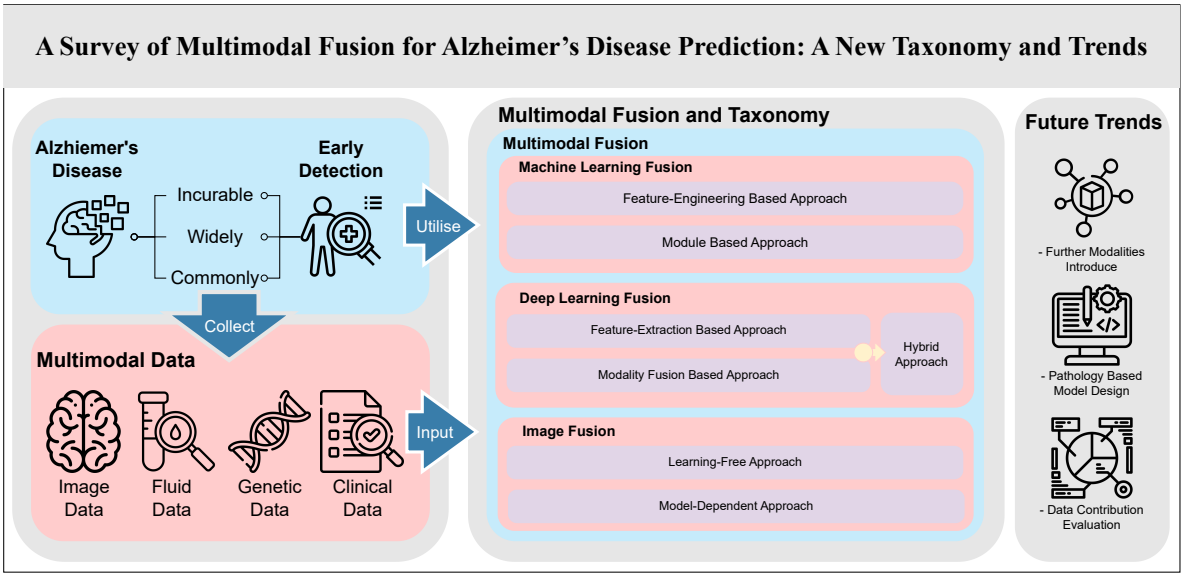
Graphical Abstract

A Survey of Multimodal Fusion for Alzheimer's Disease Prediction: A New Taxonomy and Trends

Yifan Guan,Wei Wang,Jianjun Chen,Po Yang,Jingzhou Xu,Jun Qi



Highlights

A Survey of Multimodal Fusion for Alzheimer's Disease Prediction: A New Taxonomy and Trends

Yifan Guan,Wei Wang,Jianjun Chen,Po Yang,Jingzhou Xu,Jun Qi

- Proposes a novel taxonomy of multimodal fusion tailored to AD.
- Reviews ML and DL fusion methods with scenario-specific guidance.
- Outlines emerging trends and clinical translation challenges.

A Survey of Multimodal Fusion for Alzheimer's Disease Prediction: A New Taxonomy and Trends*

Yifan Guan[a,1], Wei Wang[a], Jianjun Chen[a], Po Yang[b], Jingzhou Xu[a] and Jun Qi[a,*]

[a]Xi'an Jiaotong-Liverpool University, School of Advanced Technology, No. 111 Ren'ai Road, Suzhou, 215123, China
[b]University of Sheffield, Western Bank, S10 2TN Sheffield, UK

ARTICLE INFO

Keywords:
Machine Learning
Deep Learning
Alzheimer's Disease
Multimodal Analysis

ABSTRACT

Alzheimer's disease (AD) is a neurodegenerative disease, well-known for its incurability, and is common among the elderly population worldwide. Previous studies have demonstrated that early intervention positively influences disease progression, leading to increased research into pathological analysis and disease trajectory prediction through machine learning (ML) methods. Given the similarities across different neurodegenerative disorders, a diagnosis relying solely upon a single modality of data is inadequate. Consequently, current research predominantly focuses on multimodal analysis, integrating medical imaging and clinical patient information, with continuous identification of new data types potentially aiding AD diagnosis. Multimodal approaches have been explored extensively over the past two decades, with significant advances observed following the introduction of Deep Learning (DL) techniques. Deep neural networks can adaptively extract and fuse features directly from input data, significantly broadening the scope of multimodal analysis. However, earlier classification studies have primarily concentrated on traditional ML, often neglecting the rapid advancements in DL networks. This article provides a comprehensive description of the acquisition pathways based on modalities, discusses the modalities currently used for research in neuroimaging, human body fluids, and other relevant sources. Additionally, it classifies fusion methodologies utilised in both DL and traditional ML contexts, highlights existing challenges, and outlines potential directions for future research.

1. Introduction

AD is a neurodegenerative disorder and the most common form of dementia, primarily associated with abnormal protein aggregation and neuronal degeneration in the brain. As a leading cause of dementia, AD predominantly affects elderly populations, typically manifesting between the ages of 60 and 70, and is prevalent globally. According to a 2013 survey, over 30 million cases have been reported, with the number continuing to rise [192]. As of now, more than 6.9 million individuals aged 65 and older in the United States alone are affected by AD [2]. Despite extensive research, AD remains incurable and poses a substantial threat to the ageing population [188]. Nevertheless, a definitive diagnosis of AD typically occurs only after a clear cognitive decline is observed, by which stage intervention becomes far less effective [153]. Pathological studies indicate that AD progresses gradually over 10 to 15 years, characterised by the accumulation of abnormal proteins such as amyloid-β and tau [171]. Early in this process, patients may show only mild memory impairment without noticeable disruption to daily life. This stage, known as Mild Cognitive Impairment (MCI) [128], has reshaped our understanding of AD from a binary disease state to a progressive continuum. Crucially, both pharmacological and non-pharmacological treatments have shown promise in slowing disease progression when applied during the MCI phase [131]. Consequently, recent research has emphasised the early identification of biomarkers and pathological brain changes that signal the onset of AD or MCI [192]. Early detection enables clinicians to predict disease trajectories more accurately and apply targeted interventions that may delay symptom onset or reduce severity [192, 188]. Nonetheless, distinguishing AD from other types of dementia remains challenging when relying on a single data modality, as overlapping symptoms and imaging characteristics often obscure diagnostic certainty [123]. This limitation has driven the growing interest in multimodal analysis, now a major focus of modern AD research.

*This work was supported in part by National Natural Science Foundation of China (62301452), in part by Suzhou Science and Technology Development Plan (SYW2025056), in part by the Xi'an JiaoTong-Liverpool University Research Development Fund (RDF-21-01-080)

*Corresponding author
Jun.Qi@xjtlu.edu.cn (J. Qi)
ORCID(s):

[1]This is the first author footnote, but is common to third author as well.

With continued research and development, ML has emerged as a powerful analytical tool across numerous scientific domains, and its value has become increasingly evident in AD research. In medical applications, ML excels at processing complex, high-dimensional data and uncovering subtle, non-linear relationships that traditional statistical methods often fail to capture. Moreover, algorithms such as Support Vector Machines (SVM) and Artificial Neural Networks (ANN) have enabled ML models to generate explicit diagnostic criteria and decision rules, thereby enhancing interpretability and standardisation in clinical workflows[13]. These approaches have also demonstrated strong diagnostic performance in AD classification tasks. Building upon these foundations, recent advances in feature extraction and the integration of intelligent agents have further expanded the role of ML, enabling more efficient and automated analysis of medical imaging data [12]. Upon ML technology, DL methods subsequently advanced this progression by embedding comprehensive automated feature extraction directly into the analytical pipeline, substantially reducing reliance on manual feature engineering and significantly enhancing data processing efficiency [147]. Moreover, DL architectures demonstrate exceptional adaptability to large-scale data and possess heightened sensitivity to intricate multimodal correlations. This capability significantly improves diagnostic classification and predictive accuracy. Owing to their ease of use and powerful inherent feature extraction capabilities, deep neural networks have notably propelled AD diagnostic research forward, driving innovation in multimodal diagnostic systems [188].

This paper discusses the topic of AD diagnosis through ML. It integrates insights into current research in the AD domain, systematically explores two primary directions to deliver a comprehensive. Firstly, we summarise various data types commonly utilised in AD diagnosis, detailing the underlying principles of data acquisition and identifying critical diagnostic information extracted from each modality. Secondly, examines the development trajectory of multimodal methodologies in AD research, systematically categorising existing approaches into traditional ML and DL methods. We provide detailed classifications of prevalent network architectures, thoroughly analysing their roles, functional mechanisms, and variants as they appear across different studies, ultimately discussing potential future challenges and developmental directions for multimodal analysis.

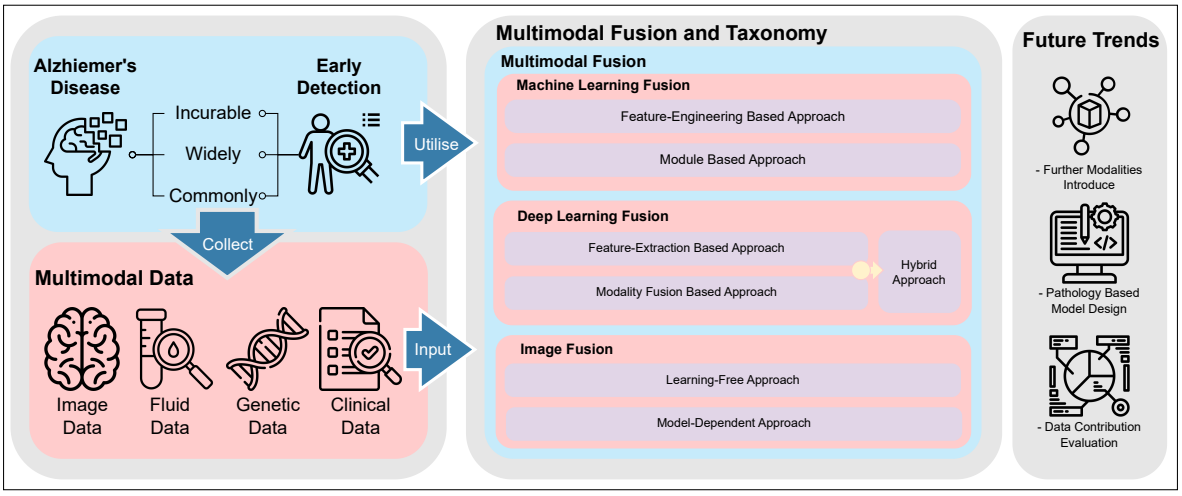


Figure 1: Overview of the scope and structure of this review. The figure reflects the logical organisation of this review, illustrating how the discussion progresses from ML to DL and image fusion paradigms. It exhibits the key components of the paper, including the modalities commonly used in AD research, the proposed taxonomy of multimodal fusion encompassing traditional ML, DL, and image fusion approaches, and a summary of future research directions.

Literature searches of this paper were performed on databases including IEEE Xplore, PubMed, ACM Digital Library, and Elsevier, supplemented by a secondary comprehensive search via Google Scholar to avoid omissions. Searches employed standard combinations of keywords and Boolean operators, strictly limited to English-language articles. According to the two primary thematic focuses of this work, two separate literature search rounds were conducted. In the first round, keywords such as "Alzheimer's Disease" and "history" were utilised to document major biomarkers identified during the historical progression of AD research, along with the chronological introduction of different

modalities and associated techniques. Subsequently, additional searches were conducted using specific modality names such as magnetic resonance imaging (MRI), Computed Tomography(CT), and positron emission tomography (PET), covering prevalent data types thoroughly. Furthermore, inspired by recent works, an additional subsection was included to discuss emerging effective modalities. The second round of searches concentrated specifically on methodological aspects of multimodal ML approaches, employing keywords including "multimodal fusion", "multimodal analysis", "Alzheimer's Disease", "machine learning", and "deep learning", alongside referencing relevant review articles. Publications identified through this second search round were restricted to those published from 2010 onward, with DL-related articles specifically limited to those from 2016 onward. This time period reflects the historical context in which multimodal approaches became systematically structured and widely adopted around 2010, thus providing an appropriate baseline for the scope of our literature survey.

This paper begins by reviewing the research background and developmental context of Alzheimer's disease AD, then extends the discussion to the multimodal data types commonly employed in study design. Building on these foundations, we analyse current multimodal fusion methods in detail. Finally, we summarise the key challenges that remain and outline potential future research directions. Figure 1 provides a concise overview of this workflow, illustrating the logical structure and core content of the review. The subsequent sections of this paper are organised as follows: section 2 presents the historical development of multimodal AD research. section 3 describes multimodal data aspects, including introducing datasets, the definition of biomarkers, and frequently used and significant modalities. section 4, representing the core content of this paper, discusses various fusion methodologies, including internal multimodal transformations within models, fusion strategies and methods, associated application scenarios, and the evaluation of resulting outputs, with a particular focus on DL networks. section 5 outlines future directions and challenges associated with fusion methodologies, especially regarding more complex modalities and clinical translation. Finally, section 6 provides a summary of the paper.

2. Historical Research

This section will retrospect the historical development of AD research, presenting the milestone discoveries and introduction of significant technologies. Due to limitations in available technology and data, multimodal methods emerged relatively late in the AD field, becoming widely recognised only after 2000. It wasn't until the appearance of deep networks and their application after 2010 that multimodal fusion became a relatively mainstream research method. Before this period, research was largely based on single-modality analysis using traditional ML methods. Figure 2 summarises the milestones of AD research, including the biomarker and modality exploration.

AD research dates back to the early 20th century when distinctive neuropathological hallmarks, particularly neurofibrillary tangles (NFTs), were initially identified and subsequently clinically recognised around the 1950s [20]. During this period, Blessed et al. established a critical link between amyloid plaque density and AD severity, significantly advancing pathological understanding [18]. Additionally, to streamline dementia diagnosis, Blessed et al. introduced the Blessed Dementia Scale (BDS), laying important groundwork for subsequent cognitive assessment scales [47]. With advancing technology, the diagnostic value of imaging modalities, including CT, MRI, and PET, was progressively validated for AD [76, 48, 36]. Around 1985, significant progress was achieved by isolating and clarifying the molecular composition of amyloid plaques and NFTs, providing essential contributions to biomarker research [62]. Soon thereafter, Weidemann et al. demonstrated the presence and properties of amyloid precursor protein (APP) in various cellular contexts and human cerebrospinal fluid (CSF), elucidating that APP cleavage generates amyloid-beta peptides, the primary constituent of plaques characteristic of AD pathology [170]. This discovery emphasised CSF's diagnostic and biomarker potential.

Up to approximately 1990, AD research predominantly centred around biomarker discovery and exploring pathological mechanisms. This era witnessed the confirmation of key biomarkers and identification of effective modalities, forming a crucial foundation for contemporary multimodal AD research [36]. The late 1990s introduced transformative innovations through ML methods. Initially, researchers applied these approaches to analyse cognitive assessment outcomes, substantially surpassing traditional dementia screening performance [106]. Concurrently, the pioneering work by deFigueiredo et al. [32] introduced ANNs for analysing single-photon emission computed tomography (SPECT) imaging data, laying the groundwork for a structured analytical workflow consisting of image registration, feature engineering, and neural network-based classification. ML demonstrated substantial advantages, particularly its ability to detect complex, non-linear features and define precise classification boundaries, enhancing diagnostic interpretability for clinical applications.

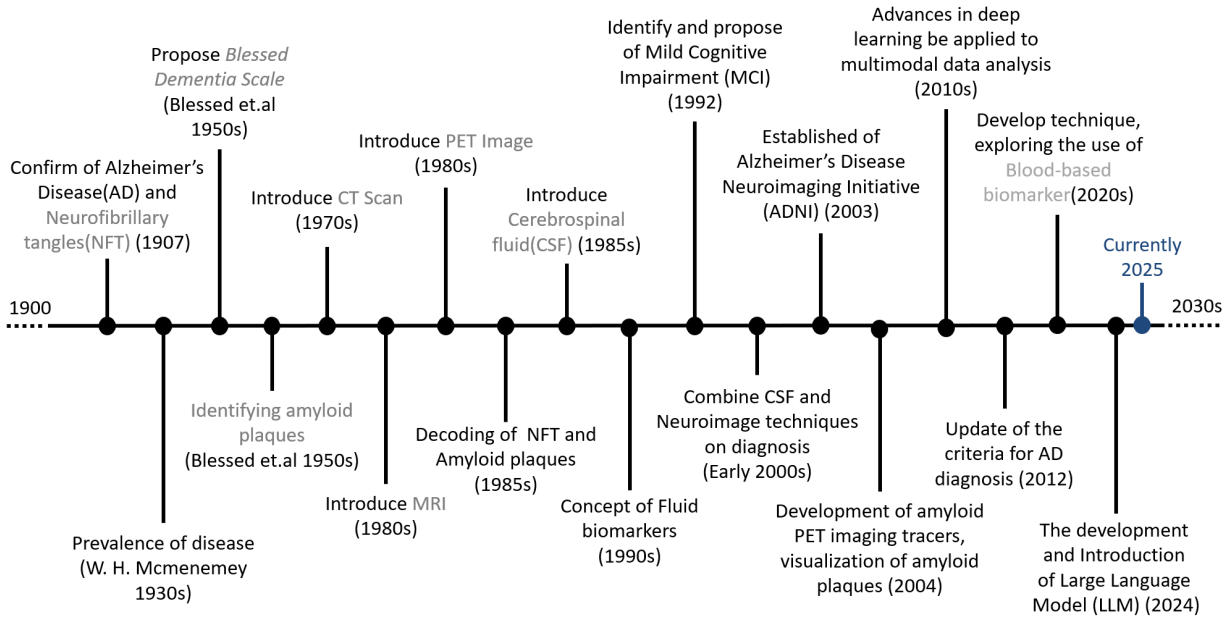


Figure 2: Historical development of AD research. From the 1950s onwards, studies initially focused on identifying key biomarkers of disease onset. Over the following decades, new imaging and analytical technologies were introduced to enhance diagnosis. After 2000, multimodal research became increasingly mainstream, and the emergence of DL after 2010 marked a new stage in multimodal AD analysis.

Nonetheless, early multimodal AD research faced substantial limitations. Although multimodal investigations existed since the early 1990s [16], several technical constraints hindered deeper exploration, including modality alignment difficulties, simplistic ML architectures incapable of sophisticated multimodal processing, and limited availability of paired multimodal data[54]. A significant turning point emerged around 2003 with the introduction of the Alzheimer's Disease Neuroimaging Initiative (ADNI) database, which markedly improved data availability and significantly reduced the barriers to image-based multimodal analyses. Simultaneously, research focus shifted from purely pathological investigations toward predicting disease progression and diagnostic staging [145]. This shift amplified research interest in image-based analyses, particularly MRI, resulting in notable improvements in AD diagnostic techniques [124]. As diagnostic tasks grew increasingly complex, single-modality approaches became insufficient due to limited informational content [123], further motivating interest in multimodal methodologies.

Despite this momentum, multimodal fusion and joint analysis initially remained underdeveloped, constrained by existing ML frameworks. Early multimodal implementations were rudimentary and lacked comprehensive analytical depth. A decisive advancement occurred with the maturation of structured multimodal fusion approaches [12] and the introduction of Multiple Kernel Learning (MKL) [144], particularly its adaptation within the AD research context [67]. These developments provided a clear methodological trajectory for multimodal fusion strategies in AD, significantly enhancing subsequent research directions.

Building upon earlier methodologies, DL models incorporate automated feature extraction processes, notably through convolutional neural networks (CNNs), substantially diminishing the reliance on manual preprocessing and significantly enhancing analytical efficiency for imaging data [147]. This automated and flexible feature extraction capability further enables DL architectures to accommodate more complex or heterogeneous data modalities, thereby considerably advancing the scope and depth of multimodal research in AD. Shen et al. [139] verified the outstanding performance of DL methods on large-scale datasets, particularly those featuring extensive imaging data. This series of studies has demonstrated that DL is both practical and effective, making it a widely recognised and influential field in medical imaging research.

3. Multimodal Data

3.1. Public Dataset

3.1.1. ADNI

To advance the effective treatment of AD, the ADNI database was established in 2004 under the leadership of Michael W. Weiner [171]. This open-access dataset is supported by a public-private partnership, funded by the National Institute on Ageing, the National Institutes of Health Foundation, the Alzheimer's Association, and numerous companies. The ADNI dataset has evolved through five phases, with the current iteration being ADNI-4. Each phase of the study recruits new participants for data collection and assessment. Currently, the dataset includes over 2,500 samples available for researchers. The dataset primarily comprises longitudinal data from participants across three stages: normal control (NC), MCI, and AD. It includes several of the most commonly used data types: clinical and cognitive assessments, MRI, PET, CSF biomarkers, blood biomarkers, and genetic data, along with additional data such as demographic information and lifestyle records that may impact the progression of AD [171]. The establishment of the ADNI dataset introduced an important standard by defining a unified protocol across different sites and platforms to ensure the consistent quality of collected images and CSF samples. This improvement has facilitated data sharing and increased the reuse rate, which is particularly significant for DL methods that require large sample sizes. The dataset can be accessed by: https://adni.loni.usc.edu/, and register for an IDA account for the data downloading application.

3.1.2. AIBL

The Australian Imaging, Biomarkers and Lifestyle Study of Ageing (AIBL) is a large-scale longitudinal cohort study initiated in 2006 by the Commonwealth Scientific and Industrial Research Organisation (CSIRO) in collaboration with several Australian institutions, including the University of Melbourne and the Austin Health Research Institute. Its primary goal is to investigate how biological, lifestyle, and genetic factors influence the onset and progression of AD. The AIBL dataset comprises approximately 1,100 Australian participants, including individuals with NC, MCI, and AD. The dataset provides multimodal data, such as T1-weighted MRI, PET imaging, blood biomarkers, apolipoprotein E(APOE) genotype, and comprehensive neuropsychological assessments. In addition to these medical and imaging data, AIBL also includes extensive lifestyle and behavioral data covering diet, physical activity, education, and mental health, enabling researchers to investigate modifiable risk factors associated with cognitive decline. Notably, AIBL is closely aligned with the ADNI initiative, following compatible imaging and clinical protocols to facilitate data harmonization and cross-cohort research. A subset of the AIBL data is also openly accessible through the ADNI Data Sharing Infrastructure (LONI-IDA), expanding the international reach of the ADNI and enabling large-scale meta-analyses across diverse populations. The AIBL dataset has become one of the most comprehensive multimodal resources for studying aging, disease prevention, and early diagnosis of AD, providing valuable support for statistical and DL models in this field. This dataset can be accessed and downloaded from the URL, which shares the same account as ADNI: https://aibl.csiro.au/adni/index.html

3.1.3. NACC

The National Alzheimer's Coordinating Center (NACC) is a large-scale, standardised database established in 1999 under the support of the National Institute on Aging (NIA) in the United States. It is managed by the University of Washington, NACC integrates clinical, neuropathological, genetic, and imaging data collected from more than 30 Alzheimer's Disease Research Centers (ADRCs) nationwide. The core objective is to provide a harmonised platform for sharing and analysing Alzheimer's disease data across multiple research sites and studies. The NACC dataset currently includes data from over 45,000 participants, spanning cognitively NC, MCI, and AD subjects, as well as individuals with other types of dementia. The dataset consists of the Uniform Data Set (UDS), which is a comprehensive collection of clinical, cognitive, functional, and behavioural assessments, alongside neuroimaging data (MRI, PET), genetic information (APOE, Genome-Wide Association Study), and post-mortem neuropathological results. Each participant undergoes regular follow-up evaluations, enabling longitudinal tracking of cognitive decline and neuropathological progression [15].

Through its standardised data collection protocols and continuous data sharing, NACC serves as a cornerstone for multi-centre and longitudinal Alzheimer's research. It enables researchers to study disease heterogeneity, validate biomarkers, and develop generalisable diagnostic and predictive models. Its integration with other major initiatives such as ADNI and AIBL further enhances its scientific utility, supporting cross-cohort analyses and advancing global Alzheimer's research collaboration. It can be accessed from the URL: https://naccdata.org/requesting-data/nacc-data

3.1.4. OASIS

Open Access Series of Imaging Studies (OASIS) is another open-source database, led by the Washington University Neuroimaging Laboratory, which aims to promote the public sharing of neuroimaging data. It was established later than ADNI, but it has also become one of the important databases for studying neurodegenerative diseases and normal ageing. Participants are recruited from the community, all submit initial blood samples, and undergo regular cognitive tests, neuroimaging examinations, and lumbar punctures. Currently, the database has developed to OASIS-3, with a total of about 1,400 subjects, including all participants in OASIS-2 [1]. The data content includes MRI, PET, blood data, cognitive assessments, and genetic information. The OASIS dataset emphasises data openness and accessibility, and covers subjects in various states such as NC, MCI, and AD. Its highly diverse samples and multimodal data types provide rich resources for cross-sectional and longitudinal studies of neurodegenerative diseases. To access and download the dataset, the following URL is available: https://sites.wustl.edu/oasisbrains/

3.1.5. IXI

The IXI dataset is part of the Information eXtraction from Images (IXI) project, which began in 2005 and was jointly released by University College London (UCL) and Imperial College London. This dataset does not contain any AD or other clinical labels. It provides high-quality MRI scans covering a broad age range and is designed to support research on imaging algorithms such as segmentation, registration, synthesis, normalisation, and brain structural analysis. The dataset comprises approximately 581 healthy volunteers with no history of neurological disorders. Data were collected at Guy's Hospital, Hammersmith Hospital, and the Institute of Psychiatry. The IXI dataset primarily includes comprehensive MRI modalities, such as T1-weighted, T2-weighted, Proton Density (PD), DTI, and Magnetic Resonance Angiography (MRA), accompanied by basic demographic metadata including age, gender, and scanning site. It is openly accessible and does not require registration or ethical approval. The dataset can be freely downloaded from the official project website: https://brain-development.org/ixi-dataset/

3.2. Biomarkers and Available Modality

Biomarkers, or biological markers, are measurable indicators that can be objectively assessed to reflect normal biological processes, pathological changes, or responses to therapeutic interventions [74]. Their definition is primarily based on their sensitivity, specificity, and clinical applicability to a particular disease. Establishing biomarkers for AD during the preclinical stage is a crucial prerequisite for early intervention, making it a significant focus in biomedical research [108]. Currently, biomarkers in the AD field can be broadly classified into two categories: fluid biomarkers and neuroimaging biomarkers. Fluid biomarkers are measurable substances found in bodily fluids such as blood, urine, and CSF. In contrast, neuroimaging biomarkers rely on imaging technologies to provide anatomical and functional insights, including indicators such as brain atrophy and metabolic alterations, typically presented in visual formats. The classification of biomarkers is largely determined by the measurement method or medium, which is distinguished by being viewed as different data modalities. Figure 3 shows the available and popular modality categories in AD research. For instance, tau protein detected in CSF is considered a fluid biomarker. Therefore, different modalities can be regarded as carriers or observation mediums for the corresponding biomarkers. For example, traits observed through imaging data from MRI and PET are classified as "imaging biomarkers," while information obtained from fluids such as CSF and blood falls under "fluid biomarkers." Given the current research trend towards minimising harm to patients, the concept of peripheral biomarkers has been introduced. This category is a subset of fluid biomarkers, derived from peripheral tissues. Hence, peripheral biomarkers are a subset of fluid biomarkers, with CSF, due to its invasive nature, classified as a fluid biomarker but not a peripheral biomarker.

The basis for AD diagnosis and its specificity originates from the identification of NFTs [20]. Subsequent research has established links between AD and other traits such as amyloid plaques [18], brain atrophy [76], white matter changes [46], and metabolic alterations [36], leading to the development of a diagnostic strategy that is not solely dependent on NFTs. This strategy also introduced imaging techniques such as CT, MRI, and PET to assist in observation and diagnosis. At the current stage, aside from post-mortem examinations, there remains no method that can definitively confirm an AD diagnosis [20]. In recent years, research on fluid biomarkers, particularly those found in blood and saliva, has made remarkable progress, offering new avenues for the clinical diagnosis of AD. However, the majority of work in the DL field still predominantly focuses on image data.

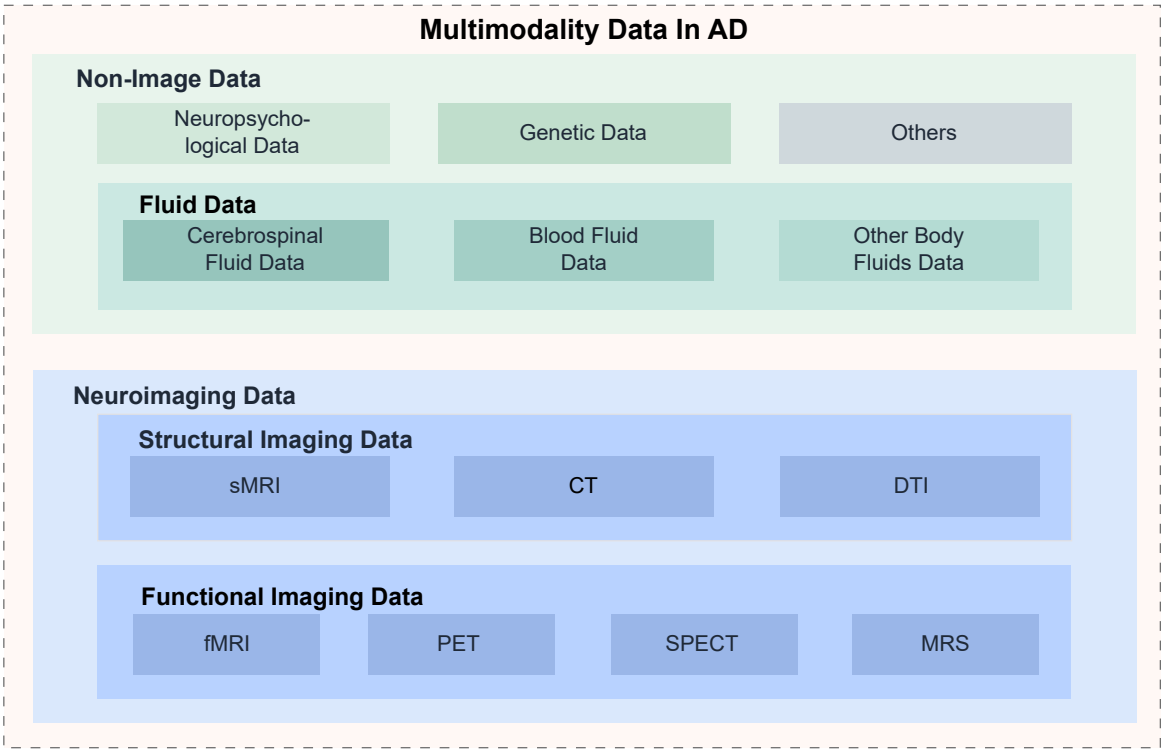


Figure 3: Overview of data modalities used in AD research. Data are broadly divided into image and non-image modalities. Neuroimaging modalities, such as MRI, PET mainly capture the structural and functional characteristics of the brain, while non-image data, including genetic, fluid, and clinical measures, provide complementary biological and cognitive information typically represented in tabular form.

3.3. Image Data

The diagnosis of AD fundamentally relies on histopathological confirmation, typically through post-mortem analysis to identify inherent biomarkers. The clinical uncertainty surrounding AD diagnosis has driven researchers to seek additional supportive measures. Neuroimaging technologies have expanded the scope of observable and summarised information, offering a window into brain changes and providing new biomarkers for AD diagnosis [171]. Currently, various brain imaging techniques are employed to study the changes associated with AD progression. These changes are broadly categorised into structural and functional types. Structural imaging focuses on providing higher spatial resolution and detailing anatomical features such as the shape, size, and integrity of organs, tissues, and other anatomical structures. Functional imaging, on the other hand, captures physiological processes within the body, including organ and tissue function as well as neurological and metabolic activities [158]. For ease of analysis, this paper classifies different techniques based on this framework rather than by the type of technology used.

3.3.1. Structural Image

Structural imaging focuses on depicting morphological alterations, including anatomical structure, regional volume, and other tissue-level changes that accumulate over time. Because such structural alterations usually occur later than molecular or metabolic abnormalities, they are less effective for early detection. However, once present, they tend to be highly persistent and quantifiable, providing valuable longitudinal biomarkers of disease progression. Consequently, structural imaging is particularly sensitive to the middle and late stages of AD, offering robust indicators of cortical atrophy, ventricular enlargement, and other macroscopic degenerative patterns. The commonly used structural imaging techniques include CT, structural MRI (sMRI), and Diffusion Tensor Imaging (DTI).

CT. CT was the earliest technique employed for AD detection, though it was not considered a standard diagnostic tool for AD [158]. Its main contribution was the introduction of a novel, non-invasive method for detecting structural changes in the brain, such as atrophy and cortical sulcus enlargement, which helped differentiate AD from other forms of dementia while reducing harm to the patient [76]. However, these changes typically occur in the later stages of AD, making CT useful primarily for confirming dementia types in already affected individuals rather than for early diagnosis. Despite this, CT remains widely used in some regions due to its lower cost and broad applicability.

sMRI. MRI is currently one of the most popular imaging techniques, initially garnering attention due to its ability to provide clearer and more detailed images compared to CT [48]. With the demonstrated correlation between white matter changes and AD progression, MRI, which offers more precise grey-white matter contrast [46], replaced CT as a crucial tool for AD detection. The discovery of the relationship between hippocampal and medial temporal lobe atrophy and cognitive deficits solidified MRI's role as a key medium in AD research [52].

The foundation of MRI lies in Nuclear Magnetic Resonance (NMR), a phenomenon where atomic nuclei with an odd number of protons or neutrons absorb and emit electromagnetic radiation in a magnetic field [90]. MRI machines generate a strong magnetic field that aligns the magnetic moments of hydrogen nuclei in the direction of the field. The machine then emits radio-frequency (RF) pulses at specific frequencies, causing the protons to deviate and rotate around the magnetic field axis. When the pulse is turned off, the protons relax and return to their initial state, releasing energy that the MRI receiver coils detect as RF signals, which are then used to create an image [19]. The magnetic field gradient, influenced by the equipment, applies additional magnetic fields along the X, Y, and Z axes, enabling the machine to selectively affect the resonance frequency in specific directions, thereby acquiring slice-by-slice images. Finally, mathematical techniques are used to convert the frequency and phase information in the signals into spatial information, resulting in detailed images.

Early MRI technology focused on anatomical analysis, which remains the primary application of sMRI. sMRI mainly captures detailed anatomical imaging. In current applications, the most direct indicators of a patient's condition include hippocampal atrophy, cortical thinning, and ventricular volume changes. Numerous studies have shown that sMRI can be used to estimate the progression from MCI to AD by observing hippocampal, temporal lobe, and entorhinal cortex volume and tissue damage [53]. Moreover, structural brain changes have been shown to correlate with stages of NFT deposition [172]. sMRI includes several imaging types, each focusing on different targets:

I) T1 and T2 weighted images. The definitions and differences between these imaging types stem from the relaxation times during the MRI proton relaxation process. T1-weighted imaging is based on T1 relaxation, also known as longitudinal relaxation, representing the time required for protons to realign with the magnetic field. Tissues with higher water content have longer T1 times. T2-weighted imaging, conversely, is based on T2 relaxation (transverse relaxation), which is the time it takes for protons in the transverse plane to lose phase coherence with each other, leading to signal decay. Contrast with T1-weighted images, areas with higher water content (such as CSF) appear brighter in T2-weighted images, while tissues with shorter T1 times (such as fat and bone marrow) appear brighter in T1-weighted images [133]. Hence, T1 and T2 are typically used in AD diagnosis to observe different targets. While T1 focuses on overall anatomical structure, assessing symptoms like hippocampal atrophy, T2 is more suitable for detecting tissue abnormalities, particularly fluid-related anomalies, and is often used to identify white matter lesions [53].

II) Fluid-Attenuated Inversion Recovery (FLAIR). In addition to T1 and T2, FLAIR imaging is another technique primarily used to suppress signals from fluid tissues such as CSF while maintaining the visibility of other fluid tissues. This technique relies on the RF pulse inversion recovery method, which inverts the net magnetisation vector of hydrogen protons. As these protons relax and realign with the main magnetic field, the differing T1 relaxation times of various tissues result in the nullification of signals from fluid tissues like CSF, without affecting the normal expression of signals from other tissues [81]. Compared to T1 and T2-weighted imaging, FLAIR is less commonly used in AD but is particularly useful for capturing subtle information that may be masked or obscured by fluid characteristics, often focusing on vascular changes and detecting mixed dementia to ensure accurate diagnosis [31].

DTI. DTI can be considered an extension of MRI technology, designed to capture the directional movement of water molecules within tissues, primarily used for imaging white matter tracts in the brain. Initially, DTI was commonly employed for stroke-related diseases closely associated with white matter integrity. Researchers later explored its potential in neurodegenerative diseases, discovering its exceptional utility in aiding AD diagnosis. The core principle of DTI involves water diffusion in the brain, where the random movement of water molecules within biological tissues is influenced by tissue structure. In white matter, water diffusion is anisotropic (FA), influenced by myelin and axonal structures, resulting in diffusion along the direction of nerve fibres [84]. DTI can therefore assess white matter integrity

and damage by measuring indices such as FA and mean diffusivity (MD). In AD detection, subtle structural changes in white matter often precede brain atrophy, making DTI a feasible technique for early diagnosis.

3.3.2. Functional Image

Functional imaging focuses on capturing dynamic changes in the brain, such as variations in metabolism, neural activity, and network connectivity. These signals are typically high-frequency and sensitive, reflecting short-term physiological responses such as compensation and network reorganisation. Because functional alterations often emerge earlier in the disease trajectory, they provide valuable biomarkers for early detection and subtype differentiation of AD. However, functional imaging data are also more susceptible to noise and external interference, including motion artefacts, physiological fluctuations, and environmental factors. Therefore, rigorous data denoising, preprocessing, and standardisation are essential to ensure reliable interpretation and reproducibility. Functional imaging commonly used in AD research includes functional MRI (fMRI), PET, magnetic resonance spectroscopy (MRS), SPECT, and various complex molecular imaging technologies.

fMRI. Contrary to sMRI, which derives from early MRI applications and conceptual transfers, fMRI represents a distinct imaging type designed to measure brain activity across different periods and states. Under the same hardware conditions, fMRI completes the functional transition by observing blood oxygen level-dependent (BOLD) signals, contrasting with sMRI methods. The core of fMRI is the study of haemodynamics, which can be simply understood as the correlation between neuronal activity and oxygen consumption, leading to increased local deoxyhaemoglobin. The body compensates for this by augmenting blood flow to active areas, resulting in a greater influx of oxygenated haemoglobin [56]. Research led by Seiji Ogawa [118] revealed that changes in blood oxygenation affect local magnetic fields, detectable via MRI. To accurately capture BOLD signals, fMRI employs echo-planar imaging (EPI) technology, enabling rapid imaging to capture dynamic changes in BOLD signals in real-time. In comparison to the repetition time (TR) of sMRI, fMRI typically operates within a TR of 1 to 3 seconds, shorter than T2 and T1 times, facilitating the capture of dynamic rather than static information. Additionally, the RF pulse frequency is adjusted based on TR, with smaller angles compared to conventional MRI settings, which are usually less than 90 degrees, to minimise signal saturation and enable continuous data collection at relatively short TRs. These subtle variations in BOLD signals are used to reflect neuronal activity, where areas of increased signal typically indicate heightened brain activity.

PET. In contrast to MRI's functional classification, PET categorises more intricately based on specific detection targets. Common types in AD research include Fluorodeoxyglucose-PET (FDG-PET), Amyloid PET, and Tau PET [117]. PET imaging relies heavily on radioactive tracers, molecules tagged with positron-emitting radioactive isotopes designed to mimic natural biomolecules like glucose and neurotransmitters. Each PET tracer targets specific biological processes or molecules within the body, generally administered via injection. Following this, the tracer naturally accumulates in regions of interest (ROI) driven by blood flow and biochemical properties. The emitted positrons from the radioactive isotopes interact with electrons in tissues, resulting in annihilation events that produce pairs of gamma photons travelling in opposite directions. The PET apparatus detects these gamma rays around the patient and reconstructs a three-dimensional image of tracer distribution based on the source of the rays [114]. Similar to MRI images, brightness can reflect the intensity of activity in corresponding areas, with brighter regions indicating a higher frequency of molecular annihilation events, thereby corresponding to metabolic activities or responses in those areas.

Current AD diagnostic research employs tracers corresponding to the aforementioned PET types: FDG, Amyloid, and Tau PET, where each of them concerns on specific element. FDG is one of the most widely used tracers, accumulating in cells with high metabolic activity, such as active regions in the brain. This tracer is primarily utilised to assess brain metabolic levels and detect regions of metabolic slowdown due to AD. Amyloid tracers specifically bind to amyloid plaques, visualising amyloid deposits in the brain, which is one of the key features of AD detection. Tau tracers bind to tau protein aggregates, representing another critical characteristic of AD, named NFTs, which can aid in understanding the extent of pathological spread in the brain [117].

MRS. The fundamental principle of MRS is similar to that of MRI, both relying on magnetic resonance principles. However, MRS focuses more on detecting and quantifying specific metabolites within tissues rather than providing detailed anatomical images. Its detection depends on the proton precession process influenced by the magnetic field, wherein protons experience slight frequency variations based on their chemical environment. These variations, known as chemical shifts, occur due to the influence of electron clouds on the magnetic field, resulting in protons exhibiting precession frequencies that slightly deviate from the baseline frequency. MRS can detect these chemical shifts; since different metabolites and compounds possess unique characteristic chemical shift patterns, MRS can differentiate various metabolites [156]. Changes in these metabolites often appear before visible structural changes on MRI,

thus holding significant importance for early diagnosis [57]. Additionally, examining metabolites aids in exploring the biochemical basis of AD, such as neurotransmitter systems and glial activation. However, the subtlety of these changes presents challenges for detection sensitivity, and their interpretation necessitates substantial expertise, which constitutes one of the higher barriers to utilising MRS.

SPECT. SPECT is a type of nuclear imaging technology, primarily differing in the tracers used and the imaging approach. SPECT employs tracers that emit single gamma photons instead of positrons. A gamma camera detects these photons, scanning around the patient and reconstructing images based on tracer locations. Common SPECT tracers for AD detection include two types: Technetium-99m (Tc-99m), Iodine-123 (I-123) [69]. Tc-99m is used to monitor regional cerebral blood flow. Upon injection, the tracer crosses the blood-brain barrier and accumulates in brain tissue proportionate to local blood flow, reflecting cerebral blood volume, which is often related to neuronal activity and function, indirectly indicating neuronal activity status. Notably, reduced blood flow in specific brain regions (e.g., parietal and temporal lobes) serves as an effective diagnostic feature for AD. I-123 similarly assesses cerebral blood flow based on its uptake in various brain regions [132].

3.4. Fluid Data

Fluid biomarkers represent a significant category distinct from imaging biomarkers. As the name suggests, they are primarily obtained by collecting bodily fluids and analysing their components for the diagnosis and progression analysis of AD. Compared to the complex types and techniques of imaging biomarkers, fluid biomarkers are simpler, with the predominant fluids being CSF and blood [134]. As research advances, saliva is also emerging as a promising subject for future studies [115]. Currently, except CSF, blood is used for the AD prediction task [162]. Their targets are quite similar, primarily focusing on biomarkers that directly reflect pathological changes associated with AD, such as amyloid beta and tau tangles. Because these biomarkers directly indicate the pathological changes of AD, fluid biomarkers can provide a more precise and effective reflection of disease progression compared to imaging biomarkers and related data [64].

CSF. CSF serves as a crucial foundation for fluid biomarkers and is a key reason for the subsequent development of fluid biomarkers. As a fluid biomarker in AD research, CSF is regarded as the gold standard due to its capability to detect biochemical changes directly occurring in the human brain. Circulating around the brain and spinal cord, CSF can directly reflect biochemical changes relevant to AD pathology. It enables the detection of various biomarkers and indicators, including APP Aβ42, tau tangles, neurofilament light chain (NfL), myo-inositol (ml) levels, and YKL-40 [119]. While the data obtained is highly reliable, CSF collection relies on lumbar puncture, which is more invasive than other methods.

Blood. With advancing research, the potential of blood in AD diagnosis has gradually been recognised. Analyses typically focus on plasma and serum independently [65]. For example, Ujiie et al. [154] found an increase in p97 levels in serum, a phenomenon exclusive to AD patients, highlighting the possibility of identifying previously unrecognised specific biomarkers associated with AD within blood. In addition to new biomarkers, blood can also be used to investigate effective biomarkers present in CSF. However, compared to CSF, the complexity of blood components results in lower stability, though it offers advantages in being less invasive and requiring only standard blood collection methods, thus enhancing its applicability and practical value.

Other Body Fluids. Furthermore, saliva and urine also hold research value. Similar to the differences and advantages of blood relative to CSF, both saliva and urine possess entirely non-invasive characteristics, providing significant advantages in routine screening and small-scale studies [115]. However, like blood, these fluids are less stable, with their analyses influenced by daily factors such as diet and lifestyle, making interpretation complex; thus, they remain in the early stages of research and are challenging to apply in practice.

Compared to imaging biomarkers, the advantage of fluid biomarkers, particularly those derived from CSF, lies in their ability to directly reflect biochemical processes and detect proteins or hormonal changes highly relevant to the aetiology of AD. Nonetheless, current DL methodologies predominantly focus on image data and imaging biomarkers. This is largely due to the inherent requirements of DL, which necessitate substantial sample sizes; imaging data typically offers greater volume than fluid biomarker data collected from sources like CSF. Additionally, advancements in specific image processing networks, such as CNNs, facilitate the analysis of imaging data within DL frameworks.

3.5. Genetic and Neuropsychological Data

Beyond imaging and fluid biomarkers, research into the genetic and genomic associations with AD has persisted since the disease was first identified. Currently, key genetic data relevant to AD include APOEε4, Presenilin 1/2

Table 1
Common Genetic Biomarker with Influence

Gene	Association	Impact	Relevance
APOEε4	Strong risk factor for late-onset AD	Increases amyloid-beta deposition and decreases clearance	Used for risk prediction; not definitive for diagnosis
PSEN1	Causes early-onset familial AD	Alters amyloid-beta production, leading to plaque buildup	Near certainty of early AD with mutations
PSEN2	Causes early-onset familial AD (less common)	Alters amyloid-beta production	Near certainty of early AD with mutations
APP	Causes early-onset familial AD (rare)	Increases production of aggregation-prone Aβ42	Near certainty of early AD with mutations
TREM2	Increases risk of late-onset AD	Impairs microglial response to amyloid and neuroinflammation	Highlights immune system's role in AD pathology

(PSEN1 / PSEN2), APP, and trigger receptor expressed in myeloid cells 2 (TREM2) [157]. Table 1 presents the functions and impacts of different genetic data types.

In addition to the aforementioned categories, neuropsychological assessment is also a crucial evaluative metric. Such assessments typically encompass several components: direct interaction between the clinician or expert and the patient (including conversations, questions, and answers), the execution of various tasks, and the completion of self-report questionnaires, thereby evaluating a range of cognitive functions in the individual being tested. To ensure the validity of the assessments, they are standardised, guaranteeing reliable and stable testing and evaluation across different environments and individuals. The scope of assessment is broad, covering memory capacity, attention, language function, visual abilities, and executive function, thus providing a comprehensive and objective evaluation outcome. Table 2 shows the contents of some of the more mainstream evaluation scales and compares the differences in their presentation. Given that AD is a longitudinal study, related assessments must be repeated over time to monitor changes in patient metrics, ensuring the tracking of AD's real-time progression [11].

Commonly used assessment tools in the field of AD include the Mini-Mental State Examination (MMSE) [11], Montreal Cognitive Assessment (MoCA) [75], and the Alzheimer's Disease Assessment Scale-Cognitive Subscale (ADAS-Cog) [125]. These primarily focus on a range of cognitive-related issues and tests, though they do not encompass all potential evaluative studies. The MMSE mainly assesses general cognitive function, while the MoCA focuses on the detection of MCI; the ADAS-Cog is specifically designed for evaluating cognitive functions related to AD. Table 2 below presents the assessment items and methodologies associated with these three tools.

3.6. Other Modalities Data

As research advances, an increasing number of physiological markers have been identified as associated with AD, or as potential aids in AD diagnosis. Park et al. [121] confirmed the effectiveness of virtual reality (VR)-based functional indicators (such as hand movement velocity, scanning path length, and completion time) as biomarkers. They integrated these indicators with MRI data in a multimodal analysis, achieving superior diagnostic results compared to using MRI alone. Similarly, previous research [135] has established a link between abnormal eye movements and AD progression. Inspired by these findings, Yin et al. [179] employed eye-tracking data for cloud-based screening of abnormal eye movements via Internet of Things (IoT), successfully identifying AD. Collectively, these studies and associated clinical assessments indicate a clear relationship between external symptomatic manifestations and disease progression. Consequently, pathological diagnosis no longer solely depends on traditional methods, opening critical avenues for future research and more accessible clinical application.

Table 2
Common Mental Test Evaluate Factor

Cognitive Domains	Mini-Mental State Examination (MMSE)	Montreal Cognitive Assessment (MoCA)	Alzheimer's Disease Assessment Scale-Cognitive Subscale (ADAS-Cog)
Focus	Assesses general cognitive function	Detect mild cognitive impairment (MCI) and assess various cognitive functions.	Evaluate cognitive function in Alzheimer's disease
Orientation	Questions about time and place	Null	Null
Attention	Simple arithmetic tasks	Tasks like serial subtraction and repeating sentences	Tasks assessing sustained attention and concentration
Calculation		Null	Null
Memory	Immediate and delayed recall of words	Recall of words after a delay	Recall of words and stories
Language	Naming objects, following commands, and repetition	Verbal fluency tasks and comprehension	Word-finding, naming, and following commands
Visual-Spatial Skills	Copying a simple design	Drawing a clock or connecting numbers and letters	Tasks that involve drawing and visual problem-solving.
Executive Function	Null		Null
Naming		Identifying animals in pictures	

4. Multimodal Fusion

While the last part discussed the various modalities available for AD research, this part shifts focus toward how these heterogeneous data sources can be effectively combined through multimodal fusion frameworks. Multimodal Fusion holds a wide range of definitions in different research fields and scenes, typically representing a kind of approach to analysing multimodal data. It should first be clarified that the concept of "fusion" discussed in this paper mainly serves the topic of multimodal analysis in ML, rather than generalised image or data fusion techniques. The application of multimodal data fusion in the field of ML primarily focuses on the joint analysis of data, with the goal of integrating the distinct information expressed by multiple modalities. This fusion leverages the complementary nature of the data to enable more comprehensive and precise analysis for downstream tasks. Multimodal research has broad applicability across various industries, and even within the medical domain, numerous specialised branches exist depending on the target condition or application. Previous reviews have typically provided comprehensive analyses from a macro perspective, summarizing multimodal approaches according to criteria such as learning paradigms (supervised, unsupervised, or semi-supervised), data types, and traditional fusion strategies [35, 190]. Despite these reviews offering detailed and systematic frameworks in a wide range of aspects, these works still struggle to deliver precise guidance regarding the applicability or practical utility of specific models to targeted diseases due to inherent task-specific differences. Here, our work narrows the scope explicitly to multimodal research related to AD. By closely examining mainstream approaches within the AD domain, we systematically summarise these studies, which used network modules and architectural strategies, and classify them according to their practical applications.

4.1. A Refined Framework for Classifying Fusion Methods

Fundamental fusion strategies were originally defined within traditional ML frameworks, categorizing multimodal integration according to the stage at which fusion occurs. Atrey et al. [12] were among the first to clearly categorise fusion methods into early (feature-level) fusion and late (decision-level) fusion, and a combined approach known as hybrid fusion. This classification standard has been widely recognised and adopted, serving as the foundation

for many subsequent studies. In addition, some early-stage surveys generally regarded neural networks as a minor subcategory within multimodal fusion methods in traditional ML [13], where they carried the classification forward into the evaluation of various DL architectures. It is undeniable that, while DL has rapidly gained prominence since around 2010 due to its convenience and powerful data analysis capabilities, traditional ML continues to hold unique advantages, especially in medical contexts that require high interpretability of data and features. However, given the rapid growth and specialised functions of contemporary deep neural networks, this topic warrants independent analysis.

Recent works began to explicitly separate it from traditional ML within multimodal analysis. For example, Elazab et al. [44] provided a comprehensive analytical framework that examined Alzheimer's disease diagnosis from both ML and DL perspectives, distinguishing between single- and multimodal studies. Their survey offered a detailed summary of existing network architectures and methodological approaches, while highlighting the transition toward multimodal strategies and the key roles of feature selection, regularisation, and model optimisation in this process. While this separation marks a clear step forward, it also reveals the limitations of current taxonomies. A key divergence between DL and ML challenges the traditional classification framework, that is DL approaches eliminate the need for manual feature extraction. DL methods inherently automate feature extraction within neural network architectures, directly affecting subsequent fusion modules and downstream analytical tasks. Consequently, the fusion outcomes are no longer exclusively determined by fusion modules alone; feature extraction mechanisms within networks also play a critical role, challenging traditional conceptions of multimodal fusion processes.

Following traditional categorisation schemes, both feature extraction and subsequent fusion steps would be broadly classified as "feature-level fusion." Yet, this classification is overly coarse and neglects the specific architectural designs and functionalities distinctive to DL. Similar limitations apply to traditional ML frameworks, where the developed architectures have increasingly expanded beyond mere classification, assuming additional roles and complexities. Building upon this observation, this paper introduces a novel, more fine-grained classification scheme for multimodal fusion. It redefines fusion according to the operational roles performed within model architectures, aiming to systematically analyse and distinguish existing fusion strategies and architectures within both traditional and DL contexts from a refined and comprehensive perspective.

4.2. Traditional Machine Learning Fusion

Traditional ML methods rarely use the raw representation of data directly as model input. Instead, they rely on manually engineered feature extraction and selection processes, heavily dependent on expert domain knowledge [43]. For example, in medical imaging analysis, ROIs are often manually defined, and in clinical data analysis, highly relevant variables are selectively identified through expert input. As a result, traditional ML models typically involve limited network components, with the major workload focused on utilizing algorithms, such as SVM, Random Forest, and Logistic Regression, to analyse these manually prepared feature sets and subsequently produce the final outcomes. This situation lasted until about 2010, with the introduction of multimodal analysis [67]. It causes modalities applied in AD research to become more complex, and the demand for data processing increased; traditional network architectures struggled to meet the growing analytical needs and challenges posed by the data. Consequently, more diverse architectures were proposed to address these complex data issues and feature selection strategies, leading to a situation where earlier classification methods could no longer fully reveal the variety of contemporary network architectures. Baltrusaitis et al. [13] further expanded this classification: Because the initial classification standard does not depend on specialised network architectures, they were labelled as Model-Agnostic Approaches. In contrast, fusion methods that include specialised feature-processing modules are categorised as Model-Based Approaches. Notable examples of these include MKL, Multi-Task Learning (MTL), and Graphical Networks(GN). This framework provides intuitive and effective guidance, offering significant value to our work. In fact, even within Model-Based Approaches, it is still possible to divide methods into early and late stages based on the fusion phase and objects involved [148, 42]. Therefore, in this paper, we will not discuss this categorisation separately but will integrate it as a fusion strategy or stage in a unified discussion with other methods. Based on this, the section will expand on the current widely accepted classification methods, exploring variations of models within different categories, providing more detailed classification criteria, and analysing the advantages and disadvantages of each approach.

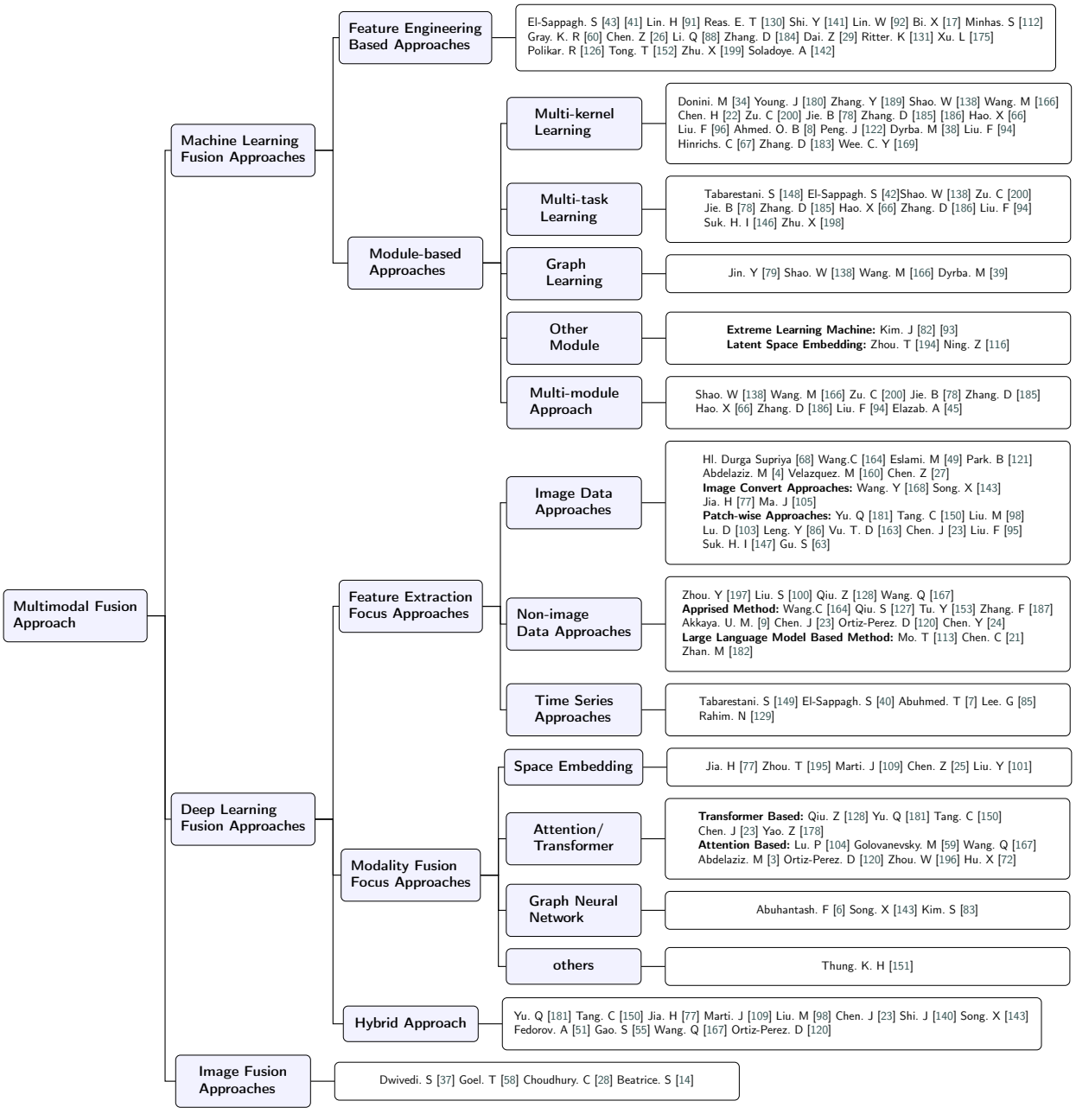


Figure 4: Taxonomy of multimodal fusion methods. The first-level classification divides approaches into three main categories: ML–based fusion, DL–based fusion, and image-level fusion. ML fusion methods are further organised by the complexity and autonomy of their frameworks, progressing from feature-engineering-based models to modular and hybrid architectures. DL fusion methods are classified according to their core functional modules, highlighting distinct design focuses such as feature extraction, modality interaction, and hybrid integration. Image fusion methods, presented as a complementary category, focus on pixel-level synthesis to generate composite images that integrate information across modalities.

Table 3
ML Based Multimodal Analysis Method: Part I

Author	Year	Modalities	Approach	Applied Module	Task	Evaluation Matrix(%)		
						SEN	SPE	ACC
Tabarestani, S [148]	2020	MRI, PET, CSF, Cognitive assessment tests(COG)	Module-Based	MTL	Regression	NA	NA	NA
El-Sappagh, S [42]	2021	MRI, PET, Cognitive and functional assessment, Genetic	Module-Based	MTL	NC vs MCI vs AD	NA	NA	93.30
Donini M [34]	2016	MRI, Clinical Score(CS)	Module-Based	MKL	NC vs AD	NA	NA	92.38
Young, J [180]	2013	MRI, FDG-PET, CSF, Genetic	Module-Based	MKL	sMCI vs cMCI	90.00	52.40	72.20
Zhang, Y [189]	2021	sMRI, fMRI, PET, DTI	Module-Based	MKL	NC vs AD	NA	NA	94.58
					NC vs MCI			81.02
					MCI vs AD			89.63
					NC vs MCI vs AD			88.54
El-Sappagh, S [43]	2021	CS, disorder medication history	FE-Based	Null	AD vs NC vs pMCI vs sMCI	NA	NA	90.85
El-Sappagh, S [41]	2022	MRI, CSF, CS neuropsychological battery markers (NSB)	FE-Based	Null	AD vs NC vs pMCI vs sMCI	84.88	NA	84.95
Chen, Z [26]	2024	sMRI, PET	FE-Based	Null	NC vs AD	83.65	96.12	93.40
					NC vs MCI	65.54	63.40	64.64
					sMCI vs pMCI	59.80	91.42	82.39
Jin Y [79]	2016	MRI, PET, CS, Genetic	Module-Based	GN	Regression	NA	NA	NA
Shao, W [138]	2020	MRI, FDG-PET	Module-Based	MKL,MTL,GN	NC vs AD	94.08	90.44	92.51
					NC vs EMCI	86.09	78.55	82.53
					EMCI vs LMCI	83.84	63.26	75.48
Wang, M [166]	2023	MRI, PET, Single Nucleotide Polymorphism (SNP)	Module-Based	MKL,GN	NC vs AD	97.83	94.68	96.48
					NC vs LMCI	88.46	73.16	81.35
					EMCI vs LMCI	93.36	52.15	76.38
Kim, J [82]	2018	MRI, FDG-PET, CSF	Module-Based	Extreme Learning Machine(ELM)	NC vs MCI	75.00	91.92	87.09
Lin, H [91]	2023	MRI, DTI, Genetic	FE-Based	Null	SCD vs NC	NA	NA	83.13
Reas, E T [130]	2023	MRI, CS, Genetic	FE-Based	Null	Regression	NA	NA	NA
Park, B [121]	2024	MRI, VR	FE-Based	Null	NC vs MCI	100.00	90.90	94.40
					AD vs NC	95.10	96.54	95.95
Zu, C [200]	2016	MRI, FDG-PET	Module-Based	MKL,MTL	NC vs MCI	84.95	70.77	80.26
					ncMCI vs cMCI	66.74	71.43	69.78
Shi, Y [141]	2019	MRI, PET, CSF	FE-Based	Null	NC vs AD	95.37	94.67	95.00
					NC vs MCI	84.75	73.02	80.71
Chen, H [22]	2022	fMRI, sMRI, DTI	Module-Based	MKL	SCD vs NC	80.56	97.14	88.73
Jie, B [78]	2015	MRI, FDG-PET	Module-Based	MKL,MTL	NC vs AD	94.90	95.00	95.03
					NC vs MCI	85.86	66.54	79.27
					ncMCI vs cMCI	64.65	71.79	68.94
Li, Q [88]	2017	sMRI, PET	FE-Based	Null	NC vs AD	99.25	95.61	97.36
					NC vs MCI	75.00	80.70	77.66
Zhang, D [185]	2012	MRI, FDG-PET, CSF	Module-Based	MKL,MTL	NC vs AD	NA	NA	93.30
					NC vs MCI			83.20
					ncMCI vs cMCI			73.90
					Regression			NA
Hao, X [66]	2020	MRI, FDG-PET Voxel-based Morphometry(VBM)	Module-Based	MKL,MTL,GN	NC vs AD	98.43	96.73	97.60
					NC vs MCI	94.04	66.15	84.47
					ncMCI vs cMCI	67.44	85.54	77.76
Lin, W [92]	2021	MRI, PET, CSF, Genetic	FE-Based	Null	NC vs AD vs MCI	NA	NA	66.70
					NC vs AD vs pMCI vs sMCI			57.30
Zhang, D [186]	2011	MRI, FDG-PET, CSF	Module-Based	MKL	NC vs AD	93.00	93.30	93.20
					NC vs MCI	81.80	66.00	76.40
					ncMCI vs cMCI	NA	NA	NA
Bi, X [17]	2020	fMRI, SNP	FE-Based	Null	NC vs AD	NA	NA	90.00
Liu, F. [96]	2013	MRI, CSF	Module-Based	MKL	NC vs pMCI	93.26	87.49	90.56

Table 4
ML Based Multimodal Analysis Method: Part II

Author	Year	Modalities	Approach	Applied Module	Task	Evaluation Matrix(%)		
						SEN	SPE	ACC
Minhas, S [112]	2017	MRI, CS	FE-Based	Null	sMCI vs pMCI	85.45	92.31	84.29
Polikar, R [126]	2010	MRI, PET, EEG	FE-Based	Null	NC vs AD	NA	NA	85.50
Hinrichs, C. [67]	2009	MRI, PET	Module-Based	MKL	NC vs AD	78.52	81.76	81.00
Zhang, D [183]	2011	MRI, PET, CSF	Module-Based	MKL	NC vs AD	93.10	92.50	92.80
					NC vs MCI	86.50	84.90	85.70
Tong, T [152]	2017	MRI, FDG-PET, CSF, Genetic	FE-Based	Null	NC vs AD	88.90	94.70	91.80
					NC vs MCI	85.10	67.10	79.50
					NC vs AD vs MCI	NA	NA	60.20
Wee, C. Y [169]	2012	rs-fMRI, DTI	Module-Based	MKL	NC vs MCI	NA	NA	96.30
Dyrba, M [39]	2020	MRI, FDG-PET, AV45-PET	Module-Based	GN	Regression	NA	NA	NA
Zhou, T [194]	2019	MRI, PET, Genetic	Module-Based	Latend Space Embedding	NC vs AD vs sMCI vs pMCI	NA	NA	70.20
Zhu, X [199]	2014	MRI, PET, CSF	FE-Based	Null	NC vs AD	95.70	98.60	95.90
					NC vs MCI	98.00	60.10	82.00
					ncMCI vs cMCI	48.50	94.40	72.60
					Regression	NA	NA	NA
Zhu, X [198]	2015	MRI, PET	Module-Based	MTL	NC vs AD vs ncMCI vs cMCI	NA	NA	61.06
Ning, Z [116]	2021	MRI, PET	Module-Based	Latent Space Embedding	NC vs AD	95.70	97.50	96.50
					NC vs MCI	NA	NA	79.90
					sMCI vs pMCI	NA	NA	82.40
Gray, K. R [60]	2013	MRI, PET, CSF, Genetic	FE-Based	Null	NC vs AD	87.90	90.00	89.00
					NC vs MCI	77.50	67.90	74.60
					sMCI vs pMCI	57.10	58.70	58.00
Dai, Z [29]	2012	MRI, rs-fMRI	FE-Based	Null	NC vs AD	87.50	90.91	89.47
Ritter, K [131]	2015	MRI, PET, CS, Genetic	FE-Based	Null	sMCI vs pMCI	NA	NA	73.00
Xu, L [175]	2015	MRI, PET	FE-Based	Null	NC vs AD	95.60	94.00	94.80
					NC vs MCI	66.40	82.10	74.50
					sMCI vs pMCI	74.10	81.50	77.80
Dyrba, M [38]	2015	MRI, rs-fMRI, DTI	Module-Based	MKL	NC vs AD	NA	NA	85.00
Liu, F [94]	2014	MRI, PET	Module-Based	MKL,MTL	NC vs AD	94.71	94.04	94.37
					NC vs MCI	84.85	67.06	78.80
					sMCI vs pMCI	67.83	70.00	67.83
Suk, H. I [146]	2014	MRI, PET, CSF	Module-Based	MTL	NC vs AD	94.00	98.33	95.27
					NC vs MCI	88.89	67.33	80.07
					MCI vs AD	46.67	94.00	74.60
					ncMCI vs cMCI	59.00	88.00	72.02
Soladoye, A [142]	2025	Multi-modal CS	FE-Based	Null	NC vs AD	94.00	NA	95.00
Lin, W [93]	2020	MRI, PET, CSF, Genetic	Module-Based	ELM	sMCI vs pMCI	85.45	92.31	84.29
Ahmed, O. B [8]	2017	MRI, DTI, CSF	Module-Based	MKL	NC vs AD	82.92	97.20	90.20
					NC vs MCI	71.58	86.05	79.42
					MCI vs AD	65.62	81.33	76.63
Zhang, D [184]	2011	MRI, PET, CSF	FE-Based	Null	NC vs AD	85.70	94.30	89.60
Peng, J [122]	2019	MRI, PET, SNP	Module-Based	MKL	NC vs AD	97.30	94.90	96.10
					NC vs MCI	85.60	69.80	80.30
					MCI vs AD	65.90	82.70	76.90

Compared to early ML methods, the increasing complexity of data and depth of analysis has resulted in an expanding volume of data that the models must process, creating a computational burden. Hence, feature selection, one of the key steps in the traditional ML workflow, has been emphasised. Initially, a manual step in traditional ML, feature selection has become increasingly difficult to manage manually as data volumes grow. Consequently, the function of feature selection gradually shifts to relying on network architecture, which also improves the development of the network structure. The complexity of networks means that current research typically does not rely on just one decision-making strategy but instead adopts multiple methods simultaneously. As such, simple definitions are insufficient to

classify the current approaches. In this context, we categorise networks based on their structure and the specific tasks they are designed to address: Feature Engineering (FE) based network, used to call the methods that follow traditional classification strategies, where feature extraction and selection depend on manual efforts; Module-based network, which includes feature selection, fusion, and other processes. Table 3and Table 4 display recent works, containing their methods and model performance as reference.

4.2.1. Feature Engineering-Based Approaches

FE-Based Approaches contains the methods that require complete handcrafted work on the data preprocessing, including feature extraction and selection. Traditional classification strategies(Feature-level, decision-level, and hybrid fusion) are performed more intuitively and clearly in this category. The FE-based approach does not involve complex network modules or architectures for additional feature processing or analysis, but is primarily influenced by manual feature extraction techniques and the subsequent work of classifiers. Early fusion participates by integrating the extracted features (for example, by directly concatenating or using simple mathematical methods for correlation), while late fusion occurs at the output stage, independently analysing and making judgments for each modality, then combining the results of these analyses for the final decision. As stated in [13], these methods do not depend on network architectures and do not impose any specific requirements on subsequent classifiers or regressors, making them the most intuitive and easily applied fusion methods in research. Figure 5 shows a general process of the FE-based ML.

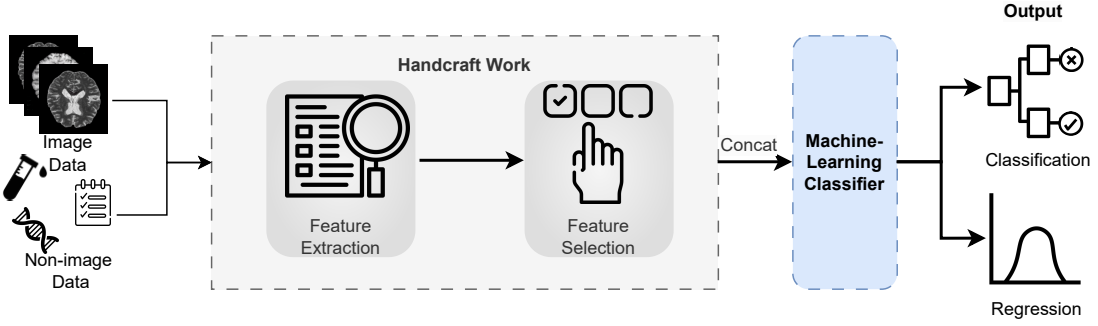


Figure 5: Feature-engineering-based ML framework for multimodal analysis. In early-stage AD research, both feature extraction and feature selection depend on handcrafted processes, such as manual region measurement, statistical filtering, or clinical variable selection. The ML component functions primarily as a classifier, integrating pre-engineered features from different modalities to produce the final diagnostic or predictive outcome.

Nevertheless, the simplicity inherent in traditional ML comes with challenges, notably limiting the model's capability to deeply interpret feature structures and effectively identify complementary or underlying modal relationships. To mitigate these shortcomings, researchers have pursued advanced improvements focused primarily on the feature extraction phase through manual engineering strategies. For instance, Reas et al. [130] developed a survival analysis-driven architecture employing Cox regression for feature selection and integration, thereby maintaining interpretability while improving clinical applicability. In contrast, Shi et al. [141]introduced an innovative non-linear feature extraction method by computing Pearson correlation coefficients between features and their power-transformed variants, systematically capturing non-linear interdependencies. This approach constructs a novel high-dimensional, non-linearly coupled feature space. These methods prioritise sophisticated feature processing techniques designed explicitly to enhance modality interactions or uncover hidden intra-modal characteristics, ultimately improving the expressive power of the model. Another strategy involves improving the performance of the model by modifying the classifier structure. For example, Lin et al. [92] introduced a unified data processing approach during the feature extraction phase using Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA). Based on LDA scores, they assessed disease progression indicators and proposed a novel binary Extreme Learning Machine (ELM) to perform multi-level classification using a decision-tree strategy. This method yielded reliable results in more complex three-class and four-class tasks. Additionally, Bi et al. [17] performed feature integration through correlation analysis,

followed by deepening the feature matrix with a newly improved clustering-based Evolutionary Random Forest, which also produced excellent results. Besides, recent work [142] reinforces this advantage of interpretability. They proposed a composite feature selection strategy that pipelines the feature selection process and introduces global and local cross-validation, ultimately providing effective guidance and detailed implementation planning for future work in conjunction with clinical scenarios.

At the early stage, due to the relatively shallow depth and technological limitations that hinder the adoption of more complex network architectures, most work in this area still relies on manual processes and correlation calculation methods. It is based on a mathematical theory, which requires complex handcraft to process the features. In contrast, the primary goal of the network, therefore, is simply to perform the final analysis based on the features that have been extracted. While these methods remain valuable for certain tasks, their inability to explore deeper intermodal relationships or adapt to more intricate data patterns limits their applicability in addressing more complex research questions. However, this does not mean that traditional methods have become obsolete. Compared with subsequent work, interpretability has always been a huge advantage of these methods. In disease fields that emphasize clinical analysis and pathological research, traditional ML has always been an effective competitive force.

4.2.2. Module-Based Approaches

The biggest difference from the FE-based approach is that features are no longer simply processed manually and directly input into classifiers for analysis. Instead, they are further integrated through specific network modules [8]. In the module-based approach, feature selection, which was originally a manual process, is now transferred to the network itself [185]. Compared to previous methods that used a specific architecture for data analysis, researchers have introduced composite multi-network architectures to select and integrate the extracted features [138, 200]. This not only increases the volume of data that can be processed but also enhances the ability to mine latent correlations between the data, thus improving model performance. This method breaks from the traditional use of ML methods as just feature fusion and classification tools, and instead, it focuses more on mining the data's inherent information and relationships. Figure 6 shows the transformation of the research core in the workflow. ML begins to bear a greater proportion of data analysis and turns into a complex composite structure. Currently, the most commonly used modules include MKL, MTL, and GN, which are applied at different stages, such as feature selection, feature fusion, or both tasks simultaneously.

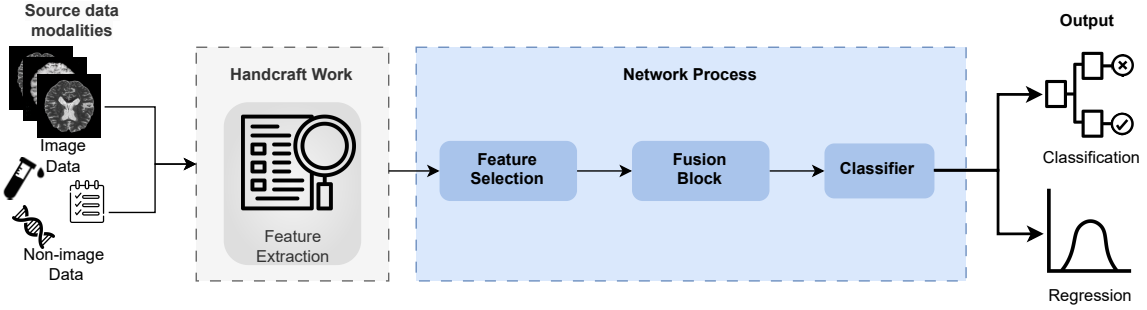


Figure 6: Module-based ML framework for multimodal analysis. In this category, ML algorithms begin to play a more active role beyond simple classification. Instead of relying solely on handcrafted features, the model incorporates functional modules to identify and combine the most relevant representations from each modality. This design allows ML to assume greater responsibility for data integration and analysis, forming a transitional stage between traditional feature-engineering workflows and modern DL architectures.

Multi-Kernel Learning. Among them, the most widely used method is MKL. Hinrichs, C. et al. [67] were the first to introduce the MKL method to multi-modal AD research, and they proposed the EasyMKL framework, which is based on robustness and the suppression of abnormal information. By assigning different modalities to different kernel spaces and distributing weights, this laid a solid foundation for subsequent multi-modal studies. The MKL method can be seen as an extension of SVM, with the key idea being the linear combination of multiple kernel functions to form a new kernel.

Specifically, a certain modality or feature is represented as an independent kernel, where different kernels focus on the feature information of their respective data and project it into a high-dimensional space. Afterward, a classifier calculates the weights and margins of the kernels to select those that contribute most to the final result. Compared to other methods, MKL can be viewed as directly integrating both feature selection and feature fusion, with a more intuitive focus on the characteristics of multi-modal data itself. It is easy to use and can be simply extended to handle varying numbers of data or modalities without the need for complex module designs. For this reason, MKL has become one of the most common multi-modal fusion modules in the AD research field.

To date, MKL has been widely adopted in various AD studies as a mainstream multi-modal fusion module, and several variations of the method have been developed. Donini, M. [34] et al. improved upon the EasyMKL framework by integrating feature selection methods, further refining the kernel design strategy. They treated each voxel and clinical feature as independent kernels and applied sparse feature selection to allocate kernel weights, selecting the most informative feature sets. This approach allows for more detailed handling of complex high-dimensional features and improves the model's sensitivity to different regions and information within modalities. It also enhances the model's interpretability and ability to process large-scale data. This method includes both feature selection and feature fusion tasks, producing a final result for classification. In contrast, Zhang, Y. [189] proposed to integrate all multi-modal features into the same kernel, using the $l_{2,1}$-norm for feature selection. Different kernels ensured the diversity of their analysis, making the model robust to various types of data distributions. Moreover, compared to traditional MKL methods, this approach focuses more on the relationships between modalities, rather than within a single modality, and, compared to the voxel and feature-based multi-kernel setups, this MKL reduces computational cost. However, this does affect the model's ability to judge the influence of a single modality, reducing interpretability. The same $l_{2,1}$-norm is also used in [96], which further proposes using Random Fourier Features (RFF) to approximate Gaussian kernels, reducing the model's computational complexity.

Clearly, MKL is flexible and easy to implement. By adjusting the number and distribution of kernels, the model can dynamically emphasise specific modalities or feature types, enabling researchers to focus on targeted objectives. In addition, since the MKL loss function is typically convex, it guarantees convergence to a global optimum, simplifying the training process [13]. However, its main limitation lies in the manual design and allocation of kernels. An inappropriate kernel choice can lead to suboptimal or even adverse multimodal fusion performance.

Multi-Task Learning. MTL shares some similarities with MKL. In fact, Baltrusaitis, T. et al. [13] categorised both MKL and MTL under kernel-based methods, proving their connection. However, this classification is somewhat misleading, as the core component of MTL is the task assignment strategy and the information sharing between related tasks, rather than focusing on the influence of a particular modality or feature. In fact, aside from kernel-based multi-task methods, multi-task architectures have shown advantages in various applications [148]. Even in AD diagnosis research, MTL frameworks focus on different aspects, such as explaining task relationships, feature selection [200], and feature learning [42]. These methods have now been successfully applied to longitudinal data, advancing research in neurodegenerative diseases [192].

Suk, H. I. et al. [146] applied MTL to the AD multi-classification task, treating each task as a cluster and learning the importance of different features for each cluster while capturing shared features important for different tasks to assist in subsequent classification. El-Sappagh, S. et al. [42] used an innovative architecture, moving away from the traditional MTL framework. They built a multi-layer network architecture that focuses on different tasks: the first layer handles the multi-classification task, while the second layer deals with the binary classification task, predicting the progression of patients from MCI to AD over three years. Like traditional MTL methods, both tasks share feature inputs and are capable of extracting shared features. Additionally, their other work applied MTL methods to multi-modal data analysis (e.g., MRI, PET) and temporal sequence data (e.g., baseline, 6 months, and 12 months). This study leveraged MTL's advantages in longitudinal research, analysing multi-modal data at different time points, and finally fused the results at the decision level for the final output [148].

In general, MTL decomposes a complex problem into several related sub-tasks, encouraging shared representations and mutual knowledge transfer. This strategy enhances both interpretability and performance, as tasks can reinforce each other through shared learning. Nevertheless, the strength of task relatedness directly determines model performanc: weakly correlated tasks may cause negative transfer, thereby degrading accuracy. Moreover, conventional MTL frameworks do not inherently address cross-modal alignment, limiting their ability to capture deeper multimodal representations.

Graphical Model. Another common method is the GN-based approach. The core of this method is to represent variables and their interdependencies through graphical structures. In AD research, this can be simply viewed as

internal nodes representing features, and edges between the nodes indicating correlations between different features [166]. These methods can be roughly divided into directed and undirected graphs, represented by Bayesian networks and Markov networks, respectively. For directed graphs, the focus is on internal causal relationships, with edges indicating the direction of influence between features. In contrast, undirected graphs focus more on the joint probability distribution of feature sets. Compared to MKL and MTL methods, GN structures offer higher interpretability, providing intuitive feedback about the dependencies between different features and parameters, which helps to better understand the relationships and interactions between multi-modal information and the underlying causes of diseases [79].

In contrast to the model improvement strategies of the former approaches, GN methods rely more on inference. Jin, Y. [79] modified the traditional Bayesian method by proposing a hybrid Bayesian network that uses Conditional Probability Distributions (CPD) to handle discrete data and Probability Density Functions (PDF) to handle continuous data, allowing the network to process both types of input information simultaneously. It also defines dependencies between variables, allowing mutual influence between different types of variables. This approach enables the handling of heterogeneous multi-modal information and tolerates noise and missing data, reducing the negative impact of data quality on the network. Dyrba, M. [39] proposed an alternative strategy based on a Gaussian Graph Network, transforming a directed graph into an undirected one. This focuses on the conditional dependencies between features, rather than causal relationships. This modification better matches the current state of AD research, as the relationships between many brain regions are not dependent on causality. This change helps capture those associations more comprehensively. Notably, Markov networks, another type of undirected graph, are seldom used in AD research because their primary function is to model sequential and regular changes. They are less effective for complex multi-modal scenarios and non-linear relationships, which limits their utility in AD research. Therefore, graph-based methods in AD research mainly focus on Bayesian networks.

Graph-based methods excel at modelling inter-relationships and population structures while maintaining high interpretability, as they explicitly represent associations among multimodal entities. Such methods are particularly advantageous for cohort-level or relational studies, where network topology itself conveys diagnostic information. However, their performance is highly dependent on graph construction: errors in similarity definition or prior selection can severely impair the model's learning and output stability. Consequently, these methods often require substantial domain knowledge to design meaningful graph priors.

Multi-Module Model. The previous section discussed the most commonly used network modules in ML methods. At the current stage of research, the integration of multiple modules for joint analysis has demonstrated its unique potential. Zhang, D. et al. [185] introduced a model that simultaneously includes both regression and classification tasks, proposing the use of MTL as a feature selection module to capture shared features across different task types. The selected features, along with genetic information, are then fused into a lower-layer multi-task module, which ultimately completes the prediction task. This architecture significantly enhances the model's expressiveness and provides an important reference for subsequent related research. This framework has also become the prototype for several models in later works [200, 78, 138]. Jie, B. et al. [78] further performed a dialectical analysis of this model, arguing that the original model lacked the ability to learn and utilise the inherent distributions of each modality. Based on this, they improved the task allocation of the multi-task module by shifting the focus from downstream tasks to sub-tasks specific to different modalities. This effectively ensured the preservation of the inherent distribution information of each modality. Similarly, based on the previous approach, the work of [200] enhanced the robustness of the model by introducing regularisation terms for labels, strengthening the important relationships between labeled themes (e.g., samples belonging to AD or MCI). This was achieved by minimizing the distance between class themes in the feature space, which in turn enhanced the associations between these themes.

Discussions of relational modelling inevitably involve graph structures. To emphasise the complex relationships between different ROI and apply these relationships to diagnostic tasks, Shao, W. et al. [138] introduced a hyper-graph during the feature selection phase. Unlike traditional graph models, hyper-graphs allow each node to connect to more than two vertices, enabling the modelling of more complex higher-order relationships. They constructed an independent hyper-graph for each modality to ensure that higher-order relationships were not disrupted, and redesigned the multi-task feature selection based on the hyper-graph. On the other hand, Hao, X. [66]focused on sample similarity, proposing the computation of a similarity matrix using the random forest method, followed by feature selection through a regularisation term, and ultimately leaving the fusion to MKL. The most recent work proposed a multimodal sparse-similarity-based feature selection strategy, which constructs modality-specific similarity matrices using RF and constrains the feature selection function through Laplacian regularisation. The selected features are then used for classification via MKL [45]. This approach jointly learns from both similarity matrices and auxiliary data, enforcing

shared sparse selection across domains, thereby allowing complex target tasks to benefit from the relevant priors learned in simpler auxiliary tasks.

4.2.3. Discussion

The classification of traditional ML is built on the development of the model framework. The gradually complex network architecture begins to be responsible for more functions in the data analysis process, and becomes a series of complete network modules. From the aforementioned method analyses, with the continuous advancements in multimodal, these network modules have gradually formed a stable application strategy.

MKL primarily addresses multimodal feature fusion, whereas graph-based architecture and MTL predominantly facilitate feature selection. The strength of MKL lies in its capacity to accommodate complex, heterogeneous multimodal inputs by effectively identifying modality-specific kernels, thereby reducing biases due to heterogeneity. Additionally, MKL employs implicit mappings into shared spaces, effectively capturing complex non-linear relationships and modality complementarity. In contrast, MTL distinctly focuses on shared multimodal feature extraction, enabling the discovery of critical biomarkers shared across multiple tasks, thus significantly enhancing interpretability and generalisation. MTL also benefits from exploiting inter-task correlations, indirectly expanding training resources and improving feature selection accuracy, especially under conditions of limited sample size. Hyper-graph methods, in comparison, excel particularly in preserving complex data structures and modelling intricate feature interactions, surpassing MTL in their capacity to represent higher-order, collective relationships among multimodal features. In summary, methods characterised by stronger feature interpretability and representational power are generally preferred for feature selection stages. Conversely, MKL has emerged as the favoured choice for feature fusion modules within ML, primarily due to its unique ability to effectively handle heterogeneous data and its inherent convenience.

Nevertheless, these methods share several limitations. The most immediate arises from their dependence on feature engineering. ML-based fusion still relies heavily on handcrafted or statistically pre-selected features, making performance sensitive to feature-extraction design and limiting scalability across datasets or imaging protocols. Similar constraints appear at the model-design level: kernel mappings or task relationships must be predefined, which restricts the capacity to capture deeper hierarchical semantics that naturally emerge in multimodal data. A further limitation concerns generalisation. Many ML-fusion frameworks are validated on small or single-centre cohorts, and their performance often deteriorates when applied to unseen or cross-site datasets.

Based on these challenges, several future directions can be outlined. First, deep-learning-based feature extraction provides a reliable path for improving the representation of input data. Integrating interpretable ML fusion with deep neural encoders could combine the strengths of both paradigms: interpretability and expressive representation. Second, a less explored direction involves automated feature and kernel construction, in which the system learns or searches for modality-specific descriptors, kernels, or projections that should enter the fusion stage. Recent sparse-similarity MKL and graph-regularised selection methods demonstrate that such automatic construction is feasible for AD cohorts and can reduce manual dependence on domain-specific feature design [45]. Finally, domain adaptation and transfer learning should also be considered. Embedding statistical harmonisation and cross-domain transfer mechanisms within ML pipelines could enhance robustness across imaging sites and demographic cohorts.

In summary, while ML-based fusion offers interpretability and flexibility, it remains constrained by limited scalability, heavy reliance on manual features, and sensitivity to data heterogeneity. Future research should aim to bridge ML and DL paradigms through hybrid, explainable, and domain-adaptive frameworks that deliver both transparency and generalisation in multimodal AD prediction.

4.3. Deep Learning Fusion

The primary distinction between traditional DL and ML arises principally from differences in their input data. DL networks can be understood as a stack of variant network modules with several functions; the ability to process raw data and extract features is inherently integrated into their complex architecture. Figure 7 displays the process and framework of the general DL method on multimodal data analysis. These networks are generally structured in hierarchical layers, where consecutive layers can be viewed as functional blocks focusing on specific tasks [177]. The last one or two layers of each block are responsible for outputting data that is then passed on to subsequent layers or modules, forming an end-to-end integrated architecture [100]. Typically, DL networks consist of three main modules: the feature extraction module, the feature fusion module and the classification module. These modules are closely linked, and different strategies or methods for feature extraction can significantly impact the final output features, thereby influencing subsequent processes like feature combination and fusion analysis. However, feature extraction

strategies are not entirely independent; they are often influenced or constrained by the nature of the features being used. Therefore, the overall process of multimodal analysis in DL is tightly interconnected, making it difficult to separate distinct stages of the process [13]. As the number of modalities and use cases increases, network architectures continue to evolve rapidly, posing challenges for classification tasks within these networks.

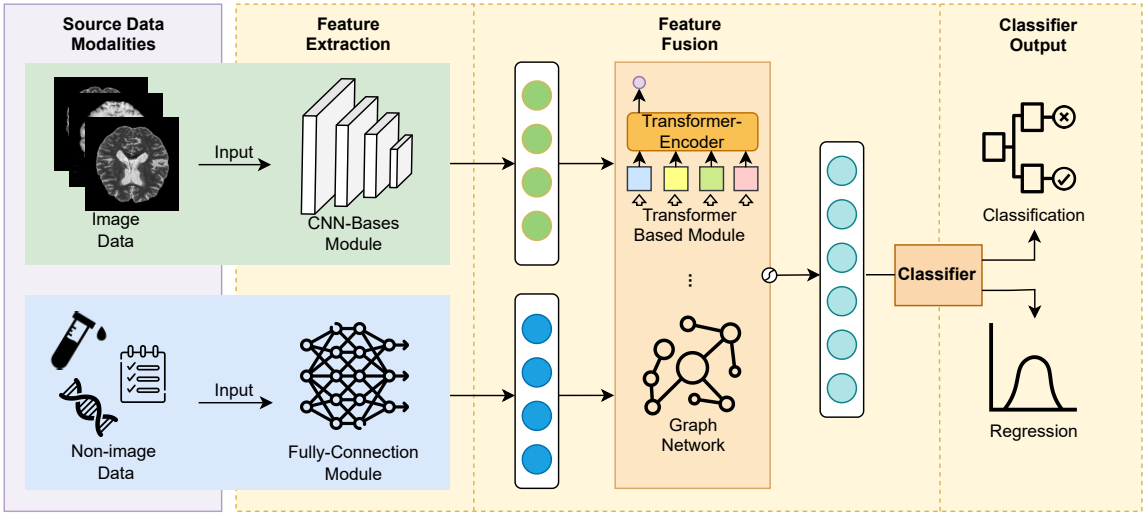


Figure 7: DL framework for multimodal analysis. DL models directly receive raw multimodal data as input and automatically perform feature extraction, fusion, and classification within a unified architecture. Unlike ML, which based frameworks and depend on handcrafted feature engineering, deep networks can hierarchically learn latent representations from both image and non-image modalities. Modules such as CNNs, transformers, and graph neural networks enable the model to capture spatial, temporal, and relational dependencies across modalities. Researchers only need to perform basic preprocessing, as the subsequent learning and decision-making processes are autonomously handled by the network.

To effectively differentiate among existing multimodal fusion methods, we focus on two critical stages of DL: feature extraction and feature fusion, and further, we propose categorizing these methods accordingly. Deep neural network architectures enable the analysis of input data using predefined or trained modules, greatly simplifying the feature extraction process. Figure 7 simply covers the situation of input data and the following data processing channels. Specifically within AD research, the relevant modalities can be broadly categorised into image and non-image data types based on their inherent data structures, guiding the feature extraction architectures toward two main categories: CNN-based networks for image data and Multilayer Perceptron (MLP)-based networks for non-image data. Feature fusion strategies, on the other hand, exhibit greater diversity. They include both conventional matrix operations (such as element-wise addition, multiplication, concatenation, and cross-product) and integration methods embedded within specialised neural network modules. Based on these distinctions, we further subdivide and systematically summarise strategies and network modules employed at different analysis stages. Adopting a "general-customised" criterion, we classify and evaluate existing network architectures to highlight their underlying principles and design philosophies.

The definition of the proposed taxonomy can be summarized here: At the feature extraction stage, methods utilizing predefined networks without modification are categorised as general, while those employing modified or specifically designed architectures are considered customised. General or standard feature extraction methods refer to the common approaches for extracting features from each modality. These methods do not require specialised techniques or significant architectural modifications aimed at enhancing feature representations. For instance, in AD research, pre-trained CNNs such as ResNet and VGG are often used for image data analysis without needing extensive adjustments based on the task at hand; Instead, considering the complexity inherent in fusion techniques, conventional matrix-based strategies are classified as general. Matrix-based strategies do not rely on complex techniques to capture or learn the relationships, dependencies, or correlations between modalities. Instead, they directly input the fused features into the downstream classifier. Whereas fusion methods involving specialised neural network modules are designated as customised, with additional subdivisions provided based on the specific network components used. This structured

classification facilitates a clearer understanding and comparison of current multimodal fusion approaches in the AD research field.

Based on these differences, DL methods can be classified according to whether they employ customised techniques or designs. We classify each study by its primary contribution, that is, the part of the pipeline where the authors claim novelty or make the largest design choice. When extraction is the main novelty and fusion is standard, we place the work under feature extraction and annotate the fusion style. When the key contribution is the way modalities are combined, we place it under fusion and record the extractors used. Table 5, Table 6, and Table 7 display the list of these multimodal research, and using "Customise" and "Normal" tags to recognise the core of model design. "Customise" at feature extraction means the model under the feature extraction focus approaches; in contrast, "Customise" at fusion strategy column means the model belongs to modality fusion focus approaches.

Table 5: DL Based Multimodal Analysis Method: Part I

Author	**Year**	**Modalities**	**Feature Extraction**	**Fusion Strategies**	**Tasks**	**Evaluation Matrix(%)**		
			(Normal / Customise)			SEN	SPE	ACC
Leng, Y. [86]	2023	sMRI, FDG-PET	Customise	Normal	NC vs AD	97.22	98.21	97.67
					NC vs SMC	72.15	85.29	81.63
Qiu, S. [127]	2022	MRI, non-image	Customise	Normal	NC vs AD vs MCI vs nADD	77.10	89.50	80.40
					AD vs ADD	75.30	75.60	77.30
Venugopalan, J. [161]	2021	MRI,SNP, Electronic Health Records (EHR)	Normal	Normal	NC vs AD vs MCI	78.00	NA	78.00
Tabarestani, S. [149]	2019	MRI, PET, CSF, CS	Customise	Normal	NC vs AD vs MCI	84.00	NA	84.00
Velazquez, M. [160]	2022	DTI, EHR	Customise	Normal	EMCI-nc vs EMCI-c	100.00	98.60	98.81
Zhou, Y. [197]	2024	Genetic	Customise	Normal	Regression	NA	NA	NA
Dwivedi, S. [37]	2022	MRI, FDG-PET	Normal	Normal	NC vs AD	97.00	97.00	97.00
					NC vs MCI	97.00	91.00	94.00
					MCI vs AD	96.00	99.00	97.50
Vu, T. D. [163]	2017	MRI, PET	Customise	Normal	NC vs AD	NA	NA	91.14
					NC vs MCI			89.20
Goel, T. [58]	2023	MRI, PET	Normal	Normal	NC vs AD vs MCI	95.33	96.17	95.89
Liu, S. [100]	2015	MRI, PET	Customise	Customise	NC vs AD	95.93	98.53	97.13
					NC vs MCI	97.91	67.04	87.24
					ncMCI vs cMCI	68.04	86.81	78.88
					NC vs AD vs ncMCI vs cMCI	53.65	85.05	57.00
Shi, J. [140]	2017	MRI, PET	Customise	Customise	NC vs AD	95.53	98.53	97.13
					NC vs MCI	97.91	67.04	87.24
					ncMCI vs cMCI	68.04	86.81	78.88
Ma, J. [105]	2022	sMRI, fMRI	Customise	Normal	NC vs AD	NA	NA	93.89
					NC vs MCI			89.83
					MCI vs AD			88.62
					ncMCI vs cMCI			84.97
					NC vs AD vs ncMCI vs cMCI			86.51
Fedorov, A. [51]	2021	MRI	Customise	Customise	NC vs AD	NA	NA	84.10
Chen, Z. [25]	2023	MRI, PET	Normal	Customise	NC vs AD	89.00	98.20	94.70
					NC vs MCI	69.70	64.00	67.10
					sMCI vs pMCI	42.50	95.30	80.20
Lee, G. [85]	2019	MRI, CSF, CS, DD	Customise	Normal	ncMCI vs cMCI	84.00	80.00	81.00
Rahim, N. [129]	2023	MRI, CS, DD	Customise	Normal	NC vs AD	92.00	NA	96.00
Abuhmed, T. [7]	2021	MRI, FDG-PET, CS, DD, NSB	Customise	Normal	NC vs AD vs MCI	88.60	NA	90.50
					Regression	NA		NA
Liu, Y. [101]	2021	MRI, PET	Normal	Customise	NC vs AD vs MCI	NA	NA	83.6

Table 6
DL Based Multimodal Analysis Method: Part II

Author	Year	Modalities	Feature Extraction (Normal / Customise)	Fusion Strategies (Normal / Customise)	Tasks	Evaluation Matrix(%)		
						SEN	SPE	ACC
Akkaya, U. M [9]	2023	Genetic	Customise	Normal	NC vs AD	83.00		83.00
					NC vs MCI	73.00	NA	77.00
					MCI vs AD	73.00		79.00
Wang, Y [168]	2018	fMRI, DTI	Customise	Normal	NC vs MCI vs AD	NA	NA	92.06
Yu, Q [181]	2024	sMRI, CS, Genetic	Customise	Customise	NC vs AD	96.00	96.00	96.00
					sMCI vs pMCI	75.00	76.00	75.00
Eslami, M [49]	2023	MRI, PET, COG, NTS, CSF,Genetic, Age	Customise	Normal	NC vs AD vs MCI			82.00
					NC vs AD vs ncMCI vs cMCI vs others	NA	NA	68.00
Tu, Y [153]	2022	MRI, CS, Genetic	Customise	Normal	NC vs AD	97.00	93.00	96.00
					sMCI vs pMCI	91.00	88.00	87.00
Abdelaziz, M. [4]	2021	MRI, PET, SNP	Customise	Normal	NC vs AD	97.78	98.76	98.22
					NC vs sMCI	92.65	93.57	93.11
					NC vs pMCI	97.72	97.58	97.35
Tang, C. [150]	2023	MRI, PET	Customise	Customise	NC vs AD	97.96	99.54	99.04
					NC vs MCI	98.52	98.61	98.57
					MCI vsAD	94.25	98.81	97.43
Thung, K. H. [151]	2017	MRI, PET, Genetic, Demographic	Normal	Customise	NC vs AD vs MCI	98.65	99.34	98.72
					NC vs AD vs MCI	NA	NA	65.8
Lu, D. [103]	2018	sMRI, FDG-PET	Customise	Normal	sMCI vs pMCI	79.69	83.84	82.93
					NC vs pNC/others	86.50	86.30	86.40
Golovanevsky, M. [59]	2022	MRI, Genetic, CS	Normal	Customise	NC vs AD vs MCI	NA	NA	96.88
Jia, H. [77]	2022	fMRI, sMRI	Customise	Customise	NC vs SMC	100.00	95.00	91.30
					NC vs MCI	93.33	90.00	92.00
					SMC vs MCI	100.00	87.50	94.44
					SMC vs AD	87.50	100.00	94.44
					MCI vs AD	90.00	100.00	95.00
Martí-Juan, G. [109]	2023	MRI, CS	Customise	Customise	NC vs AD	100.00	80.00	92.00
					Regression	NA	NA	NA
Zhang, F. [187]	2019	MRI, PET	Customise	Normal	NC vs AD	96.58	95.39	98.47
					NC vs MCI	90.11	91.82	85.74
					MCI vs AD	97.43	84.31	88.20
Liu, M. [98]	2018	MRI, FDG-PET	Customise	Customise	NC vs AD	92.05	93.94	93.26
					NC vs sMCI	81.08	84.31	82.95
					NC vs pMCI	63.07	67.31	64.04
Suk, H. I [147]	2014	MRI, PET	Customise	Normal	NC vs AD	94.65	95.22	95.35
					NC vs MCI	95.37	65.87	85.67
					sMCI vs pMCI	48.04	95.23	75.92
Abuhantash, F. [6]	2024	EHR, CS, DD	Normal	Customise	NC vs AD			99.00
					NC vs MCI	NA	NA	94.00
					MCI vs AD			93.00
					NC vs AD vs MCI			98.00
Song, X. [143]	2022	fMRI, DTI	Customise	Customise	NC vs SMC	77.20	97.50	93.20
					NC vs EMCI	82.50	95.70	91.10
					NC vs LMCI	92.70	95.70	94.20
					SMC vs EMCI	100.00	75.90	91.50
					SMC vs LMCI	100.00	79.50	95.70
					EMCI vs LMCI	93.70	89.50	92.40

Table 7
DL Based Multimodal Analysis Method: Part III

Author	Year	Modalities	Feature Extraction	Fusion Strategies	Tasks	Evaluation Matrix(%)		
			(Normal / Customise)			SEN	SPE	ACC
Yao, Z [178]	2025	MRI, PET	Normal	Customise	NC vs AD	98.80	96.90	98.30
Gu, S [63]	2025	MRI, CS	Customise	Normal	ncMCI vs cMCI	85.14	NA	86.89
Abdelaziz, M [3]	2025	MRI, PET	Normal	Customise	NC vs AD	97.35	92.71	95.25
					NC vs MCI	69.36	94.00	85.22
					sMCI vs pMCI	50.20	93.81	75.16
Ortiz-Perez, D. [120]	2025	Speech(Voice, Text)	Customise	Customise	NC vs AD	85.06	NA	87.35
Cheng, J. [23]	2025	MRI, CS	Customise	Customise	NC vs AD vs MCI	96.40	NA	97.65
Zhou, W [196]	2025	sMRI, fMRI	Normal	Customise	NC vs AD	97.34	91.48	94.19
					NC vs MCI	89.70	91.24	90.86
					MCI vs AD	92.04	89.13	90.56
					NC vs MCI vs AD	85.39	92.49	85.64
Chen, Z [27]	2025	MRI, PET	Customise	Normal	NC vs AD	94.74	93.10	95.48
					sMCI vs pMCI	70.21	80.65	78.95
Hu, X [72]	2025	MRI, PET, CS	Normal	Customise	sMCI vs pMCI	90.60	98.00	95.40
Chen, Y. [24]	2024	MRI, EEG, CS	Custmoise	Normal	NC vs MCI vs AD	88.89	NA	90.00
Zhou, T. [195]	2019	MRI, PET, SNP	Normal	Customise	NC vs AD	NA	NA	97.00
					NC vs MCI			82.00
					NC vs AD vs MCI			98.00
					NC vs AD vs sMCI vs pMCI			74.00
Massalimova, A. [110]	2021	sMRI, DTI	Normal	Normal	NC vs AD vs MCI	95.00	NA	97.00
Kim, S. [83]	2023	MRI, Genetic, CS, DD	Normal	Customise	Regression	NA	NA	NA
Wang, Q. [167]	2023	MRI, DD, APOE, Plasma	Customise	Customise	NC vs AD	95.80	97.00	96.50
					MCI vs AD	88.80	95.70	90.50
Lu P [104]	2024	MRI, SNP, Clinical	Normal	Customise	sMCI vs pMCI	88.80	85.40	87.20
Durga S HL [68]	2024	MRI, Cognitive Score(CoS), Demographic Data (DD)	Customise	Normal	NC vs AD	91.39	NA	92.65
Gao, X. [55]	2021	MRI, FDG-PET	Customise	Normal	sMCI vs pMCI	NA	NA	80.85
El-Sappagh, S [40]	2020	MRI,PET,CSF,CS	Customise	Normal	NC vs AD vs ncMCI vs cMCI	98.42	NA	92.62
Qiu Z [128]	2024	MRI, PET	Customise	Customise	NC vs AD	97.40	95.38	96.37
					NC vs MCI	73.01	73.02	73.61
					MCI vs AD	93.09	72.00	85.29
Wang, C. [164]	2024	MRI, DD, Genetic, CS	Customise	Normal	Regression	NA	NA	NA

4.3.1. Feature Extraction Approach

DL methods inherently allow neural networks to directly analyse and transform raw input data, facilitating the extraction of features beneficial for downstream analyses. How to fully exploit and extract implicit information from different modalities has been a continuously popular research topic. Methods primarily dedicated to improving multimodal analyses by deeply exploring modality-specific latent information can be categorised as modality-focused approaches. The core design principle of these methods lies in constructing effective feature extraction architectures tailored to modality characteristics, enabling the capture of highly discriminative and pathology-related features.

Given the crucial role of imaging data in AD research, CNNs have emerged as the most widely adopted architecture for image-based feature extraction. CNN architectures typically comprise convolutional and pooling layers, where convolutional layers use kernels to extract local features by convolving with input data. Each kernel generally focuses on specific types of features, and kernel size directly determines the receptive field. Pooling layers, positioned after convolutional layers, perform downsampling on feature maps, thereby reducing computational complexity and memory consumption [168]. For example, Vu et al. [163] applied CNNs to PET and MRI data, achieving accuracies of 91.1% for AD/NC classification and 89.2% for MCI/NC classification, demonstrating CNNs' robust performance in medical image analysis. However, limited by the black box in DL, feature extraction still requires further digging. In contrast

to traditional approaches that typically utilise only the features from the last layer, Wang et al. [167] captures and integrates features from all layers of the feature extraction module, apply them for downstream analysis. A similar methodology can also be observed in the work of Qiu et al. [128], where they additionally segmented features into distinct categories and utilised specialised modules corresponding to each feature type for detailed analysis.

On the other hand, AD progression is inherently longitudinal; biomarkers and clinical indicators evolve over time [192]. To effectively leverage temporal information, recurrent neural networks (RNNs) have been introduced. Originally designed for sequential data, RNN architectures retain information from previous inputs, allowing prior states to influence future outputs. Such structures enable RNNs to model local temporal dependencies effectively. For handling longer or more complex sequences, researchers have developed advanced RNN variants, such as Long Short-Term Memory (LSTM) to incorporate memory cells for better long-term information storage, and Gated Recurrent Units (GRU), which selectively gate information to mitigate gradient vanishing issues.Tabarestani et al. [149] successfully employed LSTM to analyse multimodal AD data such as psychological assessments, MRI, PET, and genetic data, and achieved 84% accuracy in a three-class classification task.

However, with ongoing advances and deeper exploration into multimodal data, pre-trained general-purpose network architectures often struggle to satisfy research-specific demands. Consequently, researchers continue to extend and refine baseline network modules, tailoring feature extraction processes to accommodate specialised requirements. We summarise the mainstream strategies into the following paths: image patch-wise method, image convert method, appraisal-based dimensionality reduction Method, and large language model-based feature Extraction.

Image Patch-wise Method One highly popular special processing method is the segmentation of images into smaller patches for analysis. Although conventional CNNs divide images into blocks during analysis, CNNs treat the image as a complete entity. In contrast, patch-wise methods treat each image block as an independent unit for analysis. This strategy, compared to working with the entire image, allows for a more focused exploration of the implicit information in different regions of the image, enabling the extraction of more detailed features. In the context of AD research, the ability to accurately identify specific brain regions is particularly significant due to the region-specific and progressive nature of the disease, making it crucial for disease detection.

Lu, D. [103] adopted this strategy and introduced a multi-scale approach to extract structural and metabolic information at different scales, thereby enhancing the model's expressive power. Although Qiu, Z. [128] did not use the same overarching approach, their research also retained unique information from different layers during the feature extraction phase and subsequently fed it into the fusion module. In the latest studies, Wang, C. [164] implemented a patch-based approach, further refining the architecture by manually segmenting the brain into different regions, processing each patch with a complete CNN block, and finally applying PCA for dimensionality reduction to predict the Clinical Dementia Rating (CDR).

Patch-wise processing targets image data by dividing a high-resolution image into smaller tiles for analysis. This strategy reduces GPU memory use and computational load, allowing models to operate at higher effective resolution. It also improves sensitivity to fine details such as tiny lesions and subtle edges, which are common in medical imaging. These gains come with trade-offs. Splitting an image can weaken the capture of global context. Tile borders may introduce artefacts, and overlapping regions can yield inconsistent predictions. Large or elongated lesions may be fragmented across tiles and partly missed. Choosing an appropriate patch size and stride is therefore critical. Typical mitigations include multi-scale or pyramid tiling, using overlap with blending, adding context padding around each tile, or combining tile-level outputs with a global branch. In addition to CNN-based patch processing, recent work has begun to use Transformer encoders purely as patch-level feature extractors, rather than as multimodal fusion modules. In this setting, the image is first divided into fixed-size patches. Each patch is then linearly projected and positionally encoded, and the Transformer is used to model local-global relationships among patches within a single modality. Because the Transformer operates on a set of patch tokens, it naturally fits the patch-wise paradigm and can capture long-range dependencies that conventional patch CNNs may miss. Importantly, in this usage the Transformer does not perform cross-modal interaction; it functions only as a stronger intra-modality encoder that produces high-quality patch embeddings to be passed to subsequent fusion or classification layers. This makes it a suitable replacement for patch CNNs when higher-level spatial reasoning is required, without changing the overall fusion pipeline.

An interesting point to note is that the shift from ML to DL is directly reflected in the increased complexity of network architectures and the reduction in manual feature engineering. The patch-wise method, in contrast, is more similar to dividing the image into multiple ROI segments and analysing them separately, which can be seen as a return to ML approaches. This raises the question, whether approximating traditional ML feature engineering within deep network design could improve model expressiveness; It warrants further discussion.

Image Convert Method Another viable approach is to convert multimodal image information during the feature extraction stage, enabling better representation of the data or aligning information and structure across different modalities. One such method involves converting the data into connectivity matrices during preprocessing, which are then used as inputs. Compared to raw image data, the transformed connectivity matrices effectively reduce noise in the original images and enhance the model's focus on relevant regions. Moreover, matrices have lower dimensionality compared to the original images, making the model more computationally efficient and easier to process, while also helping to mitigate the risk of over-fitting.

Greicius, M. [61] and colleagues explored the direct relationship between functional connectivity (FC) and structural connectivity (SC), demonstrating that FC in fMRI can reflect SC. Building on this insight, Wang, Y. [168] hypothesised that combining fMRI and DTI-derived connectivity networks could better represent brain connectivity and improve diagnostic accuracy for AD. They proposed converting DTI and fMRI images into corresponding structural connectivity networks (DTISCN) and functional connectivity networks (FCN), stacking the resulting matrices, and then inputting them into a CNN for classification. Following this approach, Ma, J. [105] focused on the overall geometric shape inherent in the data, transforming FC and SC into symmetric positive definite (SPD) matrices defined on Riemannian manifolds, thus preserving their affine invariance properties to ensure the stability of the numerical representation.

These approaches project each image into a prior-guided, common representation (for example, atlas space, tissue-type maps, cortical thickness, or PET parametric maps). The projection compresses the data and standardises scale and geometry across modalities. As a result, downstream models need less memory and compute, and the outputs are easier to interpret. Because all modalities are converted into isomorphic representations, multimodal alignment and fusion can be performed with simple operations such as concatenation or averaging. The benefits come with clear trade-offs. Strong reliance on preprocessing and template choice can introduce bias and site dependence. Conversion may discard lesion detail, texture, and other fine-grained cues, which limits representational capacity on complex tasks. These methods are therefore most suitable when data are scarce, when interpretability is a priority, or when a baseline is needed with modest resource demands. Practical mitigations include multi-resolution mappings, retaining residual or edge channels, uncertainty or quality-control flags from the conversion step, and harmonisation across sites or scanners.

Appraisal-based dimensionality reduction Method The Appraise method is primarily used for dimensionality reduction of features. Unlike direct feature integration from images, it focuses on certain aspects of the data to produce a score based on specific objectives. While the aforementioned feature processing methods primarily deal with image modalities, this method addresses the alignment of dimensions between image and non-image modalities. In practical applications, clinical and genetic data often appear in low-dimensional formats, such as tabular data, yet hold significant value for AD diagnosis. However, the independent nature of different modality expression spaces often leads to issues where the large dimensional disparity can either cause the network to overlook important information from certain modalities or excessively emphasis the influence of image modalities. Either case negatively affects the final performance of the model. Therefore, it is crucial to standardise the dimensions and project features from different modalities into a shared space for analysis.

Embedding image features in low-dimensional spaces is a popular approach, with the method of embedding directly impacting feature selection and the efficiency of representation in the lower-dimensional space. Conventional methods rely on CNN networks to directly integrate image features and preserve global information. However, other approaches can help the model focus on feature information from different perspectives. Wang, C. [164] used traditional ML dimensionality reduction techniques, such as PCA, to better preserve highly correlated information within image features. Zhang, F. [187] applied convolution to multimodal images simultaneously and then combined the features from different modalities at the end, introducing the Pearson correlation coefficient for output, ultimately producing an evaluation metric focused on the inter-modality information correlation. In contrast, [127] employed a multi-task architecture where a CNN network simultaneously outputs conventional image feature representations and an Alzheimer's (ALZ) score based on whether the patient suffers from non-AD dementia (nADD), finding that this score was more informative than typical clinical data.

These methods aim to place heterogeneous modalities in a balanced representation space. They learn scores or weights to appraise features and then apply a projection so that no single, high-dimensional source dominates the learning signal. In practice, this includes filter or embedded feature selection, sparse or low-rank projections (e.g., PCA/Partial least squares regression (PLS)/Canonical Correlation Analysis (CCA)-style bottlenecks, or their deep variants), and learned gating/attention that produces per-feature or per-channel weights. The result is a lower computational burden and improved interpretability, because the appraisal scores reveal which factors drive the

prediction. Trade-offs are clear. Down-projection can remove lesion detail and subtle texture, reducing generalisation. Supervised appraisal can also leak labels if the selector is fitted on the full dataset or outside a proper cross-validation pipeline, which inflates the apparent performance. By contrast, lifting a low-dimensional modality to a higher-dimensional space (e.g., via MLP expansion or kernel features) can enable richer interactions with high-dimensional partners, but it may inject noise, magnify variance, and destabilise training. Overall, these multi-branch designs are extensible, as new modalities can be added by simply incorporating an encoder and a projection head into the shared space. Yet each addition requires recalibration (normalisation, scaling, and loss weighting) to maintain fair contributions and robust fusion.

Large Language Model Based Feature Extraction. Compared with the earlier methods summarised in this section, the emergence of large language models (LLMs) is a more recent development, yet one that merits careful discussion. The previous subsections primarily focused on imaging-based modalities or cross-modal alignment prior to fusion. Although imaging plays a central role in AD research, non-imaging modalities, particularly textual data (such as genetic annotations, clinical narratives, and electronic health records (EHRs)), also possess substantial diagnostic value but remain relatively underexplored.

LLMs help bridge this gap. Their functionality extends far beyond traditional natural language processing (NLP) models, encompassing information retrieval, synthesis, and data generation [111]. In current AD research, LLMs are used primarily as feature extractors. They excel at processing unstructured or weakly structured data, particularly textual information. Raw text is tokenised and mapped into high-dimensional contextual embeddings, which are then treated as an additional modality alongside imaging and tabular data. Compared with classical NLP pipelines, LLMs offer broader coverage of domain-specific terminology, greater robustness to grammatical and stylistic variability, and the ability to capture longitudinal context across multiple clinical visits.

Recent studies have demonstrated the utility of LLMs in analysing clinical documents through prompt engineering and fine-tuning for dementia subtyping, improving both interpretability and classification accuracy [107, 80]. Li et al. [89] fine-tuned a foundation model to extract AD-related information from non-structured EHR data. Beyond textual data, other work has employed LLMs for speech analysis, extracting salient features from noisy recordings and identifying relevant acoustic markers. Leveraging their generative capabilities, these studies have also used LLMs to augment scarce datasets by synthesising additional speech samples [113].

While these findings highlight the potential of LLMs in AD research, their limitations, such as domain-specific bias, susceptibility to hallucination, and high computational cost, remain key challenges [102]. Given their strength in textual representation learning, future work should focus on efficiently integrating LLMs into the deep-learning feature extraction pipeline, particularly for enhancing textual and clinical modalities. Moreover, the generative abilities of LLMs offer promising directions for addressing modality imbalance and data scarcity, including synthetic data generation, anomaly correction, and cross-modal augmentation.

4.3.2. Modality Fusion Approach

Modality fusion approaches aim to optimise the performance of the fusion module within a model. AD and other neurodegenerative disorders share many pathological characteristics, which means that single-modality information is insufficient for accurate prediction or diagnosis. Compared with conventional multimodal tasks, AD research depends strongly on the complementarity and integration of information across modalities [171]. Early multimodal fusion techniques typically concatenated features from different modalities and fed them directly into a classifier network. Such methods were straightforward and often effective, yet they lacked the ability to capture the complex inter-relationships between modalities. The subsequent development of DL methods was strongly influenced by early machine-learning classification architectures, particularly the "early–late–hybrid" fusion taxonomy. Building upon this foundation, several studies adopted MTL and MKL strategies to design deep multimodal fusion frameworks, achieving notable performance improvements [151, 5]. These works demonstrate how the conceptual lineage of traditional ML continues to shape deep neural models, especially in how they separate or couple modality-specific and shared representations.

As DL matured, it gradually overcame the architectural constraints of earlier frameworks, and the boundary between feature extraction and fusion became increasingly blurred. In this work, we define a fusion module as any network block that receives two or more heterogeneous inputs simultaneously, mixes the information internally, and outputs a fused representation or decision. For instance, the Transformer architecture is widely recognised for its fusion capability, although it can also serve purely as a feature extractor for unimodal analysis. Therefore, we classify a Transformer as a fusion module only when it operates on multimodal inputs. Following this principle, we categorise fusion methods according to their interaction mechanisms among modalities. Specifically, we distinguish: Shared-embedding-based

methods, which minimise representational distance by projecting modalities into a common latent space; Attention-based methods, which model cross-modal dependencies through dynamic, data-dependent weighting; and graph-based methods, which exploit message passing and relational structures to capture spatial, temporal, or semantic relationships between modalities.

Shared embedding. In recent studies, the application of non-image modalities has become more widespread, with multimodal fusion research having a main focus on the concepts of shared space and latent space, until the welcomed multimodal fusion modules or networks are introduced. The former, inspired by correlation analysis methods like CCA, aims to align information across modalities, maintaining similar representations for further analysis, matching, or fusion. The latter, influenced by information theory and dimensionality reduction techniques, aims to reveal latent factors or features through data mapping into different dimensions. Whether in shared or latent spaces, the feature projections rely on the network's own learning, representing abstract features that make deep networks difficult to interpret, which is the major challenge in their application.

Recent work by Qiu, Z. [128] provides an intuitive framework that highlights the functions and differences between shared and latent spaces. They input the extracted features into three lower-level fusion modules: Global Attention Learning (GAL), Local Attention Learning (LAL), and Latent Space Learning (LSL). The GAL module corresponds to shared space, aiming to capture inter-modality correlations. In contrast, the LSL module interacts with different modality feature vectors through an outer product operation, creating a high-order representation that captures hidden associations between modalities and features that cannot be directly captured by shared space alone. Research by Chen, Z. [25] focuses more on LSL, introducing orthogonal latent space learning. This method adds new constraints to the projection matrix, ensuring that the information from different modalities is sufficiently distinct, thereby reducing the extra burden and risk associated with redundant data. Martí-Juan, G. [109] constructed a latent shared space in which encoded information from different modalities is projected into a shared latent space, facilitating more effective dynamic changes and cross-modality reconstruction.

Embedding-based methods aim to project heterogeneous modalities into a common representation space, thereby reducing inter-modality distance and ensuring global consistency. This unified space mitigates the imbalance that often arises when certain modalities dominate learning, preventing weaker or lower-dimensional modalities from being overlooked. Because the objective function explicitly encourages cross-modal alignment, the absence of one modality does not strongly distort the overall representation. Consequently, these methods are naturally more robust in scenarios involving missing modalities. However, the emphasis on global consistency can also be a limitation. The alignment and projection process may discard modality-specific or fine-grained features, leading the network to overlook subtle but diagnostically relevant details. Thus, while shared-embedding approaches offer strong generalisation and resilience to modality dropout, they may underperform in tasks requiring fine local discrimination. They are particularly suitable for complex training conditions, such as frequently changing modality combinations or limited sample sizes, where model stability and alignment consistency are prioritised over maximal spatial precision.

Attention based The Attention mechanism was originally applied in the image domain, where it assigns attention scores to different regions based on their relevance to the task. This allows the model to focus on important parts of the image while ignoring irrelevant features, thereby improving efficiency and accuracy [59]. The purpose is to enable the model to concentrate on specific features of an image rather than the entire image. In a multimodal context, attention scores can be computed between two modalities, enabling the integration of their information [159]. Zuo, Q. [104] utilised the attention module for feature fusion, grouping, and capturing the correlation information between different modality groups. They also introduced a non-linear gating unit before the attention module to transform modalities, ensuring that features from different modalities are projected into a shared space with consistent feature dimensions. Golovanevsky, M. [59] employed self-attention to convert raw inputs and positional information into high-dimensional representations, emphasizing the important representations within each modality and enhancing intra-modality correlations. They then used cross-attention blocks to perform cross-modality feature interactions, gradually transforming independent modality information into a joint representation.

The transformer architecture operationalises this mechanism at scale. Originally developed for natural-language processing, the Transformer organises multi-head self-attention and feed-forward layers within an encoder–decoder framework [159]. In the encoder, each input word is transformed into a vector representation through an embedding layer and processed via self-attention layers. The self-attention mechanism enables the model to consider other information from the input sequence while encoding each word, allowing it to capture dependencies between features. This mechanism improves the model's ability to understand the context of the input data. The decoder operates similarly to the encoder but also includes an additional encoder-decoder attention layer [159]. In image tasks, the encoder relies

heavily on multi-head attention to associate and select features, and because these tasks focus more on classification rather than sequence-to-sequence tasks, the transformer is typically employed using only the encoder portion of the architecture [176].

Wu, B. et al. [173] pioneered the integration of the Transformer into the residual network's final layer to replace the last convolutional layer for processing high-dimensional image features, using visual tokens to assist downstream classification tasks. This approach laid the groundwork for the development of Vision Transformers and set the stage for Transformers to be adopted as fusion modules. The work by [181] extended this framework by projecting MRI features, clinical tabular data, and genetic information into a shared space and introducing the Transformer Encoder as a multimodal fusion tool. [150] also applied the Transformer to process MRI and PET data. After performing a round of CNN feature integration and adding positional information, the different modalities were fused using cross-attention. Recent studies have further expanded the scope of input data by combining heterogeneous modalities, such as tabular, signal data and imaging inputs. Leveraging the Transformer's strong temporal modelling and cross-modal analytical capabilities, these works explicitly align MRI spatial features with EEG temporal representations, achieving promising results in multimodal AD analysis [24].

In general, attention-based methods capture cross-modal relationships through iterative query–key–value operations. The interactions between modalities occur more frequently and at a finer granularity than in other fusion strategies, enabling these methods to model subtle intra- and inter-modality associations and to achieve fine-grained alignment. This characteristic allows attention mechanisms to extract informative signals even under imbalanced sample distributions, while their inherent weighting scheme enhances interpretability and visualisation of model decisions. However, this high-frequency, fine-grained interaction comes at a cost. Attention-based models require substantial computational and memory resources, and their complexity increases exponentially with the growth of feature dimensions or the number of input tokens. Moreover, fluctuations in sample quality can lead to instability in the learned attention patterns, making data curation and preprocessing particularly critical for robust performance. On this basis, Transformers represent a structured realisation of the attention principle rather than a separate mechanism. They provide a scalable architecture for modelling both intra and inter-modality dependencies within a unified framework. Their major advantage lies in flexibility: Transformers impose few constraints on data format and can learn dependencies across diverse input types. This adaptability has made them one of the most influential architectures in multimodal fusion research. Additionally, LLM-derived embeddings can be integrated into the same attention-based fusion pipelines as imaging or tabular features, because they expose a token-level interface compatible with Transformer encoders. In such cases, the LLM itself does not perform multimodal fusion; instead, fusion is carried out by a downstream attention or Transformer block that jointly attends over image tokens, clinical tokens, and text-derived tokens. This situates LLM applications within the broader Transformer-based fusion family rather than as a separate fusion paradigm.

Graph-based fusion GNNs are increasingly being explored in AD research due to their ability to model the relationships between different data types. In contrast to other methods, Graph-based fusion methods focus on modelling interactions and relationships between information sources rather than performing direct feature-level concatenation. They represent data as a graph structure, in which nodes correspond to entities (for example, subjects, modalities, brain regions, or time points) and edges encode their spatial, temporal, or semantic relationships[174]. Through message-passing or aggregation operations, these methods propagate information along edges, allowing the model to exploit structural priors that describe anatomical connectivity, population similarity, or longitudinal progression. Such characteristics make graph approaches particularly suitable for population-level analyses and longitudinal studies, where relational patterns are central.

In the study by Abuhantash, F. [6], cognitive score similarities are used to create edges in a graph where each patient is represented as a node. This method simplifies the classification task but primarily relies on cognitive data for patient representation. On the other hand, Song, X. [143] divided the multimodal fusion process into two phases: first fusing image data, then constructing a feature matrix for each patient and using these matrices as nodes in a graph. Non-image data, like clinical information, is used to construct the graph's edges. This approach places a heavier emphasis on image data, with non-image data assisting in uncovering latent features within the images. Conversely, Kim et al. [83] take a more integrated approach, applying both image and non-image data to contribute to the construction of both nodes and edges. This fusion of both data types allows for a deeper, more comprehensive analysis of patient conditions, providing a better foundation for diagnosing AD.

A key advantage of employing GNNs as image fusion modules lies in their intrinsic interpretability, as the learned node and edge weights can be directly visualised to reveal relational importance. The design of the network

architectures inherently depends on explicit prior knowledge, resulting in an intuitive representation of structural relationships and clear tracing of neighbourhood interactions. This also allowed them to be not limited to fully supervised learning and can perform effectively in semi-supervised or weakly supervised settings, since the graph topology provides auxiliary constraints and regularisation. Nevertheless, this explicit dependence also implies that arbitrary or inadequately designed graph structures may substantially degrade model performance, limit feature representation capabilities, and hinder the extraction of implicit relationships beyond the predefined structure. Besides, they are often less effective for high-precision, voxel-level analysis, and their performance depends heavily on the design of the graph itself, which is how nodes and edges are defined and weighted. Inappropriate or noisy graph construction may propagate misleading information through the network. Consequently, graph-based fusion usually requires strong domain knowledge to construct meaningful priors and achieve stable results.

4.3.3. Hybrid Fusion

Hybrid fusion refers to the integration of both modality-focused and fusion-focused approaches within a single network architecture. In the early stages of DL research, there was no universal or mainstream network module for multi-modal fusion. Instead, various methods emerged, each offering unique and highly individualised solutions to the multi-modal fusion challenge. Liu, M. [98] and colleagues designed a cascade-level fusion based on CNNs. In this approach, patches of multi-modal images from the same position are input into a 3D CNN for feature extraction. These features are then correlated and passed through a 2D CNN to learn the inter-modal correlations. Finally, the features are fused once more before being input into the classifier. In this process, CNNs not only extract features but also play a role in modality fusion, capturing information at various stages and progressively refining the result. Similarly, Wang, Y. [168] recognised the potential of CNNs by transforming images into connectivity matrices. These matrices, representing different modalities, are then combined and fed into a CNN for feature fusion and extraction, which is used in downstream classification tasks. Shi, J. [140] attempted to introduce deep polynomial networks to AD learning, utilizing stacked network modules for feature selection and fusion. As GNNs, attention mechanisms, and transformer architectures gained traction in AD research, their powerful capabilities and adaptability became increasingly appreciated by researchers. These modern techniques have since become dominant in the choice of fusion modules, gradually unifying the selection of fusion strategies in multi-modal network architectures.

4.3.4. Robust Fusion under Missing Modalities

Ensuring robust multimodal fusion requires models that can operate reliably even when one or more input modalities are unavailable or incomplete. The missing-modality problem has long been a persistent challenge in AD research, affecting both traditional machine-learning and deep-learning frameworks. Because AD studies are inherently longitudinal, data collection must be repeated at multiple time points and across several modalities. As the importance of multimodal analysis continues to grow, the need for paired and complete datasets has become more pronounced. In practice, however, large-scale acquisition is rarely achievable: data are often corrupted, incomplete, or entirely absent due to cost constraints, site policies, human error, or acquisition failure. This reality motivates a growing body of research devoted to either reconstructing the missing information or improving model robustness under incomplete inputs. To date, most existing works can be categorised into two major directions: explicit completion of missing information and enhancement of model robustness under incomplete inputs.

This line of work aims to reconstruct the missing data using available modalities, thereby creating "simulated" complete inputs for downstream tasks. Early methods adopted simple data imputation techniques such as zero-filling or mean-value substitution [4]. Subsequent efforts introduced modality completion, reconstructing the absent modality from the existing ones. Traditional approaches often relied on atlas-based or machine-learning models, which offered interpretability but limited generalisability. The emergence of generative adversarial networks (GANs) has markedly simplified the completion process, enabling the synthesis of visually realistic and human-interpretable images. These metghods excel at capturing fine-grained cross-modal mappings and extracting detailed modality information. GFE-Mamba effectively exploits this property: rather than using the generated images as inputs to the classification network, it utilises the intermediate feature maps captured within the generative model. The classification module further incorporates a Mamba architecture to process long sequential data and to fuse clinical-scale information with latent imaging features [50]. Despite these advantages, such methods require large training datasets and remain prone to artefacts and unstable outputs, which limits their practical integration into clinical workflows.

The second direction focuses on improving the model's resilience to missing modalities, ensuring it can still learn informative representations despite incomplete inputs. Common strategies include shared latent-space learning,

knowledge distillation, and modality masking. The first approach has been discussed earlier in the DL section; knowledge distillation employs a teacher network trained with full modalities to guide a student network that operates under missing inputs, aligning them at the feature or semantic level [155]. Modality masking introduces artificial incompleteness during training by randomly omitting or occluding modality inputs, forcing the encoder to reconstruct or analyse incomplete samples, thus covering a range of real-world missingness patterns [97].

Recent work increasingly combines generative and robust strategies, forming end-to-end pipelines that directly connect the generative and predictive components. In these systems, the workflow typically follows a "missing-modality → generation → multimodal fusion → classification" paradigm, eliminating the need to manually pair generated and original samples [27]. Based on the guidance of the strategy, robustness is now often embedded as a core design principle within network architectures. For example, Hu et al. [72] synthesised PET images from MRI inputs and performed tri-modal co-attention fusion to align representations across modalities. Similar ideas have appeared in other medical domains, consistently improving performance under partial data conditions.

The emergence of diffusion-based architectures has further advanced generative solutions. Diffusion models provide greater stability, structural coherence, and higher-quality synthetic outputs. For instance, FMM-Diff [191] introduces latent reconstruction of missing modalities within the diffusion pipeline and cross-modal feature sharing to extract correlated information during synthesis. Likewise, SLaM-DiMM [136] employs multi-channel inputs to reconstruct any of four available modalities, greatly improving generation robustness. Despite their advantages, diffusion-based frameworks remain computationally demanding and complex to train, making them difficult to integrate into generation–diagnosis hybrid systems. As a result, GAN-based architectures still dominate practical hybrid designs due to their relative simplicity and efficiency.

In summary, generative approaches offer human-interpretable outputs and detailed cross-modal mappings, while iterative improvements have enhanced the stability and fidelity of generated data. However, their reliance on extensive datasets and high computational resources poses significant barriers. Moreover, the visual quality of generated images does not necessarily equate to feature-level utility, because pixel alignment and semantic alignment are not interchangeable. In contrast, robustness-driven models exhibit stronger adaptability and generalisation but still depend on the presence of informative modalities; under severe missingness, their performance deteriorates sharply. In a combined network structure, the advantages of the two methods are combined. The content generated by the generative model is usually constrained by the downstream task. The information it generates is often closer to the task-related cross-modal mapping than that of a direct generative network, thus better serving the downstream task. At the same time, the downstream network modules can also obtain more information to construct feature maps, which helps the model to master the correct feature distribution.

4.3.5. Discussion

This section gives a classification strategy for the current DL multimodal fusion approach. We firstly defined a baseline dual-/multi-branch architecture that uses pretrained backbones, with explicit registration or learned cross-modal alignment, and a simple fusion operator (concatenation, fixed/learned weighting, or element-wise gating) at a specified layer. We simply call these models "basic models". Above the "basic model", researchers will further improve on the network structure to emphasise their network core or the research point. With observation from these changes, we annotate methods as feature extraction-focused and modalities fusion–focused. This category strategy can show the consideration and research focus of those researchers.

Despite substantial progress, deep-learning-based fusion still faces several limitations. Firstly, regardless of model category, a major constraint arises from the black-box nature of DL models. Although they offer powerful representational capacity, they often lack transparency. Consequently, many studies prioritise performance gains while overlooking clinically meaningful information embedded in the input data. Secondly, data requirements and imbalance remain a persistent challenge. DL fusion typically demands large and balanced multimodal datasets; however, incomplete or small-scale cohorts frequently lead to overfitting and modality bias. Although the section Robust Fusion under Missing Modalities discusses several methods to mitigate this issue, modality incompleteness continues to represent a core bottleneck restricting broader research development. Thirdly, the high architectural complexity of deep networks substantially increases computational and memory costs. Furthermore, the private or site-specific nature of many clinical datasets hinders external validation and model deployment, thereby limiting large-scale or longitudinal applications.

To address these issues, several remedies have been proposed. For the interpretability problem, incorporating architectural priors (e.g., ROI-aware modules, anatomical graphs) and training priors (e.g., region-weighted or concept-based losses) can enhance model transparency. Uncertainty estimation and attention visualisation may further improve clinical interpretability. When introducing additional modalities, expert priors can serve as soft constraints or weak labels to guide the model toward clinically relevant regions. Looking ahead, lightweight network optimisation should remain a central focus, as efficiency is critical for clinical applicability. Simultaneously, future research should more comprehensively address model generalisation, for instance, by employing pretraining on publicly available multimodal datasets and adopting transfer learning or domain adaptation strategies to enhance robustness across different sites and imaging protocols.

4.4. Image Fusion

Image fusion aims to fuse different images into a new one, which holds the compound features and information from the source images. In this paper, we use image-level fusion to mean generating a fused, human-readable image outside the downstream task model. This differs from in-model fusion, where multimodal inputs are fused within the task network [193]. The common in-model fusion, which includes traditional ML and DL methods, typically extracts features from each image modality before fusing high-dimensional features at the model; image fusion provides direct visual data, replacing complex multimodal inputs with fused images, reduces computational cost and improves efficiency, and offers greater interpretability. In addition, an advantage in clinical medical applications is that it allows doctors to obtain more information and make faster diagnoses. Based on the technology route, we categorise them into two classes: the learning-free approach and the model-dependent approach.

4.4.1. Learning-Free Approach

Learning-free image fusion typically does not rely on data-driven parameter learning, but rather on manually designed fusion strategies. Learning-free approaches are categorised into spatial-domain and transform-domain approaches: **Spatial-domain fusion.** Spatial-domain fusion operates directly on image pixels by segmenting images into regions or blocks and merging them based on spatial information [73]. However, this approach requires feature consistency and strongly relies on the accuracy of image segmentation. which can limit complementary feature expression, potentially reducing fusion effectiveness. **Transform-domain fusion.** Transform-domain fusion converts images into another domain, fuses transformation coefficients, and reconstructs the final image using the inverse transformation. These methods are commonly based on multi-scale transformation (MST) theory, including Laplacian Pyramid (LP), Wavelet Transform (WT), Curvelet Transform (CVT), and Non-Subsampled Contourlet Transform (NSCT).

Due to its versatility, computational efficiency, and ability to capture multi-scale features, WT was the first to be applied in AD research. WT decomposes images into high and low frequency components, fuses corresponding frequency bands, and reconstructs a composite image using WT-based reconstruction techniques [61]. Dwivedi, S. [37] and Goel, T. [58] applied WT-based fusion to MRI and PET images, generating integrated images containing both anatomical and metabolic information. A recent study explored NSCT-based fusion for AD imaging, using Pulse Coupled Neural Networks (PCNN) and fuzzy rules for feature selection and integration. The fused data were then reconstructed using inverse NSCT, completing the fusion process [14]. Similar to other methods, the fusion approaches depend on manual image registration, predefined network parameters, and architectures, making them part of the preprocessing stage.

4.4.2. Model-Dependent Approach

Model-dependent approaches introduce DL methods for the feature extraction and fusion work, instead of the hand-crafted work and fusion rule design. Compared with the high cost and error-prone nature of manual feature extraction, the strong feature extraction capabilities of DL provide a reliable channel for the same task [193]. Beyond image reconstruction, these methods structurally resemble standard feature extraction networks, encompassing multimodal feature extraction, representation learning, and information integration (fusion), allowing shared architectures across tasks. Depending on the level of DL method participation and the task flow, the model-dependent approaches follow two main approaches: non-end-to-end fusion and end-to-end fusion.

End-to-End Fusion End-to-end fusion accomplishes the complete image fusion process within an independent model; that is, input the source image to the DL model, and it will give the fused target image as output which without any intermediate result output. The process can be divided into three stages: feature extraction, feature fusion,

and image reconstruction. CNN-based structure and GAN-based model are typically adopted for end-to-end image fusion [28]. Specifically, the CNN-based method converts the source image into high-dimensional features, fuses these features from different images, and finally utilises deconvolution layers to reconstruct the image [87]. For the GAN-based methods, it typically employs the CNN-based module as the generator, then introduces a discriminator and an adversarial loss to guide image generation towards matching the target distribution [165]. Furthermore, the use of discriminators is more flexible in actual model design; multiple discriminators can help the model retain more original information that is hoped to be valuable [30].

Non-End-to-End Fusion. In contrast to the end-to-end approach, Non-end-to-end means the three stages of image fusion are not done in a unified frame. Under this category, the DL method is applied for the feature extraction work in a typical situation, while the core fusion rules still rely on manual design. For instance, Amini, N. [10] used pre-trained CNNs (VGG19) for PET-MRI feature extraction and applied weighted averaging for fusion, while Do, O. et al. [33] fine-tuned VGG networks and introduced an Equilibrium Optimisation Algorithm (EOA) for adaptive fusion. These methods offer interpretability and flexibility; their performance is constrained by the effectiveness of manually designed fusion rules, which may not always yield optimal results.

4.4.3. Discussion

Image fusion methods have had a substantial impact on AD research, which can be categorised into Learning-free and model-dependent approaches. Learning-free approaches are transparent, reproducible, and annotation-efficient, offering clear interpretability and making them valuable for clinical diagnosis and early exploratory studies. However, their fixed rule design limits adaptability and the ability to capture non-linear relationships between modalities. Model-dependent approaches, on the other hand, provide superior representational power and scalability to large datasets, yet their performance strongly depends on training quality, data volume, and model stability. Compared with in-model fusion, image-level fusion offers clearer interpretability and lower computational cost, as it produces fused images with a reduced input dimension for downstream models. Nonetheless, generating a fused image inevitably causes information loss, and whether this visual simplification compromises downstream prediction accuracy remains an open question. Looking ahead, future research should focus on quantitative validation of fused images within predictive frameworks, which focus on assessing how information loss affects diagnostic accuracy, and on developing hybrid pipelines where end-to-end fusion modules can be jointly optimised with classification or regression networks. The integration of diffusion-based models may further enhance reconstruction fidelity and enable modality-aware, data-adaptive fusion, bridging the gap between interpretability and precision.

4.5. Evaluation Metrics

In this subsection, we present and discuss several widely adopted evaluation metrics within AD research. Generally, depending on the type of model outputs, downstream tasks are classified into classification tasks and regression tasks, each associated with well-established evaluation standards. These metrics serve as critical tools for assessing model performance and guiding future methodological enhancements. Calculation methods and detailed explanations of these metrics are summarised in Table 8, which contains both the classification task and regression task.

4.5.1. Evaluation Metrics for Classification Task

Classification represents a prevalent downstream task in AD research, typically aiming to categorise subjects into distinct disease stages, such as differentiating NC, MCI, and AD, or distinguishing between various subtypes of MCI. Initially, these tasks primarily explored pathological differences between patient groups. Currently, the emphasis has shifted towards accurately identifying clinical stages to enhance early diagnostic accuracy. Common evaluation metrics utilised in classification tasks include Accuracy (ACC), Sensitivity (SEN), Specificity (SPE), Precision, and Area Under the Curve (AUC). Accuracy measures the proportion of correctly predicted cases over all cases. Sensitivity, also known as recall or True Positive Rate (TPR), quantifies the model's effectiveness in identifying actual positive instances, and in contrast, has False Positive Rate (FPR). Conversely, specificity evaluates the model's accuracy in correctly excluding negative instances. Precision assesses the proportion of true positives among samples predicted as positive, reflecting prediction reliability. AUC comprehensively assesses model performance across various classification thresholds, effectively balancing sensitivity and specificity. The importance of each metric varies according to clinical context. In screening contexts, higher sensitivity and AUC values might be prioritised to minimise missed positive diagnoses, achievable through adjusting decision thresholds. Thus, researchers should carefully select or combine multiple metrics based on their specific evaluation goals.

Table 8
Popular Evaluation Metrics in AD Research: I

Metric	Formula	Best values	Description	Ref.
Classification Task				
Accuracy	$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$	Higher	Accuracy measures how many samples are correctly classified in the total samples, reflecting the overall accuracy of the model. When the category distribution is balanced, Accuracy is a good choice that can intuitively reflect the model's performance.	[37]
Sensitivity	$\text{Sensitivity} = \frac{TP}{TP+FN}$	Higher	Sensitivity measures the model's ability to detect positive samples, reflecting how many of all actual positive samples are correctly identified. In medical diagnosis scenarios, for those diseases that hold higher risks or harm, missed diagnosis (FN) is more dangerous than misdiagnosis (FP), where Sensitivity is particularly important.	[100]
Specificity	$\text{Specificity} = \frac{TN}{TN+FP}$	Higher	Specificity measures the model's ability to exclude negative samples, reflecting how many of all actual negative samples are correctly excluded. In medical scenarios, high specificity means low misdiagnosis rate (FP), which helps avoid over-diagnosis of healthy people and reduces unnecessary panic and medical burden.	[127]
Precision	$\text{Precision} = \frac{TP}{TP+FP}$	Higher	Precision measures the proportion of samples predicted by the model to be positive that are actually positive, reflecting the reliability of the model's predictions. In the AD scenarios, its focus is similar to that of SPE. Higher Precision means a more reliable prediction, and it mainly focuses on situations where normal people do not want to be misjudged as risky.	[149]
Area Under the Curve	$\text{AUC} = \int_0^1 \text{TPR}(\text{FPR})\, d(\text{FPR})$	Higher	Area Under the ROC Curve calculates the area of the performance curve under various thresholds. It cannot analyse the performance under a specific threshold, but is used to comprehensively measure the model performance under different discrimination thresholds, taking into account both sensitivity and specificity, and focusing on all sample types. Adjusting the discrimination threshold can affect the weights of sensitivity and specificity to meet the needs of different scenarios.	[152]
F1-score	$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Sensitivity}}{\text{Precision} + \text{Sensitivity}}$	Higher	F1-score is a harmonic average that takes into account both precision and recall. It measures the overall performance of the model on positive samples and is suitable for scenarios that focus on minority samples or class imbalance.	[110]
Regression Task				
Mean Squared Error	$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$	Lower	Mean Squared Error calculates the average of the squares of the differences between the predicted values and the true values. It reflects the overall level of prediction error and is more sensitive to extreme errors. It is more commonly used for mathematical analysis and visualisation of the training process.	[79]
Root Mean Squared Error	$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$	Lower	Root Mean Squared Error, calculates the square root of MSE , which also reflects the overall level of prediction error. Because it is on the same scale as the original data, it is more intuitive and easier to understand, convenient for direct comparison with actual errors, and easier for business personnel or domain experts to understand.	[148]
Mean Absolute Error	$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}\lvert y_i - \hat{y}_i \rvert$	Lower	Mean Absolute Error, which calculates the average of the absolute values of the difference between the predicted value and the true value, can better reflect the average error level of most samples. It is not affected by extreme outliers and can provide a more objective overall error level when there are outliers in the data.	[109]
Pearson correlation coefficient	$r = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}}$	Null	Pearson correlation coefficient measures the consistency between the model prediction and the true value, and is often used to observe trends in medical scenarios. However, correlation has no correlation with model prediction performance, so it is necessary to combine other criteria for joint analysis.	[130]

4.5.2. Evaluation Metrics for Regression Task

Regression tasks predict continuous outcome variables, frequently used in AD studies to quantify brain structural or metabolic parameters, cognitive scores, or progression rates of the disease. Compared to classification, regression tasks offer a more direct reflection of disease severity rather than discrete stages, facilitating understanding of longitudinal disease progression and enabling an objective evaluation of therapeutic outcomes. Standard regression metrics include Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and Pearson correlation coefficient. MSE computes the mean square difference between the predicted and true values, making it particularly sensitive to outlier predictions. RMSE, the square root of MSE, retains the same units as the original measurements, providing an intuitive interpretation of prediction errors. MAE measures average absolute differences between predicted and actual values, being more robust against extreme outliers compared to MSE or RMSE. The Pearson correlation coefficient is generally used to measure linear relationships. It is generally used to measure the consistency between model predictions and true values. Actually, as no single metric sufficiently evaluates model performance in isolation, employing multiple metrics concurrently and tailoring their selection to specific application scenarios is common, thus achieving comprehensive and meaningful performance evaluations.

5. Challenge and Future Direction

In the preceding sections, we systematically reviewed and categorised current multimodal analysis approaches for AD prediction, evaluating a wide range of methods within both traditional ML and DL frameworks. Although substantial progress has been achieved, several persistent limitations and challenges continue to hinder further development in this field.

The first challenge concerns modality missing, a long-standing issue in multimodal AD research. As network architectures become increasingly complex and the demand for large-scale training data grows, the impact of missing modalities is becoming more pronounced. Whether in longitudinal studies, multimodal fusion tasks, or datasets with unbalanced label distributions, the absence of specific modalities can lead to information bias, severely degrading model performance, particularly in learning from rare samples. Current solutions generally follow two directions: (i) generating missing data through synthesis models and (ii) designing robust architectures capable of learning from incomplete inputs. However, both strategies have limitations. Generative models are computationally demanding and prone to instability, whereas robustness-based designs often fail when key modalities are absent. Hybrid architectures combining generation and classification have recently emerged but introduce new challenges, including loss-function balancing and the potential propagation of artefacts from the generative process into downstream classifiers. Future research should therefore focus on developing more stable, lightweight, and interpretable strategies for handling incomplete multimodal data, better suited to clinical and longitudinal contexts.

The second major challenge lies in the trade-off between interpretability and learning capacity, reflecting the fundamental contrast between ML and DL paradigms. Traditional ML is inherently more interpretable due to its shallow structures and human-defined features(such as volumetric, cortical, or metabolic measures), whose contributions can be directly quantified and traced to clinical meaning. In contrast, DL models sacrifice this transparency for automatic representation learning, which enables them to extract multi-level, non-linear, and semantically rich latent features from raw multimodal data, discovering cross-modal dependencies that handcrafted features may overlook. While DL thus offers unparalleled representational power, its black-box nature limits its clinical reliability and user trust.

A promising direction is to bridge the interpretability–representation gap by integrating the strengths of both paradigms. For example, aligning ML-derived feature clusters or decision boundaries with DL feature maps can enhance visual and conceptual interpretability. The incorporation of interpretable neural operators or concept bottleneck layers could further anchor abstract network representations to clinically meaningful constructs. Alternatively, hybrid paradigms that embed interpretability constraints and expert priors within high-level semantic spaces of deep networks may achieve a synergy between performance and transparency.

Overall, future multimodal AD research should aim to construct unified frameworks that combine interpretability, robustness, and representational efficiency, ensuring that predictive models remain both powerful and clinically comprehensible. In addition, drawing upon current advancements in the field, we briefly outline several potential perspectives that can inform future studies or inspire novel strategies for DL network architecture design and performance evaluation.

5.1. Network Architecture Insights

A retrospective analysis of DL developments in AD research reveals an important observation: traditional ML methods continue to provide valuable insights into the design and construction of deep networks, sometimes even proving to be more effective in specific scenarios. For instance, slice-based feature extraction, widely adopted in DL, bears a strong resemblance to the traditional manual selection of ROI. Similarly, network segmentation around ROI regions aligns with conventional methods, allowing the model to shift its focus from the entire image to localised features, such as brain atrophy, grey and white matter volume changes, and sulcal morphology. This approach enhances the model's ability to focus on critical regions while reducing computational costs. GNNs also share structural similarities with traditional graph-based methods, with the primary difference being the shift from feature-based construction to a patient-centric approach. Additionally, conventional classifiers such as SVMs have demonstrated strong performance but are increasingly replaced by softmax classifiers due to the rising computational demands associated with larger datasets. These observations suggest that traditional ML approaches remain highly relevant and can directly inform the design of DL modules, making them an important area for future exploration.

5.2. Usable modality extension

Another important research direction involves expanding the range of available modalities. Current AD studies are heavily focused on brain imaging data, with most analyses conducted on 2D data, which can easily overlook the inherent 3D spatial information in medical images. However, 3D spatial characteristics are critical for AD diagnosis. As research into 3D image analysis and segmentation continues to gain traction, maximizing the preservation of spatial information within network architectures becomes a crucial challenge. Otherwise, the utilisation of non-imaging data is also gaining attention. Early detection and intervention are crucial in AD research. Different modalities capture distinct types of information and manifest at varying stages of the disease lifecycle. Among the biomarkers and representations currently identified as being associated with AD, CSF and PET provide the earliest and most direct indicators of pathological onset. However, both are invasive procedures, which restrict their feasibility for large-scale or routine screening. Beyond these invasive modalities, EHR data, particularly unstructured EHRs, offer an alternative, non-invasive source with substantial potential for early detection. Structured EHRs often contain limited and categorical information, whereas unstructured EHRs encompass detailed clinical narratives that may capture subtle precursors of disease onset. Yet, such information remains difficult to extract and analyse using conventional methods.

The advent of LLMs provides a transformative opportunity to address this challenge. LLMs enable the effective parsing and contextual understanding of unstructured text, allowing for the decomposition, summarization, and integration of complex clinical narratives into machine-readable representations. It also shows potential for generating and enhancing data. This capability makes LLM-integrated multimodal analysis a highly promising research direction for future AD studies, particularly for enhancing early, non-invasive detection and longitudinal disease monitoring.

Furthermore, with more modalities being introduced to assist in AD diagnosis. The expansion of available modalities brings new opportunities to the field: Recent studies have successfully incorporated speech data [137], VR data [99], and eye-tracking data [179] into ML models for AD prediction. Moreover, the former studies demonstrated the feasibility of using transfer learning to overcome challenges associated with small sample sizes in such emerging data types. These advancements offer promising opportunities for expanding the range of available modalities and provide evidence that various aspects of patient behaviour have significant potential for exploring AD pathology.

5.3. Modality Contribution Evaluation

Evaluating the contribution of each modality inside the model plays a significant role and task in the research. Building on the various modalities and technologies introduced, the available modalities in the research are increasing. However, the strategies for combining these modalities and their actual impact on diagnostic performance remain insufficiently explored. Current approaches have yet to establish the relative contribution of different modalities(both imaging and non-imaging) toward diagnostic accuracy, and only some limited initial results exploring the influences of different modalities on classification [38]. In other domains of multimodal research, modalities are often classified into primary and secondary modalities based on their contribution to downstream tasks, with constraints placed on their representation to optimise performance [70, 71]. Whether such strategies can be effectively applied to AD research and how they should be implemented remain open questions.

Additionally, regarding the optimal processing methods for each modality and the final fusion strategies, they remain uncertain. For example, with the recent introduction of Transformer models, many studies have favoured simultaneously feeding both image and non-image data into the encoder for multi-head attention computations.

However, the latest work [23] has shown that processing image features separately using Transformers while applying a separate one-dimensional attention mechanism for non-image features, followed by a simple concatenation for fusion, can achieve state-of-the-art performance. This suggests that simple fusion strategies may sometimes be more effective than complex techniques. The result suggests that feature extraction and fusion might require further discussion, especially whether the higher complexity of fusion is a positive effect on the final model performance steady. The central issue underlying this phenomenon primarily arises from a shift in network design priorities: Many researchers focus principally on architectural complexity rather than on the disease-specific problem itself, leading to insufficient interpretability. Reviewing the literature in DL research for AD, it becomes apparent that most methods rely entirely on automated feature extraction and fusion, thus ignoring a critical exploration of disease-specific attributes and mechanisms. Researchers tend to emphasise improving model performance, often overlooking the relevance to the disease itself. Among recent research, Qiu et al. [128] uniquely focused on explicitly incorporating pathological information into the feature extraction process. They classified features according to feature performance, and allowed the model to adapt network weights guided by established pathological insights and observed clinical phenomena of AD. Although the current lack of studies employing similar methodologies limits the ability to verify their effectiveness and robustness, these approaches nonetheless provide significant inspiration and valuable directions for subsequent research efforts.

6. Conclusion

AD is a complex neurodegenerative disorder with global prevalence and profound social impact. As no curative treatment currently exists, early detection and intervention remain the primary focus of modern research. Multimodal analysis has emerged as one of the most effective strategies for understanding and predicting AD, allowing integration of complementary information across structural, functional, molecular, and clinical dimensions.

This paper has provided a comprehensive overview of multimodal fusion strategies for AD prediction, covering both traditional ML and DL paradigms. We introduced and compared the modalities commonly used in AD research, summarised key biomarkers, and reviewed several major public datasets that support current investigations. Building upon this foundation, we proposed a new, fine-grained taxonomy of multimodal fusion, redefining classification from a functional perspective: namely, feature-engineering-based, module-based, and hybrid ML methods, and in DL, distinguishing between feature extraction–focused and modality fusion–focused approaches. This taxonomy offers a more interpretable framework for identifying the design logic and research focus of different architectures.

From our analysis, several key conclusions can be drawn: 1) Refined classification strategy. Conventional early/late fusion frameworks are insufficient to represent the structural and functional diversity of modern ML and DL models. The proposed taxonomy captures the fine-grained logic behind model construction and the conceptual emphasis of each design. 2) Complementarity between ML and DL. ML methods offer interpretability and transparency, whereas DL architectures provide strong representational power and adaptability to high-dimensional data. Through interpretable feature alignment, hybrid optimisation, or explainable neural operators can combinate them and achieve both performance and trustworthiness in clinical settings. 3) The importance of non-imaging data. Current research is still dominated by imaging modalities, while non-imaging data such as EHRs, genetic information, and cognitive assessments remain underexplored. Future studies should leverage LLMs and other foundation models to mine unstructured EHR data, thereby enriching multimodal representations and enabling earlier, non-invasive detection of AD. In the future, upcoming research should prioritise: (a) the development of lightweight and robust networks capable of handling missing or incomplete modalities; (b) the integration of domain priors and interpretability mechanisms; and (c) multimodal harmonisation frameworks that support cross-cohort and cross-site generalisation for clinical deployment.

In summary, although multimodal fusion for AD data has achieved remarkable progress, significant challenges remain in interpretability, generalisability, and scalability. Meanwhile, non-invasive data modalities are gaining increasing attention and interest among researchers, which may lead to a future shift in network architecture design. Future work on next-generation multimodal systems should aim to build transparent, robust, and clinically practical frameworks that integrate data diversity, domain knowledge, and deep learning, thereby advancing reliable early prediction of Alzheimer's disease.

Declaration of generative AI and AI-assisted technologies in the manuscript preparation process

During the preparation of this work, the authors used ChatGPT-5 in order to polish descriptions. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

References

[1] , 2022. Oasis-3 imaging methods & data dictionary, version 2.3. OASIS-Brains Project, Washington University in St. Louis. URL: https://sites.wustl.edu/oasisbrains/files/2024/04/OASIS-3_Imaging_Data_Dictionary_v2.3-a93c947a586e7367.pdf. data Release 2.0 (July 2022). States: "Due to anonymization participants may be included in all three datasets (OASIS-1, OASIS-2, and OASIS-3) under unique IDs.".

[2] , 2024. 2024 alzheimer's disease facts and figures. Alzheimer's & Dementia 20, 3708–3821. URL: https://alz-journals.onlinelibrary.wiley.com/doi/abs/10.1002/alz.13809, doi:https://doi.org/10.1002/alz.13809.

[3] Abdelaziz, M., Wang, T., Anwaar, W., Elazab, A., 2025. Multi-scale multimodal deep learning framework for alzheimer's disease diagnosis. Computers in biology and medicine 184, 109438.

[4] Abdelaziz, M., Wang, T., Elazab, A., 2021. Alzheimer's disease diagnosis framework from incomplete multimodal data using convolutional neural networks. Journal of biomedical informatics 121, 103863.

[5] Abrol, A., Fu, Z., Du, Y., Calhoun, V.D., 2019. Multimodal data fusion of deep learning and dynamic functional connectivity features to predict alzheimer's disease progression, in: 2019 41st annual international conference of the IEEE engineering in medicine and biology society (EMBC), IEEE. pp. 4409–4413.

[6] Abuhantash, F., Abu Hantash, M.K., AlShehhi, A., 2024. Comorbidity-based framework for alzheimer's disease classification using graph neural networks. Scientific Reports 14, 21061.

[7] Abuhmed, T., El-Sappagh, S., Alonso, J.M., 2021. Robust hybrid deep learning models for alzheimer's progression detection. Knowledge-Based Systems 213, 106688.

[8] Ahmed, O.B., Benois-Pineau, J., Allard, M., Catheline, G., Amar, C.B., Initiative, A.D.N., et al., 2017. Recognition of alzheimer's disease and mild cognitive impairment with multimodal image-derived biomarkers and multiple kernel learning. Neurocomputing 220, 98–110.

[9] Akkaya, U.M., Kalkan, H., 2023. A new approach for multimodal usage of gene expression and its image representation for the detection of alzheimer's disease. Biomolecules 13, 1563.

[10] Amini, N., Mostaar, A., 2022. Deep learning approach for fusion of magnetic resonance imaging-positron emission tomography image based on extract image features using pretrained network (vgg19). Journal of Medical Signals & Sensors 12, 25–31.

[11] Arevalo-Rodriguez, I., Smailagic, N., Roqué-Figuls, M., Ciapponi, A., Sanchez-Perez, E., Giannakou, A., Pedraza, O.L., Cosp, X.B., Cullum, S., 2021. Mini-mental state examination (mmse) for the early detection of dementia in people with mild cognitive impairment (mci). Cochrane Database of Systematic Reviews .

[12] Atrey, P.K., Hossain, M.A., El Saddik, A., Kankanhalli, M.S., 2010. Multimodal fusion for multimedia analysis: a survey. Multimedia systems 16, 345–379.

[13] Baltrušaitis, T., Ahuja, C., Morency, L.P., 2018. Multimodal machine learning: A survey and taxonomy. IEEE transactions on pattern analysis and machine intelligence 41, 423–443.

[14] Beatrice, S., Dhivviyanandnam, I., et al., 2025. An automated multimodal medical image fusion framework for alzheimer detection using deep learning .

[15] Beekly, D.L., Ramos, E.M., van Belle, G., Deitrich, W., Clark, A.D., Jacka, M.E., Kukull, W.A., et al., 2004. The national alzheimer's coordinating center (nacc) database: an alzheimer disease database. Alzheimer Disease & Associated Disorders 18, 270–277.

[16] Besson, J., Crawford, J.R., Parker, D., Ebmeier, K., Best, P., Gemmell, H., Sharp, P.F., Smith, F., 1990. Multimodal imaging in alzheimer's disease: the relationship between mri, spect, cognitive and pathological changes. The British Journal of Psychiatry 157, 216–220.

[17] Bi, X.a., Hu, X., Wu, H., Wang, Y., 2020. Multimodal data analysis of alzheimer's disease based on clustering evolutionary random forest. IEEE Journal of Biomedical and Health Informatics 24, 2973–2983.

[18] Blessed, G., Tomlinson, B.E., Roth, M., 1968. The association between quantitative measures of dementia and of senile change in the cerebral grey matter of elderly subjects. The British journal of psychiatry 114, 797–811.

[19] Brown, M.A., Semelka, R.C., 2011. MRI: basic principles and applications. John Wiley & Sons.

[20] Castellani, R.J., Rolston, R.K., Smith, M.A., 2010. Alzheimer disease. Disease-a-month: DM 56, 484.

[21] Chen, C.P., Li, J.L., 2024. Profiling patient transcript using large language model reasoning augmentation for alzheimer's disease detection, in: 2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), IEEE. pp. 1–4.

[22] Chen, H., Li, W., Sheng, X., Ye, Q., Zhao, H., Xu, Y., Bai, F., Initiative, A.D.N., 2022. Machine learning based on the multimodal connectome can predict the preclinical stage of alzheimer's disease: a preliminary study. European Radiology 32, 448–459.

[23] Chen, J., Wang, Y., Zeb, A., Suzauddola, M., Wen, Y., Initiative, A.D.N., et al., 2025a. Multimodal mixing convolutional neural network and transformer for alzheimer's disease recognition. Expert Systems with Applications 259, 125321.

[24] Chen, Y., Zhu, S., Fang, Z., Liu, C., Zou, B., Wang, Y., Chang, S., Jia, F., Qin, F., Fan, J., et al., 2024a. Toward robust early detection of alzheimer's disease via an integrated multimodal learning approach. arXiv preprint arXiv:2408.16343 .

[25] Chen, Z., Liu, Y., Zhang, Y., Li, Q., Initiative, A.D.N., et al., 2023. Orthogonal latent space learning with feature weighting and graph learning for multimodal alzheimer's disease diagnosis. Medical Image Analysis 84, 102698.

[26] Chen, Z., Liu, Y., Zhang, Y., Zhu, J., Li, Q., Wu, X., 2024b. Enhanced multimodal low-rank embedding based feature selection model for multimodal alzheimer's disease diagnosis. IEEE Transactions on Medical Imaging .

[27] Chen, Z., Wang, M., Nan, F., Yang, Y., Li, S., Zhou, M., Qi, J., Wang, H., Yang, P., 2025b. Joint image synthesis and fusion with converted features for alzheimer's disease diagnosis. Engineering Applications of Artificial Intelligence 156, 111102.

[28] Choudhury, C., Goel, T., Tanveer, M., 2024. A coupled-gan architecture to fuse mri and pet image features for multi-stage classification of alzheimer's disease. Information Fusion 109, 102415.

[29] Dai, Z., Yan, C., Wang, Z., Wang, J., Xia, M., Li, K., He, Y., 2012. Discriminative analysis of early alzheimer's disease using multi-modal imaging and multi-level characterization with multi-classifier (m3). Neuroimage 59, 2187–2195.

[30] Das, M., Gupta, D., Bakde, A., 2024. An end-to-end content-aware generative adversarial network based method for multimodal medical image fusion, in: Data Analytics for Intelligent Systems: Techniques and Solutions. IOP Publishing Bristol, UK, pp. 7–1.

[31] De Coene, B., Hajnal, J.V., Gatehouse, P., Longmore, D.B., White, S.J., Oatridge, A., Pennock, J., Young, I., Bydder, G., 1992. Mr of the brain using fluid-attenuated inversion recovery (flair) pulse sequences. American journal of neuroradiology 13, 1555–1564.

[32] Defigueiredo, R., Shankle, W.R., Maccato, A., Dick, M.B., Mundkur, P., Mena, I., Cotman, C.W., 1995. Neural-network-based classification of cognitively normal, demented, alzheimer disease and vascular dementia from single photon emission with computed tomography image data from brain. Proceedings of the National Academy of Sciences 92, 5530–5534.

[33] Do, O.C., Luong, C.M., Dinh, P.H., Tran, G.S., 2024. An efficient approach to medical image fusion based on optimization and transfer learning with vgg19. Biomedical Signal Processing and Control 87, 105370.

[34] Donini, M., Monteiro, J.M., Pontil, M., Shawe-Taylor, J., Mourao-Miranda, J., 2016. A multimodal multiple kernel learning approach to alzheimer's disease detection, in: 2016 IEEE 26th international workshop on machine learning for signal processing (MLSP), IEEE. pp. 1–6.

[35] Duan, J., Xiong, J., Li, Y., Ding, W., 2024. Deep learning based multimodal biomedical data fusion: An overview and comparative review. Information Fusion , 102536.

[36] Duara, R., Grady, C., Haxby, J., Sundaram, M., Cutler, N., Heston, L., Moore, A., Schlageter, N., Larson, S., Rapoport, S.I., 1986. Positron emission tomography in alzheimer's disease. Neurology 36, 879–879.

[37] Dwivedi, S., Goel, T., Tanveer, M., Murugan, R., Sharma, R., 2022. Multimodal fusion-based deep learning network for effective diagnosis of alzheimer's disease. IEEE MultiMedia 29, 45–55.

[38] Dyrba, M., Grothe, M., Kirste, T., Teipel, S.J., 2015. Multimodal analysis of functional and structural disconnection in a lzheimer's disease using multiple kernel svm. Human brain mapping 36, 2118–2131.

[39] Dyrba, M., Mohammadi, R., Grothe, M.J., Kirste, T., Teipel, S.J., 2020. Gaussian graphical models reveal inter-modal and inter-regional conditional dependencies of brain alterations in alzheimer's disease. Frontiers in aging neuroscience 12, 99.

[40] El-Sappagh, S., Abuhmed, T., Islam, S.R., Kwak, K.S., 2020. Multimodal multitask deep learning model for alzheimer's disease progression detection based on time series data. Neurocomputing 412, 197–215.

[41] El-Sappagh, S., Ali, F., Abuhmed, T., Singh, J., Alonso, J.M., 2022. Automatic detection of alzheimer's disease progression: An efficient information fusion approach with heterogeneous ensemble classifiers. Neurocomputing 512, 203–224.

[42] El-Sappagh, S., Alonso, J.M., Islam, S.R., Sultan, A.M., Kwak, K.S., 2021a. A multilayer multimodal detection and prediction model based on explainable artificial intelligence for alzheimer's disease. Scientific reports 11, 2660.

[43] El-Sappagh, S., Saleh, H., Sahal, R., Abuhmed, T., Islam, S.R., Ali, F., Amer, E., 2021b. Alzheimer's disease progression detection model based on an early fusion of cost-effective multimodal data. Future Generation Computer Systems 115, 680–699.

[44] Elazab, A., Wang, C., Abdelaziz, M., Zhang, J., Gu, J., Gorriz, J.M., Zhang, Y., Chang, C., 2024. Alzheimer's disease diagnosis from single and multimodal data using machine and deep learning models: Achievements and future directions. Expert Systems with Applications , 124780.

[45] Elazab, A., Wang, H., Wang, C., Zhang, Y., Gorriz, J.M., Cai, J., Chang, C., 2026. Improved alzheimer's disease diagnosis using multimodal sparse similarity feature selection and auxiliary data. Biomedical Signal Processing and Control 112, 108485.

[46] Englund, E., Brun, A., Alling, C., 1988. White matter changes in dementia of alzheimer's type: biochemical and neuropathological correlates. Brain 111, 1425–1439.

[47] Erkinjuntti, T., Hokkanen, L., Sulkava, R., Palo, J., 1988. The blessed dementia scale as a screening test for dementia. International Journal of Geriatric Psychiatry 3, 267–273.

[48] Erkinjuntti, T., Ketonen, L., Sulkava, R., Sipponen, J., Vuorialho, M., Iivanainen, M., 1987. Do white matter changes on mri and ct differentiate vascular dementia from alzheimer's disease? Journal of Neurology, Neurosurgery & Psychiatry 50, 37–42.

[49] Eslami, M., Tabarestani, S., Adjouadi, M., 2023. A unique color-coded visualization system with multimodal information fusion and deep learning in a longitudinal study of alzheimer's disease. Artificial intelligence in medicine 140, 102543.

[50] Fang, Z., Zhu, S., Chen, Y., Zou, B., Jia, F., Liu, C., Feng, X., Qiu, L., Qin, F., Fan, J., et al., 2024. Gfe-mamba: Mamba-based ad multi-modal progression assessment via generative feature extraction from mci. arXiv preprint arXiv:2407.15719 .

[51] Fedorov, A., Wu, L., Sylvain, T., Luck, M., DeRamus, T.P., Bleklov, D., Plis, S.M., Calhoun, V.D., 2021. On self-supervised multimodal representation learning: an application to alzheimer's disease, in: 2021 IEEE 18th international symposium on biomedical imaging (ISBI), IEEE. pp. 1548–1552.

[52] Fox, N., Warrington, E., Freeborough, P., Hartikainen, P., Kennedy, A., Stevens, J., Rossor, M.N., 1996. Presymptomatic hippocampal atrophy in alzheimer's disease: a longitudinal mri study. Brain 119, 2001–2007.

[53] Frisoni, G.B., Fox, N.C., Jack Jr, C.R., Scheltens, P., Thompson, P.M., 2010. The clinical use of structural mri in alzheimer disease. Nature reviews neurology 6, 67–77.

[54] Gan, H.S., Ramlee, M.H., Wang, Z., Shimizu, A., 2025. A review on medical image segmentation: Datasets, technical models, challenges and solutions. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 15, e1574.

[55] Gao, X., Shi, F., Shen, D., Liu, M., 2021. Task-induced pyramid and attention gan for multimodal brain image imputation and classification in alzheimer's disease. IEEE journal of biomedical and health informatics 26, 36–43.

[56] Glover, G.H., 2011. Overview of functional magnetic resonance imaging. Neurosurgery Clinics of North America 22, 133.

[57] Godbolt, A., Waldman, A., MacManus, D., Schott, J., Frost, C., Cipolotti, L., Fox, N., Rossor, M., 2006. Mrs shows abnormalities before symptoms in familial alzheimer disease. Neurology 66, 718–722.

[58] Goel, T., Sharma, R., Tanveer, M., Suganthan, P., Maji, K., Pilli, R., 2023. Multimodal neuroimaging based alzheimer's disease diagnosis using evolutionary rvfl classifier. IEEE Journal of Biomedical and Health Informatics .

[59] Golovanevsky, M., Eickhoff, C., Singh, R., 2022. Multimodal attention-based deep learning for alzheimer's disease diagnosis. Journal of the American Medical Informatics Association 29, 2014–2022.

[60] Gray, K.R., Aljabar, P., Heckemann, R.A., Hammers, A., Rueckert, D., Initiative, A.D.N., et al., 2013. Random forest-based similarity measures for multi-modal classification of alzheimer's disease. NeuroImage 65, 167–175.

[61] Greicius, M.D., Supekar, K., Menon, V., Dougherty, R.F., 2009. Resting-state functional connectivity reflects structural connectivity in the default mode network. Cerebral cortex 19, 72–78.

[62] Grundke-Iqbal, I., Iqbal, K., Tung, Y.C., Quinlan, M., Wisniewski, H.M., Binder, L.I., 1986. Abnormal phosphorylation of the microtubule-associated protein tau (tau) in alzheimer cytoskeletal pathology. Proceedings of the National Academy of Sciences 83, 4913–4917.

[63] GU, S.K., Purushothaman, A., et al., 2026. Alzfusionformer: Integrating multiple transformers for early alzheimer's disease detection from multi-modal data. Biomedical Signal Processing and Control 112, 108601.

[64] Gunes, S., Aizawa, Y., Sugashi, T., Sugimoto, M., Rodrigues, P.P., 2022. Biomarkers for alzheimer's disease in the current state: a narrative review. International journal of molecular sciences 23, 4962.

[65] Hansson, O., Blennow, K., Zetterberg, H., Dage, J., 2023. Blood biomarkers for alzheimer's disease in clinical practice and trials. Nature aging 3, 506–519.

[66] Hao, X., Bao, Y., Guo, Y., Yu, M., Zhang, D., Risacher, S.L., Saykin, A.J., Yao, X., Shen, L., Initiative, A.D.N., et al., 2020. Multi-modal neuroimaging feature selection with consistent metric constraint for diagnosis of alzheimer's disease. Medical image analysis 60, 101625.

[67] Hinrichs, C., Singh, V., Xu, G., Johnson, S., 2009. Mkl for robust multi-modality ad classification, in: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2009: 12th International Conference, London, UK, September 20-24, 2009, Proceedings, Part II 12, Springer. pp. 786–794.

[68] Hl, D.S., Thomas, S.M., et al., 2024. A multimodal approach integrating convolutional and recurrent neural networks for alzheimer's disease temporal progression prediction, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5207–5215.

[69] Holly, T.A., Abbott, B.G., Al-Mallah, M., Calnon, D.A., Cohen, M.C., DiFilippo, F.P., Ficaro, E.P., Freeman, M.R., Hendel, R.C., Jain, D., et al., 2010. Single photon-emission computed tomography.

[70] Hu, D., Hou, X., Wei, L., Jiang, L., Mo, Y., 2022. Mm-dfn: Multimodal dynamic fusion network for emotion recognition in conversations, in: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE. pp. 7037–7041.

[71] Hu, J., Liu, Y., Zhao, J., Jin, Q., 2021. Mmgcn: Multimodal fusion via deep graph convolution network for emotion recognition in conversation. arXiv preprint arXiv:2107.06779 .

[72] Hu, X., Shen, X., Sun, Y., Shan, X., Min, W., Su, L., Fan, X., Elazab, A., Ge, R., Wang, C., et al., 2025. Itcfn: Incomplete triple-modal co-attention fusion network for mild cognitive impairment conversion prediction, in: 2025 IEEE 22nd International Symposium on Biomedical Imaging (ISBI), IEEE. pp. 1–5.

[73] Huang, W., Jing, Z., 2007. Evaluation of focus measures in multi-focus image fusion. Pattern recognition letters 28, 493–500.

[74] Humpel, C., 2011. Identifying and validating biomarkers for alzheimer's disease. Trends in biotechnology 29, 26–32.

[75] Islam, N., Hashem, R., Gad, M., Brown, A., Levis, B., Renoux, C., Thombs, B.D., McInnes, M.D., 2023. Accuracy of the montreal cognitive assessment tool for detecting mild cognitive impairment: A systematic review and meta-analysis. Alzheimer's & Dementia 19, 3235–3243.

[76] Jacoby, R., Levy, R., 1980. Ct scanning and the investigation of dementia: a review. Journal of the Royal Society of Medicine 73, 366–369.

[77] Jia, H., Lao, H., 2022. Deep learning and multimodal feature fusion for the aided diagnosis of alzheimer's disease. Neural Computing and Applications 34, 19585–19598.

[78] Jie, B., Zhang, D., Cheng, B., Shen, D., Initiative, A.D.N., 2015. Manifold regularized multitask feature learning for multimodality disease classification. Human brain mapping 36, 489–507.

[79] Jin, Y., Su, Y., Zhou, X.H., Huang, S., Initiative, A.D.N., 2016. Heterogeneous multimodal biomarkers analysis for alzheimer's disease via bayesian network. EURASIP Journal on Bioinformatics and Systems Biology 2016, 1–8.

[80] Kashyap, A.M., Rao, D., Boland, M.R., Shen, L., Callison-Burch, C., 2025. Predicting explainable dementia types with llm-aided feature engineering. Bioinformatics 41, btaf156.

[81] Kates, R., Atkinson, D., Brant-Zawadzki, M., 1996. Fluid-attenuated inversion recovery (flair): clinical prospectus of current and future applications. Topics in Magnetic Resonance Imaging 8, 389–396.

[82] Kim, J., Lee, B., 2018. Identification of alzheimer's disease and mild cognitive impairment using multimodal sparse hierarchical extreme learning machine. Human brain mapping 39, 3728–3741.

[83] Kim, S., Lee, N., Lee, J., Hyun, D., Park, C., 2023. Heterogeneous graph learning for multi-modal medical data analysis, in: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 5141–5150.

[84] Le Bihan, D., Mangin, J.F., Poupon, C., Clark, C.A., Pappata, S., Molko, N., Chabriat, H., 2001. Diffusion tensor imaging: concepts and applications. Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine 13, 534–546.

[85] Lee, G., Nho, K., Kang, B., Sohn, K.A., Kim, D., 2019. Predicting alzheimer's disease progression using multi-modal deep learning approach. Scientific reports 9, 1952.

[86] Leng, Y., Cui, W., Peng, Y., Yan, C., Cao, Y., Yan, Z., Chen, S., Jiang, X., Zheng, J., Initiative, A.D.N., et al., 2023. Multimodal cross enhanced fusion network for diagnosis of alzheimer's disease and subjective memory complaints. Computers in Biology and Medicine 157, 106788.

[87] Li, H., Zhang, L., Jiang, M., Li, Y., 2021. Multi-focus image fusion algorithm based on supervised learning for fully convolutional neural network. Pattern Recognition Letters 141, 45–53.

[88] Li, Q., Wu, X., Xu, L., Chen, K., Yao, L., Li, R., 2017. Multi-modal discriminative dictionary learning for alzheimer's disease and mild cognitive impairment. Computer methods and programs in biomedicine 150, 1–8.

[89] Li, R., Wang, X., Berlowitz, D., Mez, J., Lin, H., Yu, H., 2025. Care-ad: a multi-agent large language model framework for alzheimer's disease prediction using longitudinal clinical notes. npj Digital Medicine 8, 541.
[90] Liang, Z., Lauterbur, P., 2000. Principles of magnetic resonance imaging (pp. 1-7).
[91] Lin, H., Jiang, J., Li, Z., Sheng, C., Du, W., Li, X., Han, Y., 2023. Identification of subjective cognitive decline due to alzheimer's disease using multimodal mri combining with machine learning. Cerebral Cortex 33, 557–566.
[92] Lin, W., Gao, Q., Du, M., Chen, W., Tong, T., 2021. Multiclass diagnosis of stages of alzheimer's disease using linear discriminant analysis scoring for multimodal data. Computers in biology and medicine 134, 104478.
[93] Lin, W., Gao, Q., Yuan, J., Chen, Z., Feng, C., Chen, W., Du, M., Tong, T., 2020. Predicting alzheimer's disease conversion from mild cognitive impairment using an extreme learning machine-based grading method with multimodal data. Frontiers in aging neuroscience 12, 77.
[94] Liu, F., Wee, C.Y., Chen, H., Shen, D., 2014a. Inter-modality relationship constrained multi-modality multi-task feature selection for alzheimer's disease and mild cognitive impairment identification. NeuroImage 84, 466–475.
[95] Liu, F., Yuan, S., Li, W., Xu, Q., Sheng, B., 2023a. Patch-based deep multi-modal learning framework for alzheimer's disease diagnosis using multi-view neuroimaging. Biomedical Signal Processing and Control 80, 104400.
[96] Liu, F., Zhou, L., Shen, C., Yin, J., 2013. Multiple kernel learning in the primal for multimodal alzheimer's disease classification. IEEE journal of biomedical and health informatics 18, 984–990.
[97] Liu, L., Liu, S., Zhang, L., To, X.V., Nasrallah, F., Chandra, S.S., 2023b. Cascaded multi-modal mixing transformers for alzheimer's disease classification with incomplete data. NeuroImage 277, 120267.
[98] Liu, M., Cheng, D., Wang, K., Wang, Y., Initiative, A.D.N., 2018. Multi-modality cascaded convolutional neural networks for alzheimer's disease diagnosis. Neuroinformatics 16, 295–308.
[99] Liu, Q., Song, H., Yan, M., Ding, Y., Wang, Y., Chen, L., Yin, H., 2023c. Virtual reality technology in the detection of mild cognitive impairment: a systematic review and meta-analysis. Ageing Research Reviews 87, 101889.
[100] Liu, S., Liu, S., Cai, W., Che, H., Pujol, S., Kikinis, R., Feng, D., Fulham, M.J., et al., 2014b. Multimodal neuroimaging feature learning for multiclass diagnosis of alzheimer's disease. IEEE transactions on biomedical engineering 62, 1132–1140.
[101] Liu, Y., Fan, L., Zhang, C., Zhou, T., Xiao, Z., Geng, L., Shen, D., 2021. Incomplete multi-modal representation learning for alzheimer's disease diagnosis. Medical Image Analysis 69, 101953.
[102] Liu, Z., Tang, L., Sun, Z., Liu, Z., Lyu, Y., Ruan, W., Xu, Y., Shan, L., Shin, J., Chen, X., et al., 2025. Ad-gpt: Large language models in alzheimer's disease. arXiv preprint arXiv:2504.03071 .
[103] Lu, D., Popuri, K., Ding, G.W., Balachandar, R., Beg, M.F., 2018. Multimodal and multiscale deep neural networks for the early diagnosis of alzheimer's disease using structural mr and fdg-pet images. Scientific reports 8, 5697.
[104] Lu, P., Hu, L., Mitelpunkt, A., Bhatnagar, S., Lu, L., Liang, H., 2024. A hierarchical attention-based multimodal fusion framework for predicting the progression of alzheimer's disease. Biomedical Signal Processing and Control 88, 105669.
[105] Ma, J., Zhang, J., Wang, Z., 2022. Multimodality alzheimer's disease analysis in deep riemannian manifold. Information Processing & Management 59, 102965.
[106] Mani, S., Shankle, W.R., Dick, M.B., Pazzani, M.J., 1999. Two-stage machine learning model for guideline development. Artificial Intelligence in Medicine 16, 51–71.
[107] Mao, C., Xu, J., Rasmussen, L., Li, Y., Adekkanattu, P., Pacheco, J., Bonakdarpour, B., Vassar, R., Shen, L., Jiang, G., et al., 2023. Ad-bert: Using pre-trained language model to predict the progression from mild cognitive impairment to alzheimer's disease. Journal of biomedical informatics 144, 104442.
[108] Márquez, F., Yassa, M.A., 2019. Neuroimaging biomarkers for alzheimer's disease. Molecular neurodegeneration 14, 21.
[109] Martí-Juan, G., Lorenzi, M., Piella, G., Initiative, A.D.N., et al., 2023. Mc-rvae: Multi-channel recurrent variational autoencoder for multimodal alzheimer's disease progression modelling. NeuroImage 268, 119892.
[110] Massalimova, A., Varol, H.A., 2021. Input agnostic deep learning for alzheimer's disease classification using multimodal mri images, in: 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), IEEE. pp. 2875–2878.
[111] Minaee, S., Mikolov, T., Nikzad, N., Chenaghlu, M., Socher, R., Amatriain, X., Gao, J., 2025. Large language models: A survey. URL: https://arxiv.org/abs/2402.06196, arXiv:2402.06196.
[112] Minhas, S., Khanum, A., Riaz, F., Khan, S.A., Alvi, A., 2017. Predicting progression from mild cognitive impairment to alzheimer's disease using autoregressive modelling of longitudinal and multimodal biomarkers. IEEE journal of biomedical and health informatics 22, 818–825.
[113] Mo, T., Lam, J.C., Li, V.O., Cheung, L.Y., 2025. Dect: Harnessing llm-assisted fine-grained linguistic knowledge and label-switched and label-preserved data generation for diagnosis of alzheimer's disease. arXiv preprint arXiv:2502.04394 .
[114] Muehllehner, G., Karp, J.S., 2006. Positron emission tomography. Physics in Medicine & Biology 51, R117.
[115] Nijakowski, K., Owecki, W., Jankowski, J., Surdacka, A., 2024. Salivary biomarkers for alzheimer's disease: a systematic review with meta-analysis. International Journal of Molecular Sciences 25, 1168.
[116] Ning, Z., Xiao, Q., Feng, Q., Chen, W., Zhang, Y., 2021. Relation-induced multi-modal shared representation learning for alzheimer's disease diagnosis. IEEE Transactions on Medical Imaging 40, 1632–1645.
[117] Nordberg, A., Rinne, J.O., Kadir, A., Långström, B., 2010. The use of pet in alzheimer disease. Nature Reviews Neurology 6, 78–87.
[118] Ogawa, S., Lee, T.M., Kay, A.R., Tank, D.W., 1990. Brain magnetic resonance imaging with contrast dependent on blood oxygenation. proceedings of the National Academy of Sciences 87, 9868–9872.
[119] Olsson, B., Lautner, R., Andreasson, U., Öhrfelt, A., Portelius, E., Bjerke, M., Hölttä, M., Rosén, C., Olsson, C., Strobel, G., et al., 2016. Csf and blood biomarkers for the diagnosis of alzheimer's disease: a systematic review and meta-analysis. The Lancet Neurology 15, 673–684.
[120] Ortiz-Perez, D., Benavent-Lledo, M., Rodriguez-Juan, J., Garcia-Rodriguez, J., Tomás, D., 2025. Cognialign: Word-level multimodal speech alignment with gated cross-attention for alzheimer's detection. arXiv preprint arXiv:2506.01890 .

[121] Park, B., Kim, Y., Park, J., Choi, H., Kim, S.E., Ryu, H., Seo, K., 2024. Integrating biomarkers from virtual reality and magnetic resonance imaging for the early detection of mild cognitive impairment using a multimodal learning approach: validation study. Journal of Medical Internet Research 26, e54538.

[122] Peng, J., Zhu, X., Wang, Y., An, L., Shen, D., 2019. Structured sparsity regularized multiple kernel learning for alzheimer's disease diagnosis. Pattern recognition 88, 370–382.

[123] Perrin, R.J., Fagan, A.M., Holtzman, D.M., 2009. Multimodal techniques for diagnosis and prognosis of alzheimer's disease. Nature 461, 916–922.

[124] Plant, C., Teipel, S.J., Oswald, A., Böhm, C., Meindl, T., Mourao-Miranda, J., Bokde, A.W., Hampel, H., Ewers, M., 2010. Automated detection of brain atrophy patterns based on mri for the prediction of alzheimer's disease. Neuroimage 50, 162–174.

[125] Podhorna, J., Krahnke, T., Shear, M., E Harrison, J., Initiative, A.D.N., 2016. Alzheimer's disease assessment scale–cognitive subscale variants in mild cognitive impairment and mild alzheimer's disease: change over time and the effect of enrichment strategies. Alzheimer's research & therapy 8, 1–13.

[126] Polikar, R., Tilley, C., Hillis, B., Clark, C.M., 2010. Multimodal eeg, mri and pet data fusion for alzheimer's disease diagnosis, in: 2010 Annual international conference of the IEEE engineering in medicine and biology, IEEE. pp. 6058–6061.

[127] 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., 2022. Multimodal deep learning for alzheimer's disease dementia assessment. Nature communications 13, 3404.

[128] Qiu, Z., Yang, P., Xiao, C., Wang, S., Xiao, X., Qin, J., Liu, C.M., Wang, T., Lei, B., 2024. 3d multimodal fusion network with disease-induced joint learning for early alzheimer's disease diagnosis. IEEE Transactions on Medical Imaging .

[129] Rahim, N., El-Sappagh, S., Ali, S., Muhammad, K., Del Ser, J., Abuhmed, T., 2023. Prediction of alzheimer's progression based on multimodal deep-learning-based fusion and visual explainability of time-series data. Information Fusion 92, 363–388.

[130] Reas, E.T., Shadrin, A., Frei, O., Motazedi, E., McEvoy, L., Bahrami, S., van der Meer, D., Makowski, C., Loughnan, R., Wang, X., et al., 2023. Improved multimodal prediction of progression from mci to alzheimer's disease combining genetics with quantitative brain mri and cognitive measures. Alzheimer's & Dementia 19, 5151–5158.

[131] Ritter, K., Schumacher, J., Weygandt, M., Buchert, R., Allefeld, C., Haynes, J.D., Initiative, A.D.N., et al., 2015. Multimodal prediction of conversion to alzheimer's disease based on incomplete biomarkers. Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring 1, 206–215.

[132] Robert, P., Darcourt, G., Koulibaly, M., Clairet, S., Benoit, M., Garcia, R., Dechaux, O., Darcourt, J., 2006. Lack of initiative and interest in alzheimer's disease: a single photon emission computed tomography study. European journal of neurology 13, 729–735.

[133] Robert, R.E., Warach, S., 1993. Magnetic resonance imaging (1). The New England journal of medicine 328, 708–716.

[134] Rosén, C., Hansson, O., Blennow, K., Zetterberg, H., 2013. Fluid biomarkers in alzheimer's disease–current concepts. Molecular neurodegeneration 8, 1–11.

[135] Rüb, U., Del Tredici, K., Schultz, C., Büttner-Ennever, J., Braak, H., 2001. The premotor region essential for rapid vertical eye movements shows early involvement in alzheimer's disease-related cytoskeletal pathology. Vision research 41, 2149–2156.

[136] Sandbhor, B., Sharma, B., Palaniappan, B., 2025. Slam-dimm: Shared latent modeling for diffusion based missing modality synthesis in mri. arXiv preprint arXiv:2509.16019 .

[137] Sarawgi, U., Zulfikar, W., Soliman, N., Maes, P., 2020. Multimodal inductive transfer learning for detection of alzheimer's dementia and its severity. arXiv preprint arXiv:2009.00700 .

[138] Shao, W., Peng, Y., Zu, C., Wang, M., Zhang, D., Initiative, A.D.N., et al., 2020. Hypergraph based multi-task feature selection for multimodal classification of alzheimer's disease. Computerized Medical Imaging and Graphics 80, 101663.

[139] Shen, D., Wu, G., Suk, H.I., 2017. Deep learning in medical image analysis. Annual review of biomedical engineering 19, 221–248.

[140] Shi, J., Zheng, X., Li, Y., Zhang, Q., Ying, S., 2017. Multimodal neuroimaging feature learning with multimodal stacked deep polynomial networks for diagnosis of alzheimer's disease. IEEE journal of biomedical and health informatics 22, 173–183.

[141] Shi, Y., Suk, H.I., Gao, Y., Lee, S.W., Shen, D., 2019. Leveraging coupled interaction for multimodal alzheimer's disease diagnosis. IEEE transactions on neural networks and learning systems 31, 186–200.

[142] Soladoye, A.A., Aderinto, N., Osho, D., Olawade, D.B., 2025. Explainable machine learning models for early alzheimer's disease detection using multimodal clinical data. International journal of medical informatics , 106093.

[143] Song, X., Zhou, F., Frangi, A.F., Cao, J., Xiao, X., Lei, Y., Wang, T., Lei, B., 2022. Multicenter and multichannel pooling gcn for early ad diagnosis based on dual-modality fused brain network. IEEE Transactions on Medical Imaging 42, 354–367.

[144] Sonnenburg, S., Rätsch, G., Schäfer, C., Schölkopf, B., 2006. Large scale multiple kernel learning. The Journal of Machine Learning Research 7, 1531–1565.

[145] Stonnington, C.M., Chu, C., Klöppel, S., Jack Jr, C.R., Ashburner, J., Frackowiak, R.S., Initiative, A.D.N., et al., 2010. Predicting clinical scores from magnetic resonance scans in alzheimer's disease. Neuroimage 51, 1405–1413.

[146] Suk, H.I., Lee, S.W., Shen, D., Initiative, A.D.N., 2014a. Subclass-based multi-task learning for alzheimer's disease diagnosis. Frontiers in aging neuroscience 6, 168.

[147] Suk, H.I., Lee, S.W., Shen, D., Initiative, A.D.N., et al., 2014b. Hierarchical feature representation and multimodal fusion with deep learning for ad/mci diagnosis. NeuroImage 101, 569–582.

[148] Tabarestani, S., Aghili, M., Eslami, M., Cabrerizo, M., Barreto, A., Rishe, N., Curiel, R.E., Loewenstein, D., Duara, R., Adjouadi, M., 2020. A distributed multitask multimodal approach for the prediction of alzheimer's disease in a longitudinal study. NeuroImage 206, 116317.

[149] Tabarestani, S., Aghili, M., Shojaie, M., Freytes, C., Cabrerizo, M., Barreto, A., Rishe, N., Curiel, R.E., Loewenstein, D., Duara, R., et al., 2019. Longitudinal prediction modeling of alzheimer disease using recurrent neural networks, in: 2019 IEEE EMBS international conference on Biomedical & Health Informatics (BHI), IEEE. pp. 1–4.

[150] Tang, C., Wei, M., Sun, J., Wang, S., Zhang, Y., Initiative, A.D.N., et al., 2023. Csagp: detecting alzheimer's disease from multimodal images via dual-transformer with cross-attention and graph pooling. Journal of King Saud University-Computer and Information Sciences

35, 101618.
[151] Thung, K.H., Yap, P.T., Shen, D., 2017. Multi-stage diagnosis of alzheimer's disease with incomplete multimodal data via multi-task deep learning, in: International Workshop on Deep Learning in Medical Image Analysis, Springer. pp. 160–168.
[152] Tong, T., Gray, K., Gao, Q., Chen, L., Rueckert, D., Initiative, A.D.N., et al., 2017. Multi-modal classification of alzheimer's disease using nonlinear graph fusion. Pattern recognition 63, 171–181.
[153] Tu, Y., Lin, S., Qiao, J., Zhuang, Y., Zhang, P., 2022. Alzheimer's disease diagnosis via multimodal feature fusion. Computers in biology and medicine 148, 105901.
[154] Ujiie, M., Dickstein, D.L., Jefferies, W.A., et al., 2002. p97 as a biomarker for alzheimer disease. Front Biosci 7, e42–7.
[155] Vadacchino, S., Mehta, R., Sepahvand, N.M., Nichyporuk, B., Clark, J.J., Arbel, T., 2021. Had-net: A hierarchical adversarial knowledge distillation network for improved enhanced tumour segmentation without post-contrast images, in: Medical imaging with deep learning, PMLR. pp. 787–801.
[156] Valenzuela, M.J., Sachdev, P., 2001. Magnetic resonance spectroscopy in ad. Neurology 56, 592–598.
[157] Van Cauwenberghe, C., Van Broeckhoven, C., Sleegers, K., 2016. The genetic landscape of alzheimer disease: clinical implications and perspectives. Genetics in medicine 18, 421–430.
[158] Varghese, T., Sheelakumari, R., James, J.S., Mathuranath, P.S., 2013. A review of neuroimaging biomarkers of alzheimer's disease. Neurology Asia 18, 239.
[159] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. Advances in neural information processing systems 30.
[160] Velazquez, M., Lee, Y., 2022. Multimodal ensemble model for alzheimer's disease conversion prediction from early mild cognitive impairment subjects. Computers in biology and medicine 151, 106201.
[161] Venugopalan, J., Tong, L., Hassanzadeh, H.R., Wang, M.D., 2021. Multimodal deep learning models for early detection of alzheimer's disease stage. Scientific reports 11, 3254.
[162] Verma, H., Srivastava, H., Singh, A., Dehraj, P., 2025. Early detection of alzheimer's disease with blood plasma proteins using svm, in: Intelligent Computing and Communication Techniques. CRC Press, pp. 32–36.
[163] Vu, T.D., Yang, H.J., Nguyen, V.Q., Oh, A.R., Kim, M.S., 2017. Multimodal learning using convolution neural network and sparse autoencoder, in: 2017 IEEE international conference on big data and smart computing (BigComp), IEEE. pp. 309–312.
[164] Wang, C., Tachimori, H., Yamaguchi, H., Sekiguchi, A., Li, Y., Yamashita, Y., Initiative, A.D.N., 2024. A multimodal deep learning approach for the prediction of cognitive decline and its effectiveness in clinical trials for alzheimer's disease. Translational psychiatry 14, 105.
[165] Wang, C., Yang, G., Papanastasiou, G., Tsaftaris, S.A., Newby, D.E., Gray, C., Macnaught, G., MacGillivray, T.J., 2021. Dicyc: Gan-based deformation invariant cross-domain information fusion for medical image synthesis. Information Fusion 67, 147–160.
[166] Wang, M., Shao, W., Huang, S., Zhang, D., 2023. Hypergraph-regularized multimodal learning by graph diffusion for imaging genetics based alzheimer's disease diagnosis. Medical Image Analysis 89, 102883.
[167] Wang, Q., Xu, R., 2023. Aanet: Attentive all-level fusion deep neural network approach for multi-modality early alzheimer's disease diagnosis, in: AMIA Annual Symposium Proceedings, p. 1125.
[168] Wang, Y., Yang, Y., Guo, X., Ye, C., Gao, N., Fang, Y., Ma, H.T., 2018. A novel multimodal mri analysis for alzheimer's disease based on convolutional neural network, in: 2018 40th annual international conference of the IEEE engineering in medicine and biology society (EMBC), IEEE. pp. 754–757.
[169] Wee, C.Y., Yap, P.T., Zhang, D., Denny, K., Browndyke, J.N., Potter, G.G., Welsh-Bohmer, K.A., Wang, L., Shen, D., 2012. Identification of mci individuals using structural and functional connectivity networks. Neuroimage 59, 2045–2056.
[170] Weidemann, A., König, G., Bunke, D., Fischer, P., Salbaum, J.M., Masters, C.L., Beyreuther, K., 1989. Identification, biogenesis, and localization of precursors of alzheimer's disease a4 amyloid protein. Cell 57, 115–126.
[171] Weiner, M.W., Veitch, D.P., Aisen, P.S., Beckett, L.A., Cairns, N.J., Cedarbaum, J., Donohue, M.C., Green, R.C., Harvey, D., Jack Jr, C.R., et al., 2015. Impact of the alzheimer's disease neuroimaging initiative, 2004 to 2014. Alzheimer's & Dementia 11, 865–884.
[172] Whitwell, J., Josephs, K., Murray, M., Kantarci, K., Przybelski, S., Weigand, S., Vemuri, P., Senjem, M., Parisi, J., Knopman, D., et al., 2008. Mri correlates of neurofibrillary tangle pathology at autopsy: a voxel-based morphometry study. Neurology 71, 743–749.
[173] Wu, B., Xu, C., Dai, X., Wan, A., Zhang, P., Yan, Z., Tomizuka, M., Gonzalez, J., Keutzer, K., Vajda, P., 2020. Visual transformers: Token-based image representation and processing for computer vision. arXiv preprint arXiv:2006.03677 .
[174] Xiang, Y., Fan, L., Tulika, S., Pang, X., Pan, Y., Zhang, H., Ji, C., 2025. Rdsa: A robust deep graph clustering framework via dual soft assignment, in: Proceedings of The 30th International Conference on Database Systems for Advanced Applications (DASFAA 2025).
[175] Xu, L., Wu, X., Chen, K., Yao, L., 2015. Multi-modality sparse representation-based classification for alzheimer's disease and mild cognitive impairment. Computer methods and programs in biomedicine 122, 182–190.
[176] Xu, P., Zhu, X., Clifton, D.A., 2023. Multimodal learning with transformers: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence 45, 12113–12132.
[177] Xu, W., Gan, H.S., Wu, S., Wang, Z., Ramlee, M.H., Hafizah, W.M., 2025. Mmknet: A multi-modal knowledge network for predicting both seen and unseen classes in medical imaging, in: 2025 28th International Conference on Computer Supported Cooperative Work in Design (CSCWD), IEEE. pp. 1599–1604.
[178] Yao, Z., Xie, W., Chen, J., Zhan, Y., Wu, X., Dai, Y., Pei, Y., Wang, Z., Zhang, G., 2025. It: An interpretable transformer model for alzheimer's disease prediction based on pet/mr images. NeuroImage 311, 121210.
[179] Yin, Y., Wang, H., Liu, S., Sun, J., Jing, P., Liu, Y., 2023. Internet of things for diagnosis of alzheimer's disease: A multimodal machine learning approach based on eye movement features. IEEE Internet of Things Journal 10, 11476–11485.
[180] Young, J., Modat, M., Cardoso, M.J., Mendelson, A., Cash, D., Ourselin, S., Initiative, A.D.N., et al., 2013. Accurate multimodal probabilistic prediction of conversion to alzheimer's disease in patients with mild cognitive impairment. NeuroImage: Clinical 2, 735–745.

[181] Yu, Q., Ma, Q., Da, L., Li, J., Wang, M., Xu, A., Li, Z., Li, W., Initiative, A.D.N., et al., 2024. A transformer-based unified multimodal framework for alzheimer's disease assessment. Computers in Biology and Medicine 180, 108979.

[182] Zhan, M., Zhao, K., Liu, G., Tang, H., 2025. A general paradigm for fine-tuning large language models in alzheimer's disease diagnosis, in: Proceedings of the AAAI Symposium Series, pp. 37–42.

[183] Zhang, D., Shen, D., 2011a. Multicost: Multi-stage cost-sensitive classification of alzheimer's disease, in: Machine Learning in Medical Imaging: Second International Workshop, MLMI 2011, Held in Conjunction with MICCAI 2011, Toronto, Canada, September 18, 2011. Proceedings 2, Springer. pp. 344–351.

[184] Zhang, D., Shen, D., 2011b. Semi-supervised multimodal classification of alzheimer's disease, in: 2011 IEEE International Symposium on Biomedical Imaging: From Nano to Macro, IEEE. pp. 1628–1631.

[185] Zhang, D., Shen, D., Initiative, A.D.N., et al., 2012. Multi-modal multi-task learning for joint prediction of multiple regression and classification variables in alzheimer's disease. NeuroImage 59, 895–907.

[186] Zhang, D., Wang, Y., Zhou, L., Yuan, H., Shen, D., Initiative, A.D.N., et al., 2011. Multimodal classification of alzheimer's disease and mild cognitive impairment. Neuroimage 55, 856–867.

[187] Zhang, F., Li, Z., Zhang, B., Du, H., Wang, B., Zhang, X., 2019. Multi-modal deep learning model for auxiliary diagnosis of alzheimer's disease. Neurocomputing 361, 185–195.

[188] Zhang, Y., Li, Y., Ma, L., 2020a. Recent advances in research on alzheimer's disease in china. Journal of clinical neuroscience 81, 43–46.

[189] Zhang, Y., Wang, S., Xia, K., Jiang, Y., Qian, P., Initiative, A.D.N., et al., 2021. Alzheimer's disease multiclass diagnosis via multimodal neuroimaging embedding feature selection and fusion. Information Fusion 66, 170–183.

[190] Zhang, Y.D., Dong, Z., Wang, S.H., Yu, X., Yao, X., Zhou, Q., Hu, H., Li, M., Jiménez-Mesa, C., Ramirez, J., et al., 2020b. Advances in multimodal data fusion in neuroimaging: Overview, challenges, and novel orientation. Information Fusion 64, 149–187.

[191] Zhong, W., Cong, C., Wang, Z., Yan, Z., Di Ieva, A., Liu, S., 2025. Fmm-diff: A feature mapping and merging diffusion model for mri generation with missing modality, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 227–236.

[192] Zhou, J., Liu, J., Narayan, V.A., Ye, J., Initiative, A.D.N., et al., 2013. Modeling disease progression via multi-task learning. NeuroImage 78, 233–248.

[193] Zhou, T., Cheng, Q., Lu, H., Li, Q., Zhang, X., Qiu, S., 2023. Deep learning methods for medical image fusion: A review. Computers in Biology and Medicine 160, 106959.

[194] Zhou, T., Liu, M., Thung, K.H., Shen, D., 2019a. Latent representation learning for alzheimer's disease diagnosis with incomplete multi-modality neuroimaging and genetic data. IEEE transactions on medical imaging 38, 2411–2422.

[195] Zhou, T., Thung, K.H., Zhu, X., Shen, D., 2019b. Effective feature learning and fusion of multimodality data using stage-wise deep neural network for dementia diagnosis. Human brain mapping 40, 1001–1016.

[196] Zhou, W., Luo, W., Gong, L., Peng, B., Group, C.C.A.C.M., 2025. Enhanced early diagnosis of alzheimer's disease with hybridca-net: A multimodal fusion approach. Expert Systems with Applications 292, 128580.

[197] Zhou, Y., Geng, P., Zhang, S., Xiao, F., Cai, G., Chen, L., Initiative, A.D.N., Lu, Q., 2024. Multimodal functional deep learning for multiomics data. Briefings in Bioinformatics 25, bbae448.

[198] Zhu, X., Suk, H.I., Lee, S.W., Shen, D., 2015. Subspace regularized sparse multitask learning for multiclass neurodegenerative disease identification. IEEE Transactions on Biomedical Engineering 63, 607–618.

[199] Zhu, X., Suk, H.I., Shen, D., 2014. A novel matrix-similarity based loss function for joint regression and classification in ad diagnosis. NeuroImage 100, 91–105.

[200] Zu, C., Jie, B., Liu, M., Chen, S., Shen, D., Zhang, D., Initiative, A.D.N., 2016. Label-aligned multi-task feature learning for multimodal classification of alzheimer's disease and mild cognitive impairment. Brain imaging and behavior 10, 1148–1159.