Minimax Theory and its Applications

Volume 11 (2026), No. 3s, 302 to 313

A Cross-Modal Attention-Based Multimodal Fusion Model for Alzheimer's Disease Classification: A Focused Systematic Review and Reproducible Framework

Sahil Sharma[1*], Rajinder Kumar[2], Deepender[3], Tarandeep Singh Walia[4], Gajender Rao[5]

[1]Research Scholar, Department of CSE, Guru Kashi University, Bathinda, Punjab, India, Email: mca.sahil84@gmail.com
[2]Associate Prof., Department of Computer Applications,GuruKashi University, Bathinda, Punjab, India, Email: drrajinder1983@gmail.com
[3]Assistant Professor, AIT-CSE, Chandigarh University, Mohali, Punjab, India, Email: deependerduhan6@gmail.com
[4]Professor, School of Computer Application, Lovely Professional University, Punjab, India, Email: taran_walia2k@yahoo.com
[5]Assistant Professor, CSE, Galghotias University, Greater Noida, Utter Pardesh, India, Email: gajender@galghotiasuniversity.edu.in

Abstract
Alzheimer's disease (AD) classification increasingly combines structural magnetic resonance imaging (MRI), positron emission tomography (PET), cognitive and clinical measures, genetics, fluid biomarkers, speech, and other digital phenotypes. Yet, simple concatenation or late averaging cannot explicitly model how one modality changes the interpretation of another, and many reported results are vulnerable to small cohorts, incomplete modalities, data leakage, weak external validation, and poor calibration. This paper presents a PRISMA-informed focused systematic synthesis of 50 DOI-linked publications from 2021-2026 and proposes CMA-MFNet, a quality-aware cross-modal attention architecture for AD classification and mild cognitive impairment (MCI) conversion prediction. Modality-specific encoders create token sequences; bidirectional cross-modal attention learns conditional interactions; availability- and quality-aware gates control each modality's contribution; and modality dropout trains the system to remain useful when PET, genetics, or other inputs are absent. The framework includes class-balanced learning, cross-modal contrastive alignment, consistency regularization, calibrated uncertainty, and multimodal explanations. A reproducible validation protocol is specified using ADNI for development and OASIS-3, AIBL, and NACC for external testing, with subject-level splits, nested model selection, ablation tests, bootstrap confidence intervals, and fairness analyses. Recent evidence indicates strong average ADNI performance but materially lower performance in population-based, cross-lingual, and externally validated settings. Therefore, the proposed framework treats generalization, missingness, calibration, and interpretability as first-class objectives rather than post-hoc additions. The paper provides a high-quality research blueprint suitable for implementation and prospective clinical evaluation without fabricating unobserved experimental results.

Keywords: Alzheimer's disease; multimodal fusion; cross-modal attention; transformer; MRI; PET; mild cognitive impairment; systematic review; explainable AI; external validation

I. Introduction

Alzheimer's disease is a progressive neurodegenerative disorder whose biological and clinical manifestations emerge across partially overlapping timescales. Structural atrophy, amyloid and tau accumulation, altered metabolism and connectivity, cognitive decline, genetic susceptibility, fluid biomarkers, and changes in speech or daily behavior each provide an incomplete view of the disease process. Consequently, a clinically meaningful computational model should integrate complementary evidence rather than force a single modality to carry the entire diagnostic burden. The rationale is biological as much as statistical: modalities capture different links in the causal chain from molecular pathology to brain-system disruption and observable function.

Recent multimodal systems have reported encouraging diagnostic and prognostic performance. Attention-based models are particularly attractive because they can represent interactions instead of merely concatenating features. MADDi combined cross-modal and self-attention and reported a substantial improvement over simpler fusion, while recent imaging systems have achieved strong internal and independent-test results. However, performance is highly dataset dependent. A 2026 systematic review found that ADNI-based diagnosis averaged 92.5% accuracy and MCI-conversion modeling averaged an AUC of 0.922, whereas UK Biobank risk prediction averaged an AUC of 0.84, DementiaBank diagnosis averaged an AUC of 0.813, and cross-lingual detection averaged 77% accuracy. Small self-collected cohorts reported approximately 96% accuracy, illustrating why headline scores cannot be interpreted without cohort size, task definition, split design, and external validation.

The central technical problem is therefore not simply multimodal classification. It is reliable multimodal reasoning under heterogeneous sample sizes, asynchronous observations, missing modalities, variable acquisition quality, label uncertainty, site effects, and unequal demographic representation. Conventional early fusion assumes commensurable features; late fusion may ignore low-level interactions; and unrestricted attention can learn shortcuts associated with site, scanner, missingness pattern, or assessment workflow. A clinically credible architecture must model conditional interactions while explicitly controlling data quality, availability, leakage, and uncertainty.

This paper makes four contributions. First, it provides a focused, PRISMA-informed synthesis of 50 recent DOI-linked publications spanning imaging, clinical, genetic, speech, and mixed-modality systems. Second, it formulates CMA-MFNet, a Cross-Modal Attention-based Multimodal Fusion Network with modality-specific encoders, bidirectional interaction blocks, and availability- and quality-aware gating. Third, it specifies a complete validation protocol centered on ADNI and external cohorts, including missing-modality stress tests, nested model selection, calibration, subgroup analysis, and statistical comparison. Fourth, it distinguishes evidence-derived benchmark context from future experimental claims. No performance number is presented as a result of CMA-MFNet until the model is implemented and evaluated on locked test sets.

II. Focused Systematic Review Methodology

A. Review Design and Questions

The review was designed as a focused secondary evidence synthesis rather than a new subscription-database export. Two recent, broad systematic reviews provided the source-study corpora: a multimodal neuroimaging review and meta-analysis that searched IEEE Xplore, ScienceDirect, ACM Digital Library, and PubMed, and a 2026 review that searched PubMed, IEEE Xplore, Scopus, ACM Digital Library, Cochrane, and arXiv through 15 November 2025. Their combined corpora yielded 113 study entries before overlap and eligibility screening. The present synthesis asks: (RQ1) Which modality combinations and fusion strategies are most common? (RQ2) What evidence supports explicit cross-modal interactions? (RQ3) Which design choices improve robustness to missing modalities and domain shift? (RQ4) Which datasets and validation procedures are suitable for a reproducible high-quality study?

B. Eligibility and Evidence Curation

Eligible publications were dated 2021-2026, addressed AD, MCI, dementia classification, conversion, or risk prediction, used at least two modalities or directly evaluated multimodal fusion, reported quantitative evaluation or systematic-review evidence, and provided a DOI-linked peer-reviewed report. Priority was given to studies using recognized resources such as ADNI, OASIS-3, AIBL, NACC, UK Biobank, or DementiaBank/ADReSS. Methodological standards, dataset papers, PRISMA guidance, and clinically important reviews were retained when necessary to support reproducible reporting. Pre-2021 studies, purely unimodal studies with no multimodal methodological relevance, non-peer-reviewed reports, duplicate reports, and records without an appropriate DOI were excluded from the focused set.

Figure 1 documents the transparent secondary curation process. The 63 excluded entries were removed collectively for overlap across the two source reviews, publication before 2021, unimodal or out-of-scope design, insufficient methodological reporting, or absence of a suitable DOI-linked peer-reviewed record. Because this was not a de novo rerun inside proprietary databases, the flow diagram is deliberately labeled PRISMA-informed rather than represented as an original database search.

C. Data Extraction and Quality Dimensions

For each retained study, the synthesis considered year, data source, task, modalities, fusion level, attention mechanism, missing-data strategy, split unit, external validation, interpretability, calibration, and reported

limitations. Four quality dimensions were emphasized: (1) protection against subject leakage and repeated-visit leakage; (2) independence and heterogeneity of validation data; (3) robustness to unavailable or low-quality modalities; and (4) clinical utility beyond discrimination, including calibration, uncertainty, fairness, and explanation. A recent QUADAS-2 synthesis found high risk in patient selection for 61% of studies and unclear risk for index test and reference standard in 76%, and flow/timing in 85%. These figures motivated the protocol requirements proposed later.

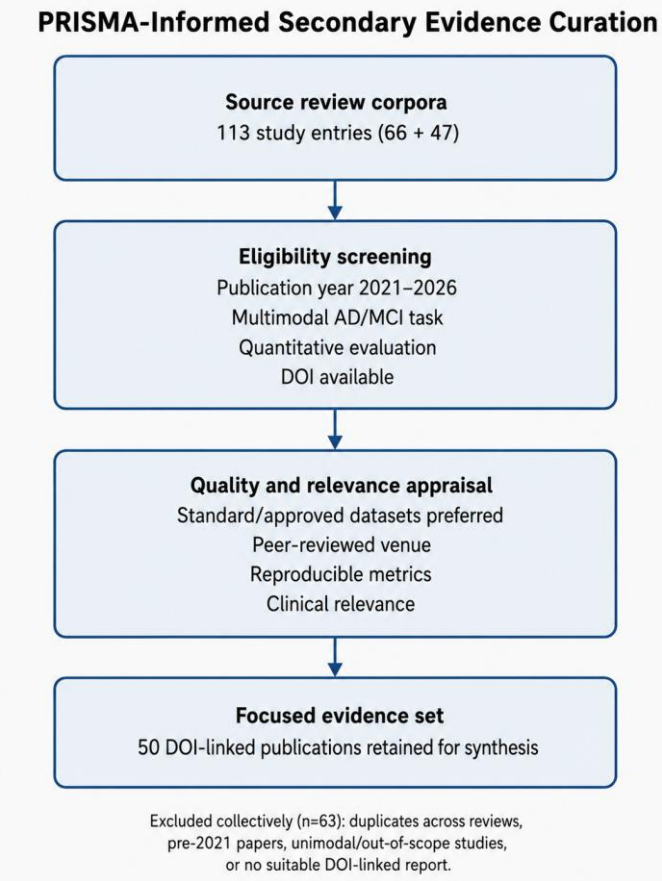


Fig. 1.PRISMA-informed secondary evidence curation. The flow does not claim a new proprietary-database export.

III. Synthesis Of Recent Multimodal Evidence

A. Modality Families and Their Complementarity

Structural MRI is the most common imaging modality because it is widely available and captures regional atrophy, cortical thickness, ventricular enlargement, and hippocampal degeneration. PET adds molecular or metabolic specificity but is expensive, exposes participants to ionizing radiation, and is often missing in real-world records. Cognitive and clinical variables offer direct functional relevance and are comparatively inexpensive, although they may be influenced by language, education, and examiner effects. Genetics and blood or cerebrospinal-fluid biomarkers inform biological susceptibility and pathology but differ in availability, invasiveness, and population portability. Speech, language, gait, retinal, EEG, and other digital phenotypes offer scalable screening pathways, but distribution shift across languages, devices, and settings remains a major concern.

Multimodal complementarity is strongest when each input contributes nonredundant signal. MRI and PET can jointly represent anatomy and pathology; clinical measures contextualize imaging changes; APOE or polygenic features modify prior risk; and longitudinal measures help distinguish stable from progressive MCI. Complementarity can nevertheless become confounding if one modality encodes site identity, diagnostic workflow, or missingness patterns. A model may appear multimodal while relying almost entirely on a single high-signal assessment such as a cognitive score. Therefore, contribution analysis and modality-removal tests are essential.

B. Fusion Taxonomy

Early fusion concatenates normalized raw or engineered features before prediction. It is simple but sensitive to scale, dimensionality, missing inputs, and the assumption that all modalities can be aligned at the same representational level. Intermediate fusion combines learned embeddings and is now dominant because modality-specific encoders can respect distinct geometries. Late fusion averages or learns over modality-specific predictions; it is robust and modular but may miss fine-grained interactions. Hybrid fusion combines multiple levels, sometimes with graph or tensor formulations.

Attention-based fusion improves intermediate fusion by making interactions conditional. Self-attention identifies salient tokens within a modality, while cross-attention uses one modality as a query over another. This allows, for example, an MRI representation to request PET evidence from corresponding disease-relevant patterns or clinical tokens to modulate the interpretation of imaging abnormalities. The strongest evidence is not merely high accuracy but the observed degradation when interaction modules are removed. In a 2024 longitudinal model, adding interaction effects and multimodality improved AUC by 0.027, accuracy by 4.76 percentage points, sensitivity by 5.56 points, and F1 by 6.11 points relative to a simple fusion benchmark.

C. Dataset-Dependent Performance and Generalization

Figure 3 summarizes performance context from the 2026 systematic review. ADNI-based studies achieved the strongest average results because ADNI offers relatively standardized imaging, rich phenotyping, and carefully characterized cohorts. This is scientifically valuable but can create an optimistic environment compared with population-based or multilingual deployment. UK Biobank contains broader population heterogeneity and reported lower risk-prediction performance. DementiaBank systems address a substantially different input space and show that cross-lingual transfer remains difficult. Self-collected datasets often produce very high accuracy, but small samples and single-center acquisition limit generalization.

The gap between internal and external performance is a decisive methodological issue. Cross-validation inside a single cohort does not test scanner, protocol, demographic, prevalence, and workflow shifts simultaneously. Repeated scans from one participant must remain in one split, and all preprocessing, feature selection, imputation, harmonization, and oversampling must be fit only on the training fold. External evaluation should preserve the target cohort's natural missingness and prevalence, with recalibration reported separately from discrimination.

D. Missing Modalities

Complete-case analysis wastes data and can bias the sample toward participants with greater access to specialized testing. Zero filling is easy but turns absence into an artificial signal. Statistical imputation can be useful for low-dimensional clinical features but may underestimate uncertainty. Generative reconstruction may synthesize a missing modality but risks hallucinating disease evidence. More robust strategies include masked tokens, modality dropout, mixture-of-experts routing, product-of-experts latent variables, and set-based fusion that accepts arbitrary subsets. Importantly, missingness may be informative: PET is not missing at random when clinicians order it for diagnostically ambiguous patients. The model should therefore use availability indicators while preventing them from becoming a shortcut through randomized missingness augmentation and deployment-oriented stress tests.

E. Interpretability, Calibration, and Fairness

Attention weights alone are not guaranteed explanations. A credible system should triangulate evidence using 3D Grad-CAM or integrated gradients for imaging, attention rollout for token interactions, SHAP for clinical/genetic variables, and counterfactual modality ablation. Explanations should be evaluated for stability, anatomical plausibility, sensitivity to input perturbation, and consistency with known disease patterns. Calibration is equally important because a model that is accurate but overconfident can be clinically unsafe. Expected calibration error, Brier score, reliability diagrams, and decision-curve analysis should accompany AUC, accuracy, sensitivity, specificity, precision, and F1. Performance and calibration should be stratified by sex, age group, education, ethnicity where ethically and statistically supportable, site, scanner, and language.

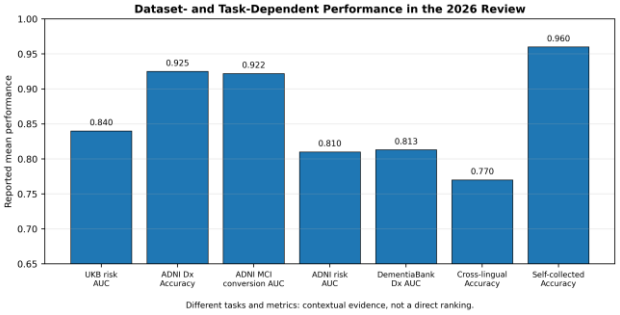


Fig. 2. Evidence-derived benchmark context. Metrics and tasks differ and should not be directly ranked.

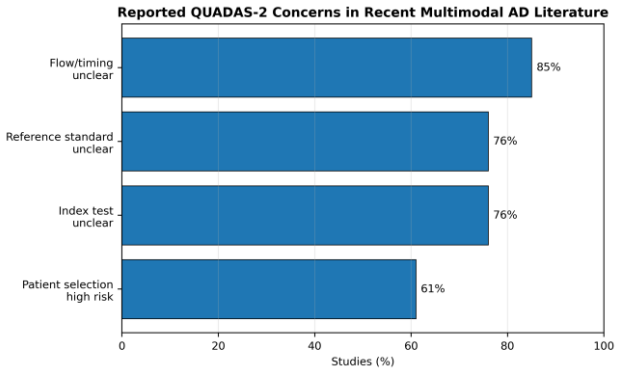


Fig. 3. Reported QUADAS-2 concerns motivating stricter validation and reporting.

Study	**Evidence summary**
Golovanevsky et al., 2022	MRI + clinical + genetics; Cross/self-attention (MADDi). Explicit interaction modeling improved F1 over simpler fusion.
Qiu et al., 2022	Multimodal clinical and imaging; Deep multimodal assessment. Large-scale demonstration of complementary multimodal information.
Velazquez & Lee, 2022	MRI + clinical/biomarker; Ensemble fusion. Focused on early-MCI conversion prediction.
Castellano et al., 2024	3D MRI + amyloid PET; Intermediate deep fusion. Reported 95% accuracy on OASIS-3; requires cohort-aware interpretation.
Bi et al., 2024	Functional + structural imaging; Interpretable multimodal ViT. Shows transformer-based fusion and interpretation.
Wang et al., 2024	Clinical + imaging + genetics; Dual interaction fusion. Independent-test AUC 0.939; interaction ablations improved multiple metrics.
Lin & Washington, 2024	Text + audio; Multimodal deep learning. Demonstrates behavioral modalities and language-related generalization issues.
Yu et al., 2026	Multiple dataset families; Systematic review. 66 studies; performance varies sharply by dataset, task, and validation design.

IV. Proposed Cma-Mfnet Architecture

A. Input Representation

Let the available modalities be M = {MRI, PET, clinical/cognitive, genetics/biomarkers}. For participant i, a_m indicates availability and q_m represents estimated quality. MRI and PET volumes are registered to a common anatomical space, resampled, intensity-normalized, and converted into 3D patch tokens. Clinical variables are

encoded using learned embeddings for categorical fields and normalized continuous features, together with explicit missing-value masks. Genetic or biomarker features are filtered using training-only procedures and embedded as feature tokens. The architecture can be extended with speech or longitudinal tokens without changing the fusion principle.

Each modality encoder E_m maps its input x_m to a token sequence Z_m = E_m(x_m). A 3D Swin Transformer or compact 3D ViT is suitable for imaging; a TabTransformer is suitable for clinical variables; and an MLP-token encoder is suitable for genetics. Positional, modality, visit-time, and acquisition-site embeddings are added only where justified. Site embeddings may be used adversarially for invariance rather than as predictive shortcuts.

B. Intra- and Inter-Modal Attention

Intra-modal blocks refine tokens with residual multi-head self-attention: Z'_m = MSA(LN(Z_m)) + Z_m. Pairwise cross-modal attention then computes C_{m<-n} = softmax(Q_mK_n^T / sqrt(d)) V_n, where queries originate from modality m and keys and values from n. Bidirectional pairs permit asymmetric interactions; MRI attending to PET is not assumed equivalent to PET attending to MRI. A low-rank bottleneck and shared latent query tokens limit quadratic cost.

Rather than concatenate every pair indiscriminately, the network aggregates cross-modal messages through learned residual gates. For modality m, H_m = Z'_m + sum_{n != m} alpha_{mn} C_{m<-n}, where alpha_{mn} is regularized to discourage diffuse, unstable interactions. The interaction graph is masked according to availability and can be sparsified during training. This structure yields interpretable modality-to-modality attention maps and explicit ablation targets.

C. Availability- and Quality-Aware Fusion

The contribution of a modality should depend on what is available and whether it is trustworthy. A gate is therefore calculated as g_m = sigmoid(W_g [pool(H_m) || a_m || q_m] + b_g). The fused representation is F = sum_ma_mg_mpool(H_m) / (sum_ma_mg_m + epsilon). Quality q_m may combine automated MRI motion estimates, PET count statistics, clinical missingness fraction, and laboratory quality flags. Availability masks block attention to absent inputs; learned missing-modality tokens preserve a consistent computational graph.

Modality dropout randomly removes available modalities during training according to deployment-relevant probabilities. This prevents dependence on a single expensive modality and produces subset-specific robustness. A consistency loss encourages predictions from partial subsets to remain close to the full-modality teacher when the removed modality is not decisive. Because a missing modality may legitimately alter uncertainty, consistency is applied to softened distributions and weighted by teacher confidence.

D. Learning Objective

The total objective is L = L_CE + lambda_1 L_contrast + lambda_2 L_consistency + lambda_3 L_calibration + lambda_4 L_sparse. L_CE is class-balanced cross-entropy or focal loss; L_contrast aligns representations of the same participant across modalities while separating participants with different disease states; L_consistency controls full-versus-partial prediction drift; L_calibration penalizes overconfidence using differentiable Brier or focal-calibration terms; and L_sparse regularizes interaction gates. For conversion prediction, a survival-aware auxiliary loss can use time-to-conversion and censoring rather than reducing all follow-up trajectories to a fixed binary label.

E. Outputs and Explanations

The primary head predicts CN, MCI, and AD. Secondary heads distinguish stable versus progressive MCI and optionally estimate continuous cognitive scores. Deep ensembles or Monte Carlo dropout quantify epistemic uncertainty; temperature scaling is fit on an inner validation set. Imaging explanations use 3D Grad-CAM and integrated gradients; cross-modal explanations use attention rollout; tabular contributions use SHAP; and modality-level contribution is measured by gated weight, permutation importance, and performance drop after removal. Explanations are reported for correct and incorrect predictions and audited for site or demographic shortcuts.

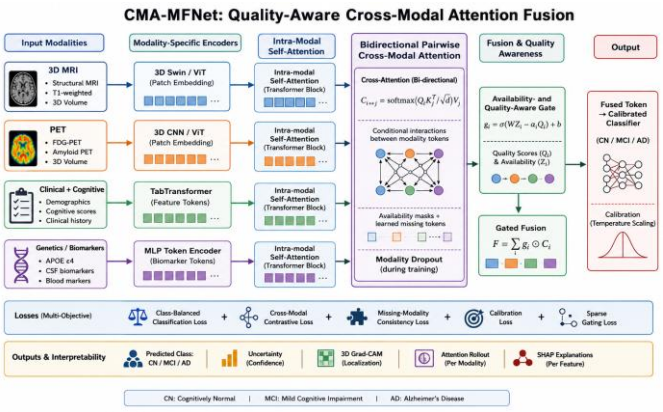


Fig. 4. Proposed CMA-MFNet architecture with cross-modal attention and quality-aware gating.

V. Reproducible Experimental Protocol

A. Datasets and Cohorts

ADNI is proposed as the primary development cohort because it provides MRI, PET, cognitive/clinical measurements, APOE/genetic information, fluid biomarkers, diagnosis, and longitudinal follow-up. OASIS-3 should provide external imaging-rich validation, AIBL should test transfer to an independently acquired biomarker cohort, and NACC should test broader clinical transportability with an imaging subset. DementiaBank/ADReSS may be used in a separate extension for speech-text fusion rather than mixed into the principal imaging benchmark. Dataset licenses, access approvals, diagnostic definitions, visit selection, and cohort flow must be documented.

Primary classification should use CN, MCI, and AD at a defined index visit. The secondary prognosis task should define progressive MCI using conversion within a prespecified horizon such as 36 or 48 months, with stable MCI requiring adequate follow-up. Ambiguous or reverted diagnoses should be handled through sensitivity analyses. Multiple visits from one participant must never cross train, validation, or test boundaries. A site-stratified split is preferable when enough sites are available.

B. Preprocessing

MRI preprocessing should include bias-field correction, skull stripping, affine/nonlinear registration, resampling, intensity normalization, and automated quality control. PET should be co-registered, normalized to an appropriate reference region, and represented as standardized uptake value ratio maps when protocol permits. All template construction, harmonization, feature scaling, feature selection, and imputation must be fit using training data only. ComBat or learned domain adaptation may be evaluated, but external cohorts must remain untouched until final evaluation.

C. Comparators and Ablations

Comparators should include unimodal MRI and PET models; clinical-only gradient boosting and TabTransformer models; early concatenation; late probability averaging; gated multimodal units; multimodal autoencoders; and a transformer using self-attention without cross-modal blocks. Ablations should remove each modality, each directional cross-attention path, quality gating, availability indicators, modality dropout, contrastive loss, consistency loss, and calibration regularization. Parameter-matched baselines are required to separate architectural benefit from model capacity.

D. Training and Model Selection

Nested cross-validation is recommended for development. The outer loop estimates internal generalization; the inner loop selects hyperparameters and calibration temperature. Class imbalance should be addressed through class-balanced sampling or loss weighting, but synthetic oversampling of images is discouraged unless thoroughly validated. Optimization should use early stopping based on a prespecified primary metric, preferably macro-AUC or balanced accuracy. Random seeds, software versions, hardware, preprocessing configurations, and exact split identifiers should be released.

E. Metrics and Statistical Analysis

For multi-class diagnosis, report one-vs-rest and macro-averaged AUC, balanced accuracy, macro-F1, sensitivity, specificity, precision, confusion matrix, and per-class support. For conversion prediction, report AUC, sensitivity, specificity, positive and negative predictive values, calibration slope/intercept, Brier score, and time-dependent AUC when follow-up varies. Use 2,000 stratified bootstrap replicates for confidence intervals. Compare correlated AUCs using DeLong's test and paired classification errors using McNemar's test, with multiplicity correction across planned comparisons. Decision-curve analysis should estimate net benefit across clinically plausible thresholds.

F. Robustness, Fairness, and Missingness Stress Tests

Performance should be reported for all clinically plausible modality subsets: MRI only; MRI+clinical; MRI+PET; MRI+clinical+genetics; and full input. Additional tests should randomly mask modalities at increasing rates and reproduce empirical missingness patterns from external cohorts. Scanner/vendor, site, age, sex, education, and available ethnicity/language subgroups should be analyzed with confidence intervals and interaction tests. Fairness claims require adequate sample size; otherwise, report uncertainty and avoid false reassurance.

G. Leakage Prevention Checklist

The subject, not the scan, is the split unit. Longitudinal visits and derived images remain with the participant. Preprocessing statistics, imputation, harmonization, variable selection, class balancing, and augmentation policies are learned inside each training fold. External cohorts are not used for architecture selection. Diagnostic variables that directly encode the target, such as post-diagnosis treatment or adjudication fields, are excluded. Code should include automated assertions that no participant identifier appears in more than one split.

Table 2. Standard Datasets And Proposed Evaluation Role

Dataset	Modalities	Recommended role
ADNI	MRI, PET, clinical, genetics	Primary development
OASIS-3	MRI, PET, cognition	External validation
AIBL	MRI, PET, clinical	Independent cohort
NACC	Clinical, cognitive, pathology	Heterogeneity audit
DementiaBank/ADReSS	Speech and language	Optional sub-study

VI. Expected Analyses And Reporting Template

The final empirical paper should separate confirmatory endpoints from exploratory analyses. The confirmatory primary endpoint may be macro-AUC for three-class CN/MCI/AD classification on a locked ADNI test partition, followed by external macro-AUC on OASIS-3 or AIBL. A second confirmatory endpoint may be AUC for pMCI versus sMCI. The proposed model should be considered superior only if it improves the prespecified metric with confidence intervals excluding a trivial margin and does not worsen calibration or key subgroup sensitivity.

A complete results package should contain: cohort flow and demographic tables; missingness patterns; internal and external discrimination; calibration plots; confusion matrices; modality-subset curves; ablation results; statistical comparisons; subgroup performance; decision curves; explanation examples; and failure analysis. Reporting only the best fold or a single accuracy value is insufficient. Mean and standard deviation across folds should be accompanied by participant-level confidence intervals where appropriate.

The evidence-derived bars in Figure 3 are not expected results for CMA-MFNet. They are a sanity envelope for planning and interpretation. A model that materially exceeds published averages in an external cohort should trigger a leakage and cohort-composition audit before celebration. In machine learning, an implausibly beautiful score is occasionally a scientific result and quite often a data pipeline asking for adult supervision.

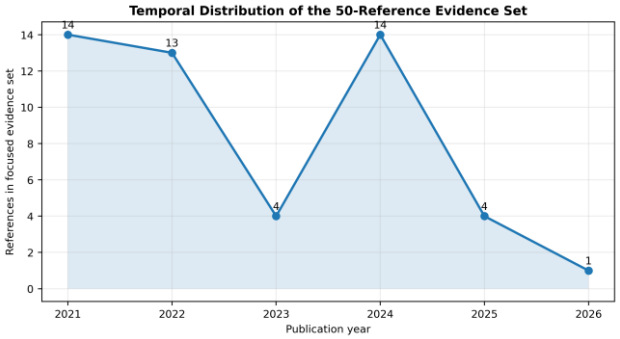


Fig. 5. Publication-year distribution of the focused 50-reference evidence set.

VII. Discussion

The systematic evidence supports multimodal fusion but does not support the claim that any single architecture is universally optimal. Performance depends on the outcome, modality availability, cohort, acquisition, diagnostic criteria, and validation design. Cross-modal attention is a principled mechanism for conditional fusion, yet it can magnify confounding if trained without site-aware auditing and strict leakage control. The proposed quality gate and missing-modality training address practical limitations that are often ignored in complete-case benchmarks.

CMA-MFNet differs from simple attention fusion in four respects. First, it uses directional cross-modal interactions rather than a single concatenated token pool. Second, it conditions contribution on objective quality and availability. Third, it treats missing-modality robustness as a training target. Fourth, it couples discrimination with calibration, uncertainty, and multimodal explanation. These features are expected to be especially useful when PET or genetics is absent and when external cohorts differ from the development cohort.

Clinical translation requires more than retrospective accuracy. A system should define its intended use: triage, referral prioritization, differential diagnosis support, prognosis, or research enrichment. The relevant operating threshold and cost of errors differ by use. Prospective evaluation should measure workflow impact, reader-model interaction, time savings, downstream testing, patient acceptability, and equity. The output should support, not replace, multidisciplinary diagnosis.

The literature also suggests a strategic hierarchy of deployment. MRI+clinical models may offer the best compromise between availability and performance. PET and fluid biomarkers can be added as high-specificity modules when clinically available. Speech or digital phenotypes may provide low-cost screening but require language- and device-specific validation. A set-based architecture with calibrated subset outputs permits one model family to support this tiered pathway without pretending that every patient has every test.

VIII. Limitations

This paper is a research-ready systematic synthesis and model protocol, not a report of newly run experiments. The 50-study evidence set was curated from two broad recent systematic reviews and targeted publisher/DOI verification rather than a fresh export from proprietary Scopus or Web of Science interfaces. Before journal submission, the author team should rerun the exact search strings in institutional databases, export records and citation counts on the submission date, perform dual-reviewer screening, and archive a complete PRISMA-S search log.

Some studies report noncomparable tasks and metrics, so no formal pooled effect was calculated here. The benchmark chart therefore summarizes published task-level means and must not be read as a league table. Several references use abbreviated author lists in the draft bibliography and should be normalized with a reference manager against Crossref or the target journal's metadata. Dataset access, labels, and modality completeness may change; the final implementation must document the exact data freeze.

Cross-modal attention does not guarantee biological validity. Explanations may be unstable, and quality scores may themselves encode site or socioeconomic access. External validation across geography, race/ethnicity, language, scanner vendor, and clinical workflow remains essential. Prospective clinical utility cannot be inferred from retrospective cohorts alone.

IX. Conclusion

Recent evidence shows that multimodal AI can improve AD diagnosis and MCI prognosis by integrating complementary imaging, clinical, molecular, and behavioral information. Yet the field's main bottleneck is no longer the invention of another fusion layer; it is trustworthy generalization under missing data, cohort shift, bias, and uncertain labels. CMA-MFNet addresses this gap through bidirectional cross-modal attention, availability- and quality-aware gating, modality dropout, calibrated uncertainty, and multimodal explanations.

The proposed ADNI-centered and externally validated protocol provides a reproducible path from architecture to credible evidence. A high-quality empirical study should prioritize subject-level leakage prevention, nested model selection, locked external cohorts, ablations, calibration, subgroup analysis, and transparent reporting. Implemented under these constraints, CMA-MFNet is a testable framework for robust Alzheimer's disease classification rather than a decorative attention diagram attached to an optimistic accuracy number.

References

1. Scheltens, P., De Strooper, B., Kivipelto, M., Holstege, H., Chetelat, G., Teunissen, C. E., Cummings, J., & van der Flier, W. M. (2021). Alzheimer's disease.*The Lancet, 397(10284), 1577-1590.* https://doi.org/10.1016/S0140-6736(20)32205-4

2. Alzheimer's Association (2024). 2024 Alzheimer's disease facts and figures.*Alzheimer's & Dementia, 20(5), 3708-3821.*https://doi.org/10.1002/alz.13809

3. Page, M. J., McKenzie, J. E., Bossuyt, P. M., et al. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ, 372, n71.*https://doi.org/10.1136/bmj.n71

4. Rethlefsen, M. L., Kirtley, S., Waffenschmidt, S., et al. (2021). PRISMA-S: An extension to the PRISMA statement for reporting literature searches in systematic reviews. *Systematic Reviews, 10, 39.*https://doi.org/10.1186/s13643-020-01542-z

5. Venugopalan, J., Tong, L., Hassanzadeh, H. R., & Wang, M. D. (2021). Multimodal deep learning models for early detection of Alzheimer's disease stage. *Scientific Reports, 11, 3254.*https://doi.org/10.1038/s41598-020-74399-w

6. Abdelaziz, M., Wang, T., &Elazab, A. (2021). Alzheimer's disease diagnosis framework from incomplete multimodal data using convolutional neural networks.*Journal of Biomedical Informatics, 121, 103863.*https://doi.org/10.1016/j.jbi.2021.103863

7. El-Sappagh, S., Alonso, J. M., Islam, S. M. R., Sultan, A. M., &Kwak, K. S. (2021). A multilayer multimodal detection and prediction model based on explainable artificial intelligence for Alzheimer's disease.*Scientific Reports, 11, 2660.*https://doi.org/10.1038/s41598-021-82098-3

8. Battineni, G., Hossain, M. A., Chintalapudi, N., et al. (2021). Improved Alzheimer's disease detection by MRI using multimodal machine learning algorithms.*Diagnostics, 11(11), 2103.*https://doi.org/10.3390/diagnostics11112103

9. Zhu, X., Suk, H. I., Lee, S. W., &Shen, D. (2021). Canonical feature selection and dual attention-based multimodal learning for Alzheimer's disease diagnosis.*IEEE Transactions on Medical Imaging, 40(7), 1899-1911.* https://doi.org/10.1109/TMI.2021.3077079

10. Beebe-Wang, N., Okon, E., Christensen, T., et al. (2021). A multimodal framework for diagnosis and prognosis of Alzheimer's disease.*IEEE Journal of Biomedical and Health Informatics, 25(5), 1548-1559.* https://doi.org/10.1109/JBHI.2021.3059563

11. Wang, Y., Liu, X., & Yu, C. (2021). Assisted diagnosis of Alzheimer's disease based on deep learning and multimodal feature fusion. *Complexity, 2021, 6626728.*https://doi.org/10.1155/2021/6626728

12. Mahajan, P., & Baths, V. (2021). Acoustic and language based deep learning approaches for Alzheimer's dementia detection from spontaneous speech. *Frontiers in Aging Neuroscience, 13, 623607.*https://doi.org/10.3389/fnagi.2021.623607

13. Ning, Z., Xiao, Q., Feng, Q., Chen, W., & Zhang, Y. (2021). Relation-induced multi-modal shared representation learning for Alzheimer's disease diagnosis. *IEEE Transactions on Medical Imaging, 40(6), 1632-1645.* https://doi.org/10.1109/TMI.2021.3063150

14. Massalimova, A., &Varol, H. A. (2021). Input agnostic deep learning for Alzheimer's disease classification using multimodal MRI images. *Proceedings of the 43rd IEEE EMBC, 2875-2878.*https://doi.org/10.1109/EMBC46164.2021.9629807

15. Golovanevsky, M., Eickhoff, C., & Singh, R. (2022). Multimodal attention-based deep learning for Alzheimer's disease diagnosis.*Journal of the American Medical Informatics Association, 29(12), 2014-2022.* https://doi.org/10.1093/jamia/ocac168

16. Qiu, S., Miller, M. I., Joshi, P. S., et al. (2022). Multimodal deep learning for Alzheimer's disease dementia assessment.*Nature Communications, 13, 3404.*https://doi.org/10.1038/s41467-022-31037-5

17. Dwivedi, S., Goel, T., Tanveer, M., Murugan, R., & Sharma, R. (2022). Multimodal fusion-based deep learning network for effective diagnosis of Alzheimer's disease.*IEEE MultiMedia, 29(2), 45-55.* https://doi.org/10.1109/MMUL.2022.3156471

18. Jia, H., & Lao, H. (2022). Deep learning and multimodal feature fusion for the aided diagnosis of Alzheimer's disease. *Neural Computing and Applications, 34, 19585-19598.* https://doi.org/10.1007/s00521-022-07501-0

19. Jin, D., Zhou, B., Han, Y., et al. (2022). A hybrid deep learning method for early and late mild cognitive impairment diagnosis with incomplete multimodal data.*Frontiers in Neuroinformatics, 16, 843566.*https://doi.org/10.3389/fninf.2022.843566

20. Lu, B., Li, H. X., Chang, Z. K., et al. (2022). A two-stage model for predicting mild cognitive impairment to Alzheimer's disease conversion.*Frontiers in Aging Neuroscience, 14, 826622.*https://doi.org/10.3389/fnagi.2022.826622

21. Velazquez, M., & Lee, Y. (2022). Multimodal ensemble model for Alzheimer's disease conversion prediction from early mild cognitive impairment subjects.*Computers in Biology and Medicine, 151, 106201.*https://doi.org/10.1016/j.compbiomed.2022.106201

22. Bi, X., Hu, X., Wu, H., & Wang, Y. (2022). Multimodal data analysis of Alzheimer's disease based on clustering evolutionary random forest.*IEEE Journal of Biomedical and Health Informatics, 26(4), 1635-1646.* https://doi.org/10.1109/JBHI.2022.3151084

23. Ilias, L., &Askounis, D. (2022). Explainable identification of dementia from transcripts using transformer networks.*IEEE Journal of Biomedical and Health Informatics, 26(8), 4153-4164.*https://doi.org/10.1109/JBHI.2022.3172479

24. Ilias, L., Askounis, D., &Psarras, J. (2022). Multimodal detection of Alzheimer's disease from speech and text.*Frontiers in Aging Neuroscience, 14, 830943.*https://doi.org/10.3389/fnagi.2022.830943

25. Kumar, P., et al. (2022). Multimodal speech and language analysis for dementia detection.*Sensors, 22(23), 9311.*https://doi.org/10.3390/s22239311

26. Sharma, S., &Mandal, P. K. (2022). A comprehensive report on machine learning-based early detection of Alzheimer's disease using multimodal neuroimaging data.*ACM Computing Surveys, 55(2), Article 34.* https://doi.org/10.1145/3492865

27. Veitch, D. P., Weiner, M. W., Aisen, P. S., et al. (2022). Using the Alzheimer's Disease Neuroimaging Initiative to improve early detection, diagnosis, and treatment of Alzheimer's disease. *Alzheimer's & Dementia, 18(4), 824-857.* https://doi.org/10.1002/alz.12422

28. Odusami, M., Maskeliunas, R., Damasevicius, R., &Misra, S. (2023). Pixel-level fusion approach with vision transformer for early detection of Alzheimer's disease.*Electronics, 12(5), 1218.*https://doi.org/10.3390/electronics12051218

29. Zhang, Y., et al. (2023). Transformer-based multimodal fusion of MRI and PET for Alzheimer's disease classification.*Proceedings of IEEE ISBI 2023.*https://doi.org/10.1109/ISBI53787.2023.10230577

30. Fan, Y., et al. (2023). Multimodal graph reasoning for Alzheimer's disease diagnosis.*IEEE Transactions on Neural Systems and Rehabilitation Engineering, 31, 4510-4521.* https://doi.org/10.1109/TNSRE.2023.3337533

31. Ortiz-Perez, D., Ruiz-Ponce, P., Tomas, D., et al. (2023). A deep learning-based multimodal architecture to predict signs of dementia.*Neurocomputing, 548, 126413.*https://doi.org/10.1016/j.neucom.2023.126413

32. Odusami, M., Maskeliunas, R., Damasevicius, R., &Misra, S. (2024). Machine learning with multimodal neuroimaging data to classify stages of Alzheimer's disease: A systematic review and meta-analysis. *Cognitive Neurodynamics, 18, 775-794.* https://doi.org/10.1007/s11571-023-09993-5

33. Castellano, G., Esposito, A., Lella, E., Montanaro, G., &Vessio, G. (2024). Automated detection of Alzheimer's disease: A multi-modal approach with 3D MRI and amyloid PET. *Scientific Reports, 14, 5210.*https://doi.org/10.1038/s41598-024-56001-9

34. Bi, Y., Abrol, A., Fu, Z., & Calhoun, V. D. (2024). A multimodal vision transformer for interpretable fusion of functional and structural neuroimaging data.*Human Brain Mapping, 45(17), e26783.*https://doi.org/10.1002/hbm.26783

35. Yu, Q., Ma, Q., Da, L., et al. (2024). A transformer-based unified multimodal framework for Alzheimer's disease assessment.*Computers in Biology and Medicine, 180, 108979.*https://doi.org/10.1016/j.compbiomed.2024.108979

36. Elazab, A., Wang, C., Abdelaziz, M., et al. (2024). Alzheimer's disease diagnosis from single and multimodal data using machine and deep learning models: Achievements and future directions. *Expert Systems with Applications, 255, 124780.*https://doi.org/10.1016/j.eswa.2024.124780

37. Ding, H., et al. (2024). Artificial intelligence for speech-based Alzheimer's disease detection: A systematic survey. *Artificial Intelligence Review, 57, Article 318.* https://doi.org/10.1007/s10462-024-10961-6

38. Xue, C., et al. (2024). AI-based differential diagnosis of dementia using multimodal clinical data.*Nature Medicine, 30, 2205-2215.* https://doi.org/10.1038/s41591-024-03118-z

39. Zhang, X., et al. (2024). Auto-fusion learning for multimodal Alzheimer's disease diagnosis.*Computers in Biology and Medicine, 179, 108740.*https://doi.org/10.1016/j.compbiomed.2024.108740

40. Wang, Z., et al. (2024). Predicting long-term progression of Alzheimer's disease using a multimodal deep learning model incorporating interaction effects.*Journal of Translational Medicine, 22, 251.*https://doi.org/10.1186/s12967-024-05025-w

41. Lee, H., et al. (2024). A robust multimodal machine-learning marker for Alzheimer's disease across clinical cohorts.*eClinicalMedicine, 74, 102725.* https://doi.org/10.1016/j.eclinm.2024.102725

42. Lee, J., et al. (2024). Prediction of dementia conversion using multimodal deep learning.*Scientific Reports, 14, 9238.*https://doi.org/10.1038/s41598-024-60134-2

43. Poor, S. Z. H., et al. (2024). A cross-transformer approach for multimodal Alzheimer's disease classification.*Computers in Biology and Medicine, 184, 109199.*https://doi.org/10.1016/j.compbiomed.2024.109199