Journal of
Dementia and Alzheimer's Disease

MDPI

Systematic Review

Counterfactual, Longitudinal, and Multimodal Explainable AI for MRI-Based Alzheimer's Diagnosis: A Structured Review

Ramisa Farha [1], Blessing Ojeme [1], Fahmi Khalifa [2] and Md Mahmudur Rahman [1,*]

[1] Department of Computer Science, Morgan State University, Baltimore, MD 21251, USA
[2] Department of Electrical and Computer Engineering, Morgan State University, Baltimore, MD 21251, USA
* Correspondence: md.rahman@morgan.edu

Abstract

Background/Objectives: Alzheimer's disease (AD) is a progressive neurodegenerative disorder for which MRI-based AI systems are increasingly used for diagnosis and prognosis. However, many published approaches remain misaligned with the requirements of trustworthy clinical use. Predicted risks are often poorly calibrated, explanations are frequently limited or non-actionable, guideline-aligned reporting is uncommon, and longitudinal prediction is inconsistently evaluated. In this paper, we conduct a PRISMA-guided structured review with scoping-style breadth of MRI-centric AI methods for AD diagnosis. This design supports a theme-based synthesis across heterogeneous study designs and is intended to summarize the current evidence base and derive practical design requirements for next-generation, clinically oriented pipelines that integrate calibrated staging, explainable outputs, and longitudinal risk modeling. **Methods:** Searches were conducted across Scopus, PubMed/PMC, and arXiv/bioRxiv (2014–2026; English; human AD/MCI imaging) and were supplemented by backward and forward snowballing. These searches yielded 2460 records. After deduplication, screening, and full-text eligibility assessment, 90 papers were included in the final synthesis. The included literature was organized into thematic streams spanning counterfactual reasoning and explainable AI (XAI), vision–language approaches for report and caption generation, longitudinal and survival-style modeling, and multimodal fusion and transformer-based methods combining MRI with clinical variables and other biomarkers. Vision–language methods were considered together with retrieval-augmented paradigms. **Results:** Key findings are that the field has shifted toward transformer architectures and multimodal fusion and shows increased interest in richer explanation mechanisms. Nevertheless, calibration metrics and robustness checks are inconsistently reported, external site-held-out validation and subgroup analyses remain relatively uncommon, and guideline-aligned structured reporting with explicit numeric provenance is rare. Vision–language and retrieval-augmented reporting methods are far more mature in general radiology than in AD MRI, highlighting a translational opportunity. **Conclusions:** Based on these findings, we recommend standardized reporting of classification calibration and longitudinal risk calibration, stronger site-held-out validation with subgroup robustness evaluation, clinically meaningful counterfactuals, and guideline-aligned reporting with reproducible numeric provenance embedded within reproducible pipelines.

Check for updates

Academic Editor: Hualou Liang

Received: 5 January 2026
Revised: 11 March 2026
Accepted: 29 April 2026
Published: 19 May 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

Keywords: Alzheimer's disease; structural MRI; systematic review; PRISMA; calibration; explainable AI; counterfactuals; vision–language; longitudinal modeling; multimodal fusion; transformers

1. Introduction

Alzheimer's disease (AD) is the leading cause of dementia, responsible for around 60–70% of all dementia cases globally. In 2021, more than 55 million people worldwide were living with dementia, with nearly 10 million new cases per year, causing a growing clinical and economic burden [1]. In the United States, an estimated 7.2 million adults aged $\geq$65 are projected to be living with Alzheimer's dementia in 2025, and total payments for individuals with Alzheimer's or other dementias are predicted to reach $384 billion in 2025 [2]. Accurate and early diagnosis is important since it determines intervention opportunities, trial eligibility, and care planning.

While neuropathologic examination (post-mortem) has traditionally been used to determine AD pathology, modern clinical procedures increasingly incorporate cognitive assessments, biomarkers, and imaging. MRI-based diagnosis and prognosis of Alzheimer's disease (AD) has rapidly adopted machine learning and deep learning to support earlier detection, risk stratification, and monitoring across the AD continuum, including cognitively normal (CN), mild cognitive impairment (MCI), and AD stages. Structural MRI remains essential for identifying neurodegeneration and atrophy patterns, and biomarker frameworks like AT(N) help standardize biological evidence for study and clinical translation. These pathways generate multimodal data (Magnetic Resonance Imaging (MRI), Positron Emission Tomography (PET) and Cerebrospinal Fluid (CSF) biomarkers, and clinical scores) on a large scale, motivating computational decision support. Consequently, AI methods for MRI-based AD diagnosis have evolved from predictive CNNs to explainable and multimodal models, including counterfactual reasoning, longitudinal risk forecasting, and fusion/transformer paradigms. However, many published systems remain misaligned with trustworthy clinical use, motivating a structured synthesis of what has been validated, what remains inconsistent, and what design requirements should guide next-generation clinically oriented pipelines.

This review synthesizes the landscape of calibrated and explainable AI methods for MRI-based AD diagnosis, focusing on four complementary streams highlighted in the title: (i) counterfactual reasoning and broader explainable AI (XAI), (ii) vision–language approaches for report/caption generation including retrieval-augmented paradigms, (iii) longitudinal modeling and trajectory forecasting, and (iv) multimodal methods that integrate MRI with clinical variables, PET, and other biomarkers. We emphasize evaluation practices that matter for trustworthy deployment, including calibration reporting, external validation, robustness and stress-testing, and subgroup analyses, and we identify recurring gaps that limit real-world adoption.

This study followed a Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA)-guided selection protocol and included 90 papers in the final synthesis. PRISMA was used to transparently report identification, de-duplication, screening, eligibility assessment, and inclusion through a standardized checklist and flow diagram, improving reproducibility of the selection process; because our synthesis is broad and theme-based (counterfactual explanation, longitudinal forecasting, and multimodal MRI-based approaches) rather than meta-analytic, this study also aligns with scoping-review reporting guidance (PRISMA-ScR) [3,4]. The database search was intentionally broad to capture both secondary evidence and primary modeling studies across counterfactual

explanation, longitudinal forecasting, and multimodal MRI-based approaches. Given that the field contains substantial synthesis work, systematic reviews and surveys remain a large portion of the included studies; however, primary modeling papers were also included using a transparent enrichment rule to ensure coverage of clinical-translation themes such as calibration reporting, counterfactual reasoning, longitudinal modeling, and vision–language/reporting.

Several recent surveys and systematic reviews have summarized AI-based Alzheimer's detection using MRI and broader neuroimaging, primarily organizing the field by model families, datasets, and reported diagnostic performance [5–10]. Complementary interpretability-focused reviews discuss common explainability toolkits such as saliency or attribution, LIME/SHAP-style explanations and their limitations for clinical use [11–14]. More specialized syntheses also highlight the rapid growth of multimodal fusion and transformer-based integration [15,16], as well as longitudinal learning challenges and progression-focused modeling [17,18], and the need for reproducibility or generalizability evaluation [19].

This review aims to move beyond theme-by-theme summaries by providing a single, clinically translation-oriented comparative framework for MRI-based Alzheimer's AI. Instead of treating calibration, explainability, multimodal learning, vision–language reporting, and longitudinal modeling as separate sections, we organize the literature around four deployment-relevant requirements: (i) probability calibration and uncertainty reporting, (ii) actionable explanations with emphasis on counterfactual reasoning, (iii) guideline-aligned communication and vision–language/reporting paradigms, and (iv) longitudinal risk and trajectory modeling. We further assess how studies support trustworthy deployment by examining evaluation practices such as external validation, robustness under distribution shift, and subgroup analyses, and we consolidate study characteristics using a consistent schema to improve cross-paper comparability. The goal is not only to synthesize methods, but to provide practical guidance for designing and evaluating clinically usable MRI-based Alzheimer's systems.

2. Methods

We conducted a structured literature review using three primary bibliographic sources: Scopus, PubMed (including PubMed Central), and preprint servers (arXiv/bioRxiv). To improve recall and reduce database-specific bias, we also performed backward and forward snowballing from seed papers identified during screening. The a priori time window spanned 2014–2026. Searches were restricted to English-language records; human-subject filters were applied where available.

Following PRISMA 2020 guidance, we designed database queries around three concept families: disease, imaging modality, and AI method family. Terms within each family were joined using OR, and families were combined using AND. Database-specific syntax (phrase matching, truncation, field tags, and subject-area filters) was used where supported.

This review combines scoping-style breadth with PRISMA-guided screening and reporting to enable a theme-based synthesis across different study designs and outcomes. Therefore, we performed a narrative thematic synthesis instead of a meta-analysis.

To ensure transparent inclusion of both secondary evidence (reviews) and primary modeling studies, we executed a two-part strategy: (i) a broad primary search that did not restrict document type (capturing reviews and primary studies together), and (ii) a targeted primary-study enrichment step using an explicit rule described below to ensure coverage of emerging themes such as vision–language/reporting, calibration reporting, and counterfactual modeling that can be underrepresented by keyword indexing.

Disease: (Alzheimer* OR ‘‘Alzheimer’s disease’’ OR AD OR ‘‘mild cognitive impairment’’ OR MCI)
Modality: (MRI OR ‘‘magnetic resonance imaging’’ OR ‘‘structural MRI’’ OR sMRI)
AI methods: (‘‘deep learning’’ OR ‘‘machine learning’’ OR ‘‘artificial intelligence’’ OR ‘‘neural network’’ OR CNN OR transformer* OR ViT OR ‘‘vision-language’’ OR multimodal OR fusion OR ‘‘survival analysis’’ OR longitudinal OR calibration OR counterfactual* OR explainab*)

Primary Scopus query (as executed; all document types):

```
TITLE-ABS-KEY ( alzheimer* OR "alzheimer’s disease" OR "mild cognitive impairment" OR mci )
AND TITLE-ABS-KEY ( mri OR "magnetic resonance imaging" OR "structural mri" OR smri )
AND TITLE-ABS-KEY ( "deep learning" OR "machine learning" OR "artificial intelligence"
  OR "neural network" OR cnn OR transformer* OR vit OR "vision-language"
  OR multimodal OR fusion OR "survival analysis" OR longitudinal
  OR calibration OR counterfactual* OR explainab* )
AND PUBYEAR > 2013 AND PUBYEAR < 2027
AND LANGUAGE ( english )
AND ( LIMIT-TO ( SUBJAREA, "NEUR" ) OR LIMIT-TO ( SUBJAREA, "MEDI" )
   OR LIMIT-TO ( SUBJAREA, "COMP" ) OR LIMIT-TO ( SUBJAREA, "ENGI" ) )
AND ( LIMIT-TO ( EXACTKEYWORD, "Alzheimer Disease" )
   OR LIMIT-TO ( EXACTKEYWORD, "Mild Cognitive Impairment" )
   OR LIMIT-TO ( EXACTKEYWORD, "Neuroimaging" )
   OR LIMIT-TO ( EXACTKEYWORD, "Magnetic Resonance Imaging" )
   OR LIMIT-TO ( EXACTKEYWORD, "Artificial Intelligence" )
   OR LIMIT-TO ( EXACTKEYWORD, "Deep Learning" )
   OR LIMIT-TO ( EXACTKEYWORD, "Machine Learning" )
   OR LIMIT-TO ( EXACTKEYWORD, "Dementia" )
   OR LIMIT-TO ( EXACTKEYWORD, "Neurodegenerative Diseases" )
   OR LIMIT-TO ( EXACTKEYWORD, "Multimodal Imaging" )
   OR LIMIT-TO ( EXACTKEYWORD, "Image Processing" )
   OR LIMIT-TO ( EXACTKEYWORD, "Image Analysis" ) )
```

Targeted primary-study enrichment rule (transparent selection). In addition to the broad primary search above, we included **primary modeling papers** using the following pre-specified rule: a primary paper was eligible if it (a) used MRI for AD/MCI diagnosis, prognosis, or related tasks, and (b) introduced or empirically evaluated at least one of the review’s target clinical-translation themes such as calibration (ECE, Brier score, reliability analysis), explainability/counterfactuals (counterfactual generation, causal/attribution reasoning beyond saliency-only reporting), longitudinal/trajectory or survival-style modeling, or vision–language/reporting (captioning, report generation, retrieval-augmented reporting, VQA). Candidate primary papers were obtained through (i) snowballing from included reviews and seminal primary works, and (ii) a focused keyword pass described below.

```
TITLE-ABS-KEY ( alzheimer* OR "alzheimer’s disease" OR mci )
AND TITLE-ABS-KEY ( mri OR "magnetic resonance imaging" OR "structural mri" )
AND TITLE-ABS-KEY ( calibration OR "expected calibration error" OR ece OR brier
  OR counterfactual* OR "causal" OR explainab* OR interpretab*
  OR "vision-language" OR caption* OR report* OR "retrieval augmented" OR rag
  OR longitudinal OR survival OR "time-to-event" )
AND PUBYEAR > 2013 AND PUBYEAR < 2027
AND LANGUAGE ( english )
```

PubMed/PMC and arXiv/bioRxiv searches were aligned to the same concept families, adapting field tags and filters to each platform (for instance, MeSH terms in PubMed and category filters in arXiv/bioRxiv). Additionally, we performed backward/forward snowballing to capture relevant papers not indexed uniformly or not surfaced by database keywording.

Inclusion.

(i) Papers focused on AD, MCI, or the AD continuum that include MRI; multimodal works were eligible if MRI was included.
(ii) **Secondary evidence:** systematic reviews, surveys, scoping reviews, and meta-analyses related to AI/ML/DL for MRI-based AD diagnosis/prognosis, including calibration, interpretability, multimodal fusion, longitudinal modeling, and vision–language/reporting.
(iii) **Primary studies:** empirical modeling papers that evaluate MRI-based AD/MCI tasks and contribute to at least one target clinical-translation theme (calibration, explainability/counterfactuals, longitudinal/survival modeling, or vision–language/reporting), as specified in the targeted enrichment rule.
(iv) Sufficient methodological detail to support structured extraction (task, model family, modalities, cohort/dataset, and evaluation setup).

Exclusion.

(i) Non-MRI domains; PET-only, genetics-only, or biomarker-only work without MRI.
(ii) Editorials, opinion pieces, short news items, or records without extractable methods/results.
(iii) General-purpose computer-vision methods not applied to AD/dementia neuroimaging.
(iv) Full text inaccessible after reasonable effort, near-duplicates, or substantial overlap with an already-included entry, retaining the most complete version.

Screening proceeded in two stages: (1) title/abstract screening and (2) full-text assessment. Counts were logged to support a PRISMA 2020 flow diagram (Figure 1) and are summarized in Table 1.

Table 1. Study selection workflow and PRISMA counts.

Stage	Count(s)	Notes
Identification	2460	Records identified across Scopus, PubMed/PMC, arXiv/bioRxiv, and snowballing.
Deduplication	780 removed; 1680 remaining	Duplicates removed to obtain unique records for screening.
Screening	1310 excluded	Title/abstract screening excluded records due to scope mismatch (non-AD, non-MRI, not AI-method-focused), non-retrievable short formats, or insufficient relevance.
Retrieval	370 sought; 42 not retrieved	Reports sought for full-text retrieval.
Eligibility	328 assessed; 238 excluded	Full texts (or detailed preprints) assessed; exclusions included PET-only/no MRI, insufficient extractable methods/results, out-of-scope focus after full-text check, overlap/duplicates, or unavailable full text.
Inclusion	90 included	Papers retained for the final synthesis.

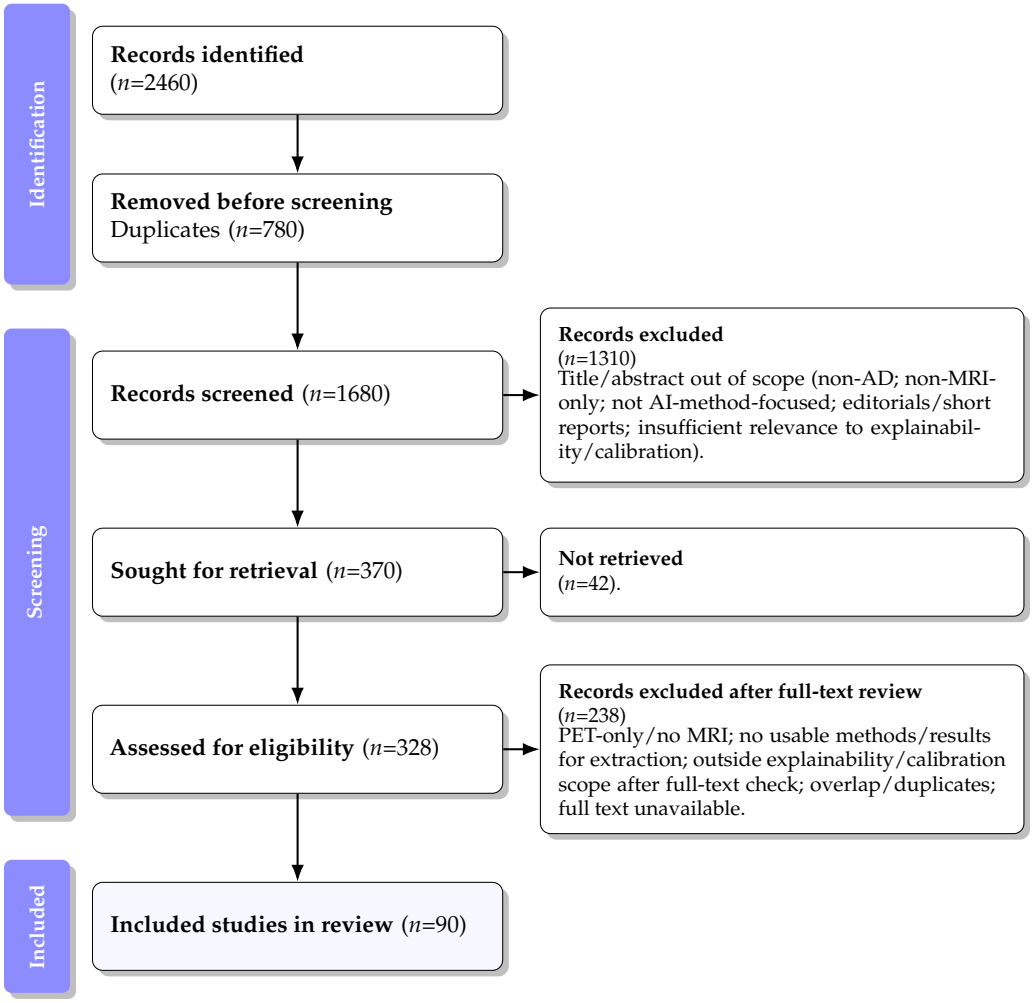


Figure 1. PRISMA 2020 flow diagram of the study selection process.

For each included paper, we extracted:

(i) Bibliographic data (authors, year, venue).
(ii) Study type (secondary evidence vs. primary modeling study).
(iii) Cohort and dataset coverage when reported including ADNI, OASIS/OASIS-3, hospital cohorts.
(iv) Modalities and preprocessing (such as bias correction, skull stripping, registration) when discussed.
(v) Tasks (diagnosis, prognosis/progression, segmentation, report/caption generation, retrieval/VQA, etc.).
(vi) Model family (CNN, transformer, survival model, graph-based, fusion architecture).
(vii) Evaluation reporting, including external validation and calibration when available such as ECE, Brier score, reliability curves.
(viii) Explainability evidence (saliency, counterfactuals, exemplar retrieval, robustness or stress-testing).
(ix) Fairness/subgroup analysis (sex, age, genotype) when reported.
(x) Code and data availability statements when provided.

Given that the included studies span heterogeneous task families, including diagnosis/staging, explainable methods, report/caption generation, and longitudinal prediction, and often lack the reporting required for standardized risk-of-bias scoring, we did not apply a formal tool such as PROBAST or QUADAS-2. Instead, we extracted a set of methodologi-

cal quality indicators as pragmatic proxies to contextualize the strength and generalizability of evidence across the included studies.

For each study, we recorded the following indicators using predefined criteria (binary or ordinal where applicable).

These indicators (Table 2) were not used to exclude studies or compute a single aggregate quality score; rather, they were used to interpret results with appropriate caution, highlight common sources of fragility such as limited external validation and incomplete calibration reporting, and derive design requirements for clinically oriented MRI-AI pipelines.

Table 2. Study-level robustness indicators.

Validation setting Internal split/cross-validation vs. external site-held-out evaluation or multicenter testing.	**Calibration reporting** Whether calibration was assessed, such as ECE/Brier/reliability curves, and whether post hoc calibration, such as temperature scaling, was used.
Domain shift handling Use of harmonization, scanner/site adjustment, domain adaptation, or robustness testing under distribution shift.	**Reproducibility signals** Availability of code, trained weights, detailed preprocessing, and sufficient implementation details to enable replication.

Common risks that could not be consistently quantified across papers included potential information leakage, for instance slice-level rather than subject-level splits, inconsistent handling of repeated visits, incomplete reporting of preprocessing and hyperparameters, and limited subgroup analyses. Where these issues were explicitly mentioned or strongly implied, we noted them qualitatively during synthesis.

Given methodological heterogeneity across the included evidence base, a quantitative meta-analysis was not feasible. We therefore performed a narrative synthesis. Papers were coded by study type (secondary evidence vs. primary studies) and by dominant theme aligned with the title: counterfactual/XAI, vision–language/reporting (including retrieval-augmented paradigms), longitudinal forecasting, and multimodal fusion/transformer approaches. Descriptive statistics (counts per theme and per year) were visualized using `pgfplots`.

We followed PRISMA 2020 recommendations for documenting information sources, eligibility criteria, and study selection. The PRISMA counts are summarized in Figure 1. Database-specific search strings (including platform-specific adaptations of the core query) can be provided in an appendix. The review was not prospectively registered.

3. Results: Included Evidence Base and Thematic Synthesis

3.1. Included Studies Overview (n = 90)

This review includes 90 papers, spanning both secondary evidence (systematic reviews, surveys, scoping reviews, and meta-analyses) and selected primary modeling papers. Primary studies were included under a transparent enrichment rule to represent emerging directions and evaluation practices that are not consistently covered in secondary literature (e.g., MRI vision–language/reporting, counterfactual generation, calibration reporting, and longitudinal prediction). Accordingly, the following overview summarizes modalities, tasks, datasets, and evaluation practices as reported across the literature.

3.1.1. Modalities

Across the 90 papers, structural MRI (predominantly T1-weighted, often MPRAGE) is the central modality for characterizing cortical and subcortical atrophy patterns relevant to AD diagnosis and staging. Many papers also discuss complementary MRI sequences,

including FLAIR for white-matter hyperintensities (WMH) and small-vessel disease markers, and resting-state fMRI for functional connectivity biomarkers; diffusion MRI and other sequences appear less consistently depending on cohort availability and study scope. In multimodal pipelines, MRI is frequently integrated with PET (amyloid/FDG), CSF biomarkers, APOE genotype, and clinical/cognitive features (MMSE, CDR), reflecting a broader trend toward multimodal phenotyping.

3.1.2. Tasks

The included literature spans a range of clinical and methodological tasks:

i. Diagnostic classification (CN vs. AD, CN/MCI/AD, and finer-grained staging).
ii. Prognosis and longitudinal prediction (conversion risk, time-to-progression, and trajectory modeling across CN→MCI→AD).
iii. Quantification/segmentation of WMH and related lesion patterns, as well as ROI-based atrophy quantification.
iv. Explainability analysis, including saliency/feature-importance approaches and counterfactual reasoning where available.
v. Vision–language tasks for neuroimaging reporting (captioning/report generation, retrieval-augmented summarization, and report simplification), noting that this direction is substantially more mature in general radiology than in AD MRI.

3.1.3. Datasets

ADNI is the most frequently referenced dataset across the included studies, commonly discussed alongside OASIS/OASIS-3 and smaller single-center cohorts. Multimodal and longitudinal works additionally reference biomarker-linked datasets enabling PET/CSF integration and progression labeling. External validation practices remain mixed: while some papers emphasize cross-cohort generalization or site-held-out evaluation, many reported results are still limited to single-dataset settings, underscoring persistent generalizability concerns.

3.1.4. Metrics and Evaluation

For classification, accuracy and ROC-AUC are the most commonly reported metrics, frequently accompanied by F1-score, sensitivity/specificity, and precision–recall measures. However, across the included studies, classification probability calibration reporting (reliability diagrams, expected calibration error (ECE), and Brier score) is less consistently reported than discrimination metrics, limiting the clinical interpretability of predicted probabilities. In this case, longitudinal risk calibration evaluates whether predicted event risks at given time horizons match observed event rates over time, while classification probability calibration evaluates whether predicted class probabilities match observed outcome rates. For longitudinal and survival-style modeling, studies typically report the C-index and time-dependent AUC; in contrast, longitudinal risk calibration of time-to-event risk estimates (e.g., horizon-specific calibration curves or observed–expected agreement over time, and time-dependent or integrated Brier-type scores) is less routinely emphasized. Specifically, the ROC-AUC, C-index, and time-dependent AUC quantify discrimination rather than evaluating whether predicted probabilities or risks over time are numerically well calibrated.

Explainability-focused work also differs widely in evaluation practice. Many rely primarily on qualitative visualizations, while others incorporate quantitative criteria such as fidelity/infidelity scores, perturbation or deletion/insertion tests, robustness or stress-testing, and plausibility constraints. This heterogeneity in evaluation protocols reduces

direct comparability across explanation methods and makes it difficult to establish consistent evidence standards.

3.2. Study Characteristics

Table 3 summarizes all 90 PRISMA-included papers using a consistent schema (type/family, method cue, task, modality, dataset, and trust reporting). Figures 2–4 provide complementary summaries of the included studies: the overall theme-level composition, the overlap of external validation and multimodal design, and the year-wise distribution of themes across 2014–2026. Three patterns stand out. First, the primary-study set remains strongly MRI-anchored (predominantly T1-weighted sMRI), while multimodal fusion becomes most common when the objective shifts toward earlier detection, conversion prediction, or richer staging (MRI+PET, MRI+genetics, MRI+MEG, MRI+fMRI, and imaging+text). Second, calibration reporting is rare: even studies that emphasize explainability or clinical trust typically omit probability calibration metrics (C0), while review/meta-analysis/commentary papers are NA by definition. Third, external validation is inconsistent: only a subset of primary papers evaluate across cohorts (for example, ADNI→AIBL/OASIS/NACC or multicenter settings), and headline accuracies should be interpreted cautiously unless split protocols and leakage controls are explicitly documented. Additionally, Figure 4 highlights the post-2022 increase in transformer/fusion and counterfactual/XAI studies.

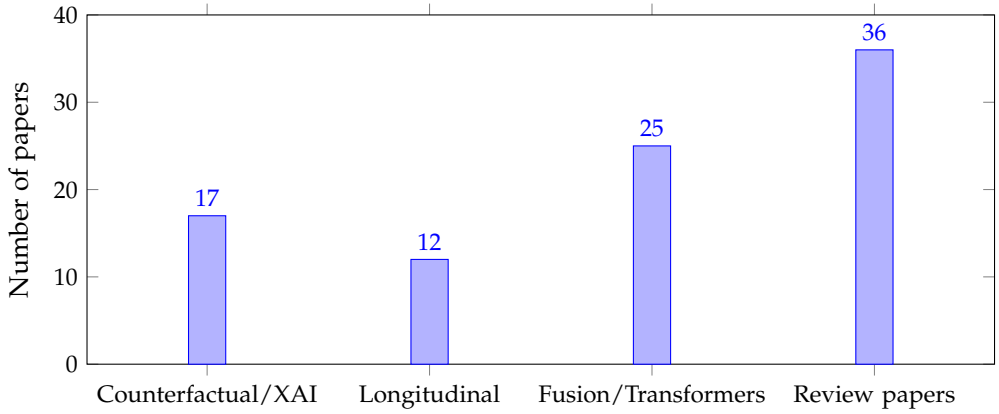


Figure 2. Distribution by dominant theme across the 90 PRISMA-included papers, showing counts for counterfactual/XAI, longitudinal modeling, fusion/transformer and other deep-learning baselines, and review papers.

Figure 3 shows how often two deployment-relevant choices appear together in the calibration-applicable primary subset: external validation (E) and multimodal design (M). Because no study in this subset reported calibration metrics (C1 = 0), the figure summarizes only the E/M overlap. Most studies fall into the "neither" group (25/44), meaning they report neither external validation nor multimodal inputs. Smaller groups report multimodal design only (10/44) or external validation only (6/44), and only a few report both (E+M: 3/44). We coded external validation (E1) when a study evaluated beyond an internal split; for example, cross-cohort testing such as ADNI→AIBL/OASIS/NACC, site-held-out testing, or multicenter evaluation. We coded multimodal design (M1) when MRI was combined with at least one non-MRI signal; for example, PET, CSF or other biofluid biomarkers, clinical or cognitive measures, demographics, or genetics such as APOE.

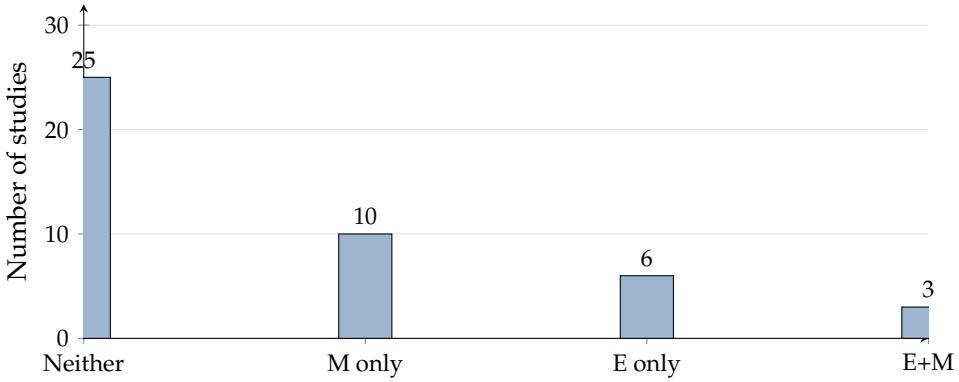


Figure 3. Overlap of external validation (E) and multimodal design (M) in the calibration-applicable primary subset ($n = 44$; reviews and NA/CNA excluded). Because no study in this subset reported calibration metrics (C1), only the E/M overlap is visualized.

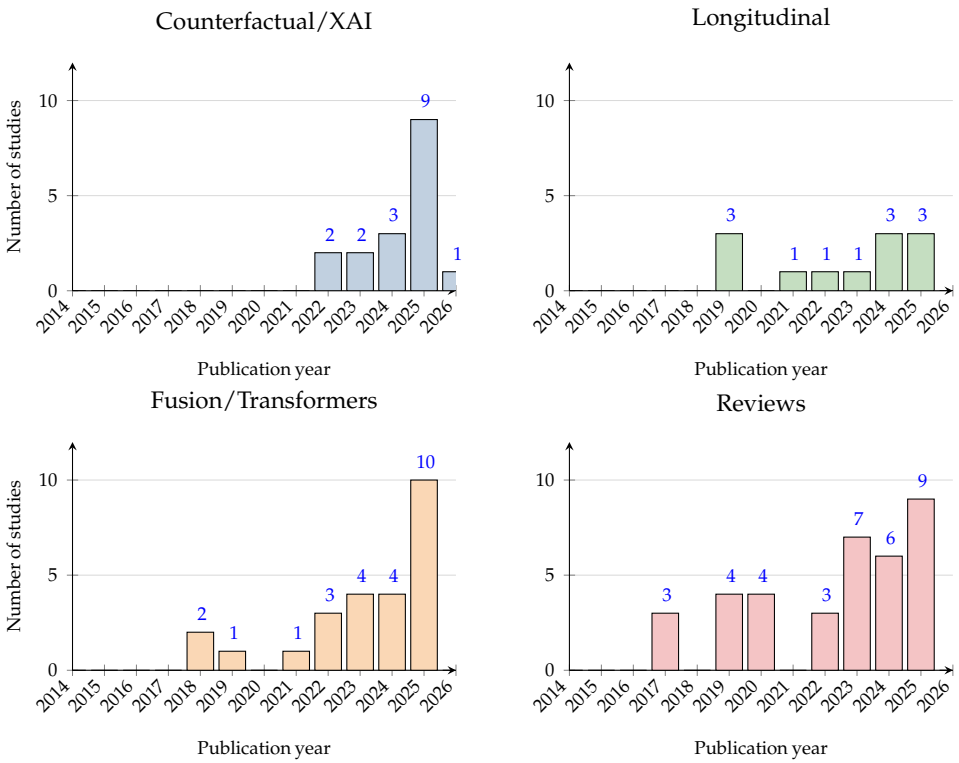


Figure 4. Per-theme year-wise distribution of included studies (2014–2026). Colors and patterns distinguish the four themes: Counterfactual/XAI, Longitudinal, Fusion/Transformers, and Reviews.

Table 3. Study characteristics.

Study	Type Family	Method (Brief)	Task	Modality	Dataset	Trust
Oh et al. (2023) [20]	Prim/XAI–CF	LEAR: CF maps + explanation-guided attention	Staging	T1 sMRI	ADNI	E0C0
Oh et al. (2025) [21]	Prim/XAI–CF	Deep CF "progression" MRI; ROI quantification	Risk/XAI	T1 sMRI	NR	E0C0
Valoor & Gangadharan (2025) [22]	Prim/XAI–CF	Case-based CF maps (U-Net + GAN)	XAI	T1 sMRI	ADNI (rep.)	E0C0
Khosroshahi et al. (2025) [11]	Review/XAI	Survey: SHAP/LIME/Grad-CAM/LRP, etc.	Review	MRI/PET	N/A	NA

Table 3. *Cont.*

Study	Type Family	Method (Brief)	Task	Modality	Dataset	Trust
Zia et al. (2022) [23]	Prim/Gen–XAI	VANT-GAN: abnormal→normal; discrepancy map	XAI	Med. img.	ADNI+ (rep.)	E0C0
Junior et al. (2024) [24]	Prim/XAI (saliency)	ResNet-50 + attention; LIME/Grad-CAM	Classif.	T1 sMRI	ADNI (rep.)	E0C0
Zhang et al. (2022) [25]	Prim/XAI (attn)	3D ResNet + self-attention; atrophy localization	Classif./Loc.	3D sMRI	NR	E0C0
Mohanraj & Sujatha (2025) [26]	Prim/Hybrid–XAI	CNN + fuzzy decision; Grad-CAM/SHAP	Severity	T1 sMRI	OASIS/ADNI (rep.)	E0C0
De Santi et al. (2024) [27]	Preprint/Proto–XAI	PIPNet3D: prototype "scoring sheet"	Classif.	3D sMRI	ADNI-1 (rep.)	E0C0
Jahan et al. (2023) [28]	Prim/Tabular + XAI	Multimodal + RF; SHAP explanations	Classif.	Multimodal	OASIS-3 (rep.)	E0C0
Jahan et al. (2025) [29]	Prim/Fed + XAI	Federated RF; SHAP; privacy-preserving	Classif.	Multimodal	OASIS-3 (rep.)	E0C0
Rahman et al. (2025) [30]	Prim/Fusion	CNN (MRI) + clinical fusion (NeuroNet-AD)	Classif.	Multimodal	ADNI + OASIS-3	E1C0
Ibrahim et al. (2025) [31]	Prim/TL + XAI	TL ensembles + saliency/Grad-CAM	Classif.	T1 sMRI	NR	E0C0
Parvin et al. (2024) [32]	Prim/Multimodal+XAI	Tabular+MRI+gene; SP-LIME/LRP	Planning	Multimodal	OASIS+gene (rep.)	E0C0
Li et al. (2025) [33]	Prim/Gen–CF	Pseudo-healthy synthesis; residual atrophy map	Atrophy CF	T1 sMRI	Multi-site (2)	E1C0
Peng et al. (2026) [34]	Prim/Gen–CF	VQ-VAE + SCM; abduction→action CFs	CF synth.	3D sMRI	ADNI+ NCANDA (rep.)	E1C0
Chitrakala & Bharathi (2025) [35]	Prim/XAI pipeline	Patch-level explainable 3D sMRI pipeline	Classif.	3D sMRI	ADNI+AIBL (rep.)	E1C0
Dong et al. (2021) [36]	Prim/Fusion	Low-rank multimodal + Laplacian regularization	Classif.	MRI+PET+CSF	ADNI (rep.)	E0C0
Madhumitha et al. (2025) [37]	Prim/Transformer–XAI	Relevance-augmented self-attention (domain priors)	Classif.	3D sMRI	NR	E0C0
Dhinagar et al. (2025) [38]	Preprint/VLM	VLM: retrieval + captioning + VQA	Multi-task	MRI+text	Mixed corpora	E0C0
Zhao et al. (2019) [39]	Prim/Trajectory	Hippocampal subfield trajectories (FreeSurfer)	Trajectory	T1 sMRI	NR	E0CNA
Somu et al. (2024) [40]	Prim/Longitudinal	Multi-site atrophy trajectory vs. amyloid/tau	Trajectory	T1 sMRI	Multi-site (rep.)	E1CNA
Wei et al. (2023) [41]	Prim/Longitudinal	Mixed-effects + survival for aMCI→AD	Trajectory	T1 sMRI	ADNI (rep.)	E0CNA
Petrella et al. (2024) [42]	Prim/Causal modeling	Personalized causal biomarker cascade forecasting	Forecast	Multimodal	ADNI (NR)	E0CNA

Table 3. *Cont.*

Study	Type Family	Method (Brief)	Task	Modality	Dataset	Trust
Mieling et al. (2025) [43]	Prim/Tabular+XAI	XGBoost staging/conversion + SHAP	Conv./Cls.	sMRI-der.+tab.	ADNI (rep.)	E0C0
Chen et al. (2022) [44]	Prim/CNN+Attn	Multiview-slice attention + 3D CNN	Classif.	T1 sMRI	ADNI-1/2 (rep.)	E0C0
Zhao et al. (2022) [45]	Prim/Gen+Prog	Score prediction + future MRI synthesis (GAN)	Progression	T1 sMRI	ADNI (GO/2)	E0C0
Barisano et al. (2024) [46]	Prim/Quant biomarker	Automated PVS/WMH; longitudinal risk association	Assoc.	sMRI+FLAIR	Multicentre	E1CNA
Sheikh-Bahaei et al. (2019) [47]	Prim/Biomarker	Microvascular signature (longitudinal)	Biomarker	OCTA(+PET)	NR	E0CNA
Wu et al. (2025) [48]	Prim/Segmentation	WMH weak supervision + rating prediction	Segm.	FLAIR	Multi-cohort (rep.)	E1CNA
Akan et al. (2024) [49]	Preprint/Transformer	ViT embeddings + Bi-LSTM over slices	Classif.	3D sMRI	ADNI (rep.)	E0C0
Akan et al. (2025) [50]	Preprint/ViViT	ViViT-style transformer over slice sequence	Classif.	3D sMRI	ADNI (rep.)	E0C0
Mandal & Shukla (2018) [51]	Position/Perspective	Argument for metabolic+structural+ behavior fusion	Position	Multimodal	N/A	NA
Leng et al. (2023) [52]	Prim/Fusion	Cross-enhanced fusion (sMRI+FDG-PET)	Classif.	sMRI+FDG-PET	ADNI (rep.)	E0C0
Liu et al. (2018) [53]	Prim/Fusion	Cascaded CNNs (local patches → aggregator)	Classif.	sMRI+PET	ADNI (rep.)	E0C0
Liu et al. (2024) [54]	Prim/Fusion	sMRI+rsMEG cross-attention fusion	Classif.	sMRI+MEG	BioFIND (rep.)	E0C0
Jin et al. (2025) [55]	Preprint/Causal+LLM	Causal intervention + LLM summaries (front-door)	Classif.	Imaging text	ADNI+NACC (rep.)	E1C0
Zhang et al. (2022) [56]	Prim/Transformer	Fourier global network for 3D MRI	Classif.	T1 sMRI	ADNI+AIBL (rep.)	E1C0
Qiang et al. (2023) [57]	Prim/Fusion	Dual-attention CNN + clinical/genetic MLP	Classif.	Multimodal	ADNI (rep.)	E0C0
Hazarika et al. (2023) [58]	Prim/CNN	DNN classifier on MRI (baseline-style)	Classif.	T1 sMRI	ADNI (rep.)	E0C0
Hojjati et al. (2019) [59]	Prim/Graph features	sMRI + rs-fMRI graph features + SVM	Classif.	sMRI+rs-fMRI	ADNI (rep.)	E0C0
Ding et al. (2025) [60]	Preprint/GNN	GNN + Kolmogorov–Arnold networks (explainable)	Classif.	sMRI (graph)	ADNI (rep.)	E0C0
Gu et al. (2025) [61]	Prim/FusionFormer	Multi-transformer fusion for conversion; attn/SHAP	Conversion	Multimodal	ADNI-1/2 (rep.)	E0C0
Mmadumbu et al. (2025) [62]	Prim/DL pipeline	Early detection pipeline; cohort transfer reported	Classif.	Multimodal	ADNI+NACC (rep.)	E1C0
Qiu et al. (2022) [63]	Prim/Multimodal DL	Multimodal dementia assessment; multi-cohort testing	Assess.	Multimodal	NACC+OASIS (rep.)	E1C0

Table 3. *Cont.*

Study	Type Family	Method (Brief)	Task	Modality	Dataset	Trust
Jumaili et al. (2025) [64]	Prim/Ensemble	Deep ensemble; external tests on OASIS/ ADNI	Classif.	T1 sMRI	Private+ OASIS+ ADNI	E1C0
Sheikh et al. (2025) [65]	Prim/Lightweight+XAI	MobileNet/EfficientNet + Grad-CAM++	Classif.	T1 sMRI	ADNI (rep.)	E0C0
Ali et al. (2025) [66]	Review/GNN	Review of GNNs for AD (uni/multimodal)	Review	Neuro imaging	N/A	NA
Saikia & Kalita (2024) [5]	Review/DL (MRI)	Review: MRI-based DL for AD detection	Review	MRI	N/A	NA
Alsubaie et al. (2024) [6]	Sys. rev./DL+ML	Systematic review of DL/ML for AD	Review	Neuro imaging	N/A	NA
Kaur et al. (2024) [7]	Sys. rev./DL+ML	PRISMA SLR: DL/ML for AD prediction	Review	Neuro imaging	N/A	NA
Pan et al. (2019) [67]	Prim/sMRI+Genetics	CNN ensemble + GWAS-guided biomarkers	Early DX	sMRI+gen	NR	E0C0
Agarwal et al. (2021) [68]	Sys. rev./Transfer	Review: transfer learning for AD neuroimaging	Review	Neuro imaging	N/A	NA
Shukla et al. (2023) [69]	Review/Pipelines	Review: AD detection pipelines and gaps	Review	MRI	N/A	NA
Franciotti et al. (2023) [70]	Prim/Tabular–ML	RF vs. GB vs. XGB (voting); multimodal biomarkers + feature selection	Conversion	Multimodal	ADNI (rep.)	E0C0
Muydinov (2025) [71]	Review/DL (MRI)	Review: DL for MRI-based AD detection	Review	MRI	N/A	NA
Battineni et al. (2024) [8]	Meta-anal./MRI	Systematic review + meta-analysis (MRI)	Review	MRI	N/A	NA
Jo et al. (2019) [72]	Review/DL	Review: DL for diagnosis + prognosis	Review	Neuro imaging	N/A	NA
Raza et al. (2025) [73]	Review/Multimodal DL	Review: multimodal DL challenges/future	Review	Multimodal	N/A	NA
Arya et al. (2023) [74]	Sys. rev./ML+DL	Systematic review: ML/DL for AD diagnosis	Review	Neuro imaging	N/A	NA
Alsubaie et al. (2025) [75]	Prim/CNN	Attention-driven CNN + multi-activation fusion	Classif.	T1 sMRI	ADNI (rep.)	E0C0
Wang et al. (2025) [76]	Review/Biomarkers	Review: biomarkers + multimodal integration	Review	Multimodal	N/A	NA
Zhang et al. (2025) [15]	Review/Fusion	Review: multimodal fusion DL (MRI/PET, etc.)	Review	Multimodal	N/A	NA
Guo et al. (2025) [16]	Meta-anal./ Transformers	Meta-analysis: transformer multimodal fusion	Review	Multimodal	N/A	NA
Deshpande et al. (2024) [77]	Review/Multimodal DL	Review: integration of multimodal DL	Review	Multimodal	N/A	NA
Odusami et al. (2024) [78]	Meta-anal./Multimodal	Meta-analysis: multimodal neuroimaging ML	Review	Multimodal	N/A	NA

Table 3. *Cont.*

Study	Type Family	Method (Brief)	Task	Modality	Dataset	Trust
Thulasimani et al. (2024) [79]	Review/Datasets+Opt	Review: datasets/ optimization/ algorithms	Review	Neuro imaging	N/A	NA
Aberathne et al. (2023) [17]	Review/Longitudinal	Review: longitudinal MRI/PET ML	Review	MRI+PET	N/A	NA
Zhou et al. (2021) [80]	Review/Progression	Review: MRI biomarkers for progression prediction	Review	MRI	N/A	NA
Naik et al. (2020) [81]	Review/ML trends	Review: ML classifiers + multimodal trends	Review	Multimodal	N/A	NA
Martí-Juan et al. (2020) [18]	Survey/Longitudinal	Survey: longitudinal learning challenges	Review	Neuro imaging	N/A	NA
Forouzannezhad et al. (2019) [82]	Survey/fMRI	Survey: fMRI analysis methods for AD	Review	fMRI	N/A	NA
Sarica et al. (2017) [83]	Sys. rev./RF	Systematic review: RF on neuroimaging	Review	Neuro imaging	N/A	NA
Rathore et al. (2017) [84]	Review/Features	Review: feature extraction + pipelines	Review	Neuro imaging	N/A	NA
Basanta-Torres et al. (2025) [9]	Sys. rev./T1 MRI	Systematic review: T1 MRI AD classification	Review	T1 sMRI	N/A	NA
Altwijri et al. (2023) [85]	Prim/CNN	EfficientNet dementia staging; heavy preprocessing	Classif.	T1 sMRI	Kaggle (rep.)	E0C0
Zhao et al. (2024) [86]	Prim/ViT+CNN	ViT-equipped CNN for 3D MRI diagnosis	Classif.	3D sMRI	ADNI (rep.)	E0C0
Ebrahimighahnavieh et al. (2020) [10]	Sys. rev./DL	Systematic review: DL across modalities	Review	Neuro imaging	N/A	NA
Khojaste-Sarakhsi et al. (2022) [87]	Survey/DL	Survey: architectures/inputs for AD DL	Review	Neuro imaging	N/A	NA
Illakiya & Karthik (2023) [88]	Review/DL trends	Review: DL trends + future perspectives	Review	Neuro imaging	N/A	NA
Sharma et al. (2022) [89]	Survey/DL	Survey: DL for AD (CNN/RNN/generative)	Review	Neuro imaging	N/A	NA
Akhavan Aghdam et al. (2025) [19]	Survey/Repro+Gen.	Survey: reproducibility/generalizability evaluation	Review	Neuro imaging	N/A	NA
Ottaviani & Monacelli (2024) [90]	Commentary	Commentary: critique of risk prediction model	Commentary	Multimodal	N/A	NA
Weiner et al. (2017) [91]	Review/ADNI program	Program review: ADNI publications and progress	Review	Multimodal	ADNI	NA
Saif et al. (2024) [12]	Review/XAI	Review: XAI methods for Alzheimer detection	Review	Neuro imaging	N/A	NA
Vimbi et al. (2024) [13]	Sys. rev./LIME+SHAP	Systematic review: LIME/SHAP in AD	Review	Neuro imaging	N/A	NA
Bibi et al. (2025) [92]	Review/XAI (broad)	Review: XAI for brain disease diagnosis	Review	Neuro imaging	N/A	NA

Table 3. *Cont.*

Study	Type Family	Method (Brief)	Task	Modality	Dataset	Trust
Wang et al. (2023) [93]	Review/GANs	Review: GANs in neuroimaging and clinical neuroscience	Review	Neuro imaging	N/A	NA
Abrol et al. (2019) [94]	Prim/Fusion	sMRI + dynamic FC fusion for MCI→AD	Conversion	sMRI+fMRI	NR	E0C0
Martin et al. (2023) [14]	Sys. rev./Interp. ML	Systematic review: interpretable ML for dementia	Review	Multimodal	N/A	NA

Trust code: E1/E0 = external validation reported/not reported; C1/C0 = calibration reported/not reported; CNA = calibration not applicable (trajectory/association/non-probabilistic); NA = not applicable (review/meta-analysis/commentary).

3.3. Thematic Synthesis by Approach Type

3.3.1. Scope of the Comparative Review

The final included studies comprise 90 PRISMA-included studies, spanning primary methodological contributions and secondary synthesis. Secondary literature represents a substantial portion of the included studies, with review/survey/meta-analysis papers accounting for 36/90 studies, reflecting continued consolidation of methods and evidence in MRI-based Alzheimer's AI [5–19,66,68–74,77–84,87–89,92,93]. In addition, selected perspective articles and initiative-level syntheses are included to situate technical developments within broader discussions of biomarkers, translational adoption, and risk modeling [51,76,90,91]. The remaining 54/90 studies report primary empirical methods. Across these primary contributions, recent work is characterized by increasingly expressive neural architectures and multimodal integration [52–54,56,57,63], interpretability methods aimed at producing clinically meaningful explanations beyond saliency visualization [20–23,27,32–35,37], and progression-aware frameworks that model disease dynamics over time [39–43,45–48,59,61,94].

Each included paper is assigned a single dominant theme according to its central methodological contribution. When a study spans multiple themes (a multimodal model accompanied by SHAP or Grad-CAM), classification is determined by the primary technical contribution rather than auxiliary analyses [28,29,31,65]. MRI-oriented vision–language retrieval, captioning, and VQA remain relatively sparse and are therefore discussed as a sub-direction within fusion/transformer baselines [38]. This limited coverage reflects the current evidence base: vision–language reporting for AD MRI remains an emerging research gap rather than an established theme in the included literature. Despite rapid growth of report generation and retrieval-augmented reporting in general radiology, AD MRI pipelines rarely target guideline-aligned narrative outputs as a first-class endpoint. Few studies report systematic fidelity evaluation, explicit evidence provenance linking statements to imaging/biomarker signals, or calibrated confidence estimates for generated statements. Consequently, the MRI-focused evidence remains limited in both volume and evaluation depth. Accordingly, we frame vision–language/reporting as a translational opportunity and outline concrete reporting- and evaluation-oriented requirements, including structured templates, provenance, and reliability testing under cohort shift, in the synthesis below.

3.3.2. Method Families and Representative Approaches

The included studies are organized into four methodological themes (see Figure 2 and Tables 4 and 5). The discussion below synthesizes each theme by method subfamilies and the evidence they contribute.

Table 4. Theme-level synthesis of included primary research studies.

Theme (Primary)	Included Works	Typical Data	Strengths/Limitations
Counterfactual & XAI (17/90)	[20–29,31–35,37,65]	Mostly T1 sMRI (ADNI/OASIS-like); some multimodal (tabular/genetic); occasional federated settings	Moves beyond saliency toward actionable evidence (counterfactual deltas, pseudo-healthy synthesis, prototypes, atlas-prior attention, AR patch selection). Gaps: explanation reliability under domain shift; calibration and clinician-centered evaluation often missing.
Longitudinal modeling (12/90)	[39–43,45–48,59,61,94]	Longitudinal sMRI; vascular markers; multimodal time series; conversion labels; occasional non-brain imaging biomarkers	Captures disease dynamics and conversion risk; supports biologically plausible biomarkers and "what-if" forecasting. Gaps: missingness handling and harmonization reporting vary; horizon-specific calibration and site-held-out robustness remain uneven.
Fusion/transformers & baselines (25/90)	[30,36,38,44,49–58,60,62–64,67,75,76,85,86,90,91]	MRI ± PET/CSF ± clinical/genetic; public + clinical cohorts; transformer/global operators; graph models; vision–language directions	Stronger representation learning and fusion; global operators and transformer-style sequence modeling; some multi-cohort benchmarking. Gaps: reproducibility controls, fairness/subgroup checks, and calibrated decision support under-reported; very high reported metrics often require careful split scrutiny.

XAI = explainable artificial intelligence; **sMRI** = structural magnetic resonance imaging; **PET** = Positron Emission Tomography; **CSF** = Cerebrospinal Fluid.

Table 5. Theme-level synthesis of included **review/survey/meta-analysis** papers.

Theme (Reviews)	Included Works	Typical Data	Strengths/Limitations
MRI/Neuroimaging DL reviews & SLRs	[5–7,9,10,68,69,71,72,74,83,84,87–89]	MRI-heavy; often ADNI/OASIS; sometimes multimodal	Broad synthesis of datasets/model families/metrics; documents protocol heterogeneity and common preprocessing choices. Gaps: uneven bias assessment; inconsistent standards for external validation and calibration.
Fusion/transformers/ GNN/longitudinal specialized reviews	[8,15–19,66,70,73,77–82]	Multimodal neuroimaging; GNNs; fMRI connectivity; longitudinal designs; biomarker integration	Deep dives by modality/model family; emphasizes translation barriers and cross-cohort concerns. Gaps: harmonization and robust external validation remain uneven; reproducibility often not enforced.
XAI/interpretability reviews	[11–14,92]	Mixed neuroimaging + clinical ML literature	Strong taxonomies and clinical framing for interpretability; highlights evaluation and trust bottlenecks. Gaps: limited evaluation of explanation reliability under shift; weak integration with calibrated decision support.
Generative neuroimaging reviews	[93]	Neuroimaging synthesis/augmentation/ harmonization	Maps the generative landscape and risks/utility for clinical neuroscience. Gaps: weak clinical validity constraints and downstream utility standards; leakage concerns.

Review, Survey, and Meta-Analysis Synthesis

The review-layer literature repeatedly converges on a consistent assessment: although high discrimination is frequently reported, clinical translation is limited by dataset concentration (especially ADNI), heterogeneous preprocessing and evaluation protocols, uneven external validation, and inconsistent reporting of robustness and calibration. Broad MRI/neuroimaging reviews and systematic surveys summarize the progression from CNN-

and transfer-learning pipelines toward attention, transformer, and hybrid designs, while emphasizing that cross-study comparability remains weak without standardized benchmarks and clearer reporting [5,6,10,69,71,72,74,87–89]. PRISMA-style synthesis further documents how study designs and outcomes vary: one systematic literature review screened hundreds of records and highlighted the dominance of deep learning and single-modality neuroimaging alongside wide spreads in reported accuracy [7], transfer-learning-specific synthesis identifies benefits for early diagnosis but flags heavy reliance on narrow datasets and limited biomarker diversity [68], and systematic review/meta-analysis evidence emphasizes substantial heterogeneity across staging schemes, cohorts, and study settings [8]. Targeted multiclass T1-weighted MRI synthesis similarly reports strong but widely varying performance ranges and stresses that methodological heterogeneity and ADNI dependence restrict generalizability [9].

Specialized reviews focus on modality/model subareas where methodological choices strongly affect outcomes. Multimodal and fusion reviews argue that combining MRI with PET, CSF, clinical, and genetic factors often improves accuracy, yet missing-modality handling, harmonization, class imbalance, and multicenter validation are persistent barriers [15,73,77,78,81,89]. Transformer-focused meta-analytic synthesis reports high pooled diagnostic performance for transformer-based multimodal fusion and highlights the continuing gap between performance reporting and clinically interpretable deployment evidence [16]. Dataset and optimization reviews emphasize that dataset characteristics, tuning practices, and pipeline decisions can dominate performance differences and should be standardized or transparently documented [79]. Progression- and longitudinal-focused surveys stress that repeated-measures learning is necessary for forecasting and trial-relevant endpoints, but missingness, irregular follow-up, and heterogeneous measurements remain core obstacles [17,18,70,80]. Functional MRI surveys further catalog connectivity analysis families like ICA and graph-theoretic approaches and motivate multimodal designs that incorporate early functional network alterations [82]. Methodology-specific synthesis includes a systematic review of random-forest-based neuroimaging classification highlighting robustness properties and frequent use for conversion prediction [83], and feature-extraction-oriented synthesis that organizes neuroimaging representations by modality and feature family [84]. Reproducibility-focused evaluation demonstrates that reimplemented open-source pipelines often generalize substantially worse across cohorts than originally reported [19]. Graph-neural-network surveys consolidate graph construction strategies and emphasize that cross-cohort generalizability is the principal bottleneck for GNN-based AD diagnosis [66].

Interpretability-oriented reviews consistently argue that explanation validity is not established by heatmaps alone, and that stability under perturbations/domain shift, clinician-centered evaluation, and integration into calibrated decision support remain uneven [11–14,92]. Generative neuroimaging synthesis outlines GAN applications to augmentation, translation, and harmonization while emphasizing risks related to leakage, clinical validity constraints, and unclear downstream utility standards [93].

Counterfactual Reasoning and XAI (17/90)

Interpretability-focused primary works increasingly move beyond post hoc saliency toward counterfactual generation, structurally grounded explanations, and model families with built-in transparency. LEAR introduces an iterative training loop in which multi-way counterfactual maps are generated and then reused as explanation-guided attention to reinforce MRI-based AD staging, reporting gains in multiclass accuracy and improved agreement with longitudinal change-map proxies relative to common saliency baselines [20]. Quantitative counterfactual pipelines further operationalize interpretability by converting

counterfactual "progression" MRIs into ROI-wise volumetric deltas and an AD-relatedness index intended to provide measurable, individualized risk signals rather than purely visual overlays [21]. Case-based counterfactual systems generate minimal morphology edits that flip the decision to support more actionable clinical interpretation [22]. Discrepancy-based attribution reframes explanation as abnormal→normal translation, using difference maps to localize disease evidence with anatomical plausibility [23].

Architectural interpretability is also pursued via explainable 3D residual self-attention designs that jointly target diagnosis and atrophy localization [25], and prototype-based 3D reasoning (PIPNet3D) that predicts via sparse voting over learned prototypes, enabling human inspection and clinically guided pruning [27]. Disentangled pseudo-healthy synthesis explicitly separates identity-preserving "healthy" counterfactuals from abnormal residual maps that localize individualized atrophy while improving classification [33]. Causal counterfactual generation appears through latent structural causal modeling with VQ-VAE compression, enabling abduction–action–prediction in latent space for fast counterfactual MRI synthesis [34]. Patch-level and region-prior approaches aim to encourage anatomically coherent evidence, including AR-enhanced patch selection tied to structural change cues [35], and relevance-augmented self-attention that injects atlas-driven priors into transformer attention with minimal parameter overhead [37].

Applied clinical pipelines frequently rely on post hoc XAI toolkits layered on strong predictors, including multiclass transfer-learning models explained with LIME/Grad-CAM [24], hybrid CNN–fuzzy pipelines augmented with Grad-CAM and SHAP [26], multimodal feature-based prediction interpreted with SHAP [28], federated multimodal prediction retaining SHAP explanations under decentralized training, [29] transfer-learning ensembles paired with saliency/Grad-CAM [31], graph/knowledge-style multimodal interpretability using SP-LIME/LRP across imaging, tabular, and genomic signals [32], and lightweight early stage detection models explained using Grad-CAM++ variants [65].

Longitudinal Modeling and Progression Forecasting (12/90)

Longitudinal studies shift the emphasis from cross-sectional staging to disease dynamics, conversion risk, and biomarker trajectories. Structural trajectory work quantifies hippocampal and subfield atrophy across preclinical and clinical stages and links regional decline to memory measures, supporting earlier staging signals [39]. Multi-site longitudinal modeling associates atrophy acceleration with amyloid- and tau-positivity while reporting distinct age-dependent dynamics for the two biomarker processes [40]. Region-wise nonlinear mixed-effects modeling maps aMCI→AD atrophic trajectories and connects age and APOE effects to time-to-conversion via survival analysis [41]. Mechanistic modeling complements statistical trajectories through personalized causal biomarker-cascade modeling fit to individual longitudinal profiles, supporting forecasting and counterfactual "what-if" exploration [42].

Predictive ML studies show that feature importance can be horizon-specific [43]. End-to-end joint modeling integrates clinical forecasting with auxiliary image generation by predicting future cognitive scores while synthesizing future MRI from baseline scans [45]. Early stage identification combines sMRI morphometry and rs-fMRI connectivity to distinguish converters from non-converters [59]. Transformer-era progression modeling integrates MRI with cognitive/ROI features for conversion prediction with interpretable attention/feature analyses [61].

Clinically relevant longitudinal risk layers include vascular-marker and lesion-burden modeling [46,48]. Retinal OCTA microvascular measures are explored as longitudinal signatures of preclinical amyloid status [47]. Multimodal progression prediction further

demonstrates that fusing sMRI representations with dynamic functional connectivity features can improve 3-year conversion prediction [94].

Fusion/Transformers and Other Deep-Learning Baselines (25/90)

The largest primary category improves discrimination through stronger architectures, richer multimodal integration, and explicit inductive biases for missing modalities and long-range dependencies. Robust multimodal learning under missing/noisy modalities is addressed via high-order Laplacian-regularized low-rank representation with modality-complete ensembling [36]. Slice/volume hybridization remains practical: multiview slice attention fused with 3D CNN features emphasizes informative slices while retaining global volumetric context [44]. PET+MRI fusion advances include cross-enhanced interaction modeling for sMRI+FDG-PET [52], and cascaded multimodality CNNs that aggregate patch-level 3D features across modalities without requiring segmentation or rigid registration [53]. Beyond PET, cross-attention fusion of sMRI and rsMEG demonstrates complementary structural–electrophysiological information for early classification [54]. Tabular+imaging+genetics integration appears via dual-attention CNN and MLP designs [57]. Cross-modal causal prediction introduces causal intervention/adjustment to reduce spurious correlations [55].

Transformer-era baselines treat 3D MRI as sequences: ViT+BiLSTM hybrids capture inter-slice dependencies [49], and ViViT-style models model slice stacks as "videos" to learn long-range structure [50]. Global-operator alternatives include 3D Fourier-domain networks with learnable frequency filters [56]. Graph-based diagnosis includes KAN-enhanced GCN designs targeting nonlinear region interactions [60]. Multimodal cross-attention frameworks fuse MRI with metadata for multiclass staging and report external validation [30]. Clinical-assessment pipelines evaluate multimodal fusion across cohorts and against clinician benchmarks [63].

Strong baseline-heavy pipelines remain common and help contextualize reported performance [58,62,64,67,75,85,86]. Vision–language modeling appears as an early multimodal direction spanning retrieval, captioning, classification, and VQA [38]. Finally, broader perspectives and initiative-level syntheses complement these baselines by emphasizing emerging non-invasive biomarkers and multimodal integration pathways, critical evaluation of long-horizon risk modeling, and ADNI-driven progress toward improved clinical trials [51,76,90,91].

3.3.3. Structured Synthesis of Findings

Three cross-cutting patterns emerge across Tables 4 and 5:

i. **Primary research concentrates on stronger fusion/transformer baselines, actionable interpretability, and progression-aware modeling.** The dominant primary efforts emphasize multimodal fusion and global/transformer operators [30,36,38,44,49,50,52–57,60,63,75,85,86], counterfactual and interpretable explanation mechanisms beyond saliency [20–23,27,32–35,37,65], and longitudinal/progression modeling targeting conversion risk and biomarker dynamics [39–43,45–48,59,61,94].

ii. **Calibration, leakage-resistant evaluation, and robust external validation are still under-reported.** Even when papers report high accuracy, many do not clearly document subject-level splitting, site-held-out testing, or calibration (reliability curves/ECE). This limitation is repeatedly highlighted across systematic reviews and meta-analyses [8,9,14,16,19]. As a practical consequence, "headline" metrics, especially extremely high accuracies reported in small or non-subject-split settings, should be interpreted cautiously in comparative synthesis [7,9,62].

iii. **Clinical translation demands more than labels and heatmaps.** Across both primary and secondary studies, there remains a gap between high-performing predictors and clinically grounded decision support: standardized evaluation, clinician-centered interpretability, reproducibility artifacts, and evidence-linked reporting remain uneven [11,12,14,19].

4. Comparative Analysis

4.1. Theme Distributions

Figure 2 summarizes the thematic composition of the 90 PRISMA-included studies. Review/survey/meta-analysis papers constitute 36 studies, while the 54 primary studies are distributed across fusion/transformer and other deep-learning baselines (25), counterfactual/XAI (17), and longitudinal modeling/progression forecasting (12).

The publication timeline tends toward 2023–2026, reflecting (i) wider adoption of transformer-era or other global operators for 3D sMRI [49,50,56], (ii) expansion of multimodal fusion beyond MRI+PET to incorporate cognitive scores, genetics, and other non-imaging variables [30,36,57,63], and (iii) growing emphasis on counterfactual and structurally grounded interpretability frameworks [20–22,33–35]. This temporal shift is visualized in Figure 4. Earlier work more often emphasizes classical neuroimaging features and conventional ML pipelines, or early deep-learning systems under constrained evaluation setups [39,59,72]. Multi-branch fusion designs and cross-cohort ambitions become more frequent in the post-2022 literature, although robust external validation remains inconsistent [9,19,63].

A parallel trend is increasing specialization in secondary syntheses: beyond broad CNN-era MRI surveys [5,10,72,74,87,88], recent reviews increasingly target specific methodological families (GNNs, transformers, longitudinal learning, and interpretability toolkits) and consistently identify reproducibility and generalizability as major translation bottlenecks [14,16,18,19,66]. This synthesis layer provides context for interpreting reported primary-study performance as contingent on cohort composition and evaluation design [8,9,19].

All included papers are summarized using a consistent schema in Table 3.

4.2. Model Families vs. Typical Outcomes

(i) **Transformers vs. CNNs (3D sMRI staging and diagnosis).** Transformer-era approaches treat a 3D volume as a sequence of tokens or frames, enabling explicit modeling of nonlocal dependencies across slices and long-range structure [49,50]. Variants that incorporate anatomical priors into attention attempt to improve both interpretability and clinical plausibility of highlighted evidence such as atlas-informed relevance bias [37]. Global-operator baselines provide an alternative route to nonlocal modeling by operating in frequency space, often reporting competitive discrimination while also enabling coarse localization [56]. However, many transformer-era papers report very high accuracy under limited or unclear external-validation settings, so comparative conclusions depend strongly on whether subject-level splitting, preprocessing standardization, and cohort shift are appropriately handled [9,19].
In contrast, CNN families remain strong and widely used, particularly when architectural inductive bias is exploited for neuroimaging particularly in multiview slice attention fused with volumetric context [44]. Lightweight attention-driven CNN variants also continue to report high AUC/accuracy on standard datasets [75]. Transfer-learning pipelines on curated MRI datasets frequently report extremely high multiclass performance [85,86], but these results require careful interpretation because

reported metrics can be sensitive to data curation choices and evaluation protocol details [9,19].

(ii) **Fusion vs. MRI-only (multimodal gains and common omissions).** Across primary studies, multimodal fusion typically improves discrimination over MRI-only models when non-imaging signals (cognitive scores, demographics, genetics) or complementary imaging (PET) are integrated [30,36,52,53,57,63]. Modern fusion designs include cross-attention between MRI and metadata [30], patch-level cross-enhanced fusion for sMRI+FDG-PET [52], and modality-complete ensembling under missing/noisy modalities via structured representation learning [36]. Beyond PET, multimodal fusion also extends to electrophysiology: cross-attention fusion of sMRI with rsMEG demonstrates complementary structural–functional information for early classification [54]. Despite these gains, the comparative literature remains limited by inconsistent reporting of missing-modality handling, harmonization steps, and site-held-out evaluation, even though these factors largely determine whether fusion models translate beyond a single cohort [15,19,78]. Multicohort benchmarking is present in a subset of stronger clinical-assessment systems [30,63], but remains the exception rather than the norm [9,19].

(iii) **Explainability families (from saliency to counterfactual structure).** The XAI segment shows a clear qualitative progression: saliency-first explanations are increasingly supplemented or replaced by counterfactual, prototype-based, and decomposition-based explanations intended to be more actionable. Counterfactual frameworks generate explicit "change maps" or morphology edits that flip the predicted class, supporting a stronger clinical narrative than gradient heatmaps [20–22]. Disentangled pseudo-healthy synthesis provides individualized residual maps that isolate abnormal components from identity-preserving structure [33]. Latent causal counterfactual modeling further formalizes interventions (abduction–action–prediction) in compressed representations to generate plausible counterfactual MRIs efficiently [34]. In parallel, prototype-based 3D networks support transparent sparse voting, enabling inspection of prototypical evidence and pruning of clinically implausible prototypes [27].
Nevertheless, much applied work still uses post hoc toolkits (Grad-CAM/LIME/SHAP) layered on top of strong predictors [24,26,28,31,65], including federated settings that keep raw data decentralized while still producing feature-attribution explanations [29]. Reviews consistently emphasize that explanation validity is not established by visualization alone, and that stability under perturbation/domain shift and clinician-centered evaluation remain uneven [11–14,92].

(iv) **Longitudinal modeling (progression-aware evidence vs. operational barriers).** Longitudinal studies consistently show that temporal context supports more clinically meaningful risk stratification than snapshot staging. Structural MRI trajectory work quantifies hippocampal/subfield decline across stages and links regional atrophy to cognitive measures [39], while multi-site analysis associates atrophy acceleration with amyloid/tau positivity under distinct aging dynamics [40]. Region-wise mixed-effects modeling further connects atrophic trajectories to time-to-conversion and covariates such as APOE and age [41]. Mechanistic biomarker cascade modeling supports individualized forecasting and counterfactual "what-if" exploration [42].
Predictive ML contributions highlight horizon-specific feature importance such as hippocampal vs. entorhinal dominance across different conversion tasks, cautioning against universal biomarker rankings [43]. Joint modeling that predicts future cognitive scores while synthesizing future MRI uses auxiliary generation to improve forecasting [45]. Progression prediction also benefits from multimodal temporal signals, including dynamic functional connectivity fused with deep structural rep-

resentations [94]. At the same time, reviews emphasize that missingness, irregular follow-up schedules, and cohort heterogeneity create major barriers to fair comparison and robust deployment [17,18,70,80].

4.3. Evaluation and Reporting Gaps

Across both MRI-only and multimodal pipelines, calibration and uncertainty reporting remain inconsistent. Systematic reviews and reproducibility analyses repeatedly note that models often report high discrimination without reliability characterization (ECE/Brier/reliability diagrams) or without documenting thresholding strategies aligned with clinical operating points [8,9,14,16,19]. This limitation complicates comparative synthesis because two models with similar AUC may behave very differently under distribution shift or when deployed with fixed decision thresholds [14,19].

External validation and cohort shift handling are also uneven. A subset of multimodal systems explicitly validate across cohorts and benchmark against clinician performance [30,63], and several works incorporate multi-site or multi-cohort data in training [35,56]. However, many studies remain effectively single-cohort, and reporting often omits site-held-out designs, harmonization details, or scanner/site subgroup performance [9,19]. Accordingly, extremely high headline metrics should be interpreted in light of protocol details such as subject-level splitting, independence of acquisitions, and preprocessing standardization [7,9,62,64,85].

For interpretability, fidelity and robustness are inconsistently evaluated. Some counterfactual pipelines compare against longitudinal change proxies or evaluate plausibility via discrepancy or residual structure [20,23,33], but cross-paper comparability remains limited because evaluation criteria vary and clinician-in-the-loop studies are rare [11,14]. Consequently, review-level conclusions converge on the need for standardized interpretability evaluation, calibrated decision support, and reproducibility artifacts to bridge the gap between accuracy demonstrations and clinical translation [11,12,14,19].

Finally, MRI-oriented vision–language modeling remains sparse in the present included studies. A single preprint demonstrates retrieval, captioning, classification, and VQA via contrastive alignment and retrieval-augmented decoding [38], but standardized outcomes for guideline-aligned narratives with explicit evidence provenance, calibrated confidence, and counterfactual sensitivity are not yet established across AD MRI vision–language pipelines.

5. Cross-Study Comparison

5.1. Modalities and Domains

Across the 90-paper included studies, structural MRI (predominantly T1-weighted) remains the backbone modality for diagnosis, staging, representation learning, and explanation generation across all themes [20–27,33–35,37,39–46,48–50,52,53,56–58,60,61,63,65,67,75,85,86]. Multimodal extensions repeatedly integrate PET (often FDG-PET), CSF/biofluid biomarkers, genetics (APOE), and clinical/neuropsychological measures through structured fusion, cross-attention, or feature-level integration [15,16,30,36,52,53,57,61,63,67,70,73,76–78,81,89]. In the primary set, multimodal learning is framed not only as performance engineering but as a practical response to the multifactorial nature of Alzheimer's disease (heterogeneous etiologies and comorbid risks) [51,70,76].

Vascular and small-vessel disease markers appear as clinically important risk layers. Large-cohort pipelines quantify perivascular spaces (PVS) and white matter lesions/hyperintensities (WMH), linking lesion burden and vascular signatures to dementia risk and accelerated brain atrophy [46]. Weakly supervised WMH segmentation coupled with visual rating prediction further supports scalable burden quantifi-

cation without dense manual masks, with downstream gains for MCI discrimination and conversion prediction [48]. Functional markers appear more selectively: resting-state fMRI connectivity features complement sMRI for early stage identification and converter/non-converter separation [59], while fusion with rsMEG demonstrates complementary structural–electrophysiological information in an early diagnosis setting [54]. A complementary non-brain imaging direction uses retinal OCTA microvascular measures as a longitudinal signature of preclinical amyloid status [47].

Overall, the multimodal evidence supports the view that AD prediction benefits from complementary domains, yet few primary works explicitly formalize causal pathways or connect multimodal inference to standardized, guideline-aligned reporting outputs [14,19,55,70].

5.2. Algorithms and Analytical Frameworks

The included studies spans classical ML, deep learning, and hybrid analytical designs. CNN-based pipelines remain common, including attention-augmented and multiview slice-attention designs that combine 2D slice cues with volumetric 3D context [44,58,65,75]. Transformer-era approaches increasingly treat 3D MRI as sequences (ViT feature extraction + sequence modeling) or as slice-stacks analogous to videos (ViViT-style), enabling explicit modeling of long-range dependencies [49,50]. Global-operator alternatives capture nonlocal structure via frequency-domain processing (3D Fourier networks), offering a distinct route to long-range dependency modeling [56]. Graph-based learning appears both in reviews and in primary methods: GNN surveys consolidate graph construction and training strategies [66], and KAN-enhanced GCN variants aim to model nonlinear region interactions while reporting salient brain-region contributions [60].

Multimodal learning is implemented through a range of fusion families, from structured low-rank representation under missing/noisy modalities [36], to PET+MRI cross-enhanced fusion and cascaded multimodal CNNs that avoid segmentation or rigid registration requirements [52,53], to cross-attention between MRI and metadata for multiclass staging with external validation [30]. A clinically oriented multimodal assessment pipeline explicitly benchmarks against clinician performance across cohorts, illustrating the direction of travel from model-centric accuracy toward deployment-relevant evaluation [63]. Causal-motivated multimodal prediction is an emerging thread, using causal intervention/adjustment concepts to reduce shortcut learning across imaging and structured clinical summaries [55].

Interpretability is a major differentiator across primary research. Counterfactual families generate explicit morphologic edits or pseudo-healthy reconstructions to support actionable "what must change" explanations rather than purely gradient-derived saliency [20–22,33,34]. Prototype-based 3D voting models provide built-in transparency by linking predictions to sparse prototype evidence [27], and atlas-prior attention introduces clinically motivated relevance bias into transformer attention to improve anatomical coherence [37]. In parallel, many applied pipelines still rely on post hoc toolkits (Grad-CAM/LIME/SHAP) layered on CNNs or multimodal predictors [24,26,28,31,65], including federated settings where decentralization is combined with feature attribution [29]. Across reviews, interpretability is consistently framed as incomplete without stability testing under perturbation/domain shift and clinician-centered evaluation [11–14,92].

Longitudinal and progression modeling draws on statistical trajectories, survival-linked analysis, and deep forecasting. Hippocampal and subfield trajectory mapping [39] multi-site biomarker-associated atrophy acceleration [40] and region-wise nonlinear mixed-effects trajectories linked to time-to-conversion [41] provide biologically grounded progression evidence. Mechanistic biomarker cascade modeling supports individualized

forecasting and "what-if" exploration [42]. Predictive ML and multimodal temporal fusion approaches further demonstrate that horizon-specific modeling can change which features dominate conversion prediction [43,45,94].

5.3. Datasets and Evaluation Rigor

Across themes, **ADNI** remains the most frequently used backbone dataset for both primary modeling and review-layer synthesis, with frequent supplementation by **OASIS/OASIS-3**, **AIBL**, **NACC**, and smaller local cohorts, as well as curated public MRI collections [9,19,20,30,33,36,39–41,43–45,49,50,52,53,56,57,59,61,63,65,70]. Review papers repeatedly highlight that dataset concentration, heterogeneous preprocessing, and inconsistent split design limit cross-paper comparability and inflate uncertainty about real-world performance [5–9,19].

External validation is present in a subset of stronger studies (notably some multimodal staging systems and clinical assessment pipelines) [30,63], and multi-cohort evaluation is occasionally used to argue robustness [35,56]. However, reproducibility-focused evaluation shows that reimplemented open-source pipelines can generalize substantially worse across cohorts than originally reported, underscoring the fragility of single-cohort claims [19]. As a result, review-layer conclusions consistently emphasize that comparative synthesis should weight study design clarity (subject-level splitting, leakage controls, cohort shift handling) alongside reported accuracy [9,19].

Calibration and explicit uncertainty quantification remain under-reported across primary studies, despite repeated calls in systematic reviews and interpretability syntheses [8,9,14,16,19]. This gap is particularly salient for high-capacity multimodal and transformer-heavy models, where overconfidence under shift is a known deployment risk and where clinical operating points depend on well-calibrated probabilities rather than rank-based discrimination alone [14,19].

5.4. What Works and Where It Struggles

Across the included studies, three consistent "what works" patterns emerge. First, multimodal fusion (MRI combined with PET, clinical scores, genetics, or structured biomarkers) often improves discrimination and is more aligned with the multifactorial nature of AD, especially for multiclass staging and conversion-focused tasks [15,30,36,52,53,57,61,63,70,78]. Second, longitudinal modeling provides more clinically plausible risk stratification than static baselines, revealing trajectory- and horizon-specific feature importance and linking imaging change to biomarkers and time-to-conversion [39–43,45,46,48,94]. Third, counterfactual and structurally grounded interpretability can produce more actionable rationales than generic heatmaps by tying decisions to explicit morphologic edits, pseudo-healthy residual maps, or sparse prototype evidence [20–23,27,33–35,37].

Where the literature struggles is also consistent across themes. A majority of primary systems do not report calibration or robust uncertainty characterization [8,9,14,16,19], often lack site-held-out designs or scanner/site subgroup analyses [9,19], and rarely evaluate explanation stability under perturbations or distribution shift with clinician-centered criteria [11,12,14]. Vision–language capability for brain MRI remains sparse in this included studies: one preprint demonstrates retrieval/captioning/VQA via contrastive alignment and retrieval-augmented decoding [38], but standardized, guideline-aligned narrative reporting with explicit provenance, calibrated confidence, and counterfactual sensitivity is not yet established as a common outcome across AD MRI pipelines [14,19].

Accordingly, the comparative evidence supports the conclusion that the main open translation gap is not only achieving high discrimination, but producing reproducible,

calibrated, and clinically interpretable decision support that links outputs to evidence and remains stable under cross-cohort deployment conditions [14,19].

6. Discussion

This study indicates that MRI-based AI for Alzheimer's disease has advanced rapidly in discrimination-oriented modeling, while translation into clinic-ready decision support remains constrained by evaluation rigor, uncertainty characterization, and standardized reporting. We synthesize the evidence into four methodological themes (Figure 2), which are often studied in isolation and assessed under non-comparable protocols [5–9,14,19].

Evidence from secondary synthesis: comparability and generalizability are persistent bottlenecks. Across systematic reviews and meta-analyses, reported performance is repeatedly judged difficult to compare across studies because cohorts, preprocessing, class definitions, and evaluation splits vary substantially [5–9]. Reproducibility-focused evaluation further shows that reimplemented open-source pipelines may generalize substantially worse across cohorts than originally reported, reinforcing the need to interpret single-cohort results cautiously and to prioritize protocol transparency [19]. Meta-analytic evidence also underscores between-study heterogeneity in staging schemes and population characteristics, implying that pooled outcomes can conceal clinically meaningful variability [8,16,78].

Representation learning: architectural progress is clear, but deployment-grade evidence lags. Primary studies reflect an evolution from CNN-centric pipelines toward transformer-era sequence modeling and nonlocal operators for 3D MRI [49,50,56]. In parallel, CNN and transfer-learning approaches remain widely used and often competitive, particularly when neuroimaging structure is exploited via multiview slice attention, attention-augmented designs, and tailored preprocessing [44,58,75,85,86]. However, across both primary and secondary studies, strong discrimination is not consistently accompanied by robust uncertainty characterization, site-held-out evaluation, or cross-cohort validation, limiting confidence in behavior under dataset shift [8,9,14,19].

Multimodal fusion: consistent gains, uneven handling of missingness and heterogeneity. Multimodal pipelines support the view that AD is multifactorial: integrating MRI with PET, clinical/neuropsychological measures, and genetics can improve discrimination and staging [15,30,52,53,57,77,78,89]. Primary work addresses fusion through cross-attention between MRI and metadata [30], cross-enhanced PET–MRI interaction modeling [52], cascaded multimodality CNN aggregation without heavy manual preprocessing [53], and structured low-rank learning designed to tolerate missing/noisy modalities [36]. A subset of studies moves toward more clinically grounded evaluation by benchmarking across cohorts and, in some settings, against clinician performance [30,63]. Nonetheless, review papers repeatedly emphasize that missing-modality handling, harmonization across centers/scanners, and multicenter external validation are not uniformly enforced, which remains a central barrier to clinical translation [15,19,73,78,81].

Interpretability: movement toward actionable explanations, limited reliability assessment. Interpretability-focused primary work increasingly shifts from post hoc saliency toward counterfactual and structurally grounded explanations that better align with clinical reasoning. Counterfactual generators and case-based editors provide explicit morphology changes or change maps intended to support actionable interpretation [20–22], while discrepancy-based attribution frames evidence localization as abnormal→normal translation [23]. Complementary families pursue built-in transparency via prototype-based voting [27], pseudo-healthy disentanglement with individualized residual atrophy mapping [33], latent causal counterfactual modeling enabling intervention in compressed representations [34], patch-level evidence selection [35], and atlas-prior attention to promote anatomically coherent transformer focus [37]. However, interpretability reviews consis-

tently argue that explanation validity is not established by visually plausible maps alone: stability under perturbations and domain shift, clinician-centered evaluation, and coupling explanations with calibrated risk thresholds remain inconsistent and rarely standardized across studies [11–14,92].

Longitudinal modeling: stronger biological grounding, continued protocol and reporting gaps. Longitudinal studies reinforce that temporal context is essential for forecasting and trial-relevant endpoints. Hippocampal/subfield trajectory analyses link regional decline to cognition and staging [39], multi-site modeling connects atrophy acceleration to amyloid/tau positivity [40], and mixed-effects trajectory modeling relates regional decline to time-to-conversion [41]. Mechanistic biomarker cascade modeling further motivates individualized forecasting and "what-if" exploration [42]. Predictive ML evidence indicates horizon-specific feature dominance, suggesting that static biomarker rankings may not generalize across conversion horizons [43]. Clinically relevant risk layers also emerge through vascular-marker quantification (PVS/WMH) and lesion burden modeling [46,48], and a complementary direction explores longitudinal retinal OCTA microvascular measures as non-brain biomarkers of preclinical signatures [47]. Despite these advances, missing data, irregular follow-up schedules, heterogeneous measurements, and inconsistent validation are repeatedly flagged as major barriers, and horizon-specific calibration or subgroup diagnostics remain infrequently reported [9,17–19,70,80]. **Primary unresolved axis: calibration, robustness under shift, and clinical communication standards.** A cross-cutting limitation across themes is the under-reporting of calibration and uncertainty quantification, despite repeated emphasis in systematic syntheses [8,9,14,16,19]. This gap is consequential because clinical decision support depends on reliable probability estimates, stable operating points, and transparent failure modes under cohort shift [14,19]. Fairness and subgroup auditing for instance, age, sex, genotype, and site/scanner are also rarely systematic in the primary set, even though cohort shift and demographic heterogeneity are central deployment risks [14,19]. Finally, MRI-oriented vision–language work remains sparse: while retrieval/captioning/VQA capability has been demonstrated [38], standardized, guideline-aligned reporting with explicit evidence provenance and calibrated statement confidence is not yet a common outcome target in AD MRI pipelines [14,19].

Implications and practical priorities. Overall, the included studies suggests that the next translation step is less about incremental architecture changes and more about standardizing evaluation and reporting: subject-level leakage-resistant splits; multicenter and site-held-out validation; explicit calibration reporting (reliability curves and Brier/ECE); uncertainty-aware decision thresholds; subgroup auditing; and reproducibility artifacts that enable independent re-evaluation [9,14,19]. Progress on these dimensions would substantially improve cross-study comparability and clinical credibility, enabling more meaningful synthesis across the rapidly expanding methodological landscape [5,8,14,19].

Limitations in the literature. Across the included studies, several recurring limitations emerge that directly map to the focus of this review on calibration, explainability, and clinically relevant MRI-based AD diagnosis across counterfactual, vision–language, longitudinal, and multimodal paradigms:

i. **Calibration is rarely treated as a first-class objective.** Despite frequent reporting of high discrimination, explicit calibration and uncertainty quantification (reliability diagrams, ECE, Brier score, confidence intervals, or threshold-selection rationale) remain under-reported across MRI-only, multimodal, and progression-oriented pipelines [8,9,14,16,19]. This limits the interpretability of predicted probabilities as risk estimates and complicates deployment in settings where operating points and clinical decision thresholds matter [14,19]. Reported near-ceiling accuracies in some stud-

ies further motivate careful scrutiny of subject-level splitting and leakage-resistant evaluation when calibration is absent [7,9,62].

ii. **Explainability is often visual or local, not validated for reliability.** The shift from saliency maps toward counterfactual and structurally grounded explanations is a clear trend [20–23,33,34,37], and interpretable architectures such as prototypes, atlas-prior attention, patch-level evidence selection provide more inspectable rationales [27,35,37]. However, explanation fidelity and stability under perturbations, preprocessing variation, and cohort shift are rarely evaluated with standardized protocols, and clinician-centered validation remains uneven [11–14,92].

iii. **Counterfactual modeling lacks common benchmarks and outcome-grounded validation.** Counterfactual methods produce actionable artifacts such as minimal morphologic edits, pseudo-healthy synthesis, ROI delta summaries that can align with clinical reasoning [20–23,33]. Yet, counterfactual plausibility constraints, sensitivity to confounding factors, and validation against longitudinal outcomes or independent cohorts are inconsistently established, making cross-paper comparison difficult [14,21,42,43].

iv. **Vision–language for brain MRI remains sparse and not reporting-oriented.** Brain MRI vision–language capabilities (retrieval/captioning/VQA) appear only sparsely in the primary set [38]. Consequently, structured reporting-style outputs with explicit evidence provenance (linking statements to imaging/biomarker signals), calibrated statement confidence, and failure-mode characterization are not yet common endpoints in AD MRI pipelines [14,19].

v. **Longitudinal modeling is clinically motivated but methodologically constrained.** Longitudinal studies show the value of repeated measures for progression forecasting and conversion-risk modeling [39–43,45,94], and vascular-marker quantification adds clinically relevant risk context [46,48]. Nevertheless, irregular follow-up schedules, missing data, and heterogeneous measurement intervals remain persistent obstacles [17,18,70,80], while horizon-specific evaluation and calibration over prediction horizons are not consistently reported [8,18,19,41,43].

vi. **Multimodal fusion improves discrimination, but missingness and harmonization are unevenly addressed.** Multimodal methods commonly report gains by integrating MRI with PET, CSF, clinical measures, and genetics [15,30,52,53,57,63,77,78,89], and some explicitly tackle missing/noisy modalities via structured representation learning [36]. However, harmonization reporting, multicenter external validation, and robustness to site/scanner shift remain inconsistent, limiting confidence in real-world generalization [15,19,73,78,81].

vii. **Dataset concentration and protocol heterogeneity limit reproducibility and comparability.** Heavy reliance on a small number of public cohorts (especially ADNI) persists across themes [5,9,19,68], while preprocessing and evaluation choices vary substantially across studies [5–9]. Reproducibility-focused work indicates that reimplemented pipelines may generalize substantially worse than originally reported, reinforcing the need for transparent protocols and robust cross-cohort evaluation [19].

viii. **Subgroup robustness and fairness auditing are rarely systematic.** Detailed subgroup performance and calibration by age, sex, APOE status, site/scanner strata are not consistently reported, despite clear demographic and acquisition heterogeneity in clinical practice [14,19,90]. Federated and hospital-oriented multimodal settings could enable stronger equity-aware evaluation, but fairness metrics and subgroup diagnostics remain uncommon [19,29,63].

7. Study Limitations

This review has several limitations. First, the search strategy was restricted to a limited set of bibliographic databases and supplemented with preprints, which may have omitted relevant studies indexed in other sources or reported in gray literature. Second, the selected time window (2014–2026) emphasizes contemporary MRI-based AI methods, but may under-represent earlier neuroimaging and statistical learning work that informed current practice. Third, restricting inclusion to English-language publications introduces potential language bias. Fourth, study screening and data extraction were performed by a single reviewer, which increases the possibility of selection bias and extraction error. Fifth, the synthesis is qualitative and structured by method families; no formal risk-of-bias tool was applied and no quantitative meta-analysis was performed, limiting the ability to pool performance estimates or formally explain between-study heterogeneity. Sixth, theme assignment used a dominant-category coding scheme to avoid double counting, which may simplify studies that contribute substantially to multiple method families.

Despite these constraints, the integrated synthesis across counterfactual/XAI, vision–language approaches for brain MRI, longitudinal and progression modeling, and multimodal fusion/transformer methods is sufficient to identify consistent methodological gaps in calibration, robustness, and explanation reliability.

8. Conclusions

This structured review indicates that MRI-based AI for Alzheimer's disease has advanced rapidly in both modeling capability and interpretability across four dominant streams: counterfactual/XAI, longitudinal and progression modeling, multimodal fusion/transformer-based diagnosis, and an emerging vision–language direction for neuroimaging-to-text reporting. Across the primary literature, modeling has clearly shifted from CNN-heavy pipelines toward stronger global and transformer-era representations, and explainability has progressed beyond saliency toward more actionable mechanisms such as counterfactual generation, prototype-based reasoning, and anatomically guided attention. Longitudinal studies further reinforce that repeated measures and progression-aware objectives provide more clinically meaningful risk stratification than static staging alone.

At the same time, the evidence base remains misaligned with deployment requirements. Across themes, calibration and uncertainty reporting are inconsistently treated as first-class outcomes; external/site-held-out validation remains uneven; and subgroup robustness like age, sex, genotype, and scanner/site strata is rarely systematic. For interpretability, explanation plausibility is often shown visually, but fidelity and stability under perturbation or domain shift are not evaluated under common, comparable standards. Finally, MRI-oriented vision–language systems are not yet routinely framed or evaluated as guideline-aligned reporting tools with explicit evidence provenance and calibrated statement confidence.

Overall, the main translation gap is no longer model capacity—it is evaluation and communication discipline. The field would benefit most from standardized, leakage-resistant protocols; routine calibration reporting (reliability analysis alongside discrimination); multicenter validation with harmonization-aware modeling; systematic subgroup diagnostics; and reproducibility artifacts that enable independent re-evaluation. Closing these gaps would move MRI-based AD AI from strong predictors toward clinically credible decision support that remains reliable under real-world distribution shift.

Author Contributions: Conceptualization, R.F. and M.M.R.; methodology, R.F.; investigation, R.F.; data curation, R.F.; formal analysis, R.F.; writing—original draft preparation, R.F.; writing—review and editing, R.F., M.M.R., F.K. and B.O.; visualization, R.F.; supervision, M.M.R. and F.K.; funding acquisition, M.M.R. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported by the National Science Foundation (NSF) grant (ID: 2131307), "CISE-MSI: DP: IIS: III: Deep Learning-Based Automated Concept and Caption Generation of Medical Images Towards Developing an Effective Decision Support."

Data Availability Statement: No new data were created or analyzed in this study. Data sharing is not applicable to this article.

Acknowledgments: The authors thank collaborators and the maintainers of the public datasets referenced in this review for enabling reproducible research. The authors acknowledge the use of an AI-based language tool for proofreading and improving the clarity of the manuscript. The authors reviewed and approved all content.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

AD, Alzheimer's disease; MCI, mild cognitive impairment; CN, cognitively normal; NC, normal control; MRI, magnetic resonance imaging; sMRI, structural magnetic resonance imaging; fMRI, functional magnetic resonance imaging; rs-fMRI, resting-state functional magnetic resonance imaging; PET, positron emission tomography; FDG, fluorodeoxyglucose; FDG-PET, fluorodeoxyglucose positron emission tomography; CSF, cerebrospinal fluid; APOE, apolipoprotein E; PVS, perivascular spaces; WMH, white-matter hyperintensities; OCTA, optical coherence tomography angiography; AI, artificial intelligence; ML, machine learning; DL, deep learning; CNN, convolutional neural network; RNN, recurrent neural network; ViT, Vision Transformer; VAE, variational autoencoder; VQ-VAE, vector-quantized variational autoencoder; GAN, generative adversarial network; XAI, explainable artificial intelligence; CF, counterfactual; LIME, Local Interpretable Model-Agnostic Explanations; SP-LIME, Submodular Pick LIME; SHAP, SHapley Additive exPlanations; Grad-CAM, Gradient-weighted Class Activation Mapping; LRP, Layer-wise Relevance Propagation; VLM, vision–language model; VQA, visual question answering; LLM, large language model; PRISMA, Preferred Reporting Items for Systematic Reviews and Meta-Analyses; ADNI, Alzheimer's Disease Neuroimaging Initiative; AIBL, Australian Imaging, Biomarkers and Lifestyle Study; OASIS, Open Access Series of Imaging Studies; NACC, National Alzheimer's Coordinating Center; ROC, receiver operating characteristic; AUC, area under the ROC curve; ECE, expected calibration error.

References

1. World Health Organization. Dementia. Fact Sheet. 2025. Available online: https://www.who.int/news-room/fact-sheets/detail/dementia (accessed on 31 December 2025).
2. Alzheimer's Association. 2025 Alzheimer's disease facts and figures. *Alzheimer's Dement.* **2025**, *21*, e70235. [CrossRef]
3. Page, M.J.; McKenzie, J.E.; Bossuyt, P.M.; Boutron, I.; Hoffmann, T.C.; Mulrow, C.D.; Shamseer, L.; Tetzlaff, J.M.; Akl, E.A.; Brennan, S.E.; et al. The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ* **2021**, *372*, n71. [CrossRef]
4. Tricco, A.C.; Lillie, E.; Zarin, W.; O'Brien, K.K.; Colquhoun, H.; Levac, D.; Moher, D.; Peters, M.D.J.; Horsley, T.; Weeks, L.; et al. PRISMA Extension for Scoping Reviews (PRISMA-ScR): Checklist and Explanation. *Ann. Intern. Med.* **2018**, *169*, 467–473. [CrossRef] [PubMed]
5. Saikia, P.; Kalita, S.K. Alzheimer Disease Detection Using MRI: Deep Learning Review. *SN Comput. Sci.* **2024**, *5*, 507. [CrossRef]
6. Alsubaie, M.G.; Luo, S.; Shaukat, K. Alzheimer's Disease Detection Using Deep Learning on Neuroimaging: A Systematic Review. *Mach. Learn. Knowl. Extr.* **2024**, *6*, 464–505. [CrossRef]
7. Kaur, A.; Mittal, M.; Bhatti, J.S.; Thareja, S.; Singh, S. A systematic literature review on the significance of deep learning and machine learning in predicting Alzheimer's disease. *Artif. Intell. Med.* **2024**, *154*, 102928. [CrossRef]
8. Battineni, G.; Chintalapudi, N.; Amenta, F. Machine Learning Driven by Magnetic Resonance Imaging for the Classification of Alzheimer Disease Progression: Systematic Review and Meta-Analysis. *JMIR Aging* **2024**, *7*, e59370. [CrossRef] [PubMed]

9. Basanta-Torres, S.; Rivas-Fernández, M.Á.; Galdo-Alvarez, S. Artificial Intelligence for Alzheimer's disease diagnosis through T1-weighted MRI: A systematic review. *Comput. Biol. Med.* **2025**, *197*, 111028. [CrossRef]
10. Ebrahimighahnavieh, A.; Luo, S.; Chiong, R. Deep learning to detect Alzheimer's disease from neuroimaging: A systematic literature review. *Comput. Methods Programs Biomed.* **2020**, *187*, 105242. [CrossRef]
11. Khosroshahi, M.T.; Morsali, S.; Gharakhanlou, S.; Motamedi, A.; Hassanbaghlou, S.; Vahedi, H.; Pedrammehr, S.; Kabir, H.M.D.; Jafarizadeh, A. Explainable Artificial Intelligence in Neuroimaging of Alzheimer's Disease. *Diagnostics* **2025**, *15*, 612. [CrossRef]
12. Saif, F.H.; Al-Andoli, M.N.; Wan Bejuri, W.M.Y. Explainable AI for Alzheimer Detection: A Review of Current Methods and Applications. *Appl. Sci.* **2024**, *14*, 10121. [CrossRef]
13. Vimbi, V.; Shaffi, N.; Mahmud, M. Interpreting artificial intelligence models: A systematic review on the application of LIME and SHAP in Alzheimer's disease detection. *Brain Inform.* **2024**, *11*, 10. [CrossRef] [PubMed]
14. Martin, S.A.; Townend, F.J.; Barkhof, F.; Cole, J.H. Interpretable machine learning for dementia: A systematic review. *Alzheimer's Dement.* **2023**, *19*, 2135–2149. [CrossRef] [PubMed]
15. Zhang, R.; Sheng, J.; Zhang, Q.; Wang, J.; Wang, B. A review of multimodal fusion-based deep learning for Alzheimer's disease. *Neuroscience* **2025**, *576*, 80–95. [CrossRef]
16. Guo, H.; Yang, Z.; Zhang, G.; Lv, L.; Zhao, X. Meta analysis of the diagnostic efficacy of transformer-based multimodal fusion deep learning models in early Alzheimer's disease. *Front. Neurol.* **2025**, *16*, 1641548. [CrossRef] [PubMed]
17. 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*, 2134–2140. [CrossRef]
18. Martí-Juan, G.; Sanroma-Guell, G.; Piella, G. A survey on machine and statistical learning for longitudinal analysis of neuroimaging data in Alzheimer's disease. *Comput. Methods Programs Biomed.* **2020**, *189*, 105348. [CrossRef]
19. Akhavan Aghdam, M.; Bozdag, S.; Saeed, F.; Alzheimer's Disease Neuroimaging Initiative. Machine-learning models for Alzheimer's disease diagnosis using neuroimaging data: Survey, reproducibility, and generalizability evaluation. *Brain Inform.* **2025**, *12*, 8. [CrossRef]
20. Oh, K.; Yoon, J.S.; Suk, H.I. Learn-Explain-Reinforce: Counterfactual Reasoning and its Guidance to Reinforce an Alzheimer's Disease Diagnosis Model. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 4843–4857. [CrossRef]
21. Oh, K.; Heo, D.W.; Mulyadi, A.W.; Jung, W.; Kang, E.; Lee, K.H.; Suk, H.I. A quantitatively interpretable model for Alzheimer's disease prediction using deep counterfactuals. *NeuroImage* **2025**, *309*, 121077. [CrossRef]
22. Valoor, A.; Gangadharan, G.R. Unveiling the decision making process in Alzheimer's disease diagnosis: A case-based counterfactual methodology for explainable deep learning. *J. Neurosci. Methods* **2025**, *413*, 110318. [CrossRef] [PubMed]
23. Zia, T.; Murtaza, S.; Bashir Bhatti, N.; Windridge, D.; Nisar, Z. VANT-GAN: Adversarial Learning for Discrepancy-Based Visual Attribution in Medical Imaging. *Pattern Recognit. Lett.* **2022**, *156*, 112–118. [CrossRef]
24. Junior, K.J.; Kouayep, S.C.; Tagne Poupi, T.A.; Kim, H.C.; Alzheimer's Disease Neuroimaging Initiative. Alzheimer's Multiclassification Using Explainable AI Techniques. *Appl. Sci.* **2024**, *14*, 8287. [CrossRef]
25. Zhang, X.; Han, L.; Zhu, W.; Sun, L.; Zhang, D. An explainable 3D residual self-attention deep neural network for joint atrophy localization and Alzheimer's disease diagnosis using structural MRI. *IEEE J. Biomed. Health Inform.* **2022**, *26*, 5289–5297. [CrossRef]
26. Mohanraj, S.; Sujatha, R. A novel CNN–fuzzy–XAI approach for Alzheimer's disease severity classification using brain MRI scans. *Cogent Eng.* **2025**, *12*, 2575105. [CrossRef]
27. De Santi, L.A.; Schlötterer, J.; Scheschenja, M.; Wessendorf, J.; Nauta, M.; Positano, V.; Seifert, C. PIPNet3D: Interpretable Detection of Alzheimer in MRI Scans. *arXiv* **2024**, arXiv:2403.18328. [CrossRef]
28. Jahan, S.; Abu Taher, K.; Kaiser, M.S.; Mahmud, M.; Rahman, M.S.; Hosen, A.S.M.S.; Ra, I.H. Explainable AI-based Alzheimer's prediction and management using multimodal data. *PLoS ONE* **2023**, *18*, e0294253. [CrossRef]
29. Jahan, S.; Adib, M.R.S.; Huda, S.M.; Rahman, M.S.; Kaiser, M.S.; Hosen, A.S.M.S.; Ghimire, D.; Park, M.J. Federated Explainable AI-Based Alzheimer's Disease Prediction with Multimodal Data. *IEEE Access* **2025**, *13*, 43435–43454. [CrossRef]
30. Rahman, S.; Rahman, M.M.; Bhatt, S.; Sundararajan, R.; Faezipour, M. NeuroNet-AD: A Multimodal Deep Learning Framework for Multiclass Alzheimer's Disease Diagnosis. *Bioengineering* **2025**, *12*, 1107. [CrossRef] [PubMed]
31. Ibrahim, N.; Elrefaei, L.; Abdel-Ghaffar, E.A. An interpretable model for the diagnosis of Alzheimer's disease using deep learning and machine learning. In Proceedings of the 2025 15th International Conference on Electrical Engineering (ICEENG), Cairo, Egypt, 12–15 May 2025; pp. 1–8. [CrossRef]
32. Parvin, S.; Nimmy, S.F.; Kamal, M.S. Convolutional neural network based data interpretable framework for Alzheimer's treatment planning. *Vis. Comput. Ind. Biomed. Art* **2024**, *7*, 3. [CrossRef]
33. Li, Z.; Zhao, K.; Chen, P.; Wang, D.; Yao, H.; Zhou, B.; Lu, J.; Wang, P.; Zhang, X.; Han, Y.; et al. Disentangled Representation Learning for Capturing Individualized Brain Atrophy via Pseudo-Healthy Synthesis. *IEEE J. Biomed. Health Inform.* **2025**, *29*, 5056–5068. [CrossRef]

34. Peng, W.; Xia, T.; De Sousa Ribeiro, F.; Bosschieter, T.M.; Adeli, E.; Zhao, Q.; Glocker, B.; Pohl, K.M. Latent Causal Modeling for 3D Brain MRI Counterfactuals. In *Deep Generative Models*; Mukhopadhyay, A., Oksuz, I., Engelhardt, S., Mehrof, D., Yuan, Y., Eds.; Lecture Notes in Computer Science; Springer: Cham, Switzerland, 2026; Volume 16128, pp. 192–201. [CrossRef]
35. Chitrakala, S.; Bharathi, U. ARE-PaLED: Augmented Reality-Enhanced Patch-Level Explainable Deep Learning System for Alzheimer's Disease Diagnosis from 3D Brain sMRI. *Symmetry* **2025**, *17*, 1108. [CrossRef]
36. Dong, A.; Li, Z.; Wang, M.; Shen, D.; Liu, M. High-Order Laplacian Regularized Low-Rank Representation for Multimodal Dementia Diagnosis. *Front. Neurosci.* **2021**, *15*, 634124. [CrossRef] [PubMed]
37. Madhumitha, V.; Padhye, S.; Madarkar, S.S.; Agrawal, S.; Mopuri, K.R. Rel-SA: Alzheimer's Disease Detection Using Relevance-augmented Self Attention by Inducing Domain Priors in Vision Transformers. In Proceedings of the 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Nashville, TN, USA, 11–12 June 2025. [CrossRef]
38. Dhinagar, N.J.; Thomopoulos, S.I.; Thompson, P.M. Leveraging a Vision-Language Model with Natural Text Supervision for MRI Retrieval, Captioning, Classification, and Visual Question Answering. *bioRxiv* **2025**. [CrossRef]
39. Zhao, W.; Wang, X.; Yin, C.; He, M.; Li, S.; Han, Y. Trajectories of the Hippocampal Subfields Atrophy in the Alzheimer's Disease: A Structural Imaging Study. *Front. Neuroinform.* **2019**, *13*, 13. [CrossRef] [PubMed]
40. Somu, S.; Zhu, A.H.; Narula, S.; Jahanshad, N.; Nir, T.M. Longitudinal Multi-Site Modeling of Brain Atrophy Trajectories Associated with Amyloid and Tau. In Proceedings of the 2024 20th International Symposium on Medical Information Processing and Analysis (SIPAIM), Antigua, Guatemala, 13–15 November 2024. [CrossRef]
41. Wei, X.; Du, X.; Xie, Y.; Suo, X.; He, X.; Ding, H.; Zhang, Y.; Ji, Y.; Chai, C.; Liang, M.; et al. Mapping cerebral atrophic trajectory from amnestic mild cognitive impairment to Alzheimer's disease. *Cereb. Cortex* **2023**, *33*, 1310–1327. [CrossRef]
42. Petrella, J.R.; Jiang, J.; Sreeram, K.; Dalziel, S.; Doraiswamy, P.M.; Hao, W.; Alzheimer's Disease Neuroimaging Initiative. Personalized Computational Causal Modeling of the Alzheimer Disease Biomarker Cascade. *J. Prev. Alzheimer's Dis.* **2024**, *11*, 435–444. [CrossRef]
43. Mieling, M.; Yousuf, M.; Bunzeck, N.; Alzheimer's Disease Neuroimaging Initiative. Predicting the progression of MCI and Alzheimer's disease on structural brain integrity and other features with machine learning. *GeroScience* **2025**, *48*, 463–487. [CrossRef] [PubMed]
44. Chen, L.; Qiao, H.; Zhu, F. Alzheimer's Disease Diagnosis with Brain Structural MRI Using Multiview-Slice Attention and 3D Convolution Neural Network. *Front. Aging Neurosci.* **2022**, *14*, 871706. [CrossRef]
45. Zhao, Y.; Ma, B.; Che, T.; Li, Q.; Zeng, D.; Wang, X.; Li, S. Multi-view prediction of Alzheimer's disease progression with end-to-end integrated framework. *J. Biomed. Inform.* **2022**, *125*, 103978. [CrossRef]
46. Barisano, G.; Iv, M.; Choupan, J.; Hayden-Gephart, M.; the Alzheimer's Disease Neuroimaging Initiative. Robust, fully-automated assessment of cerebral perivascular spaces and white matter lesions: A multicentre MRI longitudinal study of their evolution and association with risk of dementia and accelerated brain atrophy. *eBioMedicine* **2024**, *111*, 105523. [CrossRef] [PubMed]
47. Sheikh-Bahaei, N.; Sepehrband, F.; Barisano, G.; Acharya, J.; Rajamohan, A.G.; Law, M.; Toga, A.W.; Chui, C.H. Microvascular signature of preclinical Alzheimer's disease in a longitudinal cohort. *Alzheimer's Dement.* **2019**, *15*, P886. [CrossRef]
48. Wu, Y.; Dong, Z.; Li, H.B.; Chong, Y.F.; Ji, F.; Chong, J.S.X.; Tang, N.R.J.; Hilal, S.; Fu, H.; Chen, C.L.H.; et al. WMH-DualTasker: A Weakly Supervised Deep Learning Model for Automated White Matter Hyperintensities Segmentation and Visual Rating Prediction. *Hum. Brain Mapp.* **2025**, *46*, e70212. [CrossRef]
49. Akan, T.; Alp, S.; Bhuiyan, M.A.N. Vision Transformers and Bi-LSTM for Alzheimer's Disease Diagnosis from 3D MRI. *arXiv* **2024**, arXiv:2401.03132. [CrossRef]
50. Akan, T.; Alp, S.; Bhuiyan, M.S.; Disbrow, E.A.; Conrad, S.A.; Vanchiere, J.A.; Kevil, C.G.; Bhuiyan, M.A.N. Leveraging Video Vision Transformer for Alzheimer's Disease Diagnosis from 3D Brain MRI. *arXiv* **2025**, arXiv:2501.15733. [CrossRef]
51. Mandal, P.K.; Shukla, D. Brain Metabolic, Structural, and Behavioral Pattern Learning for Early Predictive Diagnosis of Alzheimer's Disease. *J. Alzheimer's Dis.* **2018**, *63*, 935–939. [CrossRef]
52. Leng, Y.; Cui, W.; Peng, Y.; Yan, C.; Cao, Y.; Yan, Z.; Chen, S.; Jiang, X.; Zheng, J.; Alzheimer's Disease Neuroimaging Initiative. Multimodal cross enhanced fusion network for diagnosis of Alzheimer's disease and subjective memory complaints. *Comput. Biol. Med.* **2023**, *157*, 106788. [CrossRef]
53. Liu, M.; Cheng, D.; Wang, K.; Wang, Y.; Alzheimer's Disease Neuroimaging Initiative. Multi-Modality Cascaded Convolutional Neural Networks for Alzheimer's Disease Diagnosis. *Neuroinformatics* **2018**, *16*, 295–308. [CrossRef]
54. Liu, Y.; Wang, L.; Ning, X.; Gao, Y.; Wang, D. Enhancing early Alzheimer's disease classification accuracy through the fusion of sMRI and rsMEG data: A deep learning approach. *Front. Neurosci.* **2024**, *18*, 1480871. [CrossRef]
55. Jin, Y.; Xiao, H.; Chu, J.; Lv, F.; Li, Y.; Li, T. Cross-modal Causal Intervention for Alzheimer's Disease Prediction. *arXiv* **2025**, arXiv:2507.13956v1. [CrossRef]

56. Zhang, S.; Chen, X.; Ren, B.; Yang, H.; Yu, Z.; Zhang, X.Y.; Zhou, Y. 3D Global Fourier Network for Alzheimer's Disease Diagnosis Using Structural MRI. In *Proceedings of the Medical Image Computing and Computer Assisted Intervention–MICCAI 2022*; Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S., Eds.; Lecture Notes in Computer Science; Springer: Cham, Switzerland, 2022; Volume 13431, pp. 34–43. [CrossRef]
57. Qiang, Y.R.; Zhang, S.W.; Li, J.N.; Li, Y.; Zhou, Q.Y.; Alzheimer's Disease Neuroimaging Initiative. Diagnosis of Alzheimer's disease by joining dual attention CNN and MLP based on structural MRIs, clinical and genetic data. *Artif. Intell. Med.* **2023**, *145*, 102678. [CrossRef]
58. Hazarika, R.A.; Maji, A.K.; Kandar, D.; Jasinska, E.; Krejci, P.; Leonowicz, Z.; Jasinski, M. An Approach for Classification of Alzheimer's Disease Using Deep Neural Network and Brain Magnetic Resonance Imaging (MRI). *Electronics* **2023**, *12*, 676. [CrossRef]
59. Hojjati, S.H.; Ebrahimzadeh, A.; Babajani-Feremi, A.; Alzheimer's Disease Neuroimaging Initiative. Identification of the Early Stage of Alzheimer's Disease Using Structural MRI and Resting-State fMRI. *Front. Neurol.* **2019**, *10*, 904. [CrossRef]
60. Ding, T.; Xiang, D.; Schubert, K.E.; Dong, L. GKAN: Explainable Diagnosis of Alzheimer's Disease Using Graph Neural Network with Kolmogorov-Arnold Networks. *arXiv* **2025**, arXiv:2504.00946v1. [CrossRef]
61. Gu, S.K.; Purushothaman, A.; Pradeep, R.; Sreeni, K. AlzFusionFormer: Integrating multiple transformers for early Alzheimer's disease detection from multi-modal data. *Biomed. Signal Process. Control* **2025**, *112*, 108601. [CrossRef]
62. Mmadumbu, A.C.; Saeed, F.; Ghaleb, F.; Qasem, S.N. Early detection of Alzheimer's disease using deep learning methods. *Alzheimer's Dement.* **2025**, *21*, e70175. [CrossRef] [PubMed]
63. Qiu, S.; Miller, M.I.; Joshi, P.S.; Lee, J.C.; Xue, C.; Ni, Y.; Wang, Y.; De Anda-Duran, I.; Hwang, P.H.; Cramer, J.A.; et al. Multimodal deep learning for Alzheimer's disease dementia assessment. *Nat. Commun.* **2022**, *13*, 3404. [CrossRef] [PubMed]
64. Jumaili, M.L.F.; Sonuç, E. ML-Driven Alzheimer's disease prediction: A deep ensemble modeling approach. *SLAS Technol.* **2025**, *32*, 100298. [CrossRef]
65. Sheikh, F.; Al Marouf, A.; Rokne, J.G.; Alhajj, R. Lightweight Deep Learning Models with Explainable AI for Early Alzheimer's Detection from Standard MRI Scans. *Diagnostics* **2025**, *15*, 2709. [CrossRef]
66. Ali, S.; Piana, M.; Pardini, M.; Garbarino, S. Graph neural networks in Alzheimer's disease diagnosis: A review of unimodal and multimodal advances. *Front. Neurosci.* **2025**, *19*, 1623141. [CrossRef]
67. Pan, D.; Huang, Y.; Zeng, A.; Jia, L.; Song, X.; Alzheimer's Disease Neuroimaging Initiative (ADNI). Early Diagnosis of Alzheimer's Disease Based on Deep Learning and GWAS. In *Human Brain and Artificial Intelligence*; Zeng, A., Pan, D., Hao, T., Zhang, D., Shi, Y., Song, X., Eds.; Communications in Computer and Information Science; Springer: Singapore, 2019; Volume 1072, pp. 52–68. [CrossRef]
68. Agarwal, D.; Marques, G.; de la Torre-Díez, I.; Franco Martin, M.A.; García Zapiraín, B.; Martín Rodríguez, F. Transfer Learning for Alzheimer's Disease through Neuroimaging Biomarkers: A Systematic Review. *Sensors* **2021**, *21*, 7259. [CrossRef]
69. Shukla, A.; Tiwari, R.; Tiwari, S. Review on Alzheimer Disease Detection Methods: Automatic Pipelines and Machine Learning Techniques. *Sci* **2023**, *5*, 13. [CrossRef]
70. Franciotti, R.; Nardini, D.; Russo, M.; Onofrj, M.; Sensi, S.L.; Alzheimer's Disease Neuroimaging Initiative; Alzheimer's Disease Metabolomics Consortium ADMC. Comparison of Machine Learning-based Approaches to Predict the Conversion to Alzheimer's Disease from Mild Cognitive Impairment. *Neuroscience* **2023**, *514*, 143–152. [CrossRef]
71. Muydinov, A. Advances in Deep Learning Techniques for Alzheimer's Disease Detection Using MRI Images Review. In Proceedings of the 8th International Conference on Future Networks & Distributed Systems (ICFNDS '24), Marrakesh, Morocco, 11–12 December 2024; pp. 266–269 [CrossRef]
72. Jo, T.; Nho, K.; Saykin, A.J. Deep Learning in Alzheimer's Disease: Diagnostic Classification and Prognostic Prediction Using Neuroimaging Data. *Front. Aging Neurosci.* **2019**, *11*, 220. [CrossRef] [PubMed]
73. Raza, M.L.; Hassan, S.T.; Jamil, S.; Hyder, N.; Batool, K.; Walji, S.; Abbas, M.K. Advancements in deep learning for early diagnosis of Alzheimer's disease using multimodal neuroimaging: Challenges and future directions. *Front. Neuroinform.* **2025**, *19*, 1557177. [CrossRef]
74. Arya, A.D.; Verma, S.S.; Chakarabarti, P.; Chakrabarti, T.; Elngar, A.A.; Kamali, A.M.; Nami, M. A systematic review on machine learning and deep learning techniques in the effective diagnosis of Alzheimer's disease. *Brain Inform.* **2023**, *10*, 17. [CrossRef]
75. Alsubaie, M.G.; Luo, S.; Shaukat, K.; Zhang, W.; Li, J. A Novel Deep Learning Approach for Alzheimer's Disease Detection: Attention-Driven Convolutional Neural Networks with Multi-Activation Fusion. *AI* **2025**, *6*, 324. [CrossRef]
76. Wang, R.; Peng, S.; Zhu, J.; Xu, Y.; Wang, M.; Zhang, L.; Qiu, Y.; Hou, D.; Wang, Q.; Liu, R. Innovations in Alzheimer's disease diagnostic technologies: Clinical prospects of novel biomarkers, multimodal integration, and non-invasive detection. *Front. Neurol.* **2025**, *16*, 1651708. [CrossRef]
77. Deshpande, P.; Kulkarni, S. Exploring Integration of Multimodal Deep Learning Approaches for Enhanced Alzheimer's Disease Diagnosis: A Review of Recent Literature. *SN Comput. Sci.* **2024**, *5*, 852. [CrossRef]

78. Odusami, M.; Maskeliūnas, R.; Damaševičius, R.; Misra, S. Machine learning with multimodal neuroimaging data to classify stages of Alzheimer's disease: A systematic review and meta-analysis. *Cogn. Neurodyn.* **2024**, *18*, 775–794. [CrossRef] [PubMed]
79. Thulasimani, V.; Shanmugavadivel, K.; Cho, J.; Veerappampalayam Easwaramoorthy, S. A Review of Datasets, Optimization Strategies, and Learning Algorithms for Analyzing Alzheimer's Dementia Detection. *Neuropsychiatr. Dis. Treat.* **2024**, *20*, 2203–2225. [CrossRef] [PubMed]
80. Zhou, Y.; Song, Z.; Han, X.; Li, H.; Tang, X. Prediction of Alzheimer's Disease Progression Based on Magnetic Resonance Imaging. *ACS Chem. Neurosci.* **2021**, *12*, 4209–4223. [CrossRef] [PubMed]
81. Naik, B.; Mehta, A.; Shah, M. Denouements of machine learning and multimodal diagnostic classification of Alzheimer's disease. *Vis. Comput. Ind. Biomed. Art* **2020**, *3*, 26. [CrossRef]
82. Forouzannezhad, P.; Abbaspour, A.; Fang, C.; Cabrerizo, M.; Loewenstein, D.; Duara, R.; Adjouadi, M. A survey on applications and analysis methods of functional magnetic resonance imaging for Alzheimer's disease. *J. Neurosci. Methods* **2019**, *317*, 121–140. [CrossRef]
83. Sarica, A.; Cerasa, A.; Quattrone, A. Random Forest Algorithm for the Classification of Neuroimaging Data in Alzheimer's Disease: A Systematic Review. *Front. Aging Neurosci.* **2017**, *9*, 329. [CrossRef] [PubMed]
84. Rathore, S.; Habes, M.; Iftikhar, M.A.; Shacklett, A.; Davatzikos, C. A review on neuroimaging-based classification studies and associated feature extraction methods for Alzheimer's disease and its prodromal stages. *NeuroImage* **2017**, *155*, 530–548. [CrossRef]
85. Altwijri, O.; Alanazi, R.; Aleid, A.; Alhussaini, K.; Aloqalaa, Z.; Almijalli, M.; Saad, A. Novel Deep-Learning Approach for Automatic Diagnosis of Alzheimer's Disease from MRI. *Appl. Sci.* **2023**, *13*, 13051. [CrossRef]
86. Zhao, Z.; Yeoh, P.S.Q.; Zuo, X.; Chuah, J.H.; Chow, C.O.; Wu, X.; Lai, K.W. Vision transformer-equipped Convolutional Neural Networks for automated Alzheimer's disease diagnosis using 3D MRI scans. *Front. Neurol.* **2024**, *15*, 1490829. [CrossRef]
87. Khojaste-Sarakhsi, M.; Haghighi, S.S.; Fatemi Ghomi, S.M.T.; Marchiori, E. Deep learning for Alzheimer's disease diagnosis: A survey. *Artif. Intell. Med.* **2022**, *130*, 102332. [CrossRef]
88. Illakiya, T.; Karthik, R. Automatic Detection of Alzheimer's Disease using Deep Learning Models and Neuro-Imaging: Current Trends and Future Perspectives. *Neuroinformatics* **2023**, *21*, 339–364. [CrossRef]
89. Sharma, S.; Mandal, P.K. A Comprehensive Report on Machine Learning-based Early Detection of Alzheimer's Disease using Multi-modal Neuroimaging Data. *ACM Comput. Surv.* **2022**, *55*, 43:1–43:44. [CrossRef]
90. Ottaviani, S.; Monacelli, F. Rethinking Dementia Risk Prediction: A Critical Evaluation of a Multimodal Machine Learning Predictive Model. *J. Alzheimer's Dis.* **2024**, *97*, 1097–1100. [CrossRef]
91. Weiner, M.W.; Veitch, D.P.; Aisen, P.S.; Beckett, L.A.; Cairns, N.J.; Green, R.C.; Harvey, D.; Jack, C.R.; Jagust, W.; Morris, J.C.; et al. Recent publications from the Alzheimer's Disease Neuroimaging Initiative: Reviewing progress toward improved AD clinical trials. *Alzheimer's Dement.* **2017**, *13*, e1–e85. [CrossRef]
92. Bibi, N.; Courtney, J.; McGuinness, K. Enhancing Brain Disease Diagnosis with XAI: A Review of Recent Studies. *ACM Trans. Comput. Healthc.* **2025**, *6*, 1–35. [CrossRef]
93. Wang, R.; Bashyam, V.; Yang, Z.; Yu, F.; Tassopoulou, V.; Chintapalli, S.S.; Skampardoni, I.; Sreepada, L.P.; Sahoo, D.; Nikita, K.; et al. Applications of generative adversarial networks in neuroimaging and clinical neuroscience. *NeuroImage* **2023**, *269*, 119898. [CrossRef]
94. Abrol, A.; Fu, Z.; Du, Y.; Calhoun, V.D. Multimodal Data Fusion of Deep Learning and Dynamic Functional Connectivity Features to Predict Alzheimer's Disease Progression. In Proceedings of the 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Berlin, Germany, 23–27 July 2019; pp. 4409–4413. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.