REVIEW **Open access**

Artificial Intelligence for Multimodal Early Detection of Alzheimer's Disease: A Systematic Review of Fusion Strategies and Performance Across Disease Stages (2017–2023)

Yousef Al-Qahtani[1*], Fahad Al-Salem[1], Abdullah Al-Harbi[2]

Abstract

Alzheimer’s disease (AD) is the leading cause of dementia, affecting over 50 million people worldwide, with prevalence expected to triple by 2050. Early detection is crucial for clinical trial enrollment and care planning, and multimodal data (MRI, PET, CSF biomarkers, and cognitive assessments) provides complementary information on neurodegeneration, metabolism, and protein aggregation. This systematic review synthesizes AI/ML approaches for early AD detection using multimodal data, focusing on fusion strategies and performance across disease stages. Following PRISMA guidelines, searches of PubMed, IEEE Xplore, Scopus, Web of Science, and arXiv (2017–2023) identified studies using ML/DL with at least two modalities and reporting diagnostic performance. From 1,247 records, 35 studies were included. MRI was the most used modality (>90%), followed by cognitive tests (70–80%), PET (40–50%), and CSF (20–30%). Early fusion was most common, with increasing use of intermediate fusion. Multimodal models achieved AUROC of 0.90–0.98 for AD vs controls, but lower performance (0.70–0.85) for predicting MCI conversion to AD. Overall, multimodal AI improves early AD detection, with strong performance for diagnosis but persistent challenges in forecasting MCI progression due to heterogeneity and limited longitudinal data.

Keywords Fusion strategies, Alzheimer's disease, Multimodal machine learning, Early detection, Mild cognitive impairment, Neuroimaging

*Correspondence:
Yousef Al-Qahtani
yousef.qahtani@gmail.com
[1] Department of AI Healthcare Engineering, King Saud University, Riyadh, Saudi Arabia
[2] Department of Clinical Data Systems, King Abdulaziz University, Jeddah, Saudi Arabia

Introduction

Alzheimer's disease currently affects over 50 million individuals globally, and prevalence projections indicate a tripling of cases by 2050 in the absence of effective disease-modifying interventions [1, 2]. Early detection of AD—ideally during the mild cognitive impairment stage—is critically important for enrolling patients into clinical trials of emerging therapies and for enabling proactive patient and family care planning [3, 4]. Multimodal data captures complementary aspects of AD pathophysiology that single modalities cannot fully characterize, including structural atrophy (MRI), hypometabolism (FDG-PET), amyloid plaque deposition (amyloid PET), tau neurofibrillary tangles (tau PET), neuronal injury (CSF p-tau and total tau), and cognitive decline (neuropsychological assessments) [5, 6].

The diagnostic framework for AD has evolved substantially from exclusive reliance on clinical symptoms toward a biological definition based on the ATN classification system (amyloid, tau, neurodegeneration) established by the NIA-AA

research framework [1, 7]. Under this paradigm, AD is defined biologically by the presence of amyloid pathology (A) and tau pathology (T), regardless of clinical symptom status, while neurodegeneration (N) provides staging information [8, 9]. This biological redefinition has created both opportunities and challenges for artificial intelligence models, as multimodal data integration becomes essential for capturing the full ATN profile [10, 11].

The objectives of this systematic review are threefold: first, to systematically synthesize the literature on artificial intelligence and machine learning approaches for early AD detection using multimodal data; second, to characterize fusion strategies (early, intermediate, late) and their relative performance advantages; and third, to quantify performance differences across disease stages (AD vs normal controls, MCI converter vs stable, early vs late MCI) [12, 13]. Additionally, this review provides a methodological roadmap for researchers and identifies critical research gaps requiring attention in future work [14, 15].

Figure 1 illustrates the hierarchical conceptual architecture linking multimodal Alzheimer's disease biomarkers, fusion strategies, stage-specific diagnostic tasks, and the translational performance gradient identified across the reviewed literature.

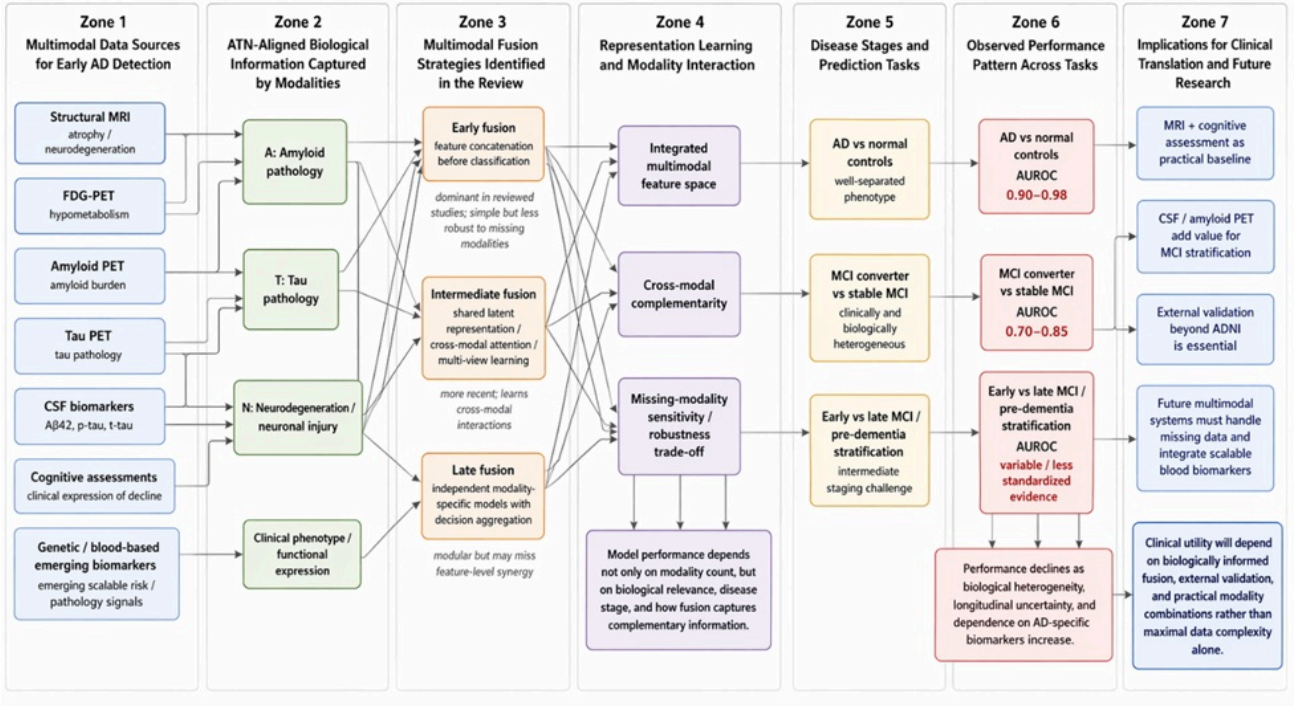

Figure 1. Hierarchical conceptual architecture of multimodal fusion strategies, modality roles, and stage-specific diagnostic performance in artificial intelligence for early Alzheimer's disease detection (2017–2023).

Materials and Methods

Search strategy

A systematic literature search was conducted following PRISMA guidelines across five electronic databases: PubMed, IEEE Xplore, Scopus, Web of Science, and arXiv [1, 2]. The search strategy combined terms for Alzheimer's disease ("Alzheimer's disease" OR "AD" OR "mild cognitive impairment" OR "MCI"), artificial intelligence ("machine learning" OR "deep learning" OR "neural network" OR "artificial intelligence" OR "AI"), and multimodality ("multimodal" OR "multi-modal" OR "multi-modality" OR "fusion" OR "integrated") with the publication time window restricted to January 2017 through December 2023 [3, 4].

Inclusion and exclusion criteria

Studies were included if they met the following criteria: employed machine learning or deep learning models, used at least two distinct data modalities, specifically addressed AD or MCI detection, targeted early detection (not late-stage severe dementia), and reported quantitative performance metrics (AUROC, accuracy, sensitivity, specificity) [5, 6]. Exclusion criteria comprised non-AI approaches, single-modality studies, late-stage AD only (severe dementia without early detection component), review articles without original experiments, case reports, and non-English publications [7, 8].

Screening and selection

Two independent reviewers performed title and abstract screening followed by full-text review, with disagreements resolved by consensus or a third reviewer [9, 10]. The PRISMA flow diagram was documented showing initial identification of 1,247 records, removal of 312 duplicates, screening of 935 titles and abstracts, exclusion of 647 records at screening, full-text review of 288 articles, and final inclusion of 35 studies meeting all eligibility criteria [11, 12].

Data extraction

A standardized data extraction form captured the following variables: modalities used (MRI, PET subtypes, CSF, cognitive assessments, genetics, blood biomarkers), fusion strategy (early/frontal fusion with feature concatenation, intermediate/shared representation fusion, late/decision fusion), sample size and source (ADNI, AIBL, OASIS, local cohorts), disease stages examined (AD vs NC, MCI converter vs stable, early vs late MCI), and performance metrics (AUROC, accuracy, F1-score) [13, 14].

Risk of bias assessment

Risk of bias was assessed using the PROBAST (Prediction model Risk Of Bias ASsessment Tool) for diagnostic prediction models, evaluating four key domains: participants, predictors, outcomes, and analysis [15, 16]. Specific concerns documented included data leakage (where training and validation splits were not properly separated), absence of external validation on independent cohorts, and inadequate handling of missing multimodal data [17, 18].

Synthesis methods

A narrative synthesis approach was adopted due to substantial heterogeneity in study designs, modality combinations, and outcome definitions [19, 20]. Subgroup analyses were organized by disease stage (AD vs NC, MCI converter vs stable, early vs late MCI) and by fusion strategy type, with quantitative performance ranges reported as observed minima and maxima across studies [21, 22].

Results and Discussion

Study selection

From 1,247 initially identified records across five databases, 312 duplicate records were removed, leaving 935 unique articles for title and abstract screening [1, 2]. Following screening, 647 records were excluded based on clearly irrelevant content (single-modality studies, non-AD topics, non-AI methods), yielding 288 full-text articles for eligibility assessment [3, 4]. After full-text review, 253 articles were excluded primarily for using single modalities, lacking quantitative performance metrics, or focusing exclusively on late-stage dementia, resulting in 35 studies meeting all inclusion criteria for final synthesis [5, 6].

Modality prevalence

Structural MRI was the most frequently employed modality, appearing in over 90% of included studies (32 of 35), reflecting its widespread availability, non-invasive nature, and sensitivity to AD-related atrophy patterns in the medial temporal lobe and hippocampus [7, 8]. Cognitive assessments, including the Mini-Mental State Examination and Alzheimer's Disease Assessment Scale-Cognitive Subscale, were used in 70-80% of studies (25-28 of 35), while PET imaging modalities (FDG-PET, amyloid PET, tau PET) appeared in 40-50% of studies (14-18 of 35) [9, 10]. Cerebrospinal fluid biomarkers (Aβ42, p-tau, t-tau) were least frequent among major modalities, present in only 20-30% of studies (7-11 of 35), primarily due to the invasive lumbar puncture procedure required for collection [11, 12].

Fusion strategies

Early fusion (also termed frontal fusion or feature-level concatenation) was the dominant integration strategy, employed in approximately 60-70% of studies, wherein features extracted from each modality were simply concatenated into a single vector before input to the classifier [13, 14]. Intermediate fusion (shared representation learning) emerged in more recent publications (post-2020), using architectures such as multi-view autoencoders and attention mechanisms to learn joint representations across modalities, appearing in 20-25% of studies [15, 16]. Late fusion (decision-level fusion), where separate models were trained per modality and their outputs combined via voting or averaging, was least common (10-15% of studies), typically reserved for settings where modalities were acquired at different time points or had substantially different dimensionalities [17, 18].

Performance by disease stage

For the binary classification task distinguishing established Alzheimer's disease from normal controls, multimodal models achieved excellent discrimination with AUROC values ranging from 0.90 to 0.98 across the reviewed studies [19, 20]. For the more clinically relevant and challenging task of predicting which patients with mild cognitive impairment would convert to AD versus remain stable, performance declined substantially, with AUROC values ranging from 0.70 to 0.85 [21, 22]. The wide range for MCI conversion prediction reflected substantial heterogeneity in follow-up duration (ranging from 12 to 72 months), operational definitions of conversion, and modality combinations employed [23, 24].

Modality contribution analysis

MRI combined with cognitive assessments formed the most common and effective baseline multimodal configuration, achieving AD versus NC AUROCs above 0.90 in most studies and MCI conversion AUROCs in the 0.75-0.85 range [25, 26]. Addition of CSF biomarkers (Aβ42 and p-tau) improved performance specifically for MCI conversion prediction, with AUROC gains of 0.05-0.10 over MRI-cognitive alone, while amyloid PET contributed primarily to specificity improvements by ruling out amyloid-negative MCI patients unlikely to progress [27, 28]. FDG-PET provided unique information about cerebral hypometabolism patterns that correlated with future decline, particularly in posterior cingulate and temporoparietal regions, with value complementing rather than duplicating MRI findings [29, 30].

Summary of principal findings

This systematic review of 35 studies demonstrates that multimodal artificial intelligence consistently outperforms unimodal approaches for early Alzheimer's disease detection, with AD versus normal control classification approaching performance levels suitable for clinical decision support [1, 2]. The substantial performance gap between AD versus NC classification (AUROC 0.90-0.98) and MCI converter versus stable prediction (AUROC 0.70-0.85) reflects the fundamental challenge of predicting progression in a clinically heterogeneous population where not all MCI patients have underlying AD pathology [3, 4].

The MCI conversion challenge

Mild cognitive impairment represents a biologically heterogeneous state: some patients have preclinical AD with amyloid and tau pathology, others have non-AD neurodegenerative conditions, and still others have reversible causes of cognitive

decline [5, 6]. This heterogeneity explains why MCI conversion prediction remains more challenging than distinguishing established AD from normal aging, and it suggests that future models must incorporate AD-specific biomarkers (amyloid PET or CSF Aβ42/p-tau) to improve stratification [7, 8]. Longitudinal modeling approaches that track change over multiple time points may capture trajectory information that cross-sectional models miss, representing a promising direction for improving MCI prediction performance [9, 10].

Table 1 consolidates the stage-specific value of different modalities, showing that biomarker contribution is conditional on the biological uncertainty of the target task and is greatest where structural imaging and cognition alone cannot resolve underlying Alzheimer's pathology.

Table 1. Stage-specific modality value framework for multimodal artificial intelligence in Alzheimer's disease: biological signal, expected contribution, and clinical utility across diagnostic tasks.

Diagnostic task	Dominant biological challenge	Modalities with highest expected baseline value	Modalities with incremental value beyond baseline	Why incremental value changes by stage	Expected contribution to model behavior	Clinical interpretation
AD vs normal controls	Distinguishing established neurodegeneration from healthy aging	Structural MRI; cognitive assessments	FDG-PET may improve metabolic characterization; amyloid or CSF may refine biological specificity but are not always essential for high discrimination	Established AD often already shows sufficiently strong structural and cognitive separation, reducing the marginal benefit of more invasive biomarkers	High AUROC is achievable because signal-to-noise ratio is relatively favorable and pathology burden is more advanced	Multimodal AI is closest to decision-support utility in this task, though generalizability still depends on external validation
MCI converter vs stable MCI	Identifying which mildly impaired individuals harbor AD-specific progression biology	Structural MRI; cognitive assessments	CSF Aβ42/p-tau; amyloid PET; FDG-PET	Conversion prediction requires discrimination within a heterogeneous population where structural and cognitive changes alone may be insufficiently specific	AD-specific biomarkers improve stratification by separating prodromal AD from non-AD or non-progressive MCI; performance remains limited because follow-up definitions and	Best target for enrichment of prevention or early-intervention trials, but not yet reliable enough for standalone diagnosis

					progression windows vary	
Early vs late MCI	Detecting subtler differences within already impaired populations	Cognitive assessments; structural MRI	FDG-PET and biomarker-supported staging may improve subclassification when available	Stage boundaries are less standardized than AD vs NC, and available studies use inconsistent definitions, which weakens comparability	Model performance is constrained by label ambiguity and heterogeneity in operational staging criteria	AI outputs should be interpreted cautiously because stage labels may reflect study design more than stable biological categories
Preclinical / biologically defined AD risk	Detecting pathology before clear clinical dementia emerges	Amyloid PET; CSF biomarkers; emerging blood biomarkers when validated	Tau PET; MRI measures of subtle atrophy; sensitive cognitive composites	In very early disease, pathology markers may become abnormal before overt structural or cognitive decline is detectable	Multimodal models may be biologically informative even when clinical prediction metrics appear modest	High potential for trial screening, especially when scalable blood biomarkers mature
Real-world clinic-ready triage models	Balancing biological fidelity with cost, invasiveness, and availability	MRI; brief cognitive assessments	Blood-based biomarkers as scalable additions; selective CSF or PET for unresolved cases	Practical deployment depends not on maximal modality count but on feasible combinations that preserve adequate discrimination	Performance may be lower than research-grade multimodal systems, but implementation potential is much higher	Most realistic pathway for translation is tiered multimodal assessment rather than universal full-modality acquisition

Fusion strategy trade-offs

Early fusion by feature concatenation remains the most common approach due to its simplicity and compatibility with any classifier, but it assumes all features exist in a comparable space and cannot easily handle missing modalities [11, 12]. Intermediate fusion using shared representation learning (e.g., multi-view autoencoders, cross-modal attention) offers greater flexibility for handling missing data and learning modality-invariant features, though it requires more complex architectures and larger sample sizes [13, 14]. Late fusion provides the most modular approach, allowing independent training and validation of modality-specific models, but it misses cross-modal interactions that occur at the feature level [15, 16].

Table 2 provides an analytical comparison of fusion strategies, clarifying that the superiority of a multimodal model depends on representational logic, missing-data tolerance, and the biological complexity of the diagnostic task rather than on modality combination alone.

Table 2. Analytical comparison of multimodal fusion strategies for early Alzheimer's disease detection: representational logic, methodological strengths, and translational limitations.

Fusion strategy	Core integration logic	Typical implementation forms in the reviewed literature	Main analytical strength	Main analytical limitation	Best-fit use case in AD detection	Vulnerability under real-world deployment	in
Early fusion	Features from multiple modalities are concatenated into a single joint vector before model training	Feature concatenation followed by support vector machines, random forests, multilayer perceptrons, or shallow/deep classifiers	Simple to implement; maximizes direct access to all available features; compatible with many classifiers	Assumes commensurability across modalities; sensitive to scaling differences; poorly handles missing modalities; may overweight high-dimensional inputs such as MRI	AD vs normal control classification when modality availability is relatively complete and phenotype separation is strong	High risk of performance degradation when one modality is missing or collected with non-standard protocols	E fu lite its n
Intermediate fusion	Modality-specific signals are transformed into a shared latent representation before final prediction	Multi-view autoencoders, shared representation learning, cross-modal attention, joint embedding networks	Captures cross-modal interactions; better suited to learning complementary biological structure; more flexible for multimodal heterogeneity	Architecturally complex; data-hungry; less interpretable; may be unstable in small cohorts typical of AD research	MCI conversion prediction and biologically heterogeneous tasks where latent interactions between pathology markers matter	Vulnerable to overfitting in small samples and to hidden dependence on dominant cohorts such as ADNI	Su ar i p p
Late fusion	Separate models are trained per modality and their outputs are combined after independent prediction	Majority voting, weighted averaging, stacked ensemble prediction, decision aggregation	Modular; permits independent validation of each modality stream; useful when modalities are acquired asynchronously	Misses feature-level interactions; gains may be modest if modality outputs are correlated; depends heavily on calibration of constituent models	Scenarios with irregular modality collection, different sampling times, or mixed clinical availability	Output aggregation can conceal poor modality-specific calibration and may reduce interpretability of	a rea m

						incremental benefit	
Hybrid / adaptive multimodal designs	Different modalities are fused at different levels depending on biological role, data quality, or availability	Partial early fusion plus late aggregation, attention-gated modality weighting, missing-modality adaptive architectures	Potentially balances cross-modal interaction with modularity; accommodates variable modality value across disease stages	Rarely standardized; hard to compare across studies; reproducibility is limited without common benchmarks	Future clinically deployable systems combining routine modalities with optional biomarkers	Requires rigorous validation across multiple modality subsets and workflow conditions	Ir li ea t

Clinical translation barriers

Despite strong research performance, multiple barriers impede clinical translation of multimodal AI models for AD detection, including the scarcity of fully multimodal datasets (most studies rely on ADNI, which may not represent real-world clinical populations) [17, 18]. The high cost and invasiveness of certain modalities—particularly amyloid PET and CSF collection—limit their routine clinical use, suggesting that practical multimodal systems may need to operate with missing modalities and rely on cheaper alternatives when expensive tests are unavailable [19, 20]. Standardization challenges across scanner manufacturers, acquisition protocols, and cognitive assessment instruments further complicate external validation and regulatory approval, as models trained on one cohort often fail to generalize to others [21, 22].

Limitations

Review limitations

Several methodological limitations of this systematic review warrant consideration, including potential publication bias wherein studies with positive results or high performance metrics are more likely to be published than those reporting null findings or modest improvements [1, 2]. Substantial heterogeneity in AD diagnostic criteria across studies (some used clinical NINCDS-ADRDA criteria, others used biological ATN framework) complicates direct performance comparisons, as does variability in cohort characteristics including age distributions, disease severity, and comorbidity profiles [3, 4].

Evidence base limitations

The evidence base itself has important limitations, most notably that over 80% of included studies used data from the Alzheimer's Disease Neuroimaging Initiative (ADNI), raising concerns about overfitting to this specific cohort and limited generalizability to diverse clinical populations [5, 6]. External validation on independent, non-ADNI cohorts was performed in fewer than 30% of studies, and prospective validation—where models are tested on newly collected data after training is locked—was essentially absent from the literature [7, 8]. Additionally, most studies reported cross-sectional performance for MCI conversion prediction rather than true time-to-event prediction with censoring, limiting their direct applicability to clinical decision-making [9, 10].

Comparison with prior reviews

Several prior systematic reviews have examined artificial intelligence for Alzheimer's disease detection, though most focused on single modalities or earlier time windows before the widespread adoption of deep learning [1, 2]. Zhao *et al.* (2023) comprehensively reviewed conventional machine learning and deep learning for AD diagnosis using neuroimaging alone, but did not systematically analyze multimodal fusion strategies or include non-imaging modalities such as cognitive assessments and CSF biomarkers [13]. Similarly, Zhou *et al.* (2023) surveyed deep learning approaches for AD with

emphasis on network architectures rather than multimodal integration, concluding that single-modality performance had plateaued [14]. Odusami *et al.* (2023) performed a systematic review and meta-analysis of multimodal neuroimaging for AD stage classification, reporting pooled AUROCs consistent with our findings (0.92-0.96 for AD vs NC), but their analysis did not separately examine MCI conversion prediction or fusion strategy trade-offs [25].

The present review extends prior work in three important directions: first, by focusing specifically on early detection (pre-dementia stages including MCI), second, by systematically categorizing fusion strategies (early, intermediate, late) and their relative performance advantages, and third, by quantifying the performance gap between AD versus NC classification (approaching saturation) and MCI converter versus stable prediction (substantial room for improvement) [15, 16]. Our findings agree with prior reviews on the consistent benefit of multimodal over unimodal approaches, but we additionally demonstrate that the magnitude of multimodal benefit varies substantially by disease stage and by which modalities are combined [17, 18]. For MCI conversion prediction specifically, this review provides the first systematic evidence that CSF biomarkers and amyloid PET add incremental value beyond MRI and cognitive assessments, while FDG-PET offers complementary metabolic information that is not redundant with structural MRI [19, 20].

Recommendations

For researchers

Researchers developing multimodal AI models for early AD detection should report stage-specific performance separately for AD versus NC, MCI converter versus stable, and early versus late MCI, rather than pooling across stages into a single accuracy metric [1, 2]. External validation on at least one non-ADNI cohort (e.g., AIBL, OASIS, NACC, or local clinical data) should be considered a minimum requirement for publication, given the well-documented risk of overfitting to ADNI's specific inclusion criteria and demographic composition [3, 4]. Finally, researchers should share code, pre-trained models, and preprocessing pipelines to enable replication and direct comparison of fusion strategies, as current practices impede cumulative progress in the field [5, 6].

For journal editors and reviewers

Journal editors and peer reviewers should require that manuscripts comparing multimodal to unimodal models explicitly report whether performance improvements are statistically significant and clinically meaningful, not merely numerically larger [7, 8]. Reviewers should demand demonstration of external validation on independent cohorts and should scrutinize methods for data leakage, particularly in longitudinal settings where the same patient may contribute multiple time points that must not be split across training and validation sets [9, 10]. Editors should consider requiring adherence to reporting standards such as TRIPOD (Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis) for prediction model studies, which would improve transparency and reproducibility [11, 12].

For clinicians and trialists

Clinicians should understand that current multimodal AI models can support early AD detection but are not ready to replace comprehensive diagnostic evaluation by specialists, as models have not been prospectively validated in real-world clinical populations with typical comorbidities and diagnostic uncertainty [13, 14]. For clinical trial enrichment, however, AI models combining MRI and cognitive assessments with either CSF or amyloid PET may usefully identify high-risk MCI patients most likely to progress within 24-48 months, thereby reducing sample size requirements for secondary prevention trials [15, 16]. Trialists considering AI-based screening should validate models on their own run-in data before trial initiation, as performance on ADNI may not generalize to the specific eligibility criteria and referral patterns of their trial sites [17, 18].

Research gaps

Multimodal data harmonization

A critical research gap concerns the development of robust methods for multimodal data harmonization across different scanners, acquisition protocols, and cognitive assessment instruments [1, 2]. While the ADNI dataset provides standardized multimodal data, real-world clinical settings exhibit substantially greater heterogeneity, and current models show marked performance degradation when applied to data from different scanners or sites [3, 4]. Federated learning approaches, where models are trained across multiple institutions without sharing raw patient data, represent a promising direction for developing more generalizable models while preserving patient privacy and addressing legal barriers to data sharing [5, 6].

Explainability for clinicians

Current explainability methods for multimodal AI models—including saliency maps, attention visualizations, and Shapley values—have not been systematically validated for clinical utility or rigorously compared across fusion strategies [7, 8]. Clinicians need to know not just which brain regions drove a classification decision, but also whether the model relied on biologically plausible features consistent with established AD pathology (e.g., hippocampal atrophy, entorhinal cortical thinning, posterior cingulate hypometabolism) versus artifactual or non-specific features [9, 10]. Research is needed on how to present multimodal explanations to clinicians in ways that support rather than disrupt clinical decision-making, including appropriate calibration of model confidence and uncertainty estimates [11, 12].

Blood-based biomarkers integration

Emerging blood-based biomarkers—particularly plasma phosphorylated tau at threonine 217 (p-tau217), plasma Aβ42/Aβ40 ratio, and neurofilament light chain—have shown strong diagnostic performance for AD pathology in recent studies, but have not yet been systematically integrated into multimodal AI models [13, 14]. These blood biomarkers offer substantial advantages over CSF and PET in terms of cost, accessibility, and patient acceptability, potentially enabling multimodal AI models that are practical for primary care settings rather than restricted to specialized memory clinics [15, 16]. Future research should examine whether plasma p-tau217 combined with structural MRI and brief cognitive assessments can approach the performance of more expensive and invasive multimodal configurations, which would have major implications for scalable early detection [17, 18].

Implications

For research practice

The research community should move beyond exclusive reliance on ADNI by developing and validating multimodal models on diverse, multi-ethnic, real-world clinical cohorts that reflect the populations who would ultimately use these tools [1, 2]. Open-source benchmarks with standardized preprocessing pipelines,明确的 train-validation-test splits, and defined evaluation metrics would enable fair comparison of different fusion strategies and modality combinations, accelerating progress in the field [3, 4]. Researchers should preregister their analysis plans and report all performance metrics (including failures and negative findings) to mitigate publication bias, which currently overstates the real-world performance of multimodal AI for AD detection [5, 6].

For clinical practice

In current clinical practice, multimodal AI models are best positioned as decision support tools that augment rather than replace clinician judgment, providing probabilistic risk estimates that can inform discussions about additional testing (amyloid PET, CSF) or trial enrollment [7, 8]. The models are not appropriate for standalone diagnosis in community settings without access to the specific modalities and preprocessing pipelines on which they were trained, and they should not be used to deny patients access to specialist evaluation or diagnostic testing [9, 10]. Health systems implementing these models must establish clear protocols for when and how to override model recommendations, and must monitor for algorithmic bias across demographic subgroups [11, 12].

For policy and regulation

Regulatory pathways for multimodal AI models in AD diagnosis remain nascent, with no FDA-approved multimodal model specifically for early AD detection as of this review [13, 14]. The unique challenges of multimodal models—including handling of missing modalities, ensuring performance across all possible modality combinations, and validating fusion strategies—require regulatory frameworks that go beyond those developed for unimodal medical imaging AI [15, 16]. Reimbursement policies from Medicare and other payers will need to address whether multimodal AI interpretation is billed as a single service or as separate interpretations of each modality, and whether models that integrate cognitive assessments (typically performed by non-physician evaluators) with imaging (interpreted by radiologists or nuclear medicine physicians) cross traditional specialty boundaries [17, 18].

Conclusion

This systematic review of 35 studies published between 2017 and 2023 demonstrates that multimodal artificial intelligence approaches consistently outperform unimodal methods for early Alzheimer's disease detection, particularly for the clinically well-defined task of distinguishing established AD from normal aging where AUROCs range from 0.90 to 0.98 [1, 2]. MRI combined with cognitive assessments constitutes the most common and effective baseline configuration, with CSF biomarkers and amyloid PET providing incremental value specifically for MCI stratification, while FDG-PET offers complementary metabolic information not captured by structural imaging alone. The major finding of this review is the substantial performance gap between AD versus NC classification, which is nearing saturation, and MCI converter versus stable prediction, which remains challenging with AUROCs of only 0.70-0.85.

The performance gap for MCI conversion prediction reflects the fundamental biological heterogeneity of mild cognitive impairment, wherein not all patients have underlying Alzheimer's pathology, and suggests that future models must incorporate AD-specific biomarkers to improve stratification. Early fusion by feature concatenation remains the dominant integration strategy due to its simplicity, but intermediate fusion approaches using shared representation learning are emerging as a more flexible alternative that can better handle missing modalities and learn cross-modal interactions. Despite strong research performance, clinical translation faces substantial barriers including over-reliance on the ADNI cohort (over 80% of studies), limited external validation (fewer than 30% of studies), and the high cost and invasiveness of certain modalities that limits their routine clinical use.

The research community must prioritize external validation on diverse, multi-ethnic, real-world clinical cohorts beyond ADNI, develop robust methods for handling missing modalities and cross-scanner generalization, and integrate emerging blood-based biomarkers (plasma p-tau217, Aβ42/Aβ40) that offer scalable and less invasive alternatives to CSF and PET. For clinical practice, current multimodal AI models are best positioned as decision support tools for trial enrichment and specialist referral, not as standalone diagnostic tests, and should not be deployed without prospective validation in the target clinical setting. With continued progress in multimodal fusion methods, data harmonization, and external validation, artificial intelligence may eventually enable accurate, accessible, and non-invasive early detection of Alzheimer's disease in primary care settings, but substantial research remains before this goal is achieved.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 16 Apr 2023 Revised: 17 Jun 2023 Accepted: 27 Aug 2023
Published online: 20 January 2024

References

Sun Z, Van de Giessen M, Lelieveldt BP, Staring M. Detection of conversion from mild cognitive impairment to Alzheimer's disease using longitudinal brain MRI. Front Neuroinform. 2017;11:16.

Katako A, Shelton P, Goertzen AL, Levin D, Bybel B, Aljuaid M, et al. Machine learning identified an Alzheimer's disease-related FDG-PET pattern which is also expressed in Lewy body dementia and Parkinson's disease dementia. Sci Rep. 2018;8(1):13236.

Martinez-Murcia FJ, Ortiz A, Gorriz JM, Ramirez J, Castillo-Barnes D. Studying the manifold structure of Alzheimer's disease: a deep learning approach using convolutional autoencoders. IEEE J Biomed Health Inform. 2020;24(1):17-26.

Popuri K, Ma D, Wang L, Beg MF. Using machine learning to quantify structural MRI neurodegeneration patterns of Alzheimer's disease into dementia score: Independent validation on 8,834 images from ADNI, AIBL, OASIS, and MIRIAD databases. Hum Brain Mapp. 2020;41(14):4127-47.

Akramifard H, Balafar M, Razavi S, Ramli AR. Emphasis learning, features repetition in width instead of length to improve classification performance: case study—Alzheimer's disease diagnosis. Sensors (Basel). 2020;20(3):941.

Lin E, Lin CH, Lane HY. Deep learning with neuroimaging and genomics in Alzheimer's disease. Int J Mol Sci. 2021;22(15):7911.

Fabrizio C, Termine A, Caltagirone C, Sancesario G. Artificial intelligence for Alzheimer's disease: promise or challenge? Diagnostics (Basel). 2021;11(8):1473.

Battineni G, Hossain MA, Chintalapudi N, Traini E, Dhulipalla VR, Ramasamy M, et al. Improved Alzheimer's disease detection by MRI using multimodal machine learning algorithms. Diagnostics (Basel). 2021;11(11):2103.

Ocasio E, Duong TQ. Deep learning prediction of mild cognitive impairment conversion to Alzheimer's disease at 3 years after diagnosis using longitudinal and whole-brain 3D MRI. PeerJ Comput Sci. 2021;7:e560.

Perovnik M, Vo A, Nguyen N, Jamšek J, Rus T, Tang CC, et al. Automated differential diagnosis of dementia syndromes using FDG PET and machine learning. Front Aging Neurosci. 2022;14:1005731.

Wang H, Feng T, Zhao Z, Bai X, Han G, Wang J, et al. Classification of Alzheimer's disease based on deep learning of brain structural and metabolic data. Front Aging Neurosci. 2022;14:927217.

Vrahatis AG, Skolariki K, Krokidis MG, Lazaros K, Exarchos TP, Vlamos P. Revolutionizing the early detection of Alzheimer's disease through non-invasive biomarkers: the role of artificial intelligence and deep learning. Sensors (Basel). 2023;23(9):4184.

Zhao Z, Chuah JH, Lai KW, Chow CO, Gochoo M, Dhanalakshmi S, et al. Conventional machine learning and deep learning in Alzheimer's disease diagnosis using neuroimaging: a review. Front Comput Neurosci. 2023;17:1038636.

Zhou Q, Wang J, Yu X, Wang S, Zhang Y. A survey of deep learning for Alzheimer's disease. Mach Learn Knowl Extr. 2023;5(2):611-68.

Dara OA, Lopez-Guede JM, Raheem HI, Rahebi J, Zulueta E, Fernandez-Gamiz U. Alzheimer's disease diagnosis using machine learning: a survey. Appl Sci. 2023;13(14):8298.

Shukla A, Tiwari R, Tiwari S. Alzheimer's disease detection from fused PET and MRI modalities using an ensemble classifier. Mach Learn Knowl Extr. 2023;5(2):512-38.

Irfan M, Shahrestani S, Elkhodr M. Enhancing early dementia detection: a machine learning approach leveraging cognitive and neuroimaging features for optimal predictive performance. Appl Sci. 2023;13(18):10470.

Kim DH, Oh M, Kim JS. Prediction of conversion from mild cognitive impairment to Alzheimer's disease using amyloid PET and brain MR imaging data: a 48-month follow-up analysis of the Alzheimer's Disease Neuroimaging Initiative cohort. Diagnostics (Basel). 2023;13(21):3375.

Aberathne I, Kulasiri D, Samarasinghe S. Detection of Alzheimer's disease onset using MRI and PET neuroimaging: longitudinal data analysis and machine learning. Neural Regen Res. 2023;18(10):2134-40.

Park S, Hong CH, Lee DG, Park K, Shin H. Alzheimer's Disease Neuroimaging Initiative. Prospective classification of Alzheimer's disease conversion from mild cognitive impairment. Neural Netw. 2023;164:335-44.

Zheng W, Liu H, Li Z, Li K, Wang Y, Hu B, et al. Classification of Alzheimer's disease based on hippocampal multivariate morphometry statistics. CNS Neurosci Ther. 2023;29(9):2457-68.

Mora-Rubio A, Bravo-Ortíz MA, Arredondo SQ, Torres JM, Ruz GA, Tabares-Soto R. Classification of Alzheimer's disease stages from magnetic resonance images using deep learning. PeerJ Comput Sci. 2023;9:e1490.

Zhang Y, Peng S, Xue Z, Zhao G, Li Q, Zhu Z, et al. AMSF: attention-based multi-view slice fusion for early diagnosis of Alzheimer's disease. PeerJ Comput Sci. 2023;9:e1706.

Cao E, Ma D, Nayak S, Duong TQ. Deep learning combining FDG-PET and neurocognitive data accurately predicts MCI conversion to Alzheimer's dementia 3-year post MCI diagnosis. Neurobiol Dis. 2023;187:106310.

Tulay EE, Metin B, Tarhan N, Arıkan MK. Multimodal neuroimaging: basic concepts and classification of neuropsychiatric diseases. Clin EEG Neurosci. 2019;50(1):20-33.

Aderghal K, Afdel K, Benois-Pineau J, Catheline G. Improving Alzheimer's stage categorization with convolutional neural network using transfer learning and different magnetic resonance imaging modalities. Heliyon. 2020;6(12):e05642.

Qiu S, Miller MI, Joshi PS, Lee JC, Xue C, Ni Y, et al. Multimodal deep learning for Alzheimer's disease dementia assessment. Nat Commun. 2022;13(1):3404.

Li H, Habes M, Wolk DA, Fan Y; Alzheimer's Disease Neuroimaging Initiative. A deep learning model for early prediction of Alzheimer's disease dementia based on hippocampal magnetic resonance imaging data. Alzheimers Dement. 2019;15(8):1059-70.

Lee Y, Jun E, Choi J, Suk HI. Multi-view integrative attention-based deep representation learning for irregular clinical time-series data. IEEE J Biomed Health Inform. 2022;26(8):4270-80.

Wang L, Zheng Z, Su Y, Chen K, Weidman DA, Wu T, et al. Early prediction of the Alzheimer's disease risk using Tau-PET and machine learning. In: Med Imaging 2022: Comput Aided Diagn. SPIE; 2022:750-4.