International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

Review of Explainable AI Approaches in Multimodal Neuroimaging for Alzheimer's Disease Detection

Serra Aksoy

Institute of Computer Science, Ludwig Maximilian University of Munich (LMU), Oettingenstrasse 67, 80538 Munich, Germany

Abstract

Alzheimer's disease is the most prevalent etiology of dementia, and its early diagnosis has been deemed instrumental for early intervention and better patient management. Over the last few years, deep learning techniques have been used with neuroimaging data, and notable gains in diagnostic performance have ensued. Such models have, however, typically been faulted for being "black boxes" that, despite their performance, hinder their acceptance in clinical practice due to a lack of transparency and interpretability. Current explainability methods have typically been used as post-hoc procedures, but they often yield inconsistent or anatomically irrelevant attribution maps that clinicians find hard to trust.

In this review, progress in explainable AI for the detection of Alzheimer's is discussed, with a focus on DL methods combined with multimodal neuroimaging. Emphasis is given to XRAI, a region-based attribution method that has been demonstrated to produce more coherent and clinically interpretable explanations compared to conventional pixel-level techniques. The application of XRAI to 2D and 3D neuroimaging is considered, together with the potential of XRAI to identify anatomically relevant brain areas implicated in disease pathology. Challenges in terms of clinical uptake, integration into workflow, and standardized evaluation of explainability techniques are also discussed. The review identifies how the pairing of high-performing AI models with strong explainability methods has the potential to enable the creation of practical and reliable diagnostic tools for real-world clinical application.

Keywords: Alzheimer's disease, Explainable AI, Neuroimaging, Multimodal MRI, XRAI, Diagnostic interpretability.

1. INTRODUCTION

Alzheimer's disease (AD) is still the global leading dementia with 55 million people affected worldwide, and estimates that by the year 2050, the number can double due to aging populations [1,2]. It is an irreversible neurodegenerative disorder whereby pathological protein aggregates, such as amyloid plaques and neurofibrillary tangles, impair neuronal function and result in irreversible cognitive dysfunction, memory loss, and behavioral change [3]. The pathophysiologic characteristics of AD are sequestration of extracellular senile plaques made of amyloid-beta (Aβ) peptides and intracellular neurofibrillary tangles made of hyperphosphorylated tau proteins, which together initiate neuroinflammatory cascades and synaptic dysfunction [2].

The condition generally progresses through discrete clinical stages, starting with cognitively normal

(CN) aging, followed by mild cognitive impairment (MCI), and then finishing in advanced Alzheimer's disease (AD) dementia. Each stage represents an elevated level of cognitive decrease and functional impairment [4]. The transition interval between normal cognitive abilities and Alzheimer's disease, known as mild cognitive impairment, represents a very promising target for a therapeutic intervention, since 10-15% of MCI-diagnosed individuals are converted to Alzheimer's disease per year, compared to 1-2% conversion rates among the general aging population [1]. The heterogeneous nature of MCI presentations, ranging from amnestic MCI primarily affecting memory to non-amnestic MCI impacting other cognitive domains, complicates early diagnostic accuracy and necessitates sophisticated assessment approaches [4]. Early signs include subtle memory lapses, particularly episodic memory deficits, cognitive difficulties in executive function and language processing, and impaired learning abilities that progressively interfere with daily living activities.

The global burden of AD extends far beyond the patients themselves, profoundly impacting families, caregivers, and healthcare systems worldwide [3]. Early and accurate diagnosis is crucial for implementing effective intervention strategies, as therapeutic interventions are most beneficial when applied during the initial stages of disease progression [5].

There remains no cure for AD, but early identification makes feasible inclusion in broad treatment planning which can possibly delay disease progression and promote quality of life for patients and their families as well as enabling forward planning and making family support achievable [6]. The imperative role for early identification is again highlighted through the methods for early intervention which can possibly delay the emergence of further disabling symptomatology and attain superior patient outcomes [3].

2. Neuroimaging in the Detection of Alzheimer's Disease

Today, neuroimaging is an anchor technique in the diagnostics and tracking for AD, as the cumulative knowledge on disease's fundamental processes of pathophysiology is provided from the various types of modalities [2]. The structural magnetic resonance imaging (sMRI) provides most current non-invasive instruments for primary structure changes related to assessment for AD in an initial stage, which are primarily hippocampal atrophy and cortex thinning, mostly associated with cognitive impairment [7,8].

Specialized MRI methods have improved the sensitivity and specificity of AD detection over and above traditional structural imaging methods. Diffusion tensor imaging (DTI) allows unparalleled observation of white matter microstructural integrity through measurement of the directional diffusion of water molecules, with the ability to detect microstructural alterations in white matter integrity that can occur before overt atrophy is detectable using conventional MRI [1,2]. Functional MRI (fMRI) contributes valuable information about neural activity patterns and network connectivity, revealing characteristic alterations in brain networks that are consistently observed across AD populations [2].

The default mode network, comprising posterior cingulate cortex, precuneus, and medial prefrontal regions, shows systematic disruption in AD that manifests as reduced connectivity in posterior brain regions and increased connectivity in frontal regions, potentially as a compensatory mechanism [2]. Resting-state fMRI can detect functional brain changes that precede structural alterations, with research indicating that AD affects functional connectivity within the default mode network. However, fMRI application in clinical settings is limited due to variability in the BOLD response and challenges in longitudinal studies [2].

Positron emission tomography (PET) imaging has revolutionized the diagnostic strategy for Alzheimer's

disease (AD) by allowing the direct visualization of molecular pathological processes underlying this disorder. Fluorodeoxyglucose (FDG)-PET measures cerebral glucose consumption, showing characteristic patterns that consistently distinguish between AD patients and healthy controls, and are useful for further classification into different disease stages. The use of FDG-PET for diagnosing AD has become widely accepted, showing unmatched sensitivity and specificity to predict at-risk status for developing AD, and recognizing preclinical ATP-III-defined impairments in glucose metabolism. In a more recent development, amyloid-PET imaging with tracers like florbetapir has made it possible to visualize amyloid plaque deposition across the brain and observe directly both the spatial pattern and temporal course of AD pathology [2].

Tau-PET imaging represents the newest advance in molecular neuroimaging, with tracers such as [18F] flortaucipir showing the potential to image neurofibrinary tangle distribution in vivo in patients. Tau PET imaging has also revealed potential for the detection of AD and the differentiation from other neurodegenerative conditions, with studies showing a strong correlation between tau accumulation and cognitive decline [2]. Tau PET imaging research has enabled tau tangles in the brain, one of the hallmark features of AD pathology, to be imaged, with visual tau PET scan interpretation based on regional uptake scoring systems to ascertain tau deposition.

The integration of different neuroimaging modalities by multimodal analysis techniques has consistently demonstrated superior diagnostic performance compared with single-modality techniques. Hybrid approaches of PET with other imaging modalities, i.e., MRI, can be utilized for improving diagnostic accuracy as well as obtaining a fuller view of AD pathology [2]. Multimodal image fusion approaches can integrate structural and metabolic information from MRI and PET imaging for improved diagnosis and monitoring of AD. Sophisticated data fusion techniques enable optimal combination of complementary information from different imaging modalities, utilizing the specific strengths of each technique to provide overall assessment of brain structure, function, and molecular pathology [1].

3. Machine Learning and Deep Learning Approaches in AD Detection

The application of artificial intelligence in AD detection has seen phenomenal expansion, transitioning from traditional ML approaches to sophisticated deep learning frameworks that have transformed automatic neuroimaging analysis. Early studies were mainly based on traditional ML techniques such as Support Vector Machines (SVM), Random Forest (RF) and logistic regression, which were effective in classifying AD stages through hand-crafted features of neuroimaging data [4,6]. These methods reported modest accuracies but were inherently limited by the requirement of hand-crafted features and domain knowledge for feature selection, and required long preprocessing chains, which limited their scalability and reproducibility in various clinical settings.

Conventional ML techniques have been extensively applied in AD classification, with SVM being a popular option for binary classification problems such as AD classification from MRI images. SVM operates by identifying the hyperplane that maximally differentiates two classes of data in a high-dimensional space and is most effective when the feature space is distinct and well established. Random Forest, an ensemble learning technique that integrates numerous decision trees, has been utilized to enhance classification performance and strength, with research showing its applicability in AD detection using feature extraction and classification [3]. However, these traditional techniques rely heavily on hand-designed features, which can be time-consuming and domain-specific, and feature selection plays an important role in model performance.

The paradigm shift towards deep learning has transformed AD detection by allowing automatic feature extraction from high-dimensional neuroimaging data through hierarchical representation learning. Convolutional Neural Networks (CNNs) have emerged as the de facto architecture for medical image analysis, achieving better performance in distinguishing between healthy brains, MCI and AD-affected brains through end-to-end learning [8]. Studies have reported that CNNs can outperform traditional ML methods in AD stage classification through high-level abstract feature extraction from MRI data, and CNN models such as VGGNet and ResNet have recorded impressive performance in discriminating between healthy brains, MCI and AD-affected brains [3].

There is a higher propensity to develop high-level CNN architectures for medical image applications, and transfer learning methods have made for successful adaptation of pre-trained networks for the assessment of neuroimaging. Transfer learning, for example, fine-tuning pre-trained CNN architectures on AD databases, was successful in enhancing diagnostic performance as well as diminishing the necessity for high-annotated databases. MRI-based brain data studies offered an indication that CNNs are capable of achieving high accuracy levels. For instance, MobileNetV3 was capable of 93% accuracy in AD classification, and DenseNet121 was capable of achieving 88% accuracy [3].

Recently developed higher-level architectures further enhance the automatic AD detection current-state-of-art performance. 3D Hybrid Compact Convolutional Transformers (HCCTs) is an emerging method that simultaneously integrate the local feature extraction advantage for CNNs and long-range dependency modeling for vision transformers to effectively extract both local anatomical knowledge and global spatial associations for 3D MRI volumes [5]. The hybrid models are compact in their computations due to their compact architecture but attain higher performance in comparison to traditional CNN-based methods, achieving an accuracy value of 96.06% for multi-class classification. The process of using the self-attention technique makes the model pay attention to desired areas in the brain in a selective manner to remove unwanted information, resulting in stricter feature representation.

The deep architecture's expansion for neural networks to medical imaging is an excellent new promise for detection for AD. The neural network's architecture, YOLOv11, exhibits clear sufficiency for localization as well as for classification in parallel on 93.6% sensitiveness, 91.6% recall, and 96.7% mAP50 in multimodal fusion techniques combining T2-weighted MRI as well as DTI pictures for concurrently estimating structural as well as microstructural change in the brain [1]. The joint detection as well as localization techniques allow automatic localization detection for areas in the brain for AD pathology but still permits classification in high accuracy for severities for disease.

Ensemble methods also increased accuracy in diagnosis-making through combining various models to capitalize on their complementarities as well as decrease individual model biases. Outputs in various models can be aggregated together using ensemble methods for an improvement in the generalizability in the learning model, e.g., averaging, stacking, or boosting [3]. Ensemble methods are also said to offer increased accuracy in diagnostics compared to in individual models because they pool the individual bias as well as the predictor variances [4]. CNNs, specifically, are said to offer increased performance in the majority of medical image understanding applications, e.g., AD diagnosis, because they can capitalize on non-linear as well as hierarchical features.

Systematic reviews and meta-analysis on applications of ML in the diagnosis for AD also suggested the implementation of numerous methods on different datasets and methodological frameworks. The meta-analysis and systematic review on AD prevalence on different stages after the implementation of ML methods suggested the predominance in prevailing prevalence which varied considerably between

studies. With consideration on the prevailing rate between the AD and the cognitively normal, the estimate was 49.28% between studies, but for three-stages cognitive impairment, the estimate was 29.75%. Different participants in different studies estimated four-stages and suggested an overall total prevalence as 13.13%, but participants-based studies estimated six-stages prevalence as 23.75% [6]. The findings suggest the change in the implementation of the diagnosis on different methodological frameworks and suggest the requirement for consistent assessment processes.

Table 1 provides a comparative overview of prominent AI models that have been applied to Alzheimer's disease detection using neuroimaging data. Conventional CNNs, such as VGGNet and ResNet, remain widely used for 2D MRI analysis because of their ability to extract hierarchical image features. These models can achieve high accuracy but are limited by their need for large training datasets and their lack of interpretability. YOLOv11, an advanced object detection framework, has been adapted for multimodal inputs such as MRI and DTI, enabling simultaneous localization and classification of disease-related brain regions with strong diagnostic performance, although at a higher computational cost. The 3D Hybrid Compact Convolutional Transformer (3D HCCT) combines the strengths of CNNs and Transformers to capture both local anatomical features and global spatial dependencies in 3D MRI data, achieving the highest reported accuracy. Despite these advances, the table illustrates that trade-offs remain between performance, computational requirements, and clinical interpretability.

Table 1. Comparison of AI Models for AD Detection

Model / Architecture	Data Type	Modality	Accuracy (%)	Advantages	Limitations
CNN (VGGNet)	MRI	2D	85	High accuracy, easy training	Low interpretability
ResNet	MRI	2D	88	Deep feature extraction	Require large datasets
YOLOv11	MRI+DTI	Multimodal	93.6	Simultaneous localization and classification	High computational cost
3D HCCT	MRI	3D	96.1	Combines CNN and Transformer advantages	Complex model, high data requirements

4. Explainability in Artificial Intelligence Models for Medical Applications

Following advancements in AD detection accuracy for existing AI systems, high-level architecture's intrinsic "black box" nature, which makes decision-making processes non-transparent and inhibits clinical confidence, is the high-level bottleneck against clinical deployability in such systems. Non-transparency and interpretability of AI decision-making processes are root causes for clinical non-acceptability, most notably in high-risk clinical applications where knowledge about rationale behind recommended diagnoses is an integral part in patient safety, regulatory approvals, and ethical medical practice [3,9]. Clinicians require further knowledge about AI systems arriving at their diagnostic outputs to interpret recommendations based on clinical expertise as well as offer professional accountability in clinical decisions for patient care.

Later, Explainable AI (XAI) was developed as a preferred remedy for closing AI performance-clinical interpretability gaps. XAI gives human interpretable explanations for AI-based decisions in an endeavor to offer transparency, instill confidence, and enhance AI tool uptake in clinical use [2]. XAI is an

ensemble of methods and technologies making AI model behavior and decision-making processes understandable and interpretable to humans in an endeavor to primarily inform end-users on how an AI model makes a made decision or prediction [3]. This is critical in clinical contexts where interpretability, accountability, as well as trust-building in the decision-making process is of paramount importance. XAI methods operate alongside a variety of methodological strategies that fall under post-hoc methods; supplying explanations after the models have been deployed, and ad-hoc methods, which entail including explainability mechanisms as integral elements in the model building process.

4.1 Post-hoc Explainability Methods

The most preferred XAI methods in existing medical applications of AI are post-hoc methods due to their model-agnostic nature, as they can be directly applied on all configurations in deep learning. The majority of DL algorithms use post-hoc methods because they are easiest to apply and can be plugged in under plug-and-play setup [9]. Post-hoc explainability is founded on gradient-based methods using mathematical properties of optimization in NNs in an effort to determine input features most contributively attributing to prediction outputs.

Gradient-Weighted Class Activation Mapping (Grad-CAM) is an instance that utilizes gradient info propagated through convolutional layers for identifying spatial regions in input pictures most responsible for model outputs [8,9]. Grad-CAM was extensively adopted in AD detection studies and provided visualization-based explanations indicating brain regions important for classification outputs. Most studies indicated that Grad-CAM can indicate image regions most responsible for model outputs according to gradient info propagated in the final convolutional layer, but the technique is not effective because spatial resolution is not high, and background regions unrelated to object features are typically emphasized [2].

SHAP (SHapley Additive exPlanations) is a sophisticated technique that provides a quantitative value for each input component for the model's outputs based on game theory reasoning under the cooperative game theory [2].

SHAP is an XAI method that gained significant attention in generating model-agnostic explanation as an explanation for each prediction in terms of variation in various features, based on Shapley values in game theory for estimating the extent to which each feature is accountable for the model prediction [3]. SHAP, when adapted in AD detection problems, was highly informative in revealing important brain regions and demographically associated factors causing diagnostic decisions, even when explanation is poor, in some way, intuitive understanding for immediate clinical translation.

Local Interpretable Model-Agnostic Explanations (LIME) generates an explanation for a sample by discovering an interpretable model in the local region surrounding the prediction through simple surrogate models, which are interpretable in nature [3]. LIME generates individual predictions by locally approximating the AI model around the particular instance being predicted, determining which features had the greatest influence on the choice by varying input data and examining changes in the model's output. While LIME offers valuable insights into individual classification decisions, studies have noted limitations in its ability to capture the full complexity of deep learning models, particularly when applied to high-dimensional medical imaging data.

Layer-wise Relevance Propagation (LRP) operates through a fundamentally different mechanism by backpropagating relevance scores from model outputs to inputs layer by layer, ensuring conservation of prediction scores throughout the network architecture [2]. LRP works by propagating the class score

backward over the neural layers to the input image using LRP specific rules, with the concept of LRP being to conserve inter-neuron dependency. In AD detection applications, LRP has shown promise in generating pixel-level attribution maps that highlight anatomically relevant brain regions while avoiding some of the gradient saturation issues that plague simpler gradient-based methods.

Table 2 summarizes widely used explainable artificial intelligence (XAI) methods that have been applied to neuroimaging-based Alzheimer's disease detection. Grad-CAM is a gradient-based post-hoc technique that generates visual heatmaps to highlight brain regions contributing to a model's prediction. Although widely adopted, it provides low spatial resolution and may emphasize irrelevant background areas. LIME is another post-hoc approach that builds locally interpretable surrogate models around individual predictions, offering valuable case-specific insights but struggling with high-dimensional medical imaging data. SHAP, based on game-theoretic Shapley values, assigns numerical importance scores to input features, making it useful for quantifying the contribution of specific regions or demographic factors to model decisions, albeit with high computational cost. XRAI is a more recent region-based attribution method that overcomes the limitations of pixel-level explanations by producing coherent and anatomically meaningful attribution maps. While its application in Alzheimer's detection is still limited, XRAI shows strong potential for improving clinical interpretability by reliably highlighting disease-relevant brain regions.

Table 2. Comparison of XAI Methods in AD Applications

Method	Type	Advantages	Limitations	Using AD Applications
Grad-CAM	Post-hoc	Provides visual explanation	Low spatial resolution	Highlights brain regions on MRI
LIME	Post-hoc	Generates local explanations	Limited in high-dimensional data	Explains decision for single cases
SHAP	Post-hoc	Gives numerical feature contributions	High computational cost	Shows clinically relevant factors
XRAI	Regional	Consistent and ana-tomically meaningful	Novel, limited applications	Highlights pathological brain regions

4.2 Sophisticated Regional Attribution Techniques

XRAI (eXplanation via Regional Attributes Integration) is a considerable improvement over pixel-level attribution techniques by overcoming intrinsic limitations in coherence, stability, and clinical relevance that typify previous explainability techniques [10]. Unlike conventional techniques that generate granular pixel-level explanations in the form of often noisy and fragmented attribution maps, XRAI adopts a new region-based methodology that over-segments images into coherent anatomical regions and recursively evaluates their importance based on integrated gradient attribution scores. This methodological innovation addresses the crucial observation that while pixel-level attributions are unreliable due to gradient saturation and optimization artifacts, region-level aggregations of such scores provide more stable and clinically relevant explanations.

The XRAI technical implementation adheres to a three-step algorithmic pipeline that begins with wide image segmentation with multiple over-segmentations with different parameters to acquire diverse region proposals. Segmentation is performed by Felzenswalb's graph-based method itself with a scale parameter ranging from 50 to 1200, giving overlapping region hierarchies with filtering of segments

smaller than 20 pixels and dilation of segment masks by 5 pixels to align segment boundaries with image boundaries [10]. The attribution calculation is then performed using black and white baselines with the integrated gradients algorithm to achieve balanced invariance for dark and light image features, respectively, while also avoiding the well-known limitations of single-baseline approaches.

The most novel aspect in XRAI is the step for region selection, which is done in an iterative process. The step gives a step-by-step algorithm incrementally initializing an empty reference mask incrementally adding incrementally, after each previous, regions in an effort to optimize maximum total reference gain in unit area. The step, in each step, gives ever-inclusive rationales, inclusive of informative pathological markers together with subsidiary support evidence, and selects regions based on their salience [10]. The mathematical grounding is on the expression in which each region's XRAI attribution score is given as summation integral gradient values for each of all the pixels in each region, thereafter, generating higher attribution scores that are associated with anatomically significant brain structures.

Comparative assessment under new criteria such as Accuracy Information Curves (AIC) and Softmax Information Curves (SIC) confirmed that XRAI is the top against all baseline attribution methods under most criteria [10,11]. The criteria are systematic perturbation experiments-based and assess, according to accuracy, which methods are the top in identifying the most important regions in prediction in models. Experiments confirmed that XRAI is the top in accuracy against all the other baseline saliency methods. The experiments, on an unprecedentedly colossal scale, confirmed that it is the top in localization accuracy as well as consistency in explanation, significantly in medical imaging applications where identification of region in pathology accurately is most important.

4.3 Challenges and Limitations in Current XAI Approaches

Despite huge advancements in XAI methodology, several inherent limitations still hamper the clinical utility and deployment readiness of current explainability techniques for medical imaging applications. Arguably the most critical limitation is low specificity, with many XAI methods producing explanations that highlight anatomically irrelevant regions or background artifacts lacking clinical significance [9]. Post-hoc XAI lacks the ability to produce class-discriminative and target-specific explanation, while generally, post-hoc XAI methods are confronted with complexity in producing understandable and class discriminative attribution maps.

Another significant barrier to clinical translation is the absence of standard evaluation metrics for XAI quality assessment, as most evaluations rely on subjective visual inspection rather than quantitative measurements enabling objective comparison of different explainability methods [9]. Quality control of the XAI approaches is not typically performed, and systematic comparison among approaches is thus challenging. Technical validation and clinical validation gaps are other major obstacles to XAI adoption in clinical environments, since studies have stressed the utmost need for validation frameworks measuring not just technical performance but also clinical usefulness and workflow integration [2].

The sophistication of contemporary deep learning structures also adds to the complexity of explainability, as the task of giving intelligible interpretations gets more difficult with the rise in model complexity. Current practice demonstrates minimal deployment of deep learning algorithms in clinical practice due to the fact that DL algorithms are not transparent and trustworthy since they have an underlying black-box mechanism [9]. To ensure successful utilization, explainable artificial intelligence may be implemented to bridge the gap between medical practitioners and DL algorithms, yet no clear agreement is present regarding how XAI needs to be implemented to bridge the gap between medical

practitioners and DL algorithms for clinical adoption.

5. Results

Despite impressive gains in AI-driven AD detection reporting high accuracies in research environments, the clinical uptake of such systems is drastically low because of inherent impediments in interpretability, usability, and clinical integration. Existing AI systems are "black boxes" that yield diagnostic predictions without clinically useful explanations to enable healthcare practitioners to comprehend and verify machine logic underlying algorithmic decisions [3,9]. This lack of transparency poses considerable impediments to clinical adoption, especially among general practitioners and non-experts who need intuitive explanations to make informed diagnostic decisions in time-pressed clinical practice.

Current explainability approaches, although technically advanced, typically do not achieve the degree of clinical interpretability necessary for deployment in the real world. Classical gradient-based approaches commonly identify anatomically irrelevant areas or generate inconsistent explanations that lack clinical validation and can confuse rather than inform clinical decision-making. Furthermore, these approaches usually function independently of clinical workflow, necessitating separate analysis pipelines that are not practical for everyday clinical use where integrated decision-making support is necessary. There is no obvious agreement as to how XAI should be implemented to bridge the gap between medical practitioners and DL algorithms for clinical deployment, with systematic technical and clinical quality evaluation of XAI approaches seldom being implemented [9].

There is a key gap between research-driven AI systems and clinically deployable products that can be easily integrated into clinical workflows without the need for specialized technical experience. Existing systems generally require a high level of technical expertise for their operation and interpretation, reducing their usability by the wider healthcare community where general practitioners are key users in clinical screening applications. The lack of extensive bias detection and monitoring features is another key limitation since demographic-related prediction biases can cause systematic misdiagnosis, resulting in considerable clinical risks precluding safe deployment. The integration of multiple neuroimaging modalities for comprehensive AD assessment has shown superior diagnostic performance compared to single-modality approaches, yet current explainability frameworks are predominantly designed for single-modality analysis and lack the capability to provide unified explanations across different imaging types [1]. XRAI's region-based approach offers promising solutions to traditional pixel-level attribution limitations, demonstrating superior performance in localizing pathologically relevant regions compared to conventional methods, but its application to AD detection using neuroimaging data remains largely unexplored, particularly in the context of comprehensive clinical deployment systems.

This research addresses these critical gaps by developing a comprehensive, clinically oriented explainable AI system that integrates advanced deep learning architectures with XRAI-based explainability for both 2D and 3D neuroimaging analysis through a unified web-based clinical deployment platform. The innovation lies in creating the first systematic application of XRAI to AD detection while simultaneously developing a practical clinical interface that enables healthcare professionals to access AI-powered diagnostic capabilities with comprehensive explainable insights and integrated workflow support that bridges the gap between advanced AI research and practical healthcare deployment for trustworthy, interpretable diagnostic tools in real-world clinical settings.

Future AI systems should be designed to function as part of routine clinical practice, offering interpretable predictions that can be understood and trusted by healthcare professionals without requiring

specialized technical expertise. Combining various neuroimaging modalities with longitudinal patient data may lead to more accurate predictions of disease onset and progression, supporting earlier and more personalized interventions. There is a pressing need for objective, standardized metrics to evaluate the quality and clinical relevance of AI-generated explanations, enabling fair comparisons between methods. Future models should include mechanisms for identifying and correcting demographic or dataset-related biases to ensure fairness, safety, and generalizability across diverse populations. Developing frameworks that can handle both 2D and 3D neuroimaging data while maintaining consistent explainability will improve the practicality of AI-based tools in real-world settings.

References

1. Hechkel, W.; Helali, A. Early Detection and Classification of Alzheimer's Disease through Data Fusion of MRI and DTI Images Using the YOLOv11 Neural Network. *Front. Neurosci.* **2025**, *19*, doi:10.3389/fnins.2025.1554015.
2. Taiyeb Khosroshahi, M.; Morsali, S.; Gharakhanlou, S.; Motamedi, A.; Hassanbaghlou, S.; Vahedi, H.; Pedrammehr, S.; Kabir, H.M.D.; Jafarizadeh, A. Explainable Artificial Intelligence in Neuroimaging of Alzheimer's Disease. *Diagnostics* **2025**, *15*, 612, doi:10.3390/diagnostics15050612.
3. Hasan Saif, F.; Al-Andoli, M.N.; Bejuri, W.M.Y.W. Explainable AI for Alzheimer Detection: A Review of Current Methods and Applications. *Appl. Sci.* **2024**, *14*, 10121, doi:10.3390/app142210121.
4. Singh, S.G.; Das, D.; Barman, U.; Saikia, M.J. Early Alzheimer's Disease Detection: A Review of Machine Learning Techniques for Forecasting Transition from Mild Cognitive Impairment. *Diagnostics* **2024**, *14*, 1759, doi:10.3390/diagnostics14161759.
5. Majee, A.; Gupta, A.; Raha, S.; Das, S. Enhancing MRI-Based Classification of Alzheimer's Disease with Explainable 3D Hybrid Compact Convolutional Transformers. **2024**, doi:10.48550/ARXIV.2403.16175.
6. Battineni, G.; Chintalapudi, N.; Amenta, F. Machine Learning Driven by Magnetic Resonance Imaging for the Classification of Alzheimer Disease Progression: Systematic Review and Meta-Analysis (Preprint) 2024.
7. Katti, G.; Ara, S.A.; Shireen, A. Magnetic Resonance Imaging (MRI) - A Review. *Int. J. Dent. Clin.* **2011**.
8. Zhang, B.; Zhang, S.; Feng, J.; Zhang, S. Age-Level Bias Correction in Brain Age Prediction. *NeuroImage Clin.* **2023**, *37*, 103319, doi:10.1016/j.nicl.2023.103319.
9. De Vries, B.M.; Zwezerijnen, G.J.C.; Burchell, G.L.; Van Velden, F.H.P.; Menke-van Der Houven Van Oordt, C.W.; Boellaard, R. Explainable Artificial Intelligence (XAI) in Radiology and Nuclear Medicine: A Literature Review. *Front. Med.* **2023**, *10*, doi:10.3389/fmed.2023.1180773.
10. Kapishnikov, A.; Bolukbasi, T.; Viégas, F.; Terry, M. XRAI: Better Attributions Through Regions 2019.
11. Brima, Y.; Atemkeng, M. Saliency-Driven Explainable Deep Learning in Medical Imaging: Bridging Visual Explainability and Statistical Quantitative Analysis. *BioData Min.* **2024**, *17*, doi:10.1186/s13040-024-00370-4.