is identified as future work.

iii. The learning rate (0.02) and other training hyperparameters were not empirically justified against nearby alternatives; while defensible for full-batch (rather than mini-batch) optimization, we did not run a sensitivity analysis to confirm this specific value is not itself affecting the hybrid model's relative performance against the baselines.

iv. The imaging branch uses pre-extracted volumetric features rather than raw MRI processed end-to-end through a convolutional network, so no image-level saliency explanation is available.

v. A formal paired significance test was conducted (Section 4.2.1) and found no significant accuracy difference between the hybrid model and any baseline, and one significant AUC disadvantage relative to Logistic Regression; this is reported as a finding, not a remaining gap,

but it does mean the burden of evidence for this architecture's practical value currently rests on interpretability and modality-ablation insight rather than on demonstrated superior accuracy.

vi. Fairness auditing and external (OASIS) validation are outside the scope of this paper and are reported in the companion paper referenced in the author note above.

5.2 Conclusion

This paper demonstrates that a cross-modal attention mechanism can fuse imaging-derived, clinical, and genetic features for Alzheimer's disease prediction in a way that is interpretable and competitive with classical machine-learning baselines, while showing clearly, through a controlled modality-ablation design, that most of the achievable predictive signal in this dataset comes from clinical features alone. Fairness auditing and external generalizability of this framework are addressed in a companion paper.

Acknowledgments

Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (https://adni.loni.usc.edu/data-samples/adni-data/). As such, the investigators within the ADNI contributed to the design and implementation of ADNI And/or provided data but did not participate in the analysis or writing of this paper. A full, current list is maintained at https://adni.loni.usc.edu/data-samples/adni-data/

Data used in the external validation reported in the companion paper were obtained from the Open Access Series of Imaging Studies (OASIS) (oasis-brains.org). A full, current list is maintained at https://sites.wustl.edu/oasisbrains/home/oasis-1/

Funding

No funding for this research

Conflicts of Interest

The authors declare no conflict of interest.

Author Contributions

Conceptualization, methodology, data curation, software, formal analysis, writing and original draft, M.A.B.; A.S.A.: validation, review, editing M.A.B., A.S.A., and Y.M.M.; supervision, A.S.A., and Y.M.M.

References

Albright, J. (2019). Forecasting the progression of Alzheimer's disease using neural networks and a novel preprocessing algorithm. *Alzheimer's & Dementia: Translational Research & Clinical Interventions*, 5, 483–491. https://doi.org/10.1016/j.trci.2019.07.001

Aldoseri, A., Al-Khalifa, K. N., & Hamouda, A. M. (2023). Re-thinking data strategy and integration for artificial intelligence: Concepts, opportunities, and challenges. *Applied Sciences*, 13(12), 7082. https://doi.org/10.3390/app13127082

Alzheimer's Association. (2024). What is Alzheimer's Disease? https://alz.org

Ashendorf, L. (2019). Rey Auditory Verbal Learning Test and Rey-Osterrieth Complex Figure Test performance validity indices in a VA Polytrauma sample. *Clinical Neuropsychology*, 33(8), 1388–1402. https://doi.org/10.1080/13854046.2018.1543454

Balaji, P., Chaurasia, M. A., Bilfaqih, S. M., Muniasamy, A., & Alsid, L. E. G. (2023). Hybridized deep learning approach for detecting Alzheimer's disease. *Biomedicines*, 11(1), 149. https://doi.org/10.3390/biomedicines11010149

Bhagwat, N., Pipitone, J., Voineskos, A. N., Chakravarty, M. M., & Alzheimer's Disease Neuroimaging Initiative. (2019). An artificial neural network model for clinical score prediction in Alzheimer's disease using structural neuroimaging measures. *Journal of Psychiatry & Neuroscience*, 44(4), 246–259. https://doi.org/10.1503/jpn.180016

Bouwman, F. H., Frisoni, G. B., Johnson, S. C., Chen, X., Engelborghs, S., Ikeuchi, T., Paquet, C., Ritchie, C., Bozeat, S., Quevenco, F. C., & Teunissen, C. (2022). Clinical application of CSF biomarkers for Alzheimer's disease: From rationale to ratios. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, 14(1), e12314. https://doi.org/10.1002/dad2.12314

Dhauria, M., Mondal, R., Deb, S., Shome, G., Chowdhury, D., Sarkar, S., & Benito-León, J. (2024). Blood-based biomarkers in Alzheimer's disease: Advancing non-invasive diagnostics and prognostics. *International Journal of Molecular Sciences*, 25(20), 10911. https://doi.org/10.3390/ijms252010911

Díaz-Mardomingo, M. C., García-Herranz, S., Rodríguez, R., Venero, C., & Peraita, H. (2017). Problems in classifying mild cognitive impairment (MCI): One or multiple syndromes? *Brain Sciences*, 7, 111. https://doi.org/10.3390/brainsci7090111

El-Geneedy, M., Moustafa, H. E-D., Khalifa, F., Khater, H., & AbdElhalim, E. (2023). An MRI-based deep learning approach for accurate detection of Alzheimer's disease. *Alexandria Engineering Journal*, 63, 211–221. https://doi.org/10.1016/j.aej.2022.07.062

Felix, K., Umar, M., Guy, P., Adam, S., & Adam, M. (2025). Review of multimodal machine learning approaches in healthcare. *Information Fusion*, 114(2), 102690. https://doi.org/10.1016/j.inffus.2024.102690

Ganguli, M., Blacker, D., Blazer, D. G., Grant, I., Jeste, D. V., Paulsen, J. S., Petersen, R. C., & Sachdev, P. S. (2011). Classification of neurocognitive disorders in DSM-5: A work in progress. *American Journal of Geriatric Psychiatry*. https://doi.org/10.1097/jgp.0b013e3182051ab4

Global Burden of Disease Collaborative Network. (2020). Global Burden of Disease Study 2020 (GBD 2020) results. *Institute for Health Metrics and Evaluation (IHME)*. https://vizhub.healthdata.org/gbd-results/

Guan, T., Shang, L., Yang, P., Tan, Z., Liu, Y., Dong, C., Li, X., Hu, Z., Su, H., & Zhang, Y. (2025). Joint ensemble learning-based risk prediction of Alzheimer's disease among mild cognitive impairment patients. *The Journal of Prevention of Alzheimer's Disease*, 12(4), 100083. https://doi.org/10.1016/j.tjpad.2025.100083

Imani, M., Beikmohammadi, A., & Arabnia, H. R. (2025). Comprehensive analysis of Random Forest and XGBoost performance with SMOTE, ADASYN, and GNUS under varying imbalance levels. *Technologies*, 13(3), 88. https://doi.org/10.3390/technologies13030088

Jin, Z., Xiaohai, H., Linbo, Q., Feng, G., & Bin, W. (2022). Brain PET synthesis from MRI using a generative adversarial network for multi-modal Alzheimer's disease diagnosis. *Computer Methods and Programs in Biomedicine*, 2, 17. https://doi.org/10.1016/j.cmpb.2022.106676

Kamatham, S., Patil, P. A., Manem, M., Ahmad, H.,& Nayak, R. (2024). Pathogenesis, diagnostics, and therapeutics for Alzheimer's disease: Breaking the memory barrier. *Biomedicine & Pharmacotherapy*, 173, 114144. https://doi.org/10.1016/j.biopha.2024.114144

Kim, S., Kim, M., & Lee, J. (2024). Prognostic model for predicting Alzheimer's disease conversion using functional connectome manifolds. *Alzheimer's Research & Therapy*, 16, 217. https://doi.org/10.1186/s13195-024-01589-3

Kumar, A., Sidhu, J., & Lui, F. (2024). Alzheimer disease. *StatPearls*. https://www.ncbi.nlm.nih.gov/books/NBK499922/

Langa, K. M. (2018). Cognitive aging, dementia, and the future of an aging population. *National Academies Press (US)*. https://www.ncbi.nlm.nih.gov/books/NBK513075/

Maleki, V. S., & Forouzanfar, M. (2024). The role of AI in hospitals and clinics: Transforming healthcare in the 21st century. *Bioengineering*, 11(4), 337. https://doi.org/10.3390/bioengineering11040337

Nasreddine, Z. S., Phillips, N. A., Bédirian, V., Charbonneau, S., Whitehead, V., Collin, I., Cummings, J. L., & Chertkow, H. (2005). The Montreal Cognitive Assessment, MoCA: A brief screening tool for mild cognitive impairment. *Journal of the American Geriatrics Society*, 53(4), 695–699. https://doi.org/10.1111/j.1532-5415.2005.53221.x

Saeedi, S., Rezayi, S., Keshavarz, H. R., & Niakan, K. S. (2023). MRI-based brain tumor detection using convolutional deep learning methods and chosen machine learning techniques. *BMC Medical Informatics and Decision Making*, 23, 16. https://doi.org/10.1186/s12911-023-02114-6

World Health Organization. (2021). First WHO ministerial conference on global action against dementia: Meeting report. Geneva, *Switzerland*, 16–17 March 2015.

World Health Organization. (2023). *Global status report on the public health response to dementia*. https://www.who.int/publications/i/item/9789240033245