Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Diagnosis of Alzheimer's Disease Based on Multimodal MRI and Amyloid/Tau PET Fusion: Quantification of Modality Complementarity and Interpretability Analysis

ronghui ma
Yunnan Normal University
zhijun yang
yangzhijun@ynnu.edu.cn

Yunnan Normal University

Research Article

Keywords: Alzheimer's disease, multimodal fusion, structural MRI, amyloid PET, tau PET, machine learning, SHAP, modality complementarity

Posted Date: November 2nd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11288631/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Diagnosis of Alzheimer's Disease Based on Multimodal MRI and Amyloid/Tau PET Fusion: Quantification of Modality Complementarity and Interpretability Analysis

Ronghui Ma[1], Zhijun Yang[1*], for the Alzheimer's Disease Neuroimaging Initiative[2]

[1*]School of Information Science and Technology, Yunnan Normal University, Kunming, 650500, Yunnan, China.

[2]Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf.

*Corresponding author(s). E-mail(s): yangzhijun@ynnu.edu.cn;
Contributing authors: 2424420024@ynnu.edu.cn;

Abstract

Background: Alzheimer's disease (AD) diagnosis relies on multimodal information, yet the optimal fusion strategy and the true complementarity among modalities remain unclear. This study aimed to systematically compare different modality configurations for AD diagnosis and to quantify modality complementarity.

Methods: A total of 874 participants from the ADNI cohort (CN = 608, AD = 266) were included. A total of 39 features were extracted, comprising structural MRI (20 features), amyloid PET (5), Tau PET (10), and clinical variables (4). We employed a three-step feature selection process using variance thresholding, t-tests, and LASSO to evaluate 7 modality configurations and 5 classifiers, and

quantified modality complementarity using centered kernel alignment (CKA), Pearson correlation coefficients, and the Jaccard index.
Results: The multimodal fusion model achieved an AUC of 0.9558 in the full dataset and 0.9910 in the complete PET subset, outperforming all single-modality and dual-modality combinations. The Jaccard index for MRI and PET was 0.0000, indicating no overlap in misclassification patterns. SHAP analysis indicated that the hippocampal volume ratio (HIPPO_RATIO) was the most important feature.
Conclusions: Multimodal fusion integrating structural MRI, amyloid/Tau PET, and clinical variables provides a robust and interpretable tool for AD diagnosis. More importantly, it addresses a key technical gap in multimodal research—the lack of a quantitative framework for assessing modality complementarity—by demonstrating that MRI and PET exhibit completely non-overlapping error patterns, which constitutes the core mechanism driving the improved performance of multimodal fusion.

Keywords: Alzheimer's disease; multimodal fusion; structural MRI; amyloid PET; tau PET; machine learning; SHAP; modality complementarity

1 Introduction

Alzheimer's disease (AD) is the most common neurodegenerative disorder worldwide and one of the leading causes of cognitive decline and dementia in older adults. As the global population continues to age at an accelerating rate, the number of people with AD and related dementias is steadily increasing, making it a major public health issue worldwide and imposing a significant burden on patients' families, healthcare systems, and socioeconomic development [1, 2]. According to data from the Global Burden of Disease study, the number of people with dementia worldwide was approximately 57.4 million in 2019 and is projected to rise to about 152 million by 2050, with AD accounting for the majority of these cases [1]. In recent years, as disease-modifying treatment strategies have gradually entered the clinical application phase, accurately identifying the pathological state of AD in its early stages and establishing a reliable auxiliary diagnostic system have become key directions in current AD research [2, 3].

AD is a neurodegenerative disease with complex pathological mechanisms. Its primary pathological features include amyloid plaques formed by the abnormal deposition of extracellular amyloid-beta (Aβ) and neurofibrillary tangles formed by the aggregation of abnormally phosphorylated tau protein within cells [3]. Numerous studies have shown that the pathological process of AD involves multiple sequential stages, in which Aβ abnormalities, Tau pathology, and neurodegenerative changes exhibit distinct temporal patterns and biological significance during disease progression [4]. Aβ deposition is generally considered an early pathological event, whereas abnormal Tau aggregation and its spatial spread are more closely associated with synaptic damage, neuronal degeneration, and cognitive decline. Driven by advances in biomarker research, the concept of AD diagnosis has gradually shifted from a traditional, clinically symptom-driven model to a biologically defined model based on pathological mechanisms. The

A/T/N framework proposed by Jack et al. categorizes the core pathological processes of AD into three dimensions: A (Amyloid), T (Tau), and N (Neurodegeneration), providing an important theoretical foundation for the precise diagnosis and staging of AD [5].

In addition to pathological changes at the molecular level, AD is also accompanied by significant neurodegenerative structural changes, among which atrophy of the hippocampus, entorhinal cortex, and temporal lobe-related regions is the most typical radiological manifestation [6]. Structural magnetic resonance imaging (MRI) enables non-invasive assessment of structural changes in brain tissue and reflects AD-related neurodegenerative processes by quantifying indicators such as hippocampal volume, cortical thickness, and regional atrophy patterns; as a result, it has become one of the most widely used imaging tools in both clinical evaluation and scientific research on AD [6, 7]. In recent years, the development of automated brain structural analysis methods, such as FreeSurfer, has further improved the standardization of quantitative MRI analysis, enabling support for large-scale, multicenter AD studies [7].

However, MRI and PET imaging reflect different biological dimensions of the AD pathological process. Structural MRI primarily captures neurodegenerative changes and cumulative structural brain damage but cannot directly measure molecular pathological burdens such as Aβ or Tau; in contrast, amyloid PET and Tau PET can provide more direct molecular pathological information [4]. However, the biological characteristics of different modalities do not equate to their predictive capabilities in specific diagnostic tasks. Previous studies have shown that the diagnostic performance of imaging modalities is influenced by various factors, including disease stage, study cohort composition, sample size, image quality, and feature extraction methods. Therefore, the fact that a particular modality has an advantage in terms of pathological specificity does not imply that it will exhibit higher predictive performance in all clinical classification tasks. For example, although amyloid PET directly reflects Aβ pathology, its diagnostic performance when used alone in clinical classification tasks may be limited by the complex relationship between pathology and symptoms. Therefore, systematically evaluating the independent diagnostic contributions of different modalities and exploring their complementary relationships is crucial for establishing reliable multimodal diagnostic strategies for AD.

In addition to imaging information, clinical and genetic factors such as age, the APOE $\varepsilon 4$ genotype, and years of education can also provide important personalized risk information. Among these, APOE $\varepsilon 4$ is currently recognized as one of the major genetic risk factors for AD, while factors such as educational level can reflect an individual's cognitive reserve [8]. However, clinical variables primarily reflect an individual's genetic background and risk level and cannot directly reveal intracerebral pathological states; therefore, relying solely on clinical information for AD diagnosis still has significant limitations [8].

In recent years, machine learning methods have been widely applied in AD diagnosis. Various classifiers based on structural MRI, PET, and clinical variables—including logistic regression, support vector machines, random forests, and gradient-boosted trees—have been used to distinguish between individuals with AD and those with normal cognition [9, 10]. However, existing studies have largely focused on specific

modality combinations, and there is a lack of systematic comparisons across multiple modality configurations; the independent contributions of different modalities and optimal fusion strategies remain unclear. Furthermore, as machine learning models are increasingly applied in clinical decision support, model interpretability has become critical for clinical translation. Clinicians need to understand the rationale behind a model's decisions in order to trust and adopt AI-assisted diagnostic results [11]. In addition to discriminative ability, a model's clinical utility must be evaluated using methods such as calibration curve and decision curve analysis to quantify the net benefit of the model at different clinical thresholds [12].

Since the pathological process of AD involves molecular pathology, neurodegenerative changes, and individual risk factors, different diagnostic modalities actually reflect different dimensions of the disease process; no single modality can fully capture the complex pathological characteristics of AD [5]. For example, PET provides molecular pathological information related to Aβ and tau, while MRI reflects neurodegenerative structural changes, and clinical variables supplement information on individual risk and cognitive reserve. Therefore, multimodal fusion of information from different sources is expected to provide a more comprehensive description of the pathological state of AD and improve the stability and accuracy of auxiliary diagnostic models. However, systematic research is still lacking regarding the true complementary relationships among different modalities, the independent contributions of each modality to the final diagnostic outcome, and the optimal fusion strategy. More critically, existing multimodal studies have only indirectly inferred modality complementarity from final performance gains, without ever quantitatively verifying whether modalities actually complement each other. For example, do MRI and PET make errors on the same samples? Do their error patterns overlap? The answers to these questions directly determine the theoretical upper bound and clinical value of multimodal fusion, yet current research cannot answer them. This constitutes a clear technical need: establishing a reusable quantitative framework for modality complementarity. Based on this, this study systematically compares seven modal configurations on the ADNI multicenter cohort and introduces centered kernel alignment, Pearson correlation, and the Jaccard index to quantify modal complementarity, in order to address the aforementioned questions.

2 Materials and Methods

2.1 Study Population

All data analyzed in this study were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) public database (adni.loni.usc.edu). Data used in this study were downloaded from the ADNI database on April 10, 2026.

To focus on the binary diagnostic objective of distinguishing between normal cognition and AD dementia, this study included only participants diagnosed as cognitively normal (CN) or with AD dementia at the baseline visit [13], systematically excluding participants in the intermediate state of mild cognitive impairment (MCI). The diagnostic labels for all participants were extracted from the DXSUM dataset of the baseline visit (VISCODE = bl). Invalid samples with missing diagnostic labels were

strictly excluded during the study, and duplicate participant data were removed. Ultimately, this study included a total of 874 eligible participants, comprising 608 in the CN group and 266 in the AD group.

Based on the study design and data integrity requirements, this study established clear inclusion and exclusion criteria, as follows:

Inclusion Criteria:

1. Diagnosed with CN or AD dementia at the baseline visit, with a clear and traceable diagnosis;
2. Possession of complete, usable baseline structural MRI image data that meets the requirements for subsequent feature extraction;
3. Complete baseline clinical variable data, including age, sex, years of education, and APOE genotype as core indicators.

Exclusion Criteria:

1. Participants diagnosed with MCI at baseline, or those with missing diagnostic labels or who cannot be clearly categorized;
2. MRI image data quality does not meet standards, or segmentation using FreeSurfer software fails, making it impossible to extract valid features;
3. Key clinical variables such as age and genotype are missing, making subsequent modeling and analysis impossible.

All ADNI participants have signed written informed consent forms, and the overall ADNI study protocol has been fully approved by the institutional review boards (IRBs) of all participating study sites. Since this study relies solely on publicly available, de-identified, and anonymized data from the ADNI database for secondary analysis—and does not involve direct contact with human subjects or pose a risk of privacy breaches—no additional ethical filing or approval processes are required [14].

2.2 Data Modalities and Feature Extraction

Building on the multidimensional pathological characteristics of AD and the core concept of multimodal fusion diagnosis described earlier, this study integrates multidimensional diagnostic indicators to systematically extract four categories of core research features: structural MRI image features, amyloid PET features, tau PET features, and clinical demographic and genetic variables. The complete technical workflow for data screening, feature extraction, and modeling analysis in this study is shown in Figure 1.

2.2.1 Structural MRI

All quantitative metrics related to brain region volumes in this study's structural MRI were extracted from the UCSFFSX7 (FreeSurfer cross-sectional) dataset [15]. As the most widely recognized and extensively used automated brain segmentation tool in the field of neuroimaging, FreeSurfer's algorithmic stability and accuracy have been thoroughly validated in large-scale, multicenter cohort studies such as ADNI, ensuring the standardization and reliability of image feature extraction in this study [7, 16, 17].

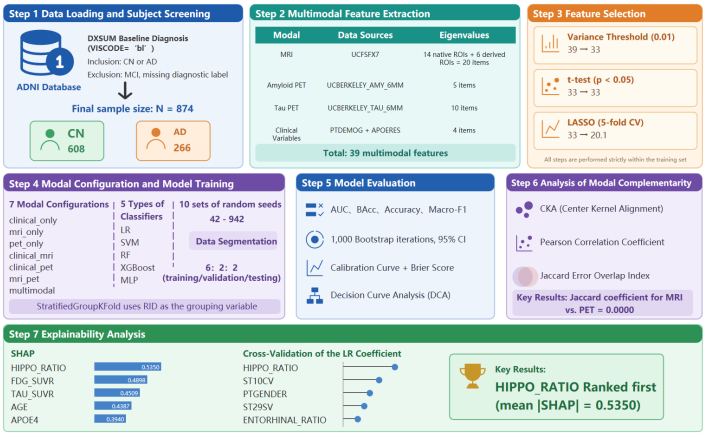


Fig. 1 Roadmap for Multimodal AD Diagnosis

To avoid redundancy in whole-brain features and focus on AD-specific pathological regions, this study drew upon classical a priori knowledge of AD pathophysiology and the pathological characteristics of the Braak staging system to select 14 regions of interest (ROIs) from the 119 ROIs in the UCSFFSX7 dataset that are highly correlated with AD onset and progression. The selection process followed three principles: First, to cover the core regions affected by AD pathology, including the entorhinal cortex and hippocampus in Braak stages I–II, the parahippocampal gyrus and middle temporal gyrus in Braak stages III–IV, the posterior cingulate in Braak stages V–VI, as well as the subcortical amygdala and thalamus; Second, coverage of different pathological mechanisms, including direct indicators of atrophy (hippocampus, entorhinal cortex, etc.), indirect indicators of atrophy (lateral ventricles), subcortical indicators (amygdala, thalamus), and normalization indicators (ICV); Third, these regions are aligned with the ADNI standard ROIs, which have been widely used in multiple ADNI studies and demonstrate good reproducibility. The rationale for selecting each ROI and its pathological significance are as follows [13, 18]:

- Hippocampus (ST29SV, ST88SV): Hippocampal atrophy is the earliest and most significant radiological biomarker to appear during the course of AD and is highly correlated with progressive memory impairment in patients;
- Entorhinal cortex (ST24CV, ST83CV): As the cortical region first affected by AD pathology in the Braak staging system, atrophy of the entorhinal cortex occurs earlier than hippocampal damage and serves as a key indicator for the early identification of AD;
- Lateral ventricles (ST37SV, ST96SV): Abnormal enlargement of the lateral ventricles serves as an indirect quantitative marker of intracerebral neuronal degeneration and brain atrophy;

- Amygdala (ST71SV): Pathological damage to the amygdala is closely associated with non-cognitive symptoms in AD patients, such as emotional disturbances and behavioral disorders;
- Thalamus (ST61SV): The thalamus is a key subcortical core structure frequently affected by AD pathology, and its morphological abnormalities can reflect the extent of central nervous system degenerative damage;
- Medial temporal gyrus (ST40CV): Atrophy of the medial temporal gyrus is directly associated with the semantic memory impairments typical of AD patients;
- Temporal pole (ST60CV): The temporal pole is one of the core brain regions affected by early pathological damage in AD, and morphological changes may indicate early-stage pathology;
- Posterior Cingulate (ST50CV): As a core node of the brain's default mode network, metabolic decline in the posterior cingulate precedes structural atrophy, making it an important marker of early pathological changes in AD;
- Gyrus fusiformis (ST26CV): The gyrus fusiformis is primarily involved in face recognition and semantic processing; its damage can lead to cognitive processing abnormalities in AD patients;
- Parahippocampal gyrus (ST44CV): The parahippocampal gyrus serves as a key functional interface connecting the hippocampus to the cerebral cortex and is crucial for memory transmission pathways; damage to this region directly affects cognitive function;
- Intracranial Volume (ST10CV): As a proxy correction variable for head circumference, it is used to standardize regional brain volume data and eliminate confounding effects from individual cranial anatomical differences [15].

To further mitigate the interference caused by individual cranial anatomical differences and extract more stable and specific AD pathological imaging signals, this study derived and constructed six composite standardized features based on the aforementioned raw ROI features. The design of these derived features was based on three objectives: First, to merge the left and right hemispheres (e.g., HIPPOCAMPUS = ST29SV + ST88SV) to reduce measurement noise in unilateral ROIs; Second, to normalize by dividing by ICV (e.g., HIPPO_RATIO = HIPPOCAMPUS / ST10CV) to eliminate individual differences in head circumference; third, to reduce multicollinearity (ST29SV and ST88SV are highly correlated; merging them avoids the impact of multicollinearity on the stability of LR coefficients). The specific calculation methods and clinical significance of the derived features are shown in Table 1.

In summary, the structural MRI module in this study ultimately included 20 valid features: 14 raw ROI features + 6 derived composite features.

2.2.2 Amyloid PET

To accurately quantify the burden of Aβ plaque deposition in the brain—a core molecular pathological feature of AD—this study conducted a quantitative analysis of amyloid PET images using the UCBERKELEY_AMY_6MM dataset. This dataset was constructed by a team at the University of California, Berkeley, and relies on FreeSurfer-standardized brain region segmentation templates to perform quantitative

Table 1 Calculation methods and clinical significance of the six derived composite MRI features.

Derived Feature	Calculation Method	Clinical Significance
HIPPOCAMPUS	ST29SV + ST88SV	Total hippocampal volume
HIPPO_RATIO	HIPPOCAMPUS / ST10CV	Normalized hippocampal volume, adjusted for head circumference
ENTORHINAL	ST24CV + ST83CV	Total entorhinal cortex volume
ENTORHINAL_RATIO	ENTORHINAL / ST10CV	Normalized entorhinal cortex volume
VENTRICLES	ST37SV + ST96SV	Total Lateral Ventricle Volume
VENTRICLES_RATIO	VENTRICLES / ST10CV	Normalized Lateral Ventricle Volume

analysis of PET images; it features high accuracy and a high degree of standardization, making it well-suited for quantitative studies of AD pathology [19].

Addressing the widespread issue of temporal leakage in PET data found in existing studies, as mentioned in the introduction, this study strictly enforced a data temporal alignment quality control strategy to eliminate bias from overestimating model performance at its source. All amyloid PET scan data were strictly limited to baseline visit (VISCODE = bl) data and sorted by scan date (SCANDATE), retaining only the first baseline scan for each subject to ensure complete temporal alignment between PET pathology data and baseline clinical diagnostic results [20, 21]. Following rigorous baseline screening, 772 valid data points were retained from the original 4,582 PET records.

Based on the diffuse characteristics of AD amyloid pathology and the ADNI standardized quantification framework, this study extracted five representative features from the UCBERKELEY_AMY_6MM dataset, covering global cortical Aβ deposition, cross-tracer standardized units, composite reference region SUVR, and two commonly used reference regions. Among these, AMY_SUVR (SUMMARY_SUVR) reflects the average Aβ deposition level across the entire cerebral cortex and is the most commonly used global quantification metric; CENTILOIDS are Aβ quantification units standardized across tracers and centers, which eliminate scale differences between different tracers; AMY_COMPOSITE (COMPOSITE_REF_SUVR) is calculated from a composite ROI comprising multiple AD-related cortical regions and is more robust than a single region; AMY_CEREBELLUM (CEREBELLUM_CORTEX_SUVR) and AMY_BRAINSTEM (BRAINSTEM_SUVR) use the cerebellar cortex and brainstem, respectively, as reference regions to reflect nonspecific binding levels; they are included as control features [22, 23]. Specific feature information is shown in Table 2.

2.2.3 Tau PET

Tau neurofibrillary tangles are a core pathological factor mediating neuronal damage and cognitive decline in Alzheimer's disease (AD). To comprehensively cover the core pathological dimension of the A/T/N diagnostic framework, this study concurrently conducted feature extraction from Tau PET scans. Quantification of Tau protein burden was performed using the UCBERKELEY_TAU_6MM dataset, applying baseline data screening rules identical to those used for amyloid PET to strictly avoid time

Table 2 Five amyloid PET features and their meanings.

Feature	Original Column Name	Meaning
AMY_SUVR	SUMMARY_SUVR	Global Cortical SUVR
CENTILOIDS	CENTILOIDS	Centiloid Standardized Units
AMY_COMPOSITE	COMPOSITE_REF_SUVR	Composite Reference Volume (SUVR)
AMY_CEREBELLUM	CEREBELLUM_CORTEX_SUVR	Cerebellar Cortex SUVR
AMY_BRAINSTEM	BRAINSTEM_SUVR	Brainstem SUVR

leakage issues. After baseline quality control screening of the original 2,327 Tau PET scan records, 617 valid baseline data points were ultimately retained.

Combining the classic Braak pathological staging criteria for AD with the patterns of tau protein infiltration in specific brain regions, this study selected 10 AD-specific cortical regions from the UCBERKELEY_TAU_6MM dataset to extract tau PET features. The selection was based on the Braak staging patterns of tau pathology: early-stage regions (Braak I–II) include the entorhinal cortex (TAU_ENTORHINAL), which is the region where tau pathology first appears; neuropathological studies have confirmed that tau PET signals in the entorhinal cortex correlate highly with the severity of postmortem NFTs in the early Braak stages; intermediate regions (Braak III–IV) include the parahippocampal gyrus (TAU_PARAHIPPOCAMPAL), fusiform gyrus (TAU_FUSIFORM), inferior temporal gyrus (TAU_INFERIOR_TEMPORAL), the middle temporal gyrus (TAU_MIDDLE_TEMPORAL), the superior temporal gyrus (TAU_SUPERIOR_TEMPORAL), and the Banks's sulcus (TAU_BANKSSTS). These are the typical regions where Tau pathology spreads outward from the medial temporal lobe; notably, Tau deposition in the fusiform gyrus is significantly correlated with cerebrospinal fluid p-Tau levels; Late-stage regions (Braak V–VI) include the precuneus (TAU_PRECUNEUS) and posterior cingulate (TAU_POSTERIOR_CINGULATE), which are hallmarks of widespread cortical diffusion of Tau pathology; changes in Tau signal in the precuneus can be observed even before the onset of symptoms in autosomal dominant AD. In addition, the meta-temporal ROI (TAU_META_TEMPORAL) is a composite ROI comprising the entorhinal cortex, fusiform gyrus, parahippocampal gyrus, and inferior temporal gyrus; it serves as a standard tool in the UC Berkeley pipeline for detecting early Tau pathology. Specific features and their corresponding Braak stages are shown in Table 3 [13, 18].

2.2.4 Clinical Variables

To compensate for the lack of information on individual genetic risk and cognitive reserve that a single imaging modality cannot capture, and to refine the multidimensional diagnostic feature system, this study included four core clinical and genetic variables covering the four core dimensions of AD risk: non-genetic risk (age), demographic characteristics (gender), cognitive reserve (years of education), and genetic risk (APOE $\varepsilon 4$). Among these, age is the strongest non-genetic risk factor for AD, with prevalence increasing exponentially with age; gender is associated with AD prevalence

Table 3 10 Tau PET features and their corresponding Braak stages.

Feature	Original Column Name	Braak Stage
TAU_META_TEMPORAL	META_TEMPORAL_SUVR	Meta-temporal ROI
TAU_ENTORHINAL	CTX_ENTORHINAL_SUVR	I-II
TAU_FUSIFORM	CTX_FUSIFORM_SUVR	III–IV
TAU_INFERIOR_TEMPORAL	CTX_INFERIORTEMPORAL_SUVR	III–IV
TAU_MIDDLE_TEMPORAL	CTX_MIDDLE_TEMPORAL_SUVR	III–IV
TAU_PARAHIPPOCAMPAL	CTX_PARAHIPPOCAMPAL_SUVR	III–IV
TAU_SUPERIOR_TEMPORAL	CTX_SUPERIORTEMPORAL_SUVR	III–IV
TAU_BANKSSTS	CTX_BANKSSTS_SUVR	III–IV
TAU_PRECUNEUS	CTX_PRECUNEUS_SUVR	V–VI
TAU_POSTERIOR_CINGULATE	CTX_POSTERIOR_CINGULATE_SUVR	V–VI

and intracranial volume; years of education is a classic proxy for cognitive reserve, reflecting an individual's ability to compensate for pathological burden; and APOE $\varepsilon 4$ is the strongest genetic risk factor for AD, with carriers facing a 3- to 15-fold increased risk. These four variables are the core clinical covariates officially recommended by the ADNI and are widely used in multimodal AD studies [5, 18]. Specific variable information, data sources, and type descriptions are provided in Table 4.

Table 4 Sources, types, and descriptions of the four clinical and genetic variables.

Variable	Source	Type	Description
AGE	PTDOB + EXAMDATE	Continuous	Baseline Age
PTGENDER	PTDEMOG	Categorical	Male = 1, Female = 0
PTEDUCAT	PTDEMOG	Continuous	Years of Education
APOE4	APOERES	Categorical	Carrying at least one $\varepsilon 4$ allele = 1

The participants' baseline age was calculated based on their date of birth and the date of the baseline examination. A reasonable age threshold was set to filter out anomalous and invalid data for individuals under 40 or over 110 years of age, ensuring the validity of the sample. APOE $\varepsilon 4$ genotype status was extracted from the APOERES dataset and binary-coded; subjects carrying at least one $\varepsilon 4$ allele were defined as positive (assigned a value of 1), while non-carriers were defined as negative (assigned a value of 0), thereby quantifying genetic susceptibility to AD.

2.3 Modality Combinations

To address the shortcomings of existing studies—namely, the limited variety of modality combinations and the lack of systematic cross-sectional comparisons—and to comprehensively quantify the independent diagnostic value and synergistic effects of different modalities and combination strategies [24, 25], this study constructed seven differentiated modality configuration schemes based on the four feature

categories mentioned above—clinical, MRI, amyloid PET, and tau PET—covering single-modality, dual-modality fusion, and full-dimensional multi-modality fusion scenarios to achieve comprehensive coverage and systematic comparison of modality combinations. The number and composition of features for each configuration scheme are shown in Table 5.

Table 5 Number and Composition of Features in the 7 Modality Configuration Schemes.

Configuration	Number of Features	Description
clinical_only	4	Clinical variables only
mri_only	20	MRI only
pet_only	15	PET Only (5 Amyloid + 10 Tau)
clinical_mri	24	Clinical + MRI
clinical_pet	19	Clinical + PET
mri_pet	35	MRI + PET
Multimodal	39	Clinical + MRI + PET

2.4 Feature Selection

After multimodal feature integration, the dataset contained a total of 39 features. Although the ratio of features to sample size (N = 874) was approximately 22:1—which does not constitute a severely high-dimensional scenario—there was some multicollinearity among the features (e.g., ST29SV was highly correlated with HIPPOCAMPUS), which could lead to unstable LR coefficients and unreliable SHAP interpretations. To reduce model dimensionality, minimize the risk of overfitting, identify the optimal subset of diagnostic features, and strictly avoid information leakage during the modeling process, this study established a three-step iterative feature selection workflow. All feature selection operations were conducted independently on the training set following each cross-validation split, without incorporating any information from the validation or test sets, fully adhering to the standardized methodological requirements for clinical modeling in bioinformatics and machine learning [26, 27].

The design logic of the three-step screening process is as follows:

1. **Variance threshold screening.** A variance threshold of 0.01 was set to exclude features with variance below this threshold. This step removes constant features with minimal intra-group variation, low information discriminability, and no contribution to disease diagnosis, thereby initially reducing the feature dimension.
2. **Univariate t-test screening.** The features selected in the preliminary screening are standardized. An independent-samples t-test is used to compare differences in feature distributions between the AD group and the CN group. Features with no statistically significant differences between groups (p-value ≥ 0.05) are excluded, retaining core features significantly associated with the AD diagnostic label.

3. **LASSO regression screening.** Based on the valid features identified in the first two steps, a Least Absolute Shrinkage and Selection Operator (LASSO) regression model was applied, and the optimal regularization parameter was determined through 5-fold cross-validation. Relying on the L1 regularization penalty mechanism, coefficients of invalid features were compressed, and coefficients of features with no diagnostic value were set to zero, thereby achieving explicit and precise feature selection. Ultimately, high-contribution features with non-zero coefficients were retained.

To verify the robustness and optimality of the three-step feature selection process and to eliminate selection bias in the feature selection scheme, this study set up five mainstream alternative feature selection strategies for a side-by-side comparison: None (no feature selection; all 39 features are included in the model), Variance (variance threshold only), T-test (t-test only), Variance+T-test (combined variance threshold and t-test), LASSO (LASSO regression only).

2.5 Model Training and Evaluation

To comprehensively compare the adaptability and performance differences of various machine learning algorithms in multimodal AD diagnosis and to identify the optimal diagnostic model, this study selected five mainstream machine learning classifiers suitable for clinical imaging data for model training. These five models cover the five most commonly used algorithmic paradigms in AD diagnosis: linear models (LR), kernel methods (SVM), bagging ensembles (RF), boosting ensembles (XGBoost), and neural networks (MLP), facilitating a systematic comparison of the suitability of different algorithmic paradigms. The core parameters and technical characteristics of each model are shown in Table 6.

Table 6 Key Parameters and Descriptions of the Five Machine Learning Classifiers

Model	Key Parameters	Description
LR	C=0.1, class_weight=balanced, max_iter=2000	L2-regularized logistic regression, a classic linear model with high interpretability
SVM	kernel=rbf, C=1.0, gamma=scale, probability=True	Radial Basis Function (RBF) Support Vector Machine, suitable for fitting nonlinear feature relationships
RF	n_estimators=300, max_depth=5, min_samples_split=5	Random Forest ensemble model, resistant to overfitting and highly robust
XGBoost	n_estimators=300, max_depth=2, learning_rate=0.03	Gradient Boosting model, with strong regularization capabilities and high fitting accuracy
MLP	hidden_layer_sizes=(64, 32), alpha=0.01, early_stopping=True	Two-layer perceptron, effectively preventing overfitting through an early stopping mechanism

To avoid random bias in a single data split and to improve the robustness and reproducibility of model evaluation results, this study employed a multi-sub-replication

holdout validation scheme for data splitting. The study set up 10 groups with different random seeds (42, 142, 242. . . 942). Using the StratifiedGroupKFold stratified grouping sampling method and based on the subject's unique identification number (RID) for grouping, the study ensured that data from the same subject belonged to only a single dataset, thereby completely avoiding the risk of cross-leakage between samples [20, 21]. For each seed, the samples were divided into training, validation, and test sets in a 6:2:2 ratio. This ratio ensures an adequate training set while reserving 20% of the samples for both the validation and test sets, thereby accommodating the sample size (N = 874) and class distribution of this study. The training set is used for model fitting and feature selection, the validation set for selecting the optimal classification threshold, and the test set for the final evaluation of model performance.

To address the small number of missing values in the data, this study adopted a standardized missing value handling process: the median of each feature was calculated based on the training set data, and missing values were imputed using SimpleImputer. The imputer was trained solely on the training set, and the results were then uniformly applied to both the validation and test sets to prevent information leakage from the validation and test sets.

Instead of using the default classification threshold of 0.5, the optimal threshold was selected by iteratively scanning the 0–1 interval in 0.01 increments on the validation set, with the objective of maximizing balanced accuracy, to accommodate the class-imbalanced sample distribution in this study.

To comprehensively and multidimensionally evaluate the model's diagnostic performance, this study selected four core evaluation metrics: AUC, balanced accuracy (BAcc), accuracy, and macro-F1 score. These metrics measure the model's overall discriminative ability, its ability to adapt to class imbalance, its overall recognition accuracy, and its comprehensive classification performance, respectively. All metrics are reported as the mean ± standard deviation across 10 sets of random seeds. Additionally, a bootstrap resampling method with 1,000 repetitions is employed to calculate the 95% confidence intervals for AUC and BAcc, thereby quantifying the stability and reliability of the evaluation results.

2.6 Calibration and Decision Curve Analysis

Relying solely on classification performance metrics cannot fully reflect the model's clinical value; therefore, this study further conducted model calibration analysis and decision curve analysis (DCA) curve analysis from two dimensions—the reliability of predicted probabilities and clinical net benefit—to refine the evaluation system for the model's clinical utility.

Calibration curves are used to assess the degree of alignment between the model's predicted probabilities and actual clinical outcomes and serve as a core metric for evaluating the reliability of model predictions. In this study, the predicted probabilities were divided into 10 equal intervals to plot the calibration curves; this 10-bin configuration is a common setting for calibration curves, striking a balance between curve smoothness and resolution. Additionally, the Brier score was calculated to quantify calibration performance; a lower Brier score indicates a smaller deviation between the

model's predicted probability and the true positive probability, and thus better calibration performance [26, 27]. Ideally, the calibration curve should largely coincide with the diagonal line, indicating that the model's predictions are accurate and reliable.

Decision curve analysis (DCA) is used to quantify the actual net clinical benefit of a model at different clinical diagnostic thresholds, providing an intuitive assessment of the model's clinical value and applicable scenarios. In this study, DCA analysis was performed across the entire threshold probability range from 0.01 to 0.99. The formula for calculating net benefit (NB) is as follows [12]:

$$NB = \frac{TP}{N} - \frac{FP}{N} \times \frac{p_t}{1 - p_t} \tag{1}$$

where TP is the number of true-positive samples, FP is the number of false-positive samples, N is the total sample size, and p_t is the preset threshold probability. In this study, the net benefit of each model was compared with two extreme clinical strategies (treat all or treat none). The wider the threshold range in which the model's net benefit outperforms the extreme strategies, the broader its clinical applicability and the higher its practical value.

2.7 Analysis of Modality Complementarity

To elucidate the core mechanisms underlying the performance improvements of multimodal fusion models and address the research gap where existing studies have been unable to quantify modality complementarity, this study utilized a PET subset cohort (N = 383) with complete amyloid PET and Tau PET data as the test set. We introduced three categories of quantitative metrics to systematically evaluate the degree of complementarity among different diagnostic modalities across three dimensions: feature similarity, predictive independence, and overlap of erroneous samples. These three dimensions progressively delve deeper from the feature level (CKA), through the prediction level (Pearson correlation coefficient), to the error level (Jaccard index), forming a comprehensive complementary assessment system spanning representation, decision-making, and error, thereby clarifying the principles underlying the benefits of multimodal fusion.

1. **Centered Kernel Alignment (CKA).** The CKA metric is used to quantify the similarity between the feature representation matrices of two modalities and intuitively reflects the information overlap rate between features of different modalities. Given two sets of feature representation matrices X and Y, the CKA formula is as follows:
$$CKA(X, Y) = \frac{HSIC(K, L)}{\sqrt{HSIC(K, K) \cdot HSIC(L, L)}} \tag{2}$$
where $K = XX^T$, $L = YY^T$, and HSIC denotes the Hilbert-Schmidt Independence Criterion [28, 29]. CKA takes values between 0 and 1; a lower value indicates lower similarity between the two sets of modal feature information and stronger information complementarity.
2. **Pearson Correlation Coefficient.** The independence of prediction results across different modalities is quantified by calculating the Pearson correlation coefficients

of the predicted probabilities from each single-modality model. A lower correlation coefficient indicates that the prediction biases of the two modalities are more independent of each other, suggesting greater potential for cross-modality complementarity.

3. **Jaccard Index.** The Jaccard index is used to precisely quantify the degree of overlap in misclassified samples across different modality models, directly verifying the complementary nature of errors between modalities. The calculation formula is as follows:

$$J = \frac{|E_1 \cap E_2|}{|E_1 \cup E_2|} \tag{3}$$

where E_1 and E_2 represent the sets of misclassified samples from the two modal models, respectively. A lower Jaccard index indicates less overlap in misclassified samples between the two models, meaning they can correct each other's diagnostic errors on different samples, resulting in significantly stronger modal complementarity.

2.8 Interpretability Analysis

Addressing the issues of insufficient interpretability in existing AI-based AD diagnostic models and the potential bias in the SHAP method regarding collinear features—which has not been adequately addressed—this study combines multiple interpretability analysis methods to systematically deconstruct the decision logic of multimodal models. Simultaneously, it validates the interpretability bias caused by feature collinearity, thereby enhancing the clinical reliability and methodological rigor of the models. When features are collinear, SHAP "distributes" the contribution of the collinear features across multiple related features, leading to an overestimation of the importance of certain features—an issue rarely discussed in existing research.

SHAP Analysis: This study utilized the SHapley Additive exPlanations (SHAP) method to analyze the interpretability of a linear regression (LR) model trained on a multimodal optimal feature set. Based on the Shapley value principle in game theory, SHAP assigns a quantified contribution value to each feature, precisely quantifying the positive or negative contribution of each feature to the model's diagnostic results for a single sample. The study calculated SHAP values on the test set using the LinearExplainer tool and generated two types of core visualizations: feature importance bar charts based on the mean absolute SHAP value, and swarm plots illustrating the relationship between feature values and SHAP value distributions.

Additionally, this study introduced LR model coefficients as a supplementary validation metric for feature importance. By cross-referencing SHAP values with regression coefficients, the study circumvented the limitations of a single explanatory method, making the feature contribution analysis results more objective and reliable.

2.9 Statistical Analysis

All data preprocessing, feature extraction, model training, performance evaluation, and visualization in this study were performed using Python 3.10, relying primarily on specialized data analysis toolkits such as scikit-learn 1.3, XGBoost, and SHAP.

The statistical methods used to assess between-group differences in baseline data for the two groups (CN group and AD group) are as follows: for continuous variables, a Welch-corrected independent two-sample t-test was used, with results presented as "mean ± standard deviation"; for categorical variables, a chi-square test was used, with results presented as frequencies and percentages. In this study, a two-sided p-value ¡ 0.05 was uniformly set as the threshold for statistical significance.

3 Results

Based on the subject selection criteria, multimodal feature extraction scheme, modal combination settings, feature selection process, and model training and evaluation methods described above, this study completed a comparative analysis of subjects with Alzheimer's disease (AD) and those with normal cognition (CN), constructed and validated a multimodal diagnostic model, and investigated the mechanisms of modal complementarity and model interpretability. It systematically elucidates the diagnostic advantages and underlying principles of the multimodal fusion model. The specific research results are as follows.

3.1 Baseline Characteristics

This study strictly adhered to predefined inclusion and exclusion criteria for subject selection. A total of 874 subjects were ultimately included, comprising 608 in the CN group (69.6%) and 266 in the AD group (30.4%). To clarify intergroup differences in baseline demographic, clinical, and imaging characteristics between the two groups and to verify that the core features selected in this study are specific to AD, this study conducted statistical comparisons between groups for all key variables. The detailed baseline characteristics of the two groups and the results of intergroup differences are shown in Table 7.

Demographic Characteristics: Baseline demographic statistics revealed significant statistical differences between the two groups in terms of age, sex, years of education, and AD susceptibility gene carrier status. Specifically, the mean age in the AD group was 75.06 ± 7.78 years, which was significantly higher than that of the CN group (72.03 ± 6.91 years). Regarding gender distribution, the AD group consisted of 147 males (55.3%) and 119 females (44.7%), while the CN group consisted of 254 males (41.8%) and 354 females (58.2%); the difference in gender distribution between the two groups was statistically significant. In terms of years of education, the AD group had 14.95 ± 2.98 years, which was significantly lower than the CN group's 16.48 ± 2.49 years. Regarding genetic characteristics, the APOE $\varepsilon 4$ carrier rate in the AD group was 0.67 ± 0.47, significantly higher than the CN group's 0.32 ± 0.47, consistent with the clinical understanding that APOE $\varepsilon 4$ is a high-risk genetic factor for AD.

MRI Characteristics: Based on the core features of AD-specific regions of interest (ROIs) screened in the previous section, a comparison of key structural imaging indicators between the two groups revealed that the AD group exhibited typical features of brain atrophy. The total hippocampal volume in the AD group was 5,930.81 ± 1,017.65 mm^3, significantly lower than that of the CN group (7,558.09 ± 874.53 mm^3); the hippocampal volume ratio, adjusted for intracranial volume, was 0.0039 ±

0.0006, significantly lower than that of the CN group (0.0051 ± 0.0007). Additionally, in the entorhinal cortex—a core brain region affected in the early stages of AD—the volume in the AD group was 2,844.23 ± 794.93 mm^3, significantly lower than the 3,690.23 ± 679.11 mm^3 in the CN group, confirming the characteristic degenerative changes in this brain region in AD patients from a structural imaging perspective.

PET Characteristics: Regarding the two core molecular pathological features of AD—Aβ plaques and tau neurofibrillary tangles—the comparison of PET quantitative indices between the two groups revealed significant differences in pathological burden. The AD group's Amyloid SUVR was 1.46 ± 0.30, significantly higher than the CN group's 1.08 ± 0.17; the corresponding Centiloids-standardized score was 81.70 ± 49.73, far higher than the CN group's 17.99 ± 29.87, suggesting significant amyloid deposition in the brains of AD patients. At the tau pathology level, the tau SUVR in the meta-temporal ROI of the AD group was 1.66 ± 0.46, significantly higher than the 1.19 ± 0.12 in the CN group, confirming the presence of marked tau protein abnormal accumulation in characteristic brain regions of AD patients.

PET Data Integrity: Since PET imaging is not a routine baseline screening modality, the dataset in this study contained some missing samples. Statistical results showed that the effective sample size for each amyloid PET feature was approximately 47.0% (411/874), while that for each tau PET feature was approximately 45.1% (394/874). To precisely investigate the diagnostic value and modality complementarity of PET modalities in subsequent analyses, this study extracted samples with complete baseline PET data for both amyloid and tau, ultimately constructing a PET subcohort comprising 383 participants for subsequent stratified comparative analysis.

Table 7 Baseline Demographic, Clinical, and Imaging Characteristics of the Study Population

variable	AD	CN	Statistic	p
AGE	75.06 ± 7.78	72.03 ± 6.91	t = 5.470	¡0.001
PTGENDER	{0.0: 119, 1.0: 147}	{0.0: 354, 1.0: 254}	χ^2 = 13.018	¡0.001
PTEDUCAT	14.95 ± 2.98	16.48 ± 2.49	t = -7.302	¡0.001
APOE4	0.67 ± 0.47	0.32 ± 0.47	t = 9.960	¡0.001
HIPPOCAMPUS	5930.81 ± 1017.65	7558.09 ± 874.53	t = -22.530	¡0.001
HIPPO_RATIO	0.0039 ± 0.0006	0.0051 ± 0.0007	t = -25.992	¡0.001
ENTORHINAL	2844.23 ± 794.93	3690.23 ± 679.11	t = -15.017	¡0.001
AMY_SUVR	1.46 ± 0.30	1.08 ± 0.17	t = 9.708	¡0.001
CENTILOIDS	81.70 ± 49.73	17.99 ± 29.87	t = 9.911	¡0.001
TAU_META_TEMPORAL	1.66 ± 0.46	1.19 ± 0.12	t = 8.028	¡0.001

3.2 Feature Selection Results

To address the issue of redundant noise in multimodal feature sets, identify a subset of highly specific AD diagnostic features, and simultaneously mitigate the risk of information leakage during the modeling process, this study strictly followed the three-step progressive feature selection process outlined in Section 2.4 (variance-threshold screening + t-test screening + LASSO regression screening) to perform layer-by-layer

screening of the 39 integrated full-dimensional multimodal features. Multiple control schemes were simultaneously established to validate the optimality of the screening strategy; the average screening results based on cross-validation with 10 random seeds are shown in Table 8.

Table 8 Stepwise screening results of the three-step feature selection process (average of 10 seeds).

Step	Method	Number of Retained Features
Initial	—	39.0
Step 1	Variance Threshold	33.0
Step 2	t-test	33.0
Step 3	LASSO	20.1

As seen in the iterative screening process, the t-test step did not further eliminate features; the number of features remained unchanged at 33.0 after screening. The primary reason is that all 39 features initially included in this study were selected based on prior knowledge of AD pathology, and the vast majority of these features exhibited statistically significant differences between the AD and CN groups, with no features lacking discernible diagnostic value. The subsequent LASSO regression screening, serving as the core dimension-reduction step, compresses the weights of redundant features while retaining effective diagnostic information, ultimately reducing the number of features from 33.0 to 20.1 and achieving efficient noise reduction and dimensionality reduction in the feature set.

Final retained features: After a three-step progressive screening process, the model retained an average of 20.1 core features across random seeds, covering the three major dimensions of MRI structural imaging, PET molecular pathology, and clinical genetics, thereby achieving the precise retention of multidimensional pathological information. Table 9 presents the features identified by LASSO screening and their corresponding regression coefficients under the first random seed, providing a visual representation of each feature's diagnostic contribution.

Feature Stability: To verify the reproducibility of the feature selection results, this study counted the frequency with which each feature was selected across 10 sets of random seeds. The results showed that a total of 13 features were consistently selected across all 10 seed sets, covering core multimodal metrics, which fully confirms the robustness and reliability of the selection results. Among these, the MRI modality included ST40CV, ENTORHINAL, ST10CV, ST26CV, ST60CV, ST71SV, and ST29SV; the PET modality included TAU_PARAHIPPOCAMPAL and TAU_ENTORHINAL; and the clinical modality included PTEDUCAT, PTGENDER, AGE, and APOE4. In addition, features such as ST96SV, TAU_PRECUNEUS, ST61SV, and HIPPOCAMPUS were also selected with high frequency, further confirming that these features are core, stable biomarkers for AD diagnosis.

Comparison of Feature Selection Strategies: To validate the optimality of the three-step progressive screening strategy used in this study, five mainstream

Table 9 Features Selected by LASSO Using the First Random Seed and Their Regression Coefficients

Rank	Feature	LASSO Coefficient	Direction
1	ST29SV (Seahorse)	-0.1716	→ CN
2	ST10CV (ICV)	+0.1439	→ AD
3	ST40CV (temporal middle gyrus)	-0.0803	→ CN
4	ENTORHINAL	-0.0694	→ CN
5	PTGENDER	+0.0651	→ AD
6	APOE4	+0.0581	→ AD
7	AGE	-0.0443	→ CN
8	ST26CV (gyrus fusiformis)	-0.0378	→ CN
9	PTEDUCAT	-0.0351	→ CN
10	TAU_ENTORHINAL	+0.0272	→ AD

single or combined screening strategies were simultaneously established as controls. A cross-comparison was conducted across three dimensions—feature reduction, model diagnostic performance, and overfitting—with detailed results for each strategy presented in Table 10.

Table 10 Performance comparison of six feature selection strategies.

Scheme	n_selected	AUC	BAcc	gap
None	39.0	0.9558 ± 0.0083	0.8851 ± 0.0112	+0.0098
Variance	33.0	0.9549 ± 0.0079	0.8837 ± 0.0119	+0.0093
T-test	37.7	0.9553 ± 0.0082	0.8848 ± 0.0139	+0.0101
Variance + T-test	33.0	0.9549 ± 0.0079	0.8837 ± 0.0119	+0.0093
LASSO	19.1	0.9557 ± 0.0072	0.8818 ± 0.0161	+0.0085
Variance + T-test + LASSO	20.1	0.9543 ± 0.0073	0.8755 ± 0.0215	+0.0096

The comparison results show that the full-feature model, which did not undergo feature selection, achieved the highest AUC (0.9558); however, it had a high feature dimension and failed to remove redundant noise, posing a potential risk of overfitting. The three-step progressive feature selection scheme adopted in this study resulted in only a slight decrease in AUC (0.0015), while reducing the number of features from 39 dimensions to 20.1 dimensions. At the same time, the model's overfitting gap was maintained at +0.0096—close to the +0.0098 of the full-feature model—effectively balancing diagnostic performance and model complexity, and significantly improving the model's generalization ability.

3.3 Multimodal Performance Comparison

Based on the seven differentiated modality combination schemes predefined in Section 2.3, to systematically compare the diagnostic performance of single-modality,

dual-modality fusion, and full-dimensional multimodal fusion, and to clarify the independent value and synergistic benefits of each modality, this study conducted stratified model training and performance comparison analyses on both the full-sample cohort (N = 874) and the PET complete data subset cohort (N = 383) to precisely investigate the impact of data completeness on the diagnostic performance of each modality.

3.3.1 Full Cohort (N = 874)

In the complete cohort of 874 subjects, the average diagnostic performance results for the seven modal configurations, validated repeatedly using 10 random seeds, are shown in Table 11.

Table 11 Diagnostic performance of the 7 modal configurations in the full-sample cohort (N = 874).

Modality	n_feat	AUC	BAcc	gap
clinical_only	4	0.7865 ± 0.0359	0.7270 ± 0.0304	+0.0026
mri_only	20	0.9277 ± 0.0113	0.8527 ± 0.0191	+0.0149
PET_only	15	0.6397 ± 0.0273	0.6511 ± 0.0222	+0.0038
clinical_mri	24	0.9478 ± 0.0102	0.8707 ± 0.0177	+0.0087
clinical_pet	19	0.8072 ± 0.0374	0.7232 ± 0.0551	+0.0223
MRI-PET	35	0.9419 ± 0.0100	0.8725 ± 0.0139	+0.0140
multimodal	39	0.9558 ± 0.0083	0.8851 ± 0.0112	+0.0098

Based on the performance data, the following key patterns can be identified: First, the full-dimensional multimodal fusion model demonstrated the best diagnostic performance, with an AUC of 0.9558, significantly outperforming all single-modality and dual-modality combinations, thereby validating the beneficial effects of integrating multidimensional pathological information; Second, the MRI modality serves as the core foundation for AD diagnosis; the AUC for the single MRI modality (0.9277) is far higher than that of the single PET modality. Performance improved significantly when clinical variables were combined with MRI (AUC increased from 0.9277 to 0.9478, +0.0201), indicating that clinical variables provide independent incremental information beyond that of MRI; Third, the independent diagnostic value of the PET modality was limited in the full dataset (AUC 0.6397), attributable to a PET data missing rate of approximately 53% (411 out of 874 valid cases), where the median imputation method diluted the specific signals of PET molecular pathology; Fourth, clinical demographic and genetic variables provide independent incremental information; although the performance of a single clinical modality is relatively low, fusing it with MRI and PET can further optimize the model's diagnostic accuracy.

3.3.2 PET Subset (N = 383)

To eliminate the interference of missing PET data on model performance and to accurately explore the pathological diagnostic value of the PET modality and modality synergy, this study replicated the model training and performance validation for seven

modality configurations in a subset cohort of 383 cases with complete PET data. The results are shown in Table 12.

Table 12 Diagnostic performance of seven modal configurations in the PET complete subset cohort (N = 383).

Modality	n_feat	AUC	BAcc	gap
clinical_only	4	0.8258 ± 0.0607	0.7355 ± 0.0558	+0.0085
mri_only	20	0.9595 ± 0.0281	0.8793 ± 0.0256	-0.0317
PET_only	15	0.9021 ± 0.0259	0.8172 ± 0.0612	+0.0220
clinical_mri	24	0.9722 ± 0.0199	0.8859 ± 0.0480	-0.0050
clinical_pet	19	0.9555 ± 0.0233	0.8138 ± 0.0814	-0.0058
MRI-PET	35	0.9856 ± 0.0128	0.9521 ± 0.0450	-0.0086
multimodal	39	0.9910 ± 0.0106	0.9397 ± 0.0269	-0.0031

The complete data results for the PET subset further refine our understanding of modality value: First, the full-dimension multimodal model achieved the best diagnostic performance in this study, with an AUC as high as 0.9910, enabling high-precision differentiation between AD and CN; Second, the performance of the PET modality improved significantly, with the AUC for the single PET modality rising from 0.6397 in the full dataset to 0.9021, directly confirming the core issue of PET signals being diluted by missing values in the full dataset; Furthermore, the performance of the MRI-PET dual-modality fusion model approached that of the full-modality model, with an AUC difference of only 0.0054 between the two, suggesting that when PET data is complete, dual-modality image fusion already possesses extremely strong diagnostic capability, and the incremental contribution of clinical variables is limited; Finally, most models in the subsets exhibited a "negative gap" phenomenon (test set AUC ¿ training set AUC). This is a common phenomenon in validation studies with small sample sizes (N = 383) and repeated holdout, likely stemming from random distribution differences between the training and test sets. It is worth noting that the absolute AUC values of these models remain very high (mri_pet 0.9856, multimodal 0.9910), indicating that the models' generalization ability has not been compromised.

3.3.3 Comparison of the Full Dataset vs. the PET Subset

To quantify the impact of PET data completeness on the diagnostic performance of various modality combinations, this study summarized the AUC differences between the two cohorts to visually illustrate the relationship between data quality and model performance. The specific comparison results are shown in Table 13.

The results of the cross-cohort comparison clearly show that modality combinations including PET are highly sensitive to data completeness. Among these, the single-modality PET approach demonstrated the greatest performance improvement (ΔAUC = +0.2623), followed by the clinical-PET dual-modality approach, while pure MRI and pure clinical modalities were minimally affected by data completeness. This result directly explains the difference in model performance between the full dataset and

Table 13 Comparison of AUCs for 7 modal configurations between the full dataset and the PET subset.

Modality	Full Dataset AUC	PET Subset AUC	ΔAUC
clinical_only	0.7865	0.8258	+0.0394
MRI_only	0.9277	0.9595	+0.0317
pet_only	0.6397	0.9021	+0.2623
clinical_mri	0.9478	0.9722	+0.0244
clinical_pet	0.8072	0.9555	+0.1483
MRI-PET	0.9419	0.9856	+0.0437
multimodal	0.9558	0.9910	+0.0352

the PET subset, fully confirming that PET molecular pathological features require complete data to realize their specific diagnostic value. This finding also suggests that in clinical practice, if a patient's PET data is incomplete, PET-based predictive results should be interpreted with caution.

3.4 Model Comparison

To identify the optimal modeling algorithm suited for the multimodal feature set in this study and to compare the differences in adaptability between linear and nonlinear models, this study conducted unified training and performance evaluation of the five mainstream machine learning models selected in Section 2.5 based on the full set of 39 multimodal features. The comparison results of the core performance metrics for each model are shown in Table 14.

Table 14 Performance comparison of five machine learning classifiers on the multimodal feature set.

Model	AUC	BAcc	gap
LR	0.9558 ± 0.0088	0.8851 ± 0.0118	+0.0098
SVM	0.9527 ± 0.0095	0.8739 ± 0.0217	+0.0298
RF	0.9367 ± 0.0097	0.8686 ± 0.0239	+0.0511
XGBoost	0.9415 ± 0.0117	0.8576 ± 0.0184	+0.0403
MLP	0.9455 ± 0.0154	0.8795 ± 0.0258	+0.0206

The model comparison results show that the logistic regression (LR) model exhibits the best overall performance, with the highest AUC and balanced accuracy among all models, as well as the smallest overfitting gap; its model robustness and generalization ability are significantly superior to those of other algorithms. The SVM model's diagnostic performance is comparable to that of LR, but it carries a higher risk of overfitting; tree-based ensemble models such as RF and XGBoost, as well as the MLP (multilayer perceptron) model, all underperformed LR on the low-dimensional, feature-refined dataset used in this study and exhibited varying degrees of overfitting.

3.5 Comparison of Full-Volume vs. Reduced MRI Data

In an earlier phase of this study, 14 core ROIs were selected based on prior pathological knowledge of AD, and 6 composite features were derived to construct a 20-dimensional reduced MRI feature set. To validate the scientific validity and necessity of this reduction strategy, this study further compared the differences in model performance and generalization ability between the full 125-dimensional raw MRI ROI features and the 20-dimensional reduced MRI features; the results are shown in Table 15.

Table 15 Performance comparison between the full (125-dimensional) and reduced (20-dimensional) MRI feature sets.

Configuration	n_feat	AUC	BAcc	gap
MRI_full	125	0.9378 ± 0.0107	0.8663 ± 0.0160	+0.0469
MRI_slim	20	0.9277 ± 0.0119	0.8527 ± 0.0201	+0.0149

Comparison results show that while the full-range MRI feature set contains more comprehensive information on brain regions, its AUC improved only slightly by 0.0101 compared to the slimmed-down feature set; however, the overfitting gap increased significantly, and generalization performance declined markedly. In contrast, the reduced MRI feature set used in this study—while resulting in only a slight loss of diagnostic accuracy—significantly reduces feature redundancy and the risk of model overfitting, achieving a better balance between diagnostic performance, model robustness, and generalization ability. This fully validates the effectiveness and scientific soundness of selecting core ROIs based on prior pathological knowledge.

3.6 Analysis of Modality Complementarity

To elucidate the core mechanisms underlying the significant performance improvement of the multimodal fusion model and to overcome the limitations of single-modality diagnosis, this study relied on the three quantitative metrics—CKA similarity, Pearson correlation coefficient, and Jaccard overlap—established in Section 2.7. Using a complete PET subset cohort, we systematically quantified the information overlap and error complementarity among clinical, MRI, and PET, to clarify the principles underlying the benefits of multimodal fusion. The quantitative results for each metric are shown in Table 16.

The results of the quantitative analysis of modality complementarity reveal a key underlying principle: the Jaccard error overlap index between MRI and PET modalities is 0.0000, indicating that there is absolutely no overlap in misclassified samples between the two modalities—no samples are misdiagnosed by both—and that they possess the ultimate error-compensating capability. This is the core reason for the breakthrough improvement in the performance of the multimodal fusion model. From an information perspective, the CKA similarity and Pearson correlation coefficients between MRI and PET are higher than those of other modality pairs, indicating a

Table 16 Modality complementarity quantified based on CKA, Pearson correlation coefficient, and Jaccard index.

Modality Pairs	CKA	Pearson r	Jaccard
clinical vs. MRI	0.1668	0.4084	0.0435
clinical vs. PET	0.1609	0.4011	0.0588
MRI vs. PET	0.3126	0.5591	0.0000

certain degree of consistency in the core decision-making logic for disease diagnosis, allowing them to mutually corroborate pathological features; in contrast, clinical modalities exhibit higher error overlap with both imaging modalities, resulting in relatively limited complementary gains.

This complementary nature can be reasonably explained from the perspective of pathological mechanisms: MRI images reflect AD-related macroscopic brain structural atrophy, while PET images capture molecular pathological abnormalities such as Aβ plaques and tau tangles. By characterizing AD at different pathological levels (macroscopic structure vs. molecular pathology), these two modalities offer distinct diagnostic advantages for subjects with different pathological phenotypes. They can compensate for the diagnostic blind spots of a single modality, ultimately enabling precise diagnosis through multimodal fusion.

3.7 Interpretability Analysis

To address the "black box" issue in machine learning models, clarify the core decision-making basis of multimodal models, and avoid interpretive biases caused by feature multicollinearity, this study combined two methods—SHAP feature importance analysis and LR coefficient validation—to explore model interpretability from the dual dimensions of feature contribution and result robustness, thereby comprehensively dissecting the model's diagnostic logic.

3.7.1 SHAP Feature Importance

Based on the optimal multimodal LR model, the SHAP algorithm was used to quantify the contribution weights of each feature to the model's diagnostic decisions, identifying the top 20 core features by contribution. The specific rankings and modal affiliations are shown in Table 17. Figure 2A shows a bar chart of feature importance sorted by average absolute SHAP values, while Figure 2B displays a swarm plot of feature values and corresponding SHAP values, revealing the direction of influence that high or low feature values have on the prediction results.

Based on the SHAP quantification results, the core decision-making feature hierarchy of the multimodal model can be clearly identified: The top 10 high-contribution features consist primarily of MRI structural images and clinical-genetic features, among which the hippocampal volume ratio (HIPPO_RATIO) has the highest contribution weight (mean |SHAP| = 0.5350). It serves as the core imaging indicator for

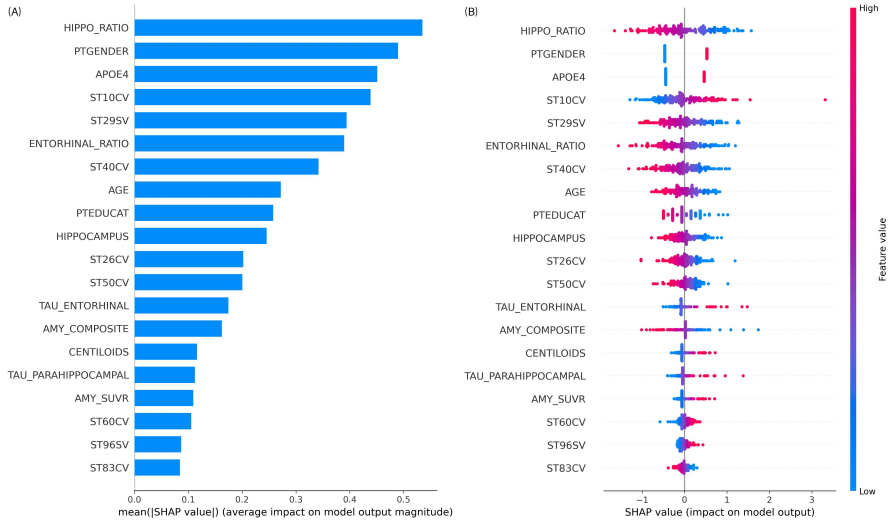


Fig. 2 Visualization of SHAP feature importance in the multimodal model: (A) SHAP bar chart showing the top 20 features sorted by mean absolute SHAP value; (B) SHAP swarm plot illustrating the relationship between feature values and SHAP values. SHAP, Shapley Additive Explanation; mean |SHAP|, mean absolute SHAP value.

the model to distinguish AD from CN, aligning closely with the pathological mechanism of characteristic hippocampal atrophy in AD. On the clinical dimension, gender (PTGENDER), the APOE4 gene, age (AGE), and years of education (PTEDUCAT) all ranked among the high-contribution features, confirming that demographic and genetic baseline characteristics provide significant auxiliary value for AD diagnosis. The overall contribution weights of PET molecular pathological features were relatively low; however, core indicators such as the combined Tau and amyloid-beta values in the entorhinal cortex consistently ranked among the top 20 features. This indicates that molecular pathological information can serve as a key supplementary diagnostic basis, further refining the model's disease discrimination capability. The overall distribution of feature contributions spans the three major modalities—clinical, MRI, and PET—fully demonstrating the core advantage of the multimodal fusion model in making collaborative decisions based on multidimensional pathological information.

3.7.2 Validation of LR Coefficient Results

To avoid result bias from a single interpretability method, we conducted cross-validation of SHAP feature contributions and their directional effects based on the regression coefficients of the optimal LR model from this study. The results show that the core features with high SHAP contributions align closely with the ranking of absolute LR coefficient values. The coefficient weights for hippocampus-related MRI features, APOE4, gender, and pathological features of the middle temporal gyrus and entorhinal cortex all rank among the highest, providing strong corroboration of the SHAP analysis results.

Table 17 Top 20 features in the multimodal model, ranked by average absolute SHAP value.

Rank	Feature	mean |SHAP|	Modal
1	HIPPO_RATIO	0.5350	MRI
2	PTGENDER	0.4898	Clinical
3	APOE4	0.4509	Clinical
4	ST10CV	0.4387	MRI
5	ST29SV	0.3940	MRI
6	ENTORHINAL_RATIO	0.3897	MRI
7	ST40CV	0.3419	MRI
8	AGE	0.2718	Clinical
9	PTEDUCAT	0.2577	Clinical
10	HIPPOCAMPUS	0.2455	MRI
11	ST26CV	0.2020	MRI
12	ST50CV	0.2005	MRI
13	TAU_ENTORHINAL	0.1743	PET
14	AMY_COMPOSITE	0.1626	PET
15	CENTILOIDS	0.1160	PET
16	TAU_PARAHIPPOCAMPAL	0.1123	PET
17	AMY_SUVR	0.1092	PET
18	ST60CV	0.1054	MRI
19	ST96SV	0.0869	MRI
20	ST83CV	0.0843	MRI

In terms of the direction of feature effects, the coefficients for features such as hippocampal volume, entorhinal cortex volume, and years of education were negative, suggesting that higher values of these indicators are associated with a higher probability of normal cognition in participants; conversely, the coefficients for APOE4 carrier status, advanced age, male gender, and high pathological burden of Aβ and Tau in the brain were positive, corresponding to an increased risk of AD, which fully aligns with the clinical and pathological patterns of AD and the findings of differences between baseline groups in this study. These consistent results fully confirm that the model's decision logic aligns with the objective mechanisms of the disease, is free from overfitting bias, and yields robust and reliable interpretable results.

4 Discussion and Conclusions

This study proposes a diagnostic method for Alzheimer's disease based on multimodal feature fusion and ensemble learning, utilizing structural MRI, amyloid PET, tau PET, and clinical variables to distinguish AD from CN. During the data preprocessing and feature extraction stages, 39 multimodal features were extracted from the ADNI multicenter cohort, and a three-step progressive feature selection process—comprising variance thresholding, t-tests, and LASSO regression—was employed for feature fusion and dimensionality reduction. For classification, the system compared five classifiers—logistic regression, support vector machines, random forests, XGBoost, and multi-layer perceptrons—and evaluated model performance using a cross-validation scheme with 10 random seeds. Given that PET data in the ADNI cohort had a missing rate of

approximately 53%, this study conducted stratified comparative analyses in both the full-sample cohort and the PET-complete subset cohort.

The experimental results show that the full-dimensional multimodal fusion model achieved an AUC of 0.9558 in the full dataset, and the AUC further improved to 0.9910 in the complete PET subset, with performance comprehensively superior to that of single-modality and dual-modality combinations. Structural MRI is the core modality supporting AD diagnosis, and the diagnostic performance of MRI alone far exceeds that of PET alone. Quantification of modality complementarity revealed that the Jaccard overlap index between MRI and PET was 0.0000, indicating that there was absolutely no overlap in misclassified samples between the two modalities—a key mechanism driving the improvement in multimodal fusion performance. The methodological contribution of this study lies not only in performance improvement, but more importantly in providing a reusable quantitative framework for modality complementarity. Future multimodal studies can use the CKA, Pearson correlation, and Jaccard index to assess the complementarity of their modality combinations, thereby guiding modality selection and avoiding blind fusion. This directly addresses the technical need in multimodal research: how to determine whether modality fusion is worthwhile and where the gains come from.

From a clinical perspective, this study provides quantitative evidence for a practical decision: whether to add PET to routine MRI. The Jaccard index of 0.0000 between MRI and PET indicates that the two modalities make errors on completely different samples—suggesting that adding PET can correct cases that MRI alone would miss. Given the high cost and limited accessibility of PET, this finding helps clinicians make evidence-based decisions about whether PET is necessary for a given patient. Specifically, for patients whose MRI-based prediction falls in the borderline probability range, adding PET may provide the most diagnostic value; for patients with highly confident MRI-based predictions, the incremental benefit of PET may be limited.

From a clinical perspective, this study provides a reliability assessment framework for AI-assisted AD diagnosis. By combining SHAP and LR coefficient analyses, calibration curves, and DCA, we demonstrated that the multimodal model not only achieves high accuracy but also produces reliable probability estimates and clinically meaningful net benefit. Moreover, we identified a general pitfall of SHAP—when features are collinear, SHAP may overestimate feature importance—which helps clinicians interpret AI outputs with appropriate caution. This finding is particularly relevant for clinical translation, as it suggests that SHAP-based explanations should be interpreted alongside domain knowledge rather than taken at face value.

This finding aligns closely with the A/T/N pathological framework proposed by Jack et al.: Amyloid PET corresponds to the A dimension, Tau PET corresponds to the T dimension, and structural MRI corresponds to the N dimension. Together, these three modalities comprehensively characterize the full spectrum of AD pathology across different levels—from molecular pathology to macroscopic degeneration. Dual validation using SHAP and LR coefficients revealed that the hippocampal standardized volume ratio (HIPPO_RATIO) is the most critical diagnostic biomarker, which aligns closely with the pathological mechanism of AD-characteristic hippocampal atrophy. Furthermore, the gender variable (PTGENDER) exhibited an unusually

high contribution in the initial SHAP analysis. Given that gender is strongly and naturally correlated with intracranial volume (ICV), it is likely that the SHAP algorithm partially attributed the indirect contribution of ICV to PTGENDER, leading to an overestimation of its importance. This finding has methodological implications beyond AD diagnosis: it reveals a general pitfall of SHAP-based interpretability—when features are collinear, SHAP may overestimate the importance of certain features. This suggests that relying solely on SHAP rankings may not fully reflect a feature's independent diagnostic value, and model interpretation must incorporate prior domain knowledge. This addresses another technical need: improving the reliability of SHAP-based interpretability in medical AI.

In practical deployment, multimodal imaging and clinical data will be collected through a standardized preprocessing workflow, followed by feature extraction, fusion, and classification, to provide clinicians with auxiliary diagnostic decision support. Furthermore, in real-world clinical practice, classifier outputs can be combined with DCA net benefit analysis to perform stratified risk assessment (prioritizing intervention for high-confidence AD patients and conducting further examinations for borderline cases). This study focuses on methodological-level quantification of modality complementarity, addressing the technical need of "how to quantify multimodal fusion mechanisms," rather than directly addressing specific clinical needs such as early AD screening or MCI progression prediction. Nevertheless, several limitations remain: it relies on the ADNI public retrospective cohort, which consists primarily of Caucasians and is ethnically homogeneous; the missing rate for PET data across the entire cohort is approximately 53%; only CN and AD patients were included, while MCI patients were excluded; and no generalization testing was conducted in independent external cohorts such as AIBL or OASIS.

Future work will involve constructing a "CN-MCI-AD" full-disease-course diagnostic model by incorporating an MCI cohort, conducting multicenter external validation in independent external cohorts such as AIBL and OASIS, and optimizing multimodal fusion by introducing deep learning architectures such as Transformers and graph neural networks. We will also investigate the method's robustness to variations in scanning equipment and population heterogeneity in real-world clinical settings. Clinical implications, such as risk stratification strategies, alert logic, and delays in continuous monitoring systems, will also be further explored.

Abbreviations

AD: Alzheimer's disease; CN: cognitively normal; MRI: magnetic resonance imaging; PET: positron emission tomography; AUC: area under the receiver operating characteristic curve; BAcc: balanced accuracy; LASSO: least absolute shrinkage and selection operator; SHAP: SHapley Additive exPlanations; CKA: centered kernel alignment; DCA: decision curve analysis; ICV: intracranial volume; APOE4: apolipoprotein E $\varepsilon 4$; ROI: region of interest; SUVR: standardized uptake value ratio; MCI: mild cognitive impairment; ADNI: Alzheimer's Disease Neuroimaging Initiative.

Declarations

Ethics approval and consent to participate

This study was conducted in accordance with the Declaration of Helsinki. This study is a secondary analysis of de-identified data from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database. The ADNI study was approved by the institutional review boards of all participating sites, and all participants provided written informed consent.

Consent for publication

Not applicable.

Availability of data and materials

The data that support the findings of this study are available from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu), but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. Data are however available from the authors upon reasonable request and with permission of ADNI.

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported by the Yunnan Province Wu Zhonghai Expert Workstation (Grant No. 202305AF150045) and the Industry Innovation Talents Project of Yunnan Xingdian Talents Support Plan (Certificate No.: YNWR-CYJS2020-017).

Authors' contributions

RM and ZY contributed to the study conception and design. RM performed the data analysis and drafted the manuscript. ZY supervised the study and critically revised the manuscript. Both authors read and approved the final manuscript.

Acknowledgements

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U19 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California.

ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

Authors' information

Not applicable.

References

[1] Nichols E, Steinmetz JD, Vollset SE, Fukutaki K, Chalek J, Abd-Allah F, et al. Estimation of the global prevalence of dementia in 2019 and forecasted prevalence in 2050: an analysis for the Global Burden of Disease Study 2019. The Lancet Public Health. 2022;7(2):e105–e125.

[2] Livingston G, Huntley J, Liu KY, Costafreda SG, Selbæk G, Alladi S, et al. Dementia prevention, intervention, and care: 2024 report of the Lancet standing Commission. Lancet. 2024;404(10452):572–628.

[3] Scheltens P, De Strooper B, Kivipelto M, Holstege H, Chételat G, Teunissen CE, et al. Alzheimer's disease. Lancet. 2021;397(10284):1577–1590.

[4] Hampel H, Hardy J, Blennow K, Chen C, Perry G, Kim SH, et al. The Amyloid-β Pathway in Alzheimer's Disease. Molecular Psychiatry. 2021;26(10):5481–5503.

[5] Jack CR, Bennett DA, Blennow K, Carrillo MC, Dunn B, Haeberlein SB, et al. NIA-AA Research Framework: Toward a biological definition of Alzheimer's disease. Alzheimers Dement. 2018;14(4):535–562.

[6] Frisoni GB, Fox NC, Jack CR, Scheltens P, Thompson PM. The clinical use of structural MRI in Alzheimer disease. Nature Reviews Neurology. 2010;6(2):67–77.

[7] Fischl B. FreeSurfer. NeuroImage. 2012;62(2):774–781.

[8] Belloy ME, Napolioni V, Greicius MD. A Quarter Century of APOE and Alzheimer's Disease: Progress to Date and the Path Forward. Neuron. 2019;101(5):820–838.

[9] Arya AD, Verma SS, Chakarabarti P, Chakrabarti T, Elngar AA, Kamali AM, et al. A systematic review on machine learning and deep learning techniques in the effective diagnosis of Alzheimer's disease. Brain Informatics. 2023;10(1):17.

[10] Jayman ZS, Lakmal HKIS. Multimodal Deep Learning for Alzheimer's Diagnosis Using PET and MRI. In: 2025 9th SLAAI International Conference on Artificial Intelligence (SLAAI-ICAI); 2025. .

[11] Ahmed F, Naz NS, Khan S, Rehman AU, Ismael WM, Khan MA. Explainable artificial intelligence (XAI) in medical imaging: a systematic review of techniques, applications, and challenges. BMC Medical Imaging. 2026;26(1):37.

[12] Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. Med Decis Making. 2006;26(6):565–574.

[13] Jack CR, Andrews JS, Beach TG, Buracchio T, Dunn B, Graf A, et al. Revised criteria for diagnosis and staging of Alzheimer's disease: Alzheimer's Association Workgroup. Alzheimers Dement. 2024;20(8):5143–5169.

[14] Weiner MW, Veitch DP, Aisen PS, Beckett LA, Cairns NJ, Green RC, et al. The Alzheimer's Disease Neuroimaging Initiative 3: Continued innovation for clinical trial improvement. Alzheimers Dement. 2017;13(5):561–571.

[15] Mieling M, Yousuf M, Bunzeck N, Weiner M, Aisen P, Petersen R, et al. Predicting the progression of MCI and Alzheimer's disease on structural brain integrity and other features with machine learning. GeroScience. 2026;48(1):463–487.

[16] Hedges EP, Dimitrov M, Zahid U, Brito Vega B, Si S, Dickson H, et al. Reliability of structural MRI measurements: The effects of scan session, head tilt, inter-scan interval, acquisition sequence, FreeSurfer version and processing stream. Neuroimage. 2022;246:118751.

[17] Kecskemeti S, Freeman A, Travers BG, Alexander AL. FreeSurfer based cortical mapping and T1-relaxometry with MPnRAGE: Test-retest reliability with and without retrospective motion correction. NeuroImage. 2021;242:118447.

[18] Braak H, Braak E. Neuropathological stageing of Alzheimer-related changes. Acta Neuropathologica. 1991;82(4):239–259.

[19] Jagust WJ, Koeppe RA, Rabinovici GD, Villemagne VL, Harrison TM, Landau SM, et al. The ADNI PET Core at 20. Alzheimer's & Dementia. 2024;20(10):7340–7349.

[20] Rosenblatt M, Tejavibulya L, Jiang R, Noble S, Scheinost D. Data leakage inflates prediction performance in connectome-based machine learning models. Nature Communications. 2024;15(1):1829.

[21] Kapoor S, Narayanan A. Leakage and the reproducibility crisis in machine-learning-based science. Patterns. 2023;4(9):100804.

[22] Landau SM, Fero A, Baker SL, Koeppe R, Mintun M, Chen K, et al. Measurement of Longitudinal β-Amyloid Change with ^{18}F-Florbetapir PET and Standardized Uptake Value Ratios. Journal of Nuclear Medicine. 2015;56(4):567–574.

[23] Klunk WE, Koeppe RA, Price JC, Benzinger TL, Devous MD, Jagust WJ, et al. The Centiloid Project: Standardizing quantitative amyloid plaque estimation by PET. Alzheimer's & Dementia. 2015;11(1):1–15.e4.

[24] Xue C, Kowshik SS, Lteif D, Puducheri S, Jasodanand VH, Zhou OT, et al. AI-based differential diagnosis of dementia etiologies on multimodal data. Nature Medicine. 2024;30(10):2977–2989.

[25] Hao X, Li Y, Wang X, Ma C, Liu R, Jiao Y, et al. Multimodal radiomics of cerebellar subregions for machine learning-driven Alzheimer's disease diagnosis. Frontiers in Aging Neuroscience. 2025;17.

[26] Keranmu N, Aizezi D, Pan X, Yang L, Liu Y, Liu J. Interpretable machine learning model based on multimodal MRI radiomics for Alzheimer's disease diagnosis. Frontiers in Aging Neuroscience. 2026;18.

[27] Chen Y, Qi Y, Hu Y, Qiu X, Qiu T, Li S, et al. Integrated cerebellar radiomic-network model for predicting mild cognitive impairment in Alzheimer's disease. Alzheimer's & Dementia. 2025;21(1):e14361.

[28] Kornblith S, Norouzi M, Lee H, Hinton G. Similarity of Neural Network Representations Revisited. In: Proceedings of the 36th International Conference on Machine Learning. Proceedings of Machine Learning Research. PMLR; 2019. p. 3519–3529.

[29] Alvarez SA. Gaussian RBF Centered Kernel Alignment (CKA) in the Large-Bandwidth Limit. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2023;45(5):6587–6593.