American Journal of Advanced Technology and Engineering Solutions

Volume: 6; **Issue:** 1
Pages: 622-649
eISSN: 3067-5146

open access

Crossref

Machine Learning and Deep Learning for Alzheimer's Disease Diagnosis: A Survey of MRI, PET, Multimodal Fusion, Transformers, Graph Neural Networks, and Explainable AI

Imran Ahmad[1], Tasfia Tarannum[2], Mohammed Majbah Uddin[3];**

[1] *Department of Business Analytics, Wichita State University, Wichita, KS, USA; Email: ixahmad1@shockers.wichita.edu*

[2] *Southern Arkansas University, Magnolia, AR, USA, Email: ttarannum3228@muleriders.saumag.edu*

[3] *School of Business & Technology, Emporia State University, Emporia, Kansas, USA; Email: udduddinmajbahthe@gmail.com*

***Corresponding author:** Mohammed Majbah Uddin*

Doi: 10.63125/vf6dkg22

Received: *11 December 2025;* ***Revised:*** *13 January 2026;* ***Accepted:*** *13 February 2026;* ***Published:*** *25 March 2026*

Abstract

Alzheimer's disease is the leading cause of dementia and remains a major challenge in neurological diagnosis because early symptoms often overlap with normal aging and other neurodegenerative disorders. In recent years, machine learning and deep learning methods have been widely applied to neuroimaging and multimodal clinical data to improve automated diagnosis, risk prediction, and disease-stage classification. This survey reviews representative studies on Alzheimer's disease diagnosis using traditional machine learning, convolutional neural networks, multimodal fusion methods, transformer-based architectures, graph neural networks, and explainable artificial intelligence. The literature shows a clear transition from handcrafted-feature pipelines to end-to-end deep models trained on magnetic resonance imaging, positron emission tomography, and multimodal data. It also suggests that multimodal systems often outperform single-modality systems because they integrate complementary structural and metabolic information, while graph and transformer methods aim to capture more global and relational disease patterns. At the same time, the field still faces major challenges, including dataset dependence, limited external validation, inconsistent evaluation settings, class imbalance, insufficient interpretability assessment, and weak evidence for clinical deployment. This survey organizes the literature by modeling paradigm and data modality, compares representative methods, and outlines future directions for robust, explainable, and clinically useful Alzheimer's disease diagnosis systems.

Keywords

Alzheimer's Disease, Machine Learning, Deep Learning, MRI, PET, Multimodal Fusion, Transformers, Graph Neural Networks, Explainable AI.

INTRODUCTION

Alzheimer's disease (AD) is the most common cause of dementia and is a major public health concern because it progressively impairs memory, cognition, and everyday functioning. Early diagnosis is clinically important because patient management, supportive care, and disease monitoring are more effective when started before severe decline (Alsubaie et al., 2024; Qiu et al., 2022). However, early-stage AD is difficult to detect because mild cognitive impairment and early Alzheimer's disease may show subtle and overlapping patterns. For this reason, neuroimaging and computational intelligence have become central to research on improved diagnostic precision (Qiu et al., 2022; Zhao et al., 2024).

Among neuroimaging modalities, structural magnetic resonance imaging (MRI) is widely used to capture brain atrophy and anatomical degeneration, while positron emission tomography (PET) provides metabolic or amyloid-related information that complements structural findings. These modalities have motivated a large body of machine learning and deep learning research that aims to distinguish healthy controls, mild cognitive impairment, Alzheimer's disease, and, in some studies, other dementia types. The field has gradually moved from conventional machine learning with handcrafted features toward convolutional neural networks, multimodal fusion models, transformer-equipped architectures, and graph-based approaches that learn richer representations directly from data (Castellano et al., 2024; Lu et al., 2018; Zhang et al., 2023).

The aim of this survey is to synthesize representative work at the intersection of Alzheimer's disease diagnosis and machine learning or deep learning, with particular emphasis on neuroimaging-based methods. Rather than listing papers sequentially, the survey explains how the field has evolved, what methodological directions have become dominant, where current methods succeed, and which challenges continue to hinder clinical translation. To do this, the survey organizes the literature into five broad groups: traditional machine learning, convolutional deep learning, multimodal fusion, transformer and hybrid attention models, and graph neural networks with explainability components (Alsubaie et al., 2024).

Scope and Selection of Representative Studies

This article is designed as a structured narrative survey rather than a formal systematic review. Its purpose is not to exhaustively catalogue every published study on Alzheimer's disease diagnosis, but to synthesize a representative body of influential work that captures the main methodological directions of the field. Because research on Alzheimer's disease diagnosis using machine learning and deep learning has grown rapidly across multiple data modalities and modeling paradigms, a carefully selected narrative approach is useful for organizing the literature into a coherent framework. In this survey, emphasis is placed on the conceptual evolution of the field, the comparative strengths of major model families, and the methodological features that are most relevant to clinical translation.

The paper set was selected to cover approximately 10 to 15 representative and influential studies spanning early MRI-based convolutional neural network models, multimodal MRI plus PET systems, attention-based fusion frameworks, graph neural networks, transformer-equipped architectures, and explainability-oriented approaches. In addition to original research articles, review papers were included when they helped support broader observations about datasets, modality trends, evaluation practices, and unresolved methodological limitations (Alsubaie et al., 2024). This combination of representative primary studies and supporting review literature makes it possible to discuss not only how individual models work, but also how the overall field has progressed from handcrafted-feature pipelines to more integrated, multimodal, and clinically aware AI systems.

The surveyed studies include work by Islam and Zhang (2018), Lu et al. (2018), Golovanevsky et al. (2022), Qiu et al. (2022), Zhang et al. (2023), Castellano et al. (2024), Hu et al. (2024), and Zhao et al. (2024). These papers were chosen because they are methodologically distinct, widely discussed in the field, and collectively adequate to support a coherent survey taxonomy. Together, they capture the transition from early CNN-based diagnosis to multimodal learning, relational modeling, transformer-based representation learning, and explainability-focused frameworks. They also represent different levels of clinical ambition, ranging from single-modality classification studies to broader multimodal diagnostic systems that attempt external validation, interpretation, or more realistic disease assessment settings.

The scope of the survey is centered primarily on neuroimaging-based Alzheimer's disease diagnosis, with particular emphasis on structural MRI, PET, and multimodal combinations of imaging and non-imaging data. This emphasis is intentional because neuroimaging remains one of the most active and clinically important areas in computational Alzheimer's disease research. At the same time, the survey does not treat imaging in isolation. Studies that integrate demographic, phenotypic, cognitive, biomarker, or claims-based variables are also considered when they contribute meaningfully to the discussion of multimodal diagnosis, graph-based modeling, or explainability. In this way, the survey reflects the broader movement of the field toward richer and more clinically realistic diagnostic frameworks.

Another important feature of this survey is that it emphasizes methodological contribution and clinical relevance rather than ranking studies only by reported accuracy. This is necessary because the literature is highly heterogeneous in terms of datasets, tasks, preprocessing choices, evaluation settings, and target diagnostic categories. Some papers focus on binary Alzheimer's disease versus healthy control classification, while others address mild cognitive impairment, progression prediction, or multi-class dementia assessment. Similarly, some studies are based on single-modality MRI alone, whereas others use multimodal imaging or integrate clinical variables. Under these conditions, direct numerical comparison can be misleading. A model that reports very high performance on a simple and highly curated task may not necessarily be more meaningful than one that addresses a harder but more clinically relevant problem. For this reason, the present survey prioritizes interpretive depth, methodological structure, and translational value over leaderboard-style comparison.

Finally, the survey is intended to provide a balanced view of both progress and limitations. It highlights the advantages of newer methods such as multimodal fusion, graph neural networks, transformer-equipped models, and explainable AI, while also acknowledging ongoing challenges such as benchmark dependence, inconsistent evaluation, missing modality handling, and limited external validation. By selecting representative studies across these themes, the survey aims to provide readers with a clear understanding of where the field has come from, where it currently stands, and which research directions are most likely to shape the next generation of Alzheimer's disease diagnosis systems.

Background: Alzheimer's Disease and Computational Diagnosis

Computational Alzheimer's disease diagnosis generally focuses on a range of prediction and classification tasks that vary in both clinical importance and methodological difficulty. Common tasks include distinguishing Alzheimer's disease from cognitively normal controls, separating Alzheimer's disease from mild cognitive impairment, performing multi-class stage classification, or predicting whether a patient with mild cognitive impairment is likely to progress to Alzheimer's disease over time. These tasks are not equally challenging. Binary classification between Alzheimer's disease and healthy controls is usually the most straightforward because the structural and functional differences between these groups are often more pronounced. In contrast, mild cognitive impairment occupies an intermediate and heterogeneous stage, which makes its classification and progression prediction much more difficult. Similarly, multi-class diagnosis is more clinically informative because it reflects disease staging more realistically, but it is also more demanding because class boundaries are less distinct and the underlying biological patterns are more overlapping. For this reason, reported performance values must be interpreted carefully, since a model that performs well on an easier binary task may still be less clinically useful than a model addressing a harder but more relevant diagnostic problem (Lu et al., 2018; Qiu et al., 2022).

The importance of computational diagnosis in Alzheimer's disease arises from the limitations of traditional clinical assessment alone. Clinical diagnosis often depends on neurological examination, cognitive testing, patient history, and neuroimaging interpretation, but early disease signs can be subtle and may overlap with normal aging or other forms of dementia. Machine learning and deep learning methods have therefore been developed to assist by identifying hidden or distributed patterns in imaging and multimodal clinical data that may not be easily captured through manual inspection alone. In this context, computational models are not intended simply to replace clinicians, but rather to support earlier detection, more consistent classification, improved risk stratification, and more informed clinical decision-making.

A major aspect of this research area is the strong role of neuroimaging. Structural MRI is commonly used because it can reveal anatomical changes such as hippocampal atrophy, cortical thinning, ventricular enlargement, and other signs of neurodegeneration. PET imaging contributes complementary metabolic or amyloid-related information, which is particularly valuable in identifying disease processes that may not yet be fully visible in structural scans. Together with demographic, cognitive, genetic, or biomarker information, these modalities form the basis of many computational Alzheimer's disease diagnosis systems. As a result, modern research increasingly treats Alzheimer's disease diagnosis as a multimodal learning problem rather than a purely image-based classification problem.

Most recent Alzheimer's disease diagnosis studies rely heavily on public datasets, especially the Alzheimer's Disease Neuroimaging Initiative, or ADNI. This dataset has become central to the field because it provides standardized neuroimaging, cognitive, and clinical data across different diagnostic groups, making it highly attractive for benchmarking machine learning and deep learning methods. More recent studies have also begun to explore OASIS-3, AIBL, and external cohort validation, which is an important development because dependence on a single benchmark can limit generalizability. Dataset choice influences many aspects of model performance, including sample size, class balance, modality availability, demographic representation, disease-stage distribution, and preprocessing practice. Consequently, variation in dataset design can strongly affect how well different models appear to perform.

Another important issue is that methodological comparisons are often complicated by inconsistent evaluation settings. Different studies may use different train-test splits, cross-validation schemes, diagnostic definitions, preprocessing pipelines, inclusion criteria, and target tasks. Some studies classify only Alzheimer's disease versus healthy control, while others include mild cognitive impairment, progression prediction, or broader dementia categories. Some rely on single-modality MRI, whereas others use PET, multimodal fusion, or additional clinical inputs. Because of this heterogeneity, performance numbers such as accuracy, sensitivity, specificity, or AUC cannot always be compared directly across papers. A high score in one study may reflect an easier task, more curated data, or less realistic evaluation conditions rather than a genuinely superior model.

Review studies have repeatedly noted that lack of standardized evaluation and overreliance on a few benchmark datasets remain central weaknesses in the literature (Alsubaie et al., 2024; Qiu et al., 2022). These issues are particularly significant for clinical translation, because a model that performs well in a controlled experimental setting may not transfer successfully across scanners, institutions, populations, or real-world workflow conditions. In addition, many published studies do not sufficiently examine robustness to class imbalance, missing modalities, demographic diversity, or domain shift. These limitations suggest that the field still faces a gap between promising computational performance and dependable clinical deployment.

Overall, the background of computational Alzheimer's disease diagnosis reflects both substantial progress and important unresolved challenges. The field has moved from relatively simple feature-based classification toward increasingly complex multimodal and deep learning systems, yet its central goals remain the same: earlier detection, more reliable diagnosis, and better support for clinical decision-making. Understanding the diversity of tasks, datasets, modalities, and evaluation settings is therefore essential for interpreting the literature fairly and for identifying which methodological advances are likely to have genuine clinical value.

Datasets and Input Modalities

Structural magnetic resonance imaging (MRI) is the most widely used input modality in Alzheimer's disease diagnosis research because it is noninvasive, clinically accessible, and highly informative for characterizing neuroanatomical degeneration. MRI enables quantitative assessment of cortical thinning, hippocampal atrophy, ventricular enlargement, gray matter loss, and other structural alterations associated with progressive neurodegeneration. These characteristics make MRI particularly valuable for computational pipelines that aim to differentiate cognitively normal controls, mild cognitive impairment, and Alzheimer's disease using morphometric and voxel-level patterns. From a machine learning perspective, MRI also provides rich spatial information that can be exploited through region-of-interest analysis, volumetric feature extraction, texture modeling, and end-to-end

deep representation learning.

Positron emission tomography (PET) provides a complementary imaging perspective by capturing metabolic and pathological processes that may not be fully visible in structural MRI alone. In Alzheimer's disease research, FDG-PET is commonly used to assess cerebral glucose metabolism, while amyloid PET provides information about amyloid-beta deposition. These modalities are especially important because functional or molecular abnormalities may emerge before pronounced structural degeneration becomes apparent. As a result, PET can strengthen diagnostic sensitivity in early-stage disease assessment and progression-related modeling. Several influential studies have shown that MRI and PET are not redundant modalities, but rather complementary sources of structural and biological information, which is a major reason why multimodal learning has become a dominant methodological trend in the field (Castellano et al., 2024; Lu et al., 2018).

Beyond neuroimaging alone, many recent Alzheimer's disease diagnosis systems incorporate non-imaging variables such as age, sex, education, APOE-related genetic risk markers, phenotypic information, biomarker profiles, clinical history, and cognitive test scores. These variables can provide important contextual information that is difficult to infer directly from imaging data. For example, demographic and cognitive features may improve discrimination in borderline or heterogeneous cases, while biomarker and genetic information may strengthen disease characterization in multimodal risk prediction settings. The integration of such heterogeneous inputs reflects the broader shift from image-only classification toward clinically informed multimodal diagnosis frameworks.

This multimodal perspective is especially important because Alzheimer's disease is not purely a structural imaging problem. It is a complex neurodegenerative disorder involving anatomical, metabolic, cognitive, and molecular dimensions. Computational models that combine these sources are therefore often better positioned to capture clinically meaningful disease signatures than models based on a single input channel. Qiu et al. (2022), for example, proposed a multimodal framework that extends beyond narrow image classification and supports successive diagnostic steps across normal cognition, mild cognitive impairment, Alzheimer's disease, and non-Alzheimer dementias. Similarly, multimodal attention-based approaches and graph neural network frameworks have demonstrated that non-imaging variables can improve predictive performance when they are fused in a principled way with imaging-derived representations (Golovanevsky et al., 2022; Qiu et al., 2022; Zhang et al., 2023).

At the dataset level, Alzheimer's disease diagnosis research relies heavily on a small number of large public resources. The Alzheimer's Disease Neuroimaging Initiative (ADNI) remains the most dominant benchmark because it provides longitudinal and cross-sectional imaging, cognitive, demographic, and biomarker data across multiple diagnostic groups. OASIS and OASIS-3 are also important because they broaden the evaluation landscape and include structural imaging together with associated clinical and cognitive information. AIBL is another relevant dataset used in studies involving multimodal and biomarker-informed modeling. These datasets differ in scale, cohort composition, modality availability, diagnostic labeling, and longitudinal structure, which means that model performance can depend substantially on the dataset used for training and evaluation.

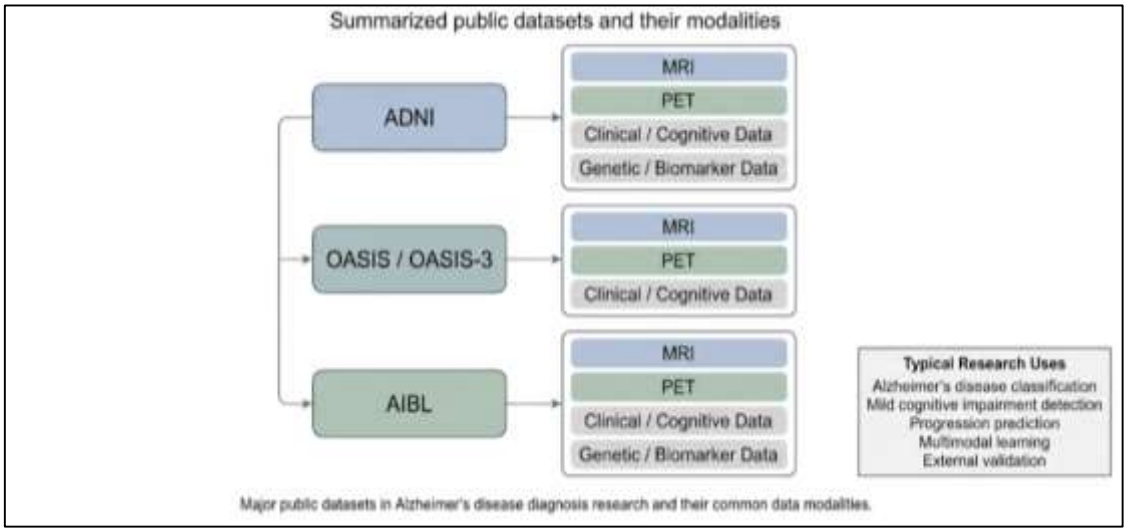
From a technical standpoint, input representation varies considerably across studies. MRI and PET may be used as raw 2D slices, full 3D volumes, region-based summaries, voxel-based maps, or features extracted after spatial normalization and anatomical parcellation. Non-imaging variables may be concatenated directly with learned image features, processed through separate network branches, or incorporated through attention-based fusion or graph-based relational modeling. The choice of input representation has major implications for what information the model can learn. Slice-based approaches may reduce computational cost but can lose global anatomical continuity, whereas volumetric pipelines preserve richer spatial context at the cost of higher memory and training demands.

A recurring methodological issue in this area is preprocessing. Alzheimer's disease studies differ in whether they apply skull stripping, intensity normalization, image registration, spatial alignment to standard atlases, tissue segmentation, bias field correction, or region-of-interest extraction before model training. These preprocessing steps are not merely technical details. They affect signal consistency, anatomical comparability, feature stability, and ultimately model performance. In multimodal settings, the challenge becomes even greater because alignment across modalities must also be handled carefully. Consequently, differences in preprocessing strategy often make comparison across studies

difficult and remain an important source of variability in reported results.

Figure 1 summarizes the major public datasets commonly used in Alzheimer's disease diagnosis research and highlights the associated imaging and non-imaging modalities that support modern multimodal modeling. The figure emphasizes that the field is increasingly driven not only by advances in model architecture, but also by the availability and integration of structurally, functionally, and clinically heterogeneous inputs.

Figure 1. Overview of common datasets and modalities in Alzheimer's disease diagnosis research



A recurring methodological issue is preprocessing. Studies differ in whether they use 2D slices, full 3D volumes, skull stripping, spatial normalization, intensity normalization, region-of-interest extraction, or handcrafted brain network construction. These differences shape what patterns the model can learn and complicate fair comparison across studies. Review evidence suggests that reproducibility and standardized preprocessing remain weak across a substantial part of the literature (Alsubaie et al., 2024).

Figure 2: Representative MRI samples from Alzheimer's disease-related diagnostic categories.
The figure shows example brain MRI scans from categories such as non-demented, very mild dementia, and mild dementia, illustrating the visual variability present in Alzheimer's disease imaging datasets.

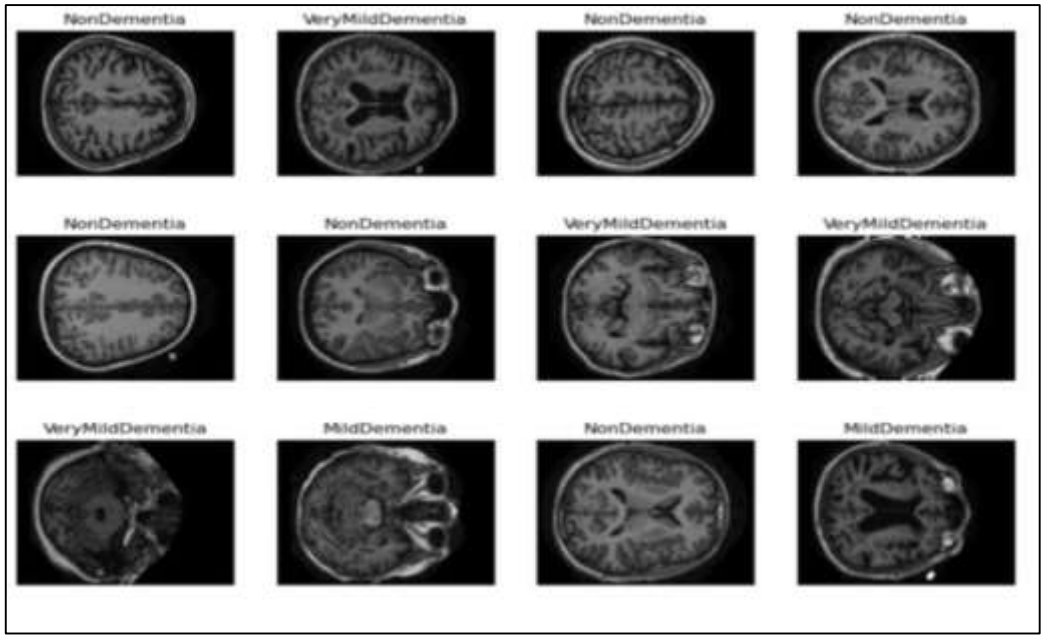


Figure 2: Representative MRI samples from Alzheimer's disease-related diagnostic categories

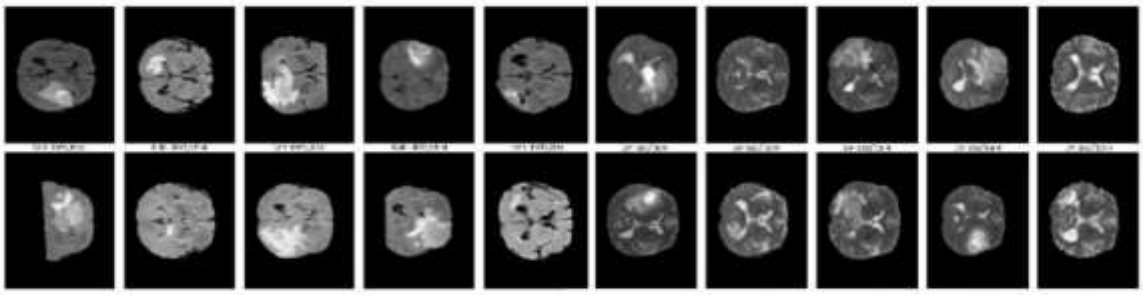
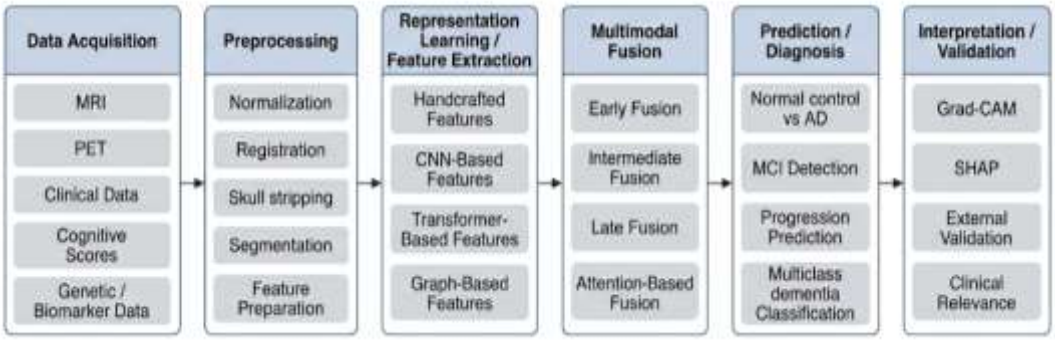


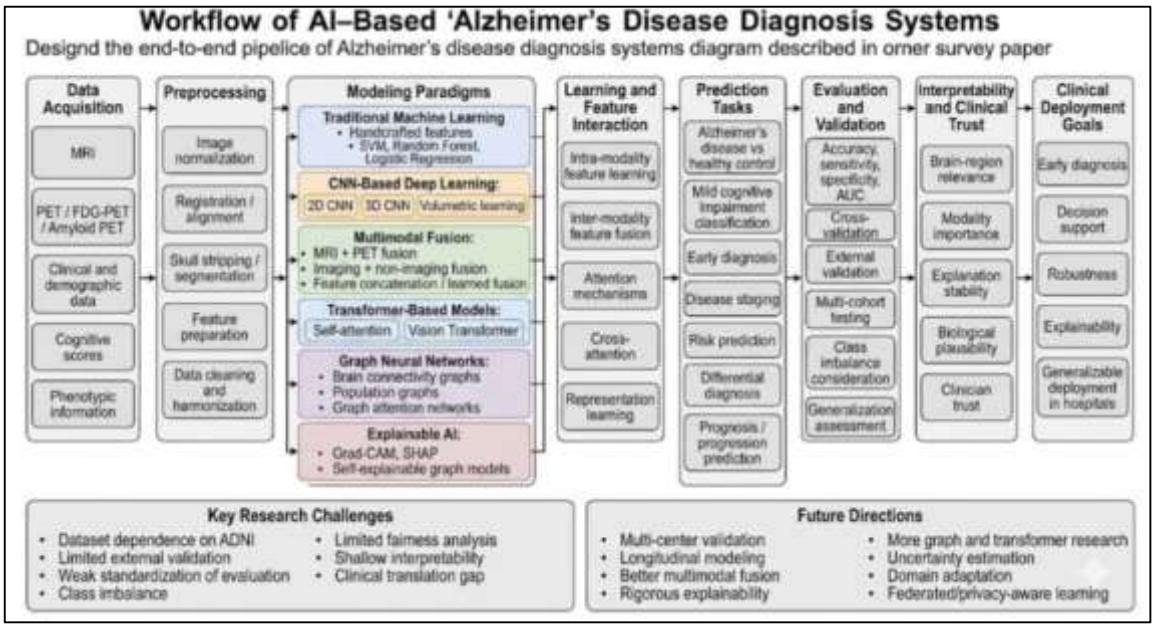
Fig. 1. Overview of Datasets

Figure 3. General workflow of machine learning and deep learning pipelines for Alzheimer's disease diagnosis



The figure illustrates a typical end-to-end pipeline used in Alzheimer's disease diagnosis systems, beginning with multimodal data acquisition and preprocessing, followed by feature learning, multimodal fusion, diagnostic prediction, and interpretation or validation.

Figure 4: Taxonomy of Machine Learning and Deep Learning Approaches for Alzheimer's Disease Diagnosis



TAXONOMY OF METHODS

Traditional machine learning

Traditional machine learning methods generally rely on handcrafted features such as cortical thickness, volumetric measurements, texture descriptors, clinical variables, or brain network statistics, which are then used by classifiers such as support vector machines, random forests, logistic regression, or boosting methods. In the early stages of Alzheimer's disease diagnosis research, these approaches were especially important because they demonstrated that computational models could detect meaningful disease-related patterns from neuroimaging and associated patient data. They provided the initial methodological foundation for automated diagnosis by showing that structural and clinical markers could be quantified, selected, and translated into predictive models. In this sense, traditional machine learning established the basic feasibility of using computational intelligence to distinguish healthy controls, mild cognitive impairment, and Alzheimer's disease from available imaging and clinical information.

One major strength of these methods is that they are often easier to interpret and less computationally demanding than deep learning systems. Because the input features are usually defined explicitly, researchers can directly examine which variables contribute to the model, such as hippocampal volume, cortical thickness in specific brain regions, or selected cognitive scores. This makes traditional machine learning attractive in smaller datasets or settings where transparency and simplicity are especially important. In addition, these methods often perform reasonably well when the feature engineering process is carefully designed and guided by clinical or neuroscientific knowledge. For this reason, they played a central role in early Alzheimer's disease research and continue to serve as useful baselines in many contemporary studies. However, these approaches also have important limitations. Their performance depends heavily on the quality of feature design, preprocessing, and expert-driven selection of relevant biomarkers. If the handcrafted features fail to capture subtle or distributed disease-related patterns, the model may miss important diagnostic information. This is particularly significant in Alzheimer's disease, where pathology is often complex, progressive, and spread across multiple interconnected brain regions. Unlike end-to-end deep learning methods, traditional machine learning does not automatically learn hierarchical feature representations from raw data. As a result, its flexibility is restricted, especially when dealing with large-scale multimodal data or complex nonlinear patterns that are difficult to express through manually engineered features alone (Lu et al., 2018; Alsubaie et al., 2024).

Another challenge is that handcrafted-feature pipelines often vary considerably across studies. Different researchers may use different preprocessing workflows, feature extraction techniques, region-of-interest definitions, or feature selection strategies, making it difficult to compare results directly. This variability reduces reproducibility and can limit the generalizability of findings across datasets or institutions. Consequently, although traditional machine learning methods remain valuable for interpretability, baseline comparison, and smaller-scale clinical studies, the broader trend in the field has shifted toward deep learning approaches that can learn richer and more adaptive representations directly from neuroimaging and multimodal data.

Convolutional neural networks

Convolutional neural networks became dominant in Alzheimer's disease diagnosis research because they can learn discriminative features directly from imaging data rather than depending entirely on manually engineered inputs. This marked an important shift in the field. Earlier machine learning approaches required researchers to define which anatomical, textural, or clinical features might be relevant before training a classifier. CNNs reduced this dependency by allowing the model to learn hierarchical representations automatically from MRI or PET images. In Alzheimer's disease, where pathological changes may be subtle, spatially distributed, and difficult to summarize through a small set of handcrafted descriptors, this ability to learn directly from raw or minimally processed imaging data became especially valuable (Islam & Zhang, 2018).

A major advantage of CNNs is their capacity to capture local spatial structure through convolutional filters and pooling operations. This makes them particularly well suited to neuroimaging tasks, where disease-related changes may appear as regional patterns of cortical thinning, ventricular enlargement,

hippocampal atrophy, or other structural abnormalities. By stacking multiple convolutional layers, CNNs can learn increasingly abstract and informative feature hierarchies, moving from lower-level image characteristics to more complex disease-related patterns. This hierarchical learning capability enabled CNN-based systems to outperform many earlier feature-engineered pipelines and established deep learning as a major methodological direction in Alzheimer's disease diagnosis.

Early CNN studies demonstrated that these models could automatically extract disease-related information from MRI without requiring all relevant features to be explicitly engineered by humans. This was an important step because it showed that the model itself could identify useful representations for distinguishing healthy controls, mild cognitive impairment, and Alzheimer's disease. The work of Islam and Zhang (2018) is representative of this stage and helped show that end-to-end deep convolutional diagnosis was both feasible and effective for Alzheimer-related neuroimaging tasks. Such studies laid the groundwork for a much broader wave of CNN-based medical imaging research and strongly influenced later developments in multimodal, volumetric, and attention-based systems.

CNNs also opened the way for deeper, more scalable, and more transferable neuroimaging pipelines. As the field progressed, researchers moved from simpler 2D slice-based models toward more sophisticated 3D CNN architectures that could better preserve anatomical context across the brain. This transition was important because Alzheimer's disease does not affect isolated slices alone, but rather involves distributed structural changes across multiple regions. CNN-based models also became more compatible with transfer learning, ensemble design, and multimodal fusion, allowing researchers to build stronger diagnostic systems that could integrate MRI, PET, and clinical features within a common deep learning framework.

Despite these strengths, CNN-based methods are not without limitations. Standard convolutions are inherently local, which means that although CNNs are powerful for spatial feature extraction, they may not fully capture broader global relationships across distant brain regions. In addition, deep CNNs often require substantial training data and careful regularization to avoid overfitting, which can be challenging in medical imaging settings where labeled datasets are relatively limited compared with natural image domains. These limitations have motivated newer developments such as multimodal fusion, transformer-equipped models, and graph neural networks. Even so, CNNs remain one of the most important foundations of modern Alzheimer's disease diagnosis research because they transformed the field from manual feature engineering toward learned representation-based modeling.

Multimodal fusion methods

Multimodal fusion methods have become one of the most important directions in Alzheimer's disease diagnosis research because they combine multiple sources of information rather than relying on a single imaging modality or isolated clinical feature set. In most studies, multimodal learning involves the integration of structural MRI and PET, although some frameworks also incorporate cognitive scores, demographic variables, genetic information, biomarkers, or other clinical measures. This approach is especially valuable in Alzheimer's disease because the disorder is complex and progressive, and no single modality is able to capture all of its relevant biological and clinical characteristics. MRI provides structural information about cortical thinning, hippocampal atrophy, ventricular enlargement, and other anatomical changes, while PET contributes metabolic or amyloid-related evidence that may reveal disease processes not fully visible in structural scans alone. These sources are therefore complementary rather than redundant, which is why multimodal frameworks often achieve stronger predictive performance and greater clinical relevance than single-modality systems (Lu et al., 2018; Castellano et al., 2024; Qiu et al., 2022).

A major strength of multimodal learning is that it supports a more comprehensive representation of disease state. Alzheimer's disease is not only a structural neurodegenerative condition, but also a metabolic, cognitive, and sometimes biomarker-driven disorder. By combining imaging and non-imaging information, multimodal systems can capture multiple dimensions of the disease simultaneously. For example, a model may use MRI to identify anatomical degeneration, PET to detect functional or pathological abnormalities, and cognitive or clinical variables to contextualize those imaging findings in relation to patient-level status. This broader representation is particularly useful in clinically challenging tasks such as distinguishing mild cognitive impairment from early Alzheimer's disease, predicting disease progression, or assessing dementia severity across multiple categories.

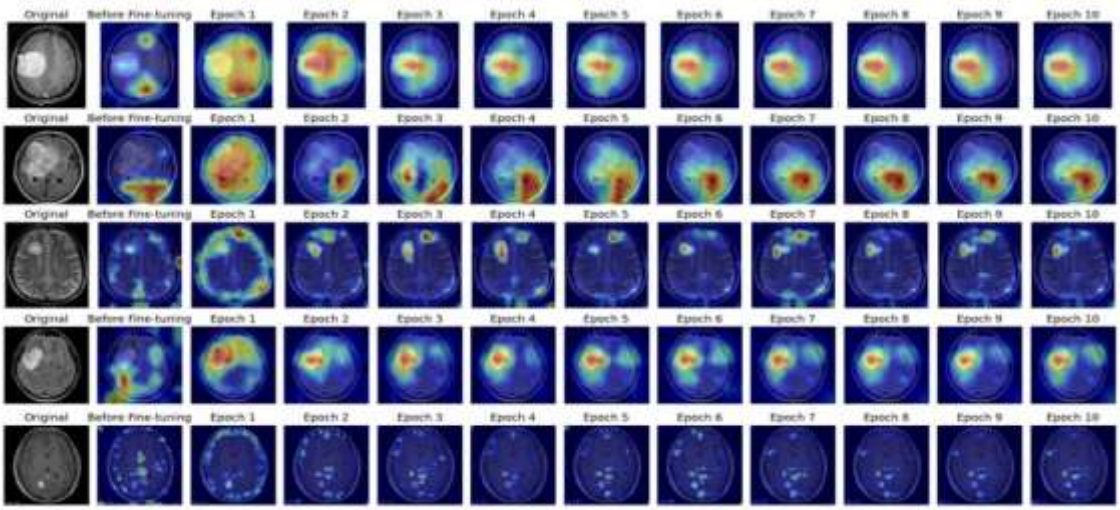
Multimodal fusion can be implemented in different ways. Some models use early fusion, where raw or low-level features from multiple modalities are combined at the input stage. Others use intermediate or deep fusion, where separate modality-specific branches first learn individual representations and are then merged at a later stage. Still others apply late fusion, where independent modality-specific predictions are combined at the decision level. More advanced approaches incorporate attention mechanisms to weight modalities adaptively, allowing the model to focus more strongly on the most informative sources for a given patient or classification task. These design choices are important because they influence how much complementary information the system can preserve and how effectively it can exploit cross-modal relationships.

The rise of multimodal fusion also reflects the broader movement of the field toward clinically realistic diagnostic systems. In real healthcare environments, clinicians do not make decisions based on one image alone. Instead, they often consider multiple imaging modalities alongside patient history, cognitive performance, laboratory findings, and biomarker evidence. Multimodal machine learning therefore mirrors clinical reasoning more closely than single-modality models. This makes such systems especially attractive for future clinical translation, as they can potentially support decision-making in a way that is more aligned with how Alzheimer's disease is actually evaluated in practice.

Despite these advantages, multimodal methods also introduce several challenges. Different modalities may vary in quality, dimensionality, scale, and availability. In many real-world datasets, not every patient has all modalities available, which creates missing-data problems that can reduce performance or complicate model design. Multimodal systems are also often more computationally complex than single-modality models, and their fusion strategies may be difficult to compare across studies. In addition, strong performance in multimodal settings may depend heavily on careful preprocessing, alignment, and balancing of information across sources. For these reasons, although multimodal fusion is one of the most promising directions in Alzheimer's disease diagnosis research, its success depends not only on combining more data, but on combining those data in a principled, robust, and clinically meaningful way.

Transformer and hybrid attention models

Transformer-based and transformer-equipped models represent a newer direction in Alzheimer's disease diagnosis research because they are designed to capture long-range dependencies and broader contextual relationships more effectively than conventional convolution alone. While CNNs are highly effective at learning local spatial patterns from MRI and PET images, they are inherently limited by the locality of convolutional kernels. In Alzheimer's disease, however, pathological changes are often distributed across multiple brain regions rather than confined to one small area. This makes global context modeling especially important. Attention mechanisms and transformer architectures address this limitation by allowing the model to weigh relationships among distant regions of the input, potentially capturing more comprehensive disease-related structural patterns (Zhao et al., 2024).



Examples of Epoch-Wise Evolution of Attention Heatmaps During Model Fine-Tuning. This figure shows the progression of model attention maps across multiple training epochs, beginning with the original MRI image and the initial attention map before fine-tuning, followed by updated heatmaps from Epoch 1 through Epoch 10. The sequence illustrates how the model's focus changes during training and how the highlighted regions gradually become more concentrated on diagnostically relevant image areas. Warmer colors indicate stronger model attention, while cooler colors represent lower contribution. This type of visualization is useful for understanding how fine-tuning affects model interpretability and whether the model increasingly attends to meaningful pathological regions over the course of training.

In practical terms, transformer-based methods aim to improve the representation of subtle neurodegenerative changes that may not be fully captured by purely local feature extraction. Hybrid CNN-transformer models are particularly attractive because they combine the strengths of both paradigms. The CNN component is usually responsible for extracting lower-level spatial and textural features from the imaging data, while the transformer component models broader contextual interactions among those learned features. This combined approach is especially relevant for Alzheimer's diagnosis because the disease affects interconnected brain structures, and clinically meaningful patterns may emerge only when relationships across regions are considered together rather than independently. As a result, these models are often viewed as a promising step beyond standard CNN pipelines.

Another important advantage of transformer and hybrid attention models is their flexibility in multimodal settings. Since Alzheimer's disease diagnosis increasingly relies on a combination of MRI, PET, cognitive variables, and other clinical information, attention mechanisms can help identify which features or modalities are more informative for a given prediction. This adaptive weighting process may improve multimodal fusion by allowing the network to focus more strongly on the most relevant information while reducing the influence of less informative inputs. In this sense, attention-based systems are not only tools for global feature modeling but also potential mechanisms for better modality integration.

Despite these advantages, transformer-based approaches also introduce significant challenges. They typically require larger training datasets, greater computational resources, and more careful optimization than standard CNNs. Medical imaging datasets are often much smaller than the large-scale benchmarks used in natural image recognition, which increases the risk of overfitting and unstable training. For this reason, many transformer-based Alzheimer's studies use hybrid designs, transfer learning strategies, or architectural constraints that preserve some of the inductive biases of CNNs. These design choices help make transformer methods more practical in limited-data settings, but they also mean that the superiority of transformers over strong CNN baselines is not yet universally established. Current evidence suggests that transformer-equipped models are promising and potentially powerful, but they should still be viewed as an emerging methodological direction rather than a settled replacement for CNN-based diagnosis systems (Zhao et al., 2024).

Graph neural networks

Graph neural networks have emerged as a particularly promising direction in Alzheimer's disease diagnosis because they are well suited to modeling the brain as a system of interconnected regions rather than as a purely regular image grid. Conventional CNN-based methods process MRI or PET data mainly through local convolutional operations, which are effective for spatial pattern recognition but may not fully capture the relational organization of the brain. In contrast, GNNs are designed for non-Euclidean data structures, making them especially appropriate when the problem involves relationships among brain regions, connections among subjects, or associations between imaging and phenotypic variables. This is important in Alzheimer's disease because neurodegeneration often involves distributed network-level changes rather than isolated abnormalities in a single location (Zhang et al., 2023; Hu et al., 2024).

In Alzheimer's disease research, graph representations can be constructed in several ways. Nodes may represent brain regions, subjects, or feature groups, while edges may represent anatomical connectivity, functional similarity, imaging-based relationships, or phenotypic associations. This flexibility allows

GNN-based models to integrate different forms of information into a unified framework. For example, multimodal GNN approaches can combine structural MRI, PET, and demographic or cognitive variables while explicitly modeling the relationships among them. Such relational modeling can be especially useful when disease-related patterns are not only spatially distributed but also interconnected across multiple biological and clinical dimensions. As a result, GNNs offer a more natural way of representing complex dependencies than models that rely entirely on grid-based image learning.

Another important strength of graph neural networks is their ability to support multimodal and explainable frameworks. In multimodal settings, GNNs can fuse heterogeneous information sources while preserving relational structure, which may help capture subtle interactions between neuroimaging findings and non-imaging variables. In explainability-focused settings, graph-based models can also provide insight into which nodes, edges, or relational patterns contribute most strongly to a prediction. This is especially relevant for clinical applications, where trust and interpretability are increasingly important. Recent work has therefore extended graph-based Alzheimer's diagnosis beyond prediction alone and toward more transparent modeling strategies, including self-explainable graph neural networks and graph attention mechanisms that highlight influential connections or regions (Zhang et al., 2023; Hu et al., 2024).

Despite these advantages, graph neural networks also introduce several methodological challenges. Their performance depends heavily on how the graph is constructed, including the definition of nodes, the choice of edges, and the criteria used to encode relationships. Different graph construction strategies may lead to different results, which raises concerns about reproducibility and comparability across studies. In addition, graph models can be computationally complex and may be sensitive to limited sample sizes, noisy connections, or inconsistent multimodal alignment. For these reasons, although GNNs are a highly promising direction for Alzheimer's disease diagnosis, they should still be regarded as an emerging and developing methodology rather than a universally established alternative to more mature CNN-based systems. Current evidence suggests that their greatest value lies in settings where relational structure, multimodal integration, and interpretability are central to the diagnostic task.

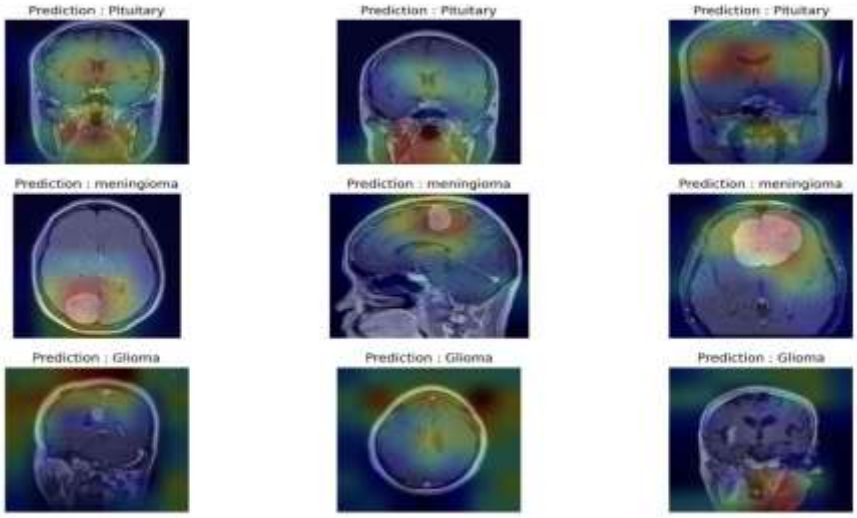
Explainable AI

Explainable artificial intelligence has become an increasingly important component of Alzheimer's disease diagnosis research because medical systems are expected not only to produce accurate predictions but also to provide reasoning that clinicians and researchers can evaluate. In high-stakes healthcare settings, a model that outputs a diagnosis without any interpretable basis is often difficult to trust, especially when the decision may influence patient monitoring, treatment planning, or further diagnostic testing. For this reason, explainability is now widely seen as a key requirement for clinically meaningful machine learning and deep learning systems rather than an optional add-on. In Alzheimer's disease research, this need is particularly strong because the disorder involves subtle, progressive, and regionally distributed brain changes that must be interpreted carefully in relation to known clinical and neurobiological patterns (Castellano et al., 2024; Hu et al., 2024).

In practice, explainability in Alzheimer's disease studies is commonly implemented through techniques such as attention visualization, Grad-CAM, SHAP, saliency maps, or model-intrinsic mechanisms that estimate the contribution of specific features, regions, nodes, or relationships. These methods are used to identify which image regions, modalities, or variables are most influential in a model's decision. For example, Grad-CAM and related visualization methods can highlight areas of an MRI or PET image that the network considers important for distinguishing Alzheimer's disease from healthy controls or mild cognitive impairment. Likewise, SHAP-based methods can estimate the relative importance of non-imaging variables such as clinical scores, demographics, or biomarkers. These tools are valuable because they offer a bridge between model predictions and clinically interpretable evidence, helping researchers assess whether the model is focusing on plausible disease-related patterns rather than irrelevant artifacts.

Explainability also plays an important role in comparing different model families. In multimodal and graph-based systems, interpretability methods can reveal which modalities, brain regions, or relational structures contribute most strongly to classification or risk prediction. This is especially relevant in

Alzheimer's disease because clinically meaningful evidence often depends on distributed patterns rather than a single dominant feature. As a result, explainability can help determine whether a model is capturing biologically reasonable structure, whether it is over-relying on confounding signals, and whether its reasoning aligns with established understanding of neurodegeneration. Recent work has therefore moved beyond simple post hoc visualization and has begun to explore more integrated approaches, such as self-explainable graph models and architecture-level attention mechanisms that make interpretability part of the model design itself (Castellano et al., 2024; Hu et al., 2024).



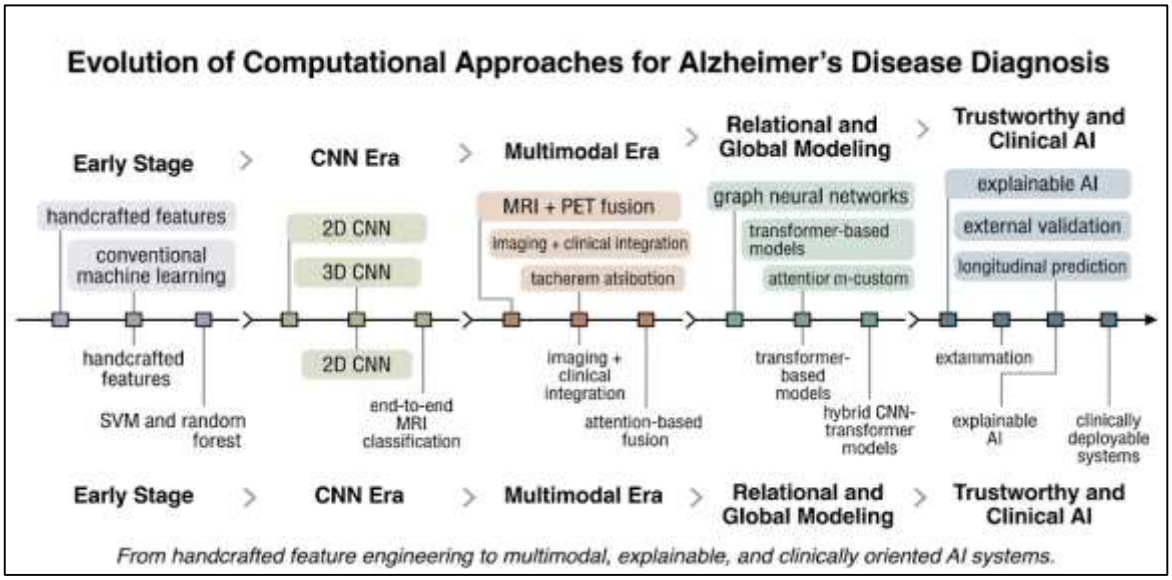
Examples of using Explainable AI: Grad-CAM-Based Visualization of Brain Tumor Classification Predictions. This figure presents Grad-CAM heatmap visualizations for multiple brain MRI images classified into tumor categories such as pituitary tumor, meningioma, and glioma. The colored overlays indicate the image regions that contributed most strongly to the model's prediction, with warmer colors representing areas of higher importance. These visual explanations help illustrate how the deep learning model focuses on tumor-relevant regions during classification and provide an interpretable view of the decision-making process. Such visualization is useful for assessing whether the model is attending to clinically meaningful structures rather than irrelevant background patterns.

Despite this progress, the rigor of explainability analyses still varies considerably across the literature. Many studies include only one or two example heatmaps or attention maps without systematically evaluating whether those explanations are stable, reproducible, clinically plausible, or consistent across datasets. In some cases, explainability is treated more as a presentation feature than as a rigorously assessed scientific component. This remains a major weakness because visual explanation alone does not guarantee trustworthy reasoning. For explainable AI to become truly useful in Alzheimer's disease diagnosis, future research must move toward more standardized evaluation of explanation quality, including stability analysis, expert validation, and cross-cohort consistency. Therefore, while explainability has become a major strength of modern Alzheimer's disease diagnosis research, it is still an evolving area that requires stronger methodological foundations before it can fully support routine clinical trust.

Beyond technical transparency, explainability also has an important practical role in supporting communication between artificial intelligence systems and clinical experts. In real diagnostic settings, clinicians are more likely to engage with a model when they can understand why it reached a particular conclusion and whether that reasoning is consistent with known disease mechanisms. This is especially important in Alzheimer's disease, where diagnostic interpretation often depends on subtle structural, metabolic, and cognitive patterns rather than a single obvious marker. Explainable systems can therefore help bridge the gap between computational performance and clinical usability by making model behavior easier to inspect, question, and validate. In this sense, the future value of explainable AI in Alzheimer's disease diagnosis is not limited to improving model interpretation alone, but also extends to increasing clinician confidence, facilitating human-AI collaboration, and supporting more

responsible translation of machine learning systems into real-world healthcare practice. At the same time, the usefulness of explainable AI will depend on whether explanation methods are evaluated with the same seriousness as predictive performance. Future studies should not only report accuracy, sensitivity, or area under the curve, but should also assess whether the explanations are stable across repeated runs, robust across datasets, and aligned with clinically meaningful brain regions or patient-level factors. This is particularly important in Alzheimer's disease because misleading or unstable explanations could create false confidence rather than genuine trust. Accordingly, stronger evaluation frameworks, closer collaboration with neurologists and radiologists, and more standardized reporting practices will be necessary to ensure that explainability contributes to both scientific rigor and clinical adoption. Under these conditions, explainable AI has the potential to become not merely a supplementary visualization tool, but a core component of trustworthy Alzheimer's disease diagnosis systems.

Figure 5 summarizes the methodological evolution of computational approaches for Alzheimer's disease diagnosis, highlighting the field's transition from handcrafted-feature machine learning to CNN-based, multimodal, graph-based, transformer-equipped, and clinically oriented explainable AI systems.



SURVEY OF REPRESENTATIVE STUDIES

Early CNN-based Alzheimer's diagnosis

A key early deep learning study was the work of Islam and Zhang (2018), who used an ensemble of deep convolutional neural networks for MRI-based Alzheimer's disease diagnosis. Their paper is important because it showed that CNN-based systems could classify Alzheimer's stages directly from MRI while reducing dependence on handcrafted feature extraction. This study helped establish deep convolutional diagnosis as a serious direction in Alzheimer's research and influenced many later MRI-based approaches (Islam & Zhang, 2018).

Another influential step was taken by Lu et al. (2018), who proposed a multimodal and multiscale deep neural network using structural MRI and FDG-PET. Their framework showed that Alzheimer's diagnosis benefits from combining structural and functional information and that multimodal modeling can improve performance, particularly in more clinically challenging progression-related settings such as identifying which mild cognitive impairment cases will convert to Alzheimer's disease. This paper is important because it bridges the transition from single-modality deep learning to more sophisticated multimodal fusion (Lu et al., 2018).

Multimodal attention and broader multimodal diagnosis

Golovanevsky et al. (2022) proposed a multimodal attention-based deep learning framework for Alzheimer's disease diagnosis. This study moved beyond simple feature concatenation and instead

used attention to model interactions among imaging, genetic, and clinical inputs. Cross-modal attention is meaningful in this setting because not all modalities contribute equally to every diagnostic case, and attention allows the network to weight information adaptively. The paper therefore represents a more mature stage of multimodal learning than earlier direct-fusion methods (Golovanevsky et al., 2022).

Qiu et al. (2022) extended the multimodal perspective further by introducing a deep learning framework for dementia assessment that could identify normal cognition, mild cognitive impairment, Alzheimer's disease, and non-Alzheimer dementias through successive diagnostic steps. This work moved beyond narrow binary classification and toward a broader clinical decision-support setting. It also included external validation across multiple cohorts and interpretability analysis, making it one of the strongest papers in terms of translational ambition (Qiu et al., 2022).

Multimodal 2D and 3D imaging frameworks

Castellano et al. (2024) proposed and evaluated classification models using 2D and 3D MRI together with amyloid PET in both uni-modal and multimodal settings. Their study is particularly useful for survey purposes because it directly compares representational choices and shows that volumetric models learn more effective representations than 2D models, while multimodal integration significantly improves performance over single-modality approaches. The study also used the OASIS-3 cohort, which is important because much of the literature remains heavily centered on ADNI (Castellano et al., 2024).

This paper supports two broader conclusions. First, full 3D anatomical context is usually more informative than isolated 2D slices when the target is neurodegeneration. Second, multimodal learning is not beneficial simply because more data are added, but because MRI and PET encode different aspects of disease biology. These conclusions recur across multiple studies and form a central theme of the modern literature (Castellano et al., 2024; Lu et al., 2018).

Graph neural networks for relational diagnosis

Zhang et al. (2023) proposed a multi-modal graph neural network for early diagnosis of Alzheimer's disease using sMRI and PET scans. The significance of this paper lies in its explicit use of brain networks and phenotypic information within a graph framework, with each modality handled by its own GNN branch and fused at multiple levels. The study argues that conventional CNN-based multimodal pipelines do not naturally incorporate both image-derived and phenotypic relational information, whereas GNNs are better suited for such non-Euclidean structures (Zhang et al., 2023).

Hu et al. (2024) introduced a self-explainable graph neural network for Alzheimer disease and related dementias risk prediction. Although this work uses claims data rather than image-only diagnosis, it remains highly relevant to the survey because it shows how graph learning and explainability can be combined within a model-intrinsic framework. Instead of relying only on post hoc saliency methods, the paper emphasizes relation importance as part of the prediction process itself, which is a meaningful development for trustworthy clinical AI (Hu et al., 2024).

Transformer-equipped and hybrid models

Transformer-based approaches are relatively newer in Alzheimer's disease diagnosis but are becoming increasingly visible. Zhao et al. (2024) proposed a vision transformer-equipped convolutional neural network for automated Alzheimer's disease diagnosis using 3D MRI scans. This work is important because it tries to combine local pattern extraction from CNNs with stronger global contextual modeling from transformer-inspired mechanisms. The motivation is clear: Alzheimer-related anatomical changes can be distributed across the brain, and attention-based mechanisms may capture long-range dependencies that conventional convolutions miss (Zhao et al., 2024).

Hybrid CNN-transformer models are promising, but they also raise practical questions. They often increase architectural complexity, computational cost, and data requirements. Since medical imaging datasets are usually much smaller than natural-image benchmarks, the success of transformer methods depends heavily on careful training design, augmentation, transfer learning, or inductive structure from CNN backbones. For this reason, transformer methods should currently be seen as promising rather than unquestionably superior replacements for CNNs in Alzheimer's diagnosis (Zhao et al., 2024; Alsubaie et al., 2024).

COMPARISON OF REPRESENTATIVE STUDIES

Figure 6. Comparative framework of major model families in Alzheimer's disease diagnosis. The figure contrasts traditional machine learning, CNN-based models, multimodal deep learning, transformer-based approaches, graph neural networks, and explainable AI-oriented systems in terms of input requirements, core strengths, principal limitations, interpretability, and clinical readiness.

Comparative Framework of Major Model Families in Alzheimer's Disease Diagnosis

	Typical Input	Key Strength	Main Limitation	Interpretability	Clinical Readiness
Traditional Machine Learning	handcrafted features	simple and efficient	limited representation learning	moderate	moderate
CNN-Based Models	MRI or PET	strong spatial learning	limited global context	low to moderate	moderate
Multimodal Deep Learning	MRI + PET + clinical	complementary information fusion	missing modality challenge	moderate	high potential
Transformer / Hybrid Attention Models	imaging volumes	global context modeling	data-hungry and complex	moderate	emerging
Graph Neural Networks	brain or subject graphs	relational modeling	graph design sensitivity	moderate to high	emerging
Explainable AI-Oriented Models	multimodal or image-based	improved transparency	explanation quality not standardized	high	high potential

Figure 6 presents a high-level comparative view of the major methodological families used in Alzheimer's disease diagnosis and highlights how their capabilities differ across important practical dimensions. Traditional machine learning methods are generally based on handcrafted features and remain attractive for their simplicity and efficiency, but they are limited in representation learning capacity. CNN-based models improve substantially on this by learning spatial patterns directly from MRI or PET data, although they may still struggle to capture broader global context. Multimodal deep learning approaches extend this capability by combining imaging and non-imaging sources, such as MRI, PET, and clinical variables, allowing them to exploit complementary disease information, but they are often challenged by missing modalities and increased system complexity. Transformer and hybrid attention models further strengthen global context modeling, while graph neural networks are especially useful for capturing relational structures among brain regions or subjects. Explainable AI-oriented models focus on improving transparency and trust, which are essential for clinical adoption, even though explanation quality is not yet fully standardized. Overall, the figure shows that the field is moving from simpler but limited models toward more expressive, multimodal, and clinically promising approaches, with a growing emphasis on interpretability and deployment readiness.

Table 1 highlights the progression of the field from image-only CNN models toward multimodal, graph-based, and explainable frameworks. It also shows that the strongest recent contributions are not only those with high reported accuracy, but those that address more realistic clinical settings, incorporate multiple data sources, or attempt more meaningful interpretation (Alsubaie et al., 2024; Qiu et al., 2022; Zhang et al., 2023).

Table 1. Comparison of representative studies in Alzheimer's disease diagnosis.

Study	Year	Modality / Data	Core Method	Main Contribution	Main Limitation
Islam & Zhang	2018	MRI	Ensemble CNN	Early influential MRI-based end-to-end CNN diagnosis	Single-modality and earlier-generation benchmark setting
Lu et al.	2018	MRI + FDG-PET	Multimodal multiscale DNN	Showed the value of multimodal fusion and progression-related prediction	Strong dependence on curated benchmark data
Golovanevsky et al.	2022	Imaging + genetic + clinical	Attention-based multimodal DL	Adaptive cross-modal weighting beyond simple concatenation	Architecturally complex and less deployment-oriented
Qiu et al.	2022	Imaging + non-imaging multimodal data	Successive-step multimodal DL	Broad dementia assessment with external validation and interpretation	High complexity and demanding multimodal pipeline
Zhang et al.	2023	sMRI + PET + phenotypic data	Multi-modal GNN	Integrates relational and phenotypic information in graph form	Graph construction choices can affect reproducibility
Castellano et al.	2024	2D/3D MRI + amyloid PET	Uni-modal and multimodal CNNs	Clear comparison of 2D vs 3D and single vs multimodal settings	Mainly centered on one cohort
Hu et al.	2024	Claims / relational health data	Self-explainable GNN	Built-in interpretability through relation importance	Not a pure neuroimaging diagnosis system
Zhao et al.	2024	3D MRI	Transformer-equipped CNN	Hybrid local-global representation learning	Transformer benefit still requires broader validation

CRITICAL DISCUSSION

Why multimodal methods matter

One of the clearest findings across the Alzheimer's disease diagnosis literature is that multimodal methods often outperform single-modality methods. This pattern is consistent with the biological complexity of the disease itself. Alzheimer's disease is not expressed through a single type of abnormality, and no single modality can capture all relevant aspects of its progression. Structural MRI and PET, for example, reflect different but complementary dimensions of the disease process. MRI is especially valuable for identifying anatomical degeneration such as hippocampal atrophy, cortical thinning, ventricular enlargement, and gray matter loss, whereas PET contributes functional or molecular information, including altered glucose metabolism and amyloid-related burden. When these modalities are combined in a principled way, the resulting representation can be substantially more informative than what either modality can provide independently (Lu et al., 2018; Castellano et al., 2024).

The advantage of multimodal learning lies not only in adding more data, but in integrating heterogeneous sources of evidence that describe different dimensions of neurodegeneration. A model that uses both structural and metabolic imaging can potentially detect disease signatures that are weak

or ambiguous when observed from only one perspective. This becomes especially important in clinically challenging settings, such as differentiating mild cognitive impairment from early Alzheimer's disease, predicting disease progression, or identifying subtle abnormalities before extensive structural damage becomes visible. In such cases, a single modality may be insufficient to capture the full disease state, whereas multimodal integration can strengthen diagnostic sensitivity and improve the robustness of prediction.

Multimodal methods also align more closely with real clinical decision-making. In practice, neurologists and other specialists rarely make a diagnosis based on only one image or one biomarker. Instead, they consider structural imaging, functional imaging, cognitive testing, demographic factors, laboratory evidence, and clinical history together. Multimodal machine learning frameworks attempt to mirror this integrative reasoning process by combining imaging and non-imaging variables into a unified computational model. This makes such methods not only technically appealing but also more clinically relevant, because they move closer to the way Alzheimer's disease is assessed in real healthcare settings.

Another important strength of multimodal systems is their ability to support richer feature interaction. In more advanced architectures, fusion is not limited to simple concatenation of input vectors. Instead, the model may learn cross-modal relationships through attention mechanisms, hierarchical feature merging, shared latent representations, or graph-based relational integration. These approaches allow the system to identify how information from one modality complements or reinforces patterns from another. For example, structural atrophy observed in MRI may become more diagnostically meaningful when considered together with reduced glucose metabolism in PET or abnormal cognitive test scores. In this sense, multimodal learning provides not merely additional variables, but a more complete and structured representation of disease-related evidence.

At the same time, multimodal learning is not automatically superior in every setting. Its benefits depend on several important conditions, including data quality, subject-level alignment across modalities, preprocessing consistency, feature compatibility, and the extent of missing data. If modalities are poorly aligned, heavily imbalanced, or inconsistently acquired, the model may not benefit from fusion and may even perform worse than a well-designed single-modality baseline. Missing modality problems are especially important in clinical datasets, where not every patient has MRI, PET, biomarkers, and cognitive assessments available at the same time. These limitations mean that the effectiveness of multimodal learning depends not only on having more information, but on having that information organized, harmonized, and fused in a robust manner.

Figure 7. Conceptual illustration of explainable AI in Alzheimer's disease diagnosis

An illustration showing how MRIs with Grad-CAM highlight brain tumor images

The literature also suggests that not all fusion strategies are equally effective. Early multimodal studies often relied on straightforward concatenation of modality-specific features, but more recent work increasingly favors learned fusion mechanisms such as attention-based integration, intermediate fusion, and graph-based relational fusion. These strategies are generally better suited to preserving complementary information and modeling nonlinear cross-modal dependencies. Even so, direct comparison across papers remains difficult because studies vary widely in task definition, dataset composition, diagnostic categories, and evaluation protocols (Golovanevsky et al., 2022; Qiu et al., 2022; Zhang et al., 2023). As a result, although the overall trend strongly supports multimodal methods, the exact degree of their advantage remains dependent on study design and implementation details.

Overall, the importance of multimodal learning in Alzheimer's disease diagnosis lies in its ability to better reflect the multidimensional nature of the disease. By integrating structural, metabolic, cognitive, and clinical information, multimodal systems offer a more comprehensive basis for prediction than models built on a single source alone. This does not eliminate the methodological challenges of fusion, missing data, or evaluation heterogeneity, but it explains why multimodal approaches have become a central focus of current research and why they are widely viewed as one of the most clinically promising directions in the field.

Why accuracy alone is not enough

A major problem in Alzheimer's disease artificial intelligence research is the overemphasis on headline accuracy as the primary indicator of model quality. Although accuracy, sensitivity, specificity, and area under the curve are useful performance metrics, they do not by themselves capture whether a model is clinically meaningful, methodologically robust, or generalizable beyond a single experimental setting. This is especially important in Alzheimer's disease diagnosis because the literature is highly heterogeneous. Different studies classify different target groups, use different train-test splits or cross-validation structures, apply different preprocessing pipelines, and may or may not include external validation. Some models are evaluated only on relatively simple binary tasks such as Alzheimer's disease versus healthy control, while others attempt more challenging and clinically valuable tasks such as progression prediction, multi-class staging, or differential diagnosis across multiple dementia types. Under these conditions, direct comparison based solely on reported accuracy can be misleading. A model that achieves very high accuracy on Alzheimer's disease versus healthy control classification may still be less clinically useful than a model with lower numerical performance that addresses a more difficult but more realistic diagnostic problem. For example, separating mild cognitive impairment from early Alzheimer's disease, predicting which patients are likely to convert from mild cognitive impairment to Alzheimer's disease, or distinguishing Alzheimer's disease from non-Alzheimer dementias are tasks of much greater practical relevance, yet they are also substantially harder. Consequently, lower performance in these settings does not necessarily reflect a weaker methodology. Rather, it may indicate that the model is being evaluated under more realistic clinical conditions. This is why survey papers should be careful not to interpret all reported numbers as directly comparable or to imply that the model with the highest published accuracy is automatically the strongest contribution (Qiu et al., 2022; Alsubaie et al., 2024).

Another reason accuracy alone is insufficient is that performance metrics may be inflated by dataset-specific characteristics. A highly curated benchmark with well-separated diagnostic groups, balanced classes, and standardized preprocessing may favor strong performance even if the resulting model is not robust to real-world variation. In contrast, models tested across multiple cohorts, different scanners, more heterogeneous patient populations, or incomplete multimodal settings often face a more difficult but more clinically meaningful evaluation environment. External validation, demographic diversity, robustness to missing data, and resistance to domain shift are therefore just as important as raw predictive performance when judging the true value of a diagnostic model.

This issue is closely related to the broader distinction between benchmark optimization and clinical readiness. A model optimized narrowly for one dataset or one simplified task may look impressive in numerical terms but remain unsuitable for actual healthcare deployment. In Alzheimer's disease diagnosis, clinically useful systems must not only classify accurately but also perform consistently across different populations, support interpretation, handle incomplete or variable input data, and maintain reliability under practical constraints. Models that sacrifice some headline performance in

exchange for greater robustness, transparency, or generalizability may therefore offer more real-world value than models tuned solely to maximize benchmark scores.

More broadly, recent medical machine learning research has also emphasized the importance of balanced, generalized, and robust predictive frameworks for healthcare applications, reinforcing the need for models that go beyond narrow benchmark optimization (Abubakkar et al., 2025). This broader perspective is highly relevant to Alzheimer's disease research, where the ultimate goal is not simply to win a performance comparison on a curated dataset, but to develop diagnostic systems that are dependable, interpretable, and clinically transferable. For this reason, future evaluation standards in Alzheimer's disease AI should place greater emphasis on external validation, task difficulty, reproducibility, robustness, and clinical applicability alongside conventional performance metrics.

Reproducibility and generalizability

Generalizability remains one of the most significant weaknesses in Alzheimer's disease diagnosis research based on machine learning and deep learning. A large proportion of published studies are developed and evaluated primarily on ADNI, which has undoubtedly played a critical role in advancing the field by providing a standardized and widely accessible benchmark. However, this strong dependence on a single dataset also introduces an important limitation. Models that perform well on one curated benchmark may not necessarily maintain the same level of performance when applied to data from different scanners, acquisition protocols, clinical populations, or healthcare environments. In other words, success on ADNI does not automatically imply readiness for broader clinical deployment.

This problem arises because benchmark datasets often differ from real-world clinical settings in several important ways. Curated research datasets may have cleaner labels, better image quality, more consistent preprocessing, and more controlled cohort composition than data encountered in routine care. Real-world clinical populations are usually more heterogeneous, with greater variation in demographics, comorbidities, disease severity, scanner vendors, imaging protocols, and completeness of multimodal information. As a result, a model trained under narrow benchmark conditions may learn dataset-specific patterns that do not generalize well beyond the original source. This creates a risk of benchmark dependence, where apparent model strength reflects familiarity with a particular dataset rather than true robustness to the broader clinical population.

Reproducibility is closely linked to this issue. Many studies differ substantially in preprocessing choices, train-test splits, cross-validation strategies, inclusion criteria, and reporting practices. Some studies provide detailed methodological descriptions and external testing, while others report only internal validation on a single cohort with limited transparency regarding preprocessing or data partitioning. This variability makes it difficult to reproduce published findings exactly or to determine whether performance differences across studies reflect true methodological improvements or simply differences in experimental setup. In addition, incomplete code release, inconsistent hyperparameter reporting, and limited access to preprocessing details further weaken reproducibility across the literature.

Another important challenge is cross-site and cross-scanner variability. Neuroimaging data acquired at different institutions often vary in resolution, contrast, scanner hardware, field strength, and acquisition protocol. These factors can introduce distribution shifts that degrade model performance when a system trained in one environment is applied to another. A model that appears highly accurate under internal validation may therefore fail to generalize when deployed in a different clinical center. This is especially concerning in Alzheimer's disease diagnosis, where subtle imaging differences can strongly influence automated predictions. Robustness to such variability should therefore be considered a central evaluation criterion rather than a secondary concern.

Qiu et al. (2022) stand out in this regard because they incorporated external validation across multiple cohorts, which provides stronger evidence that the model captures meaningful disease-related information rather than only benchmark-specific structure. Castellano et al. (2024) also broadened the dataset landscape by incorporating OASIS-3, which helps reduce overreliance on ADNI alone. Even so, routine cross-center validation remains uncommon in the field, and many published methods are still evaluated in relatively narrow experimental settings. This means that despite strong progress in model development, evidence for consistent real-world transfer remains limited.

The issue of generalizability also extends beyond imaging source alone. It includes demographic diversity, disease-stage variation, missing modality patterns, and differences in clinical workflow. A model trained primarily on one demographic distribution or one style of multimodal completeness may not perform equally well in other populations. This is particularly important in healthcare AI, where fairness, robustness, and external validity are essential for responsible use. Therefore, future Alzheimer's disease diagnosis research should move toward multi-center evaluation, stronger external validation, more transparent reporting, and reproducible experimental pipelines. Without these improvements, even highly accurate models may remain scientifically interesting but clinically fragile.

Explainability and trust

Explainability is increasingly presented as a major strength in Alzheimer's disease diagnosis research, but in practice its implementation often remains limited and uneven across the literature. Many studies now include one or two saliency maps, Grad-CAM overlays, SHAP plots, or attention visualizations as evidence that the model is interpretable. While these additions are helpful at a basic level, they do not by themselves establish that a system is truly explainable in a clinically meaningful sense. In many cases, explanation outputs are presented as illustrative figures rather than systematically evaluated components of the model. As a result, explainability is often treated more as a persuasive presentation feature than as a rigorously validated property of the diagnostic system itself (Castellano et al., 2024; Hu et al., 2024).

This limitation is especially important in Alzheimer's disease because the disorder involves subtle, progressive, and anatomically distributed changes that are not always visually obvious. A model may highlight a region of interest in an MRI or PET image, but without proper validation it remains unclear whether that highlighted region corresponds to biologically meaningful pathology, to dataset-specific artifacts, or simply to unstable internal model behavior. In other words, a visually plausible explanation is not necessarily a trustworthy explanation. True explainability requires more than the ability to generate heatmaps. It requires evidence that the explanatory signal is stable across repeated runs, consistent across cohorts, aligned with accepted clinical or neurobiological knowledge, and informative enough to support human judgment.

Trust in clinical AI depends heavily on this distinction. In healthcare, clinicians are unlikely to rely on a system simply because it achieves strong numerical performance. They also need to understand whether the model's reasoning is compatible with known disease mechanisms and whether its decisions can be challenged, inspected, and interpreted in a meaningful way. In Alzheimer's disease diagnosis, this issue is particularly sensitive because the disease often overlaps with normal aging, mild cognitive impairment, and other dementias. A model that produces a confident prediction without interpretable justification may therefore raise more concern than confidence. Explainability becomes important not only for transparency, but also for accountability, error analysis, and responsible clinical adoption.

Another difficulty is that many explainability methods are inherently post hoc. Techniques such as Grad-CAM, saliency maps, and SHAP are often applied after model training to visualize which inputs may have influenced a prediction. Although useful, these methods do not guarantee that the model itself has learned a clinically sound decision process. Post hoc explanations may sometimes appear convincing even when the underlying model is relying on spurious correlations, scanner-specific artifacts, preprocessing biases, or confounding signals. This creates an important methodological challenge: an explanation method can improve the appearance of transparency without necessarily improving the actual trustworthiness of the system. For this reason, the field increasingly recognizes that explanation quality must be evaluated independently rather than assumed from visualization alone.

From a deeper methodological perspective, explainability in Alzheimer's disease diagnosis can be considered at several levels. At the feature level, one may ask which anatomical regions, biomarkers, or cognitive variables contribute to a prediction. At the model level, one may ask how different modalities or branches interact within a multimodal architecture. At the decision level, one may ask whether the reasoning behind a classification remains consistent when the input is perturbed, when different cohorts are used, or when competing explanations are tested. A strong explainability framework should ideally address all of these levels. However, most current studies focus only on the

first level, usually by showing a few representative visualizations. Much less attention is paid to explanation robustness, cross-dataset stability, or clinician-oriented validation.

The move toward self-explainable graph models is therefore particularly notable because it suggests a shift from post hoc interpretation toward architecture-level transparency (Castellano et al., 2024; Hu et al., 2024). In such approaches, interpretability is not simply added after prediction, but is built into the structure of the model itself, for example through relation importance, node weighting, or graph attention mechanisms. This is a meaningful methodological advance because it tries to make the reasoning process more directly inspectable. In the context of Alzheimer's disease, where distributed brain-region interactions and multimodal dependencies are highly relevant, this type of built-in explainability may offer a more faithful representation of how the model reaches a conclusion. It also better supports the broader goal of trustworthy AI, which requires transparency to be structurally connected to prediction rather than visually appended afterward.

Even so, architecture-level transparency does not solve the entire trust problem. A self-explainable model may still produce explanations that are unstable, overly simplified, or difficult to align with expert clinical reasoning. Trust also depends on whether the explanations are understandable to domain experts, whether they improve human decision-making, and whether they remain meaningful across patient populations and acquisition settings. Therefore, the field must move beyond the question of whether a model can generate an explanation and instead ask whether the explanation is useful, reliable, and clinically valid. This requires evaluation frameworks that include not only model developers, but also neurologists, radiologists, and other domain experts who can assess whether the highlighted patterns correspond to accepted medical understanding.

A deeper issue is that explainability and trust are related but not identical concepts. A model may be explainable in the sense that it provides interpretable outputs, yet still not be trustworthy if its predictions are biased, unstable, or poorly generalized. Conversely, a model may be statistically robust yet still face adoption barriers if clinicians cannot understand how it reaches its decisions. In this sense, trust emerges not from explainability alone, but from the interaction of explainability with robustness, reproducibility, fairness, and external validation. For Alzheimer's disease diagnosis systems to become genuinely trusted, they must therefore provide explanations that are not only visually appealing, but also empirically grounded and integrated into a broader framework of clinical reliability.

Ultimately, the challenge for the field is to transform explainability from a mostly descriptive add-on into a rigorously evaluated scientific component of diagnostic modeling. Future studies should not only present explanation examples, but also measure explanation stability, test explanation consistency across cohorts, compare explanation outputs with known Alzheimer-related brain regions or biomarkers, and involve expert review in the assessment process. Under such standards, explainability could move beyond simple visualization and become a key mechanism for building trust, improving clinical acceptance, and supporting more responsible translation of machine learning systems into real-world Alzheimer's disease diagnosis.

Strengths and weaknesses of transformers and GNNs

Transformers and graph neural networks have attracted growing interest in Alzheimer's disease diagnosis because they attempt to address important limitations of conventional CNN-based systems. Standard CNNs are highly effective for local spatial feature extraction, but their receptive field grows only gradually and their inductive bias is centered on local neighborhood structure. In Alzheimer's disease, however, diagnostically meaningful patterns are often distributed across multiple brain regions and may involve long-range anatomical, functional, or multimodal relationships. This makes the problem inherently more complex than a purely local image recognition task. Transformers and GNNs are appealing precisely because they are designed to capture forms of structure that CNNs may not model as effectively. Transformers can represent long-range contextual dependencies through self-attention, while GNNs can encode structured relations among regions, subjects, or modalities within non-Euclidean data representations. These characteristics make both model families theoretically well suited to the distributed and relational nature of Alzheimer's disease pathology (Zhang et al., 2023; Zhao et al., 2024).

One of the main strengths of transformer-based approaches is their ability to model global context more directly than conventional convolution. In MRI-based Alzheimer's disease diagnosis, subtle changes in

one region may be more meaningful when interpreted together with alterations in distant parts of the brain. Self-attention mechanisms allow the model to weigh interactions across the entire input rather than relying solely on local kernel operations. This is especially useful when neurodegeneration affects multiple interconnected structures and when diagnostic evidence is spatially distributed rather than concentrated in a single lesion-like region. Hybrid CNN-transformer models are particularly promising because they combine the local feature extraction strength of CNNs with the global contextual modeling capacity of attention-based architectures. This allows them to retain useful inductive biases from convolution while extending the model's ability to capture long-range dependencies.

Graph neural networks offer a different but equally important advantage. Rather than treating the brain as only a regular image grid, GNNs can model it as a relational structure composed of nodes and edges. Nodes may represent brain regions, subjects, or multimodal feature groups, while edges may encode anatomical similarity, functional association, phenotypic relationships, or learned connectivity patterns. This flexibility makes GNNs particularly valuable for Alzheimer's disease diagnosis because the disorder often involves network-level degeneration and complex interactions across regions and patient-level variables. GNN-based systems can therefore provide a richer representation of disease structure than methods that process each modality independently and then concatenate the results. Their capacity to integrate relational and multimodal information also makes them attractive for explainable and clinically oriented modeling.

Despite these conceptual advantages, both transformers and GNNs introduce important methodological and practical challenges. Transformer-based models are typically more data hungry than CNNs because self-attention mechanisms involve a larger parameterization and weaker built-in assumptions about local image structure. In natural image domains, these models often succeed because they can be trained on extremely large datasets, but medical imaging datasets for Alzheimer's disease are usually much smaller. This creates a risk of overfitting, unstable optimization, and sensitivity to augmentation or initialization choices. Hybrid models partially address this problem by preserving some CNN inductive biases, but they do not eliminate it entirely. Consequently, the apparent strength of transformers in some studies may reflect careful architectural design or training strategy rather than a universally superior modeling principle.

Graph neural networks face a different set of limitations. Their performance depends heavily on graph construction choices, including how nodes are defined, how edges are generated, and how relational weights are assigned. These decisions are often nontrivial and can vary substantially across studies. A graph built from anatomical parcellation, for example, may emphasize different patterns than one constructed from similarity across subjects or multimodal feature affinity. This variability can affect both predictive performance and interpretability, and it complicates reproducibility because different graph construction strategies may lead to different conclusions. In addition, graph models can be computationally expensive and may become sensitive to noise, sparse connectivity, or inconsistent multimodal alignment. Thus, while GNNs are powerful in theory, their practical reliability depends strongly on how the graph structure is engineered or learned.

Another important issue is that the superiority of these models is not yet consistently established across the literature. Although some studies report strong results for transformer-equipped models or graph-based architectures, the evaluation settings often differ in datasets, preprocessing, target tasks, and degree of multimodal support. This makes it difficult to determine whether the observed performance gains reflect a truly better modeling paradigm or simply study-specific design choices. In many cases, strong multimodal CNN baselines remain highly competitive, especially when carefully tuned and evaluated under realistic conditions. Therefore, while transformers and GNNs represent important advances, current evidence is not yet sufficient to conclude that they are uniformly better than established CNN-based multimodal systems across all Alzheimer's disease diagnosis tasks.

These newer model families should therefore be viewed as promising but still developing directions. Their greatest value may lie not merely in replacing CNNs, but in expanding the range of structures that can be modeled. Transformers extend the field toward broader context modeling, while GNNs extend it toward relational and non-Euclidean representation learning. In the long term, their impact may be strongest in hybrid systems that combine local spatial learning, multimodal fusion, relational reasoning, and explainability within a unified diagnostic framework. For now, however, more studies

with stronger external validation, fair baseline comparison, transparent reporting, and clinically realistic evaluation are needed before transformers and GNNs can be considered consistently superior to strong multimodal CNN baselines (Zhang et al., 2023; Zhao et al., 2024).

OPEN CHALLENGES

- Dataset dependence remains strong. Public datasets are useful, but overreliance on a few benchmarks reduces confidence in real-world generalization (Alsubaie et al., 2024).
- Class imbalance and task inconsistency remain widespread. The literature mixes easier and harder tasks in ways that make direct comparison unreliable, especially when studies do not report the same diagnostic categories or evaluation protocols (Qiu et al., 2022).
- Clinical translation remains limited. High performance in experimental settings does not guarantee usefulness in routine care unless the model can handle missing modalities, scanner variation, uncertainty, and heterogeneous patient populations (Castellano et al., 2024; Qiu et al., 2022).
- Interpretability still lacks standardized evaluation. Many studies claim explainability, but few define how explanation quality should be measured or validated in collaboration with domain experts (Hu et al., 2024).

Several open challenges continue to limit the progress and clinical maturity of machine learning and deep learning for Alzheimer's disease diagnosis. One major issue is the continued dependence on a small number of public benchmark datasets. While resources such as ADNI, OASIS, and related cohorts have been essential for advancing the field, heavy reliance on these datasets reduces confidence in whether developed models will generalize effectively to real-world clinical environments. A second challenge is the widespread presence of class imbalance and task inconsistency across the literature. Some studies focus on relatively straightforward binary classification tasks, while others address more difficult and clinically meaningful problems such as progression prediction, multi-class staging, or differential diagnosis. Because diagnostic categories, evaluation protocols, and class distributions vary substantially from one study to another, direct comparison of reported performance often becomes unreliable. Clinical translation remains another major limitation. Strong results obtained under controlled experimental conditions do not automatically imply that a model will be useful in routine practice, especially if it cannot handle missing modalities, scanner variability, uncertainty, heterogeneous patient populations, and operational constraints of healthcare settings. In addition, interpretability remains insufficiently standardized. Although many studies claim some form of explainability, few provide a rigorous framework for measuring explanation quality, assessing explanation stability, or validating interpretability findings in collaboration with neurologists, radiologists, or other domain experts. Taken together, these challenges show that the field has made substantial methodological progress, but still faces important barriers before such systems can be regarded as dependable, generalizable, and clinically trustworthy (Alsubaie et al., 2024; Castellano et al., 2024; Hu et al., 2024; Qiu et al., 2022).

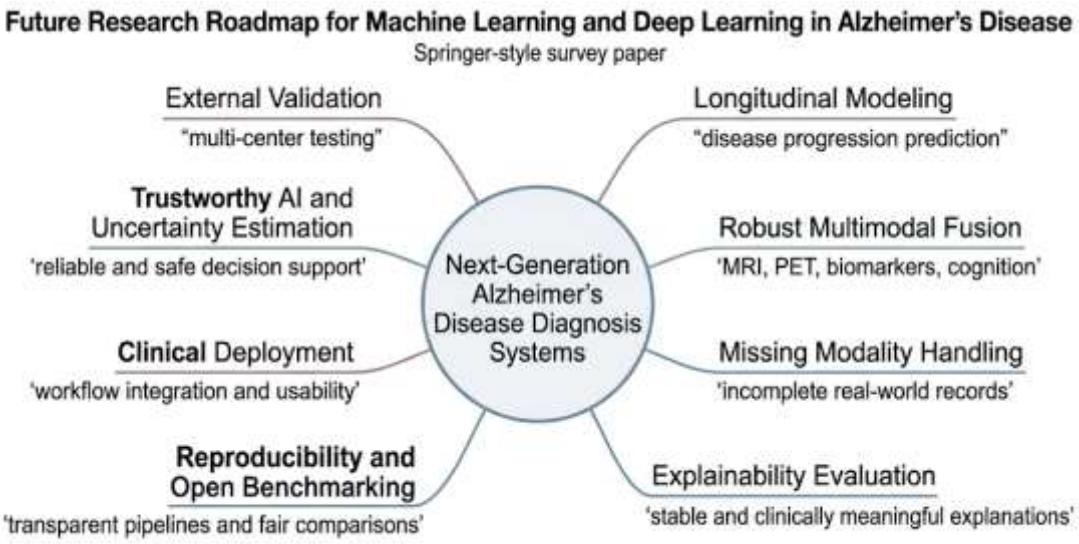


Figure 8 summarizes the major future research directions for machine learning and deep learning in Alzheimer's disease diagnosis, emphasizing the need for stronger validation, improved multimodal robustness, interpretability, reproducibility, and clinical deployment readiness.

FUTURE DIRECTIONS

- External validation across institutions and scanners should become routine, because without it even sophisticated models remain difficult to trust for practical deployment (Qiu et al., 2022).
- Multimodal integration should expand beyond MRI and PET to cognitive scores, blood biomarkers, genomics, longitudinal histories, and clinical text, with principled handling of missing data (Golovanevsky et al., 2022; Qiu et al., 2022).
- Longitudinal and prognostic modeling deserve greater emphasis because Alzheimer's disease is progressive and static single-time-point diagnosis captures only part of the clinical problem (Lu et al., 2018).
- Explainability must become more rigorous, with evaluation of stability, anatomical plausibility, and usefulness to clinicians. Architecture-level explainability may be especially promising (Hu et al., 2024).
- Reproducibility and open benchmarking require stronger reporting standards, code availability, preprocessing transparency, and fair comparison frameworks (Alsubaie et al., 2024).

Several future directions are especially important if machine learning and deep learning systems for Alzheimer's disease diagnosis are to move closer to real clinical usefulness. First, external validation across institutions, scanners, and patient populations should become a routine expectation rather than an occasional addition, because without such validation even highly sophisticated models remain difficult to trust for deployment in real healthcare settings. Models that are only tested on one benchmark dataset may perform well under controlled conditions but still fail when confronted with scanner variation, demographic differences, or site-specific workflow patterns. Second, multimodal integration should continue to expand beyond MRI and PET alone. Future systems are likely to be more clinically informative if they can also incorporate cognitive scores, blood-based biomarkers, genomics, longitudinal clinical histories, and even clinical text, while handling missing data in a principled and robust manner. This is especially important because real-world patient records are often incomplete, and practical AI systems must be able to function under such imperfect conditions (Golovanevsky et al., 2022; Qiu et al., 2022).

A third major direction is the need for stronger longitudinal and prognostic modeling. Alzheimer's disease is progressive by nature, and static single-time-point diagnosis captures only one part of the clinical problem. Future models should increasingly focus on questions such as disease progression,

risk of conversion from mild cognitive impairment to Alzheimer's disease, and temporal monitoring of neurodegenerative change. Such modeling would align more closely with the needs of clinicians, who are often concerned not only with current classification but also with how the disease is likely to evolve over time (Lu et al., 2018). Fourth, explainability must become more rigorous and scientifically grounded. Rather than relying only on isolated heatmaps or attention plots, future studies should evaluate whether explanations are stable, anatomically plausible, reproducible across cohorts, and genuinely useful to clinicians. In this regard, architecture-level explainability may be particularly promising because it embeds interpretability into the model design rather than treating it as a purely post hoc add-on (Hu et al., 2024).

Finally, reproducibility and open benchmarking should receive much stronger emphasis across the field. This includes better reporting standards, greater availability of code and preprocessing details, clearer documentation of hyperparameters and evaluation protocols, and fairer comparison frameworks across datasets and tasks. Without these improvements, it remains difficult to determine whether progress in the literature reflects genuine methodological advancement or simply differences in experimental setup. Taken together, these future directions point toward a broader shift in the field: from isolated benchmark-driven studies toward clinically robust, multimodal, transparent, reproducible, and externally validated systems that are better suited for real-world Alzheimer's disease diagnosis and monitoring (Alsubaie et al., 2024; Qiu et al., 2022).

CONCLUSION

The literature on machine learning and deep learning for Alzheimer's disease diagnosis has progressed rapidly over the past several years, moving from traditional feature-based classifiers toward more advanced CNNs, multimodal fusion systems, graph neural networks, transformer-equipped architectures, and explainable AI models. This progression reflects a broader methodological shift in the field, from manually designed pipelines that depend heavily on handcrafted features and expert preprocessing toward data-driven systems capable of learning increasingly rich, hierarchical, and multimodal representations directly from neuroimaging and associated clinical data. As a result, Alzheimer's disease diagnosis research has become not only more computationally sophisticated, but also more closely aligned with the multidimensional nature of the disease itself (Ghosh et al., 2024).

One of the clearest conclusions from the surveyed literature is that multimodal learning has emerged as a particularly valuable direction. MRI and PET provide complementary structural, metabolic, and pathological information, and their integration often produces stronger and more clinically meaningful predictive performance than single-modality systems alone. In addition, the incorporation of non-imaging variables such as cognitive scores, phenotypic information, biomarkers, and demographic features further strengthens the ability of models to represent disease heterogeneity. This trend suggests that the future of Alzheimer's disease diagnosis is likely to depend less on isolated modality-specific systems and more on integrated frameworks capable of combining multiple forms of evidence in a principled and robust manner (Reja Sweet et al., 2024).

The survey also shows that newer architectural directions, particularly graph neural networks and transformer-equipped models, are expanding the representational capacity of diagnostic systems beyond what conventional CNNs can easily capture. Transformers offer improved modeling of long-range contextual dependencies, while graph neural networks provide a natural framework for representing relational structures among brain regions, subjects, or multimodal features. These developments are especially relevant in Alzheimer's disease because the disorder is not confined to one local image region, but instead involves distributed and interconnected changes across structural, functional, and clinical dimensions. At the same time, the surveyed studies suggest that these newer architectures should still be viewed with careful balance. Although they are promising, their superiority over strong multimodal CNN baselines has not yet been consistently established across all datasets, tasks, and evaluation settings (Rakin et al., 2024).

Another major theme that emerges from the literature is the growing importance of explainability. As diagnostic models become more complex, transparency and interpretability become increasingly important for trust, accountability, and potential clinical adoption (Sharif et al., 2024). The field has begun to move from simple post hoc visualization toward more integrated forms of explainability, including attention mechanisms and self-explainable graph models. However, the survey also makes

clear that interpretability remains methodologically underdeveloped in many studies. Explanations are often presented without systematic assessment of stability, plausibility, reproducibility, or clinical utility. For explainable AI to play a truly meaningful role in Alzheimer's disease diagnosis, future work will need to evaluate it with the same seriousness as predictive performance.

Despite substantial progress, several limitations continue to prevent the field from reaching full clinical maturity. Benchmark dependence remains strong, with many studies relying heavily on a small number of public datasets, especially ADNI. External validation across institutions and cohorts is still relatively uncommon. Evaluation settings remain heterogeneous, making direct comparison across studies difficult. Reproducibility is also weakened by inconsistent preprocessing pipelines, incomplete methodological reporting, and limited code or data transparency. Together, these issues mean that many impressive research results still fall short of what would be required for dependable real-world deployment (Sharif et al., 2025).

For this reason, the next generation of Alzheimer's disease diagnosis systems will need to be judged not only by predictive accuracy, but also by robustness, generalizability, interpretability, reproducibility, and clinical relevance. Models that can perform well across diverse datasets, handle multimodal and incomplete inputs, provide trustworthy explanations, and align more closely with actual clinical workflows are likely to have the greatest long-term impact. Thus, while the field has already made remarkable progress, its most important future challenge is not simply to build more complex models, but to develop systems that are scientifically rigorous, clinically meaningful, and practically deployable. In this sense, a clinically useful next-generation Alzheimer's disease diagnosis framework will likely be multimodal, externally validated, explainable, and explicitly designed for real-world robustness rather than benchmark performance alone (Castellano et al., 2024; Qiu et al., 2022; Zhang et al., 2023; Zhao et al., 2024).

REFERENCES

[1]. Alsubaie, M. G., Luo, S., & Shaukat, K. (2024). Alzheimer's disease detection using deep learning on neuroimaging: A systematic review. Machine Learning and Knowledge Extraction, 6(1), Article 24. https://doi.org/10.3390/make6010024

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

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

[4]. Hu, X., Sun, Z., Nian, Y., Wang, Y., Dang, Y., Li, F., Feng, J., Yu, E., & Tao, C. (2024). Self-explainable graph neural network for Alzheimer disease and related dementias risk prediction: Algorithm development and validation study. JMIR Aging, 7, e54748. https://doi.org/10.2196/54748

[5]. Islam, J., & Zhang, Y. (2018). Brain MRI analysis for Alzheimer's disease diagnosis using an ensemble system of deep convolutional neural networks. Brain Informatics, 5(2), Article 2. https://doi.org/10.1186/s40708-018-0080-3

[6]. Lu, D., Popuri, K., Ding, G. W., Balachandar, R., Beg, M. F., & Alzheimer's Disease Neuroimaging Initiative. (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. https://doi.org/10.1038/s41598-018-22871-z

[7]. 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., Dwyer, B. C., Hao, H., Kaku, M. C., Kedar, S., Lee, P. H., Mian, A. Z., Murman, D. L., O'Shea, S., Paul, A. B., Saint-Hilaire, M.-H., Sartor, E. A., Saxena, A. R., Shih, L. C., Small, J. E., Smith, M. J., Swaminathan, A., Takahashi, C. E., Taraschenko, O., You, H., Yuan, J., Zhou, Y., Zhu, S., Alosco, M. L., Mez, J., Stein, T. D., Poston, K. L., Au, R., & Kolachalama, V. B. (2022). Multimodal deep learning for Alzheimer's disease dementia assessment. Nature Communications, 13, 3404. https://doi.org/10.1038/s41467-022-31037-5

[8]. Zhang, Y., He, X., Chan, Y. H., Teng, Q., & Rajapakse, J. C. (2023). Multi-modal graph neural network for early diagnosis of Alzheimer's disease from sMRI and PET scans. Computers in Biology and Medicine, 164, 107328. https://doi.org/10.1016/j.compbiomed.2023.107328

[9]. Zhao, Z., Yeoh, P. S. Q., Zuo, X., Chuah, J. H., Chow, C.-O., Wu, X., & Lai, K. W. (2024). Vision transformer-equipped convolutional neural networks for automated Alzheimer's disease diagnosis using 3D MRI scans. Frontiers in Neurology, 15, 1490829. https://doi.org/10.3389/fneur.2024.1490829

[10]. Abubakkar, M., Alsaud, F. A., Dolon, T. N., & Sharif, K. S. (2025). Prognostic modeling for hepatic disorders: A paradigm of equilibrated and generalized machine learning methodologies. *American Journal of Scholarly Research and Innovation, 4*(01), 352–362. https://doi.org/10.63125/13dazp67.

[11]. Bishnu Padh Ghosh, Touhid Imam, Nishat Anjum, Md Tuhin Mia, Cynthia Ummay Siddiqua, Kazi Shaharair Sharif, Md Munsur Khan, Md Atikul Islam Mamun, & Md Zakir Hossain. (2024). Advancing chronic kidney disease

prediction: Comparative analysis of machine learning algorithms and a hybrid model. *Journal of Computer Science and Technology Studies, 6*(3), 15–21. https://doi.org/10.32996/jcsts.2024.6.3.2

[12]. Reja Sweet, M. M., Md Arif, Uddin, A., Sharif, K. S., Tusher, M. I., Devi, S., & Islam Sarkar, M. A. (2024). Credit risk assessment using statistical and machine learning: Basic methodology and risk modeling applications. *International Journal on Computational Engineering, 1*(3), 62–67. https://doi.org/10.62527/comien.1.3.21

[13]. Rakin, A. A., Nayyem, M. N., Sharif, K. S., Hossain, A.-A., & Arafin, R. (2024). A comprehensive framework for advanced machine learning and deep learning models in cervical cancer prediction. In *2024 IEEE Asia Pacific Conference on Wireless and Mobile (APWiMob)* (pp. 1–6). https://doi.org/10.1109/APWiMob64015.2024.10792967

[14]. Sharif, K. S., Uddin, I. I., Abubakkar, M., Khan, M. M., Ahmad, I., & Uddin, M. M. (2025). DNA sequence classification: An advanced machine learning framework for accurate splice junction detection. In *2025 International Conference on Metaverse and Current Trends in Computing (ICMCTC)* (pp. 1–6). https://doi.org/10.1109/ICMCTC62214.2025.11196541