IJIRSS
International Journal of Innovative Research & Scientific Studies

ISSN: 2617-6548

URL: www.ijirss.com

International Journal of Innovative Research and Scientific Studies
ONLINE ISSN : 2617-6548
www.ijirss.com

Systematic review of explainable AI in Alzheimer's diagnosis

Bayan Al Durgham[1], Moatsum Alawida[2*], Murad Al-Rajab[2]

[1]Department of Computer Sciences and IT, College of Engineering, Abu Dhabi University, Abu Dhabi 59911, United Arab Emirates.
[2]Abu Dhabi University, Abu Dhabi 59911, United Arab Emirates.

Corresponding author: Moatsum Alawida (*Email: moatsum.alawida@adu.ac.ae*)

Abstract

This study aims to provide a comprehensive and structured understanding of Explainable Artificial Intelligence (XAI) approaches used in the diagnosis of Alzheimer's Disease (AD). It seeks to bridge the gap between emerging XAI techniques and their clinical applicability, addressing the urgent need for transparent and interpretable diagnostic tools. A systematic literature review was conducted using a structured search strategy to identify relevant studies published in the last five years. A total of 37 peer-reviewed articles were included, focusing on the application of XAI techniques—such as LIME, SHAP, Grad-CAM, and other emerging frameworks—within Machine Learning (ML) and Deep Learning (DL) models for AD diagnosis. The review reveals a growing interest in integrating XAI methods into clinical workflows, highlighting their potential to enhance diagnostic reliability and transparency. It presents a comparative analysis of major XAI frameworks, evaluating their effectiveness, interpretability, and suitability for clinical adoption. Key challenges identified include a lack of standardization across studies, limited dataset availability, and difficulties in generalizing findings. Several research gaps are noted, particularly in the consistency of XAI implementation and interpretability across different ML/DL models. XAI offers promising enhancements to AD diagnosis, but the field is still developing. Standardized methodologies, larger datasets, and improved generalization capabilities are essential for advancing clinical adoption. This review lays the groundwork for future research by identifying critical gaps and suggesting directions for the development of more interpretable and robust XAI models. The insights provided in this review can guide researchers, clinicians, and developers in selecting appropriate XAI frameworks for AD diagnosis. It also underscores the importance of interpretability in AI-driven healthcare applications, helping to foster trust and usability in real-world clinical settings.

Keywords: Alzheimer's disease (AD) Diagnosis, Deep learning in healthcare, Explainable artificial intelligence (XAI), Feature attribution techniques (SHAP, LIME), Machine learning.

DOI: 10.53894/ijirss.v8i8.10536
Funding: This study received no specific financial support.
History: Received: 12 August 2025 / **Revised:** 15 September 2025 / **Accepted:** 17 September 2025 / **Published:** 7 October 2025

Competing Interests: The authors declare that they have no competing interests.
Authors' Contributions: All authors contributed equally to the conception and design of the study. All authors have read and agreed to the published version of the manuscript.
Transparency: The authors confirm that the manuscript is an honest, accurate, and transparent account of the study; that no vital features of the study have been omitted; and that any discrepancies from the study as planned have been explained. This study followed all ethical practices during writing.
Institutional Review Board Statement: This article is a systematic review and does not involve any human participants, personal data, or intervention. Therefore, approval from an Institutional Review Board (IRB) was not required.
Publisher: Innovative Research Publishing

1. Introduction and Literature Review

AD is a clinically established chronic neurodegenerative disorder that remains challenging to diagnose and treat [1]. Since symptoms are symptomatic, it is crucial that they are diagnosed early so that treatments and management can be implemented that can potentially improve patient outcomes. The diagnostic modalities that are available are also typically cumbersome and not very sensitive, particularly in the early stage of the disease [2].

With its importance, XAI methods are increasingly applied in the field of healthcare to improve diagnostic precision and interpretability in AD assessments. XAI improves transparency in AI-driven decisions by providing transparent and understandable explanations, thereby enhancing doctors' confidence in its recommendations [1].

An obstacle of scale to AI rollout is the "black-box issue," with deep algorithms that generate diagnoses without clear articulation. It's such an interpretability deficiency that limits clinical judgment to AD diagnostics. With non-explainable reasoning, clinicians may be hesitant to adopt AI-recommended interventions, delaying the use of AI in clinical applications [2]. XAI frameworks must be developed that balance interpretability and predictive performance to overcome this obstacle.

Through an extensive examination of the current-day literature, this assessment aims to shed light on the barriers, potentialities, and destiny that help in this rapidly developing field. Through a whole assessment, this examination hopes to provide valuable data that may aid the introduction and awareness of XAI answers, advancing the assessment and management of AD to exceptional steps [3].

The 'black box' problem in AI significantly hinders the adoption of XAI in healthcare. Many algorithms, including deep neural networks, analyze large amounts of information and produce outcomes through various layers of connected nodes. Even though those models often attain surprising accuracy, it's still impossible for human beings to recognize how these algorithms function internally. The "black box" nature of AI algorithms significantly affects clinical decision-making when diagnosing AD. To protect their moves and supply sufferers with trust, clinicians want to offer clear and understandable reasons. However, medical doctors will be hesitant to believe and observe AI-driven suggestions if they are furnished without clean justifications [4]. Clinicians may additionally find it difficult to shield or explain AI-powered judgments to patients, regulatory corporations, or secure authorities on negative consequences or mistakes [5]. This, therefore, highlights the importance of creating XAI structures with a key focus on interpretability and transparency in the sale of responsibility and self-assurance in clinical practice.

The trade-off between interpretability and accuracy may be the biggest barrier to the adoption and reliability of AI models in AD analysis, as shown in Figure 1. Reasonably accurate models, mostly brilliant in their complexity and class, tend to attain better diagnostic results by picking up minute styles and correlations across various sources of facts. On the other hand, apparent elements for human alternatives are provided through interpretable models, together with rule-based structures, which promote human comprehension and self-assurance [6]. However, those patterns could lose out on predicted accuracy, especially when dealing with high-dimensional, complex information; this is normal in diagnosing AD. The key to correctly integrating AI into medical practice is finding the right combination between interpretability and accuracy. There are interesting techniques to enhance the interpretability of complex AI models without sacrificing accuracy, along with model distillation, surrogate models, and post-hoc interpretability techniques [3].

XAI answers can provide physicians with actionable insights from AI-driven suggestions, thus supporting more knowledge and assured decision-making in the analysis and management of AD [7]. XAI researchers can create and put AI models in force that prioritize patient protection, interpretability, and medical utility by utilizing their clinical insights and domain experience. This painting attempts to provide a thorough image of the cutting-edge country of XAI applications in AD analysis through a scientific literature assessment [8]. This examines interests to offer navigation for future research regions and medical translation tasks by synthesizing existing information and identifying gaps and obstacles.

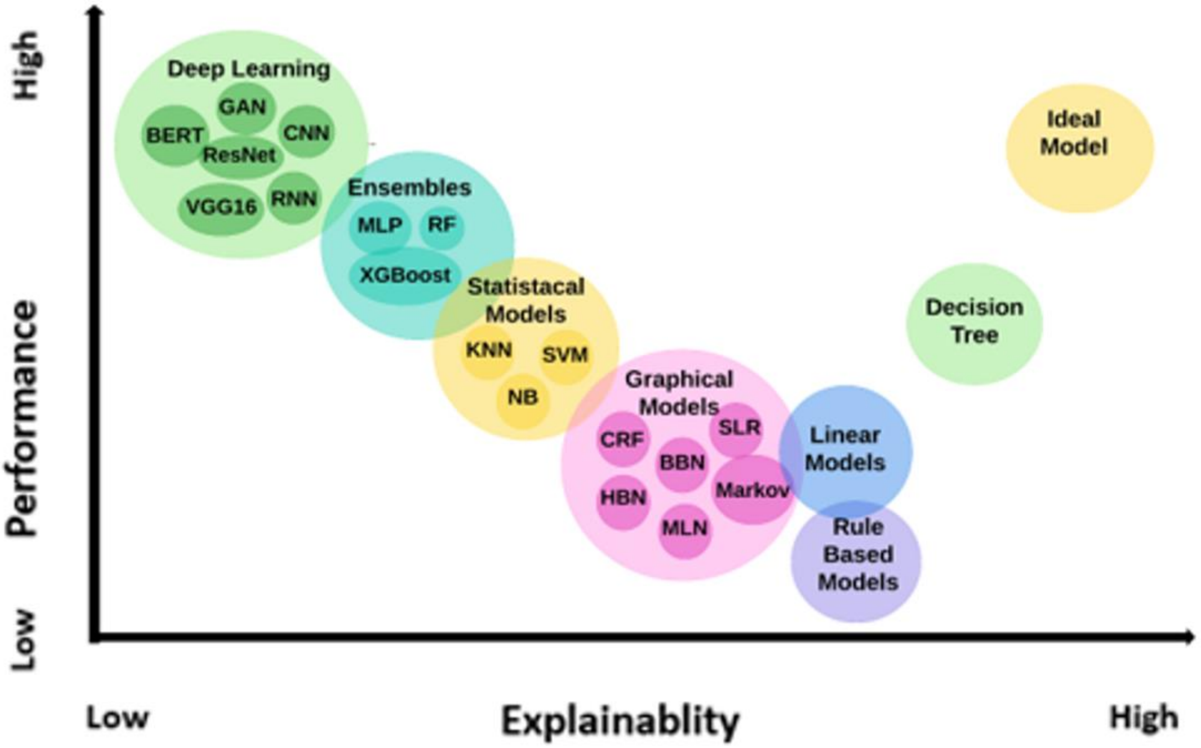


Figure 1.
Trade-off between interpretability and accuracy in XAI-based Alzheimer's diagnosis, highlighting the need for explainability in clinical adoption Viswan, et al. [1].

Figure 1 illustrates the use of version-agnostic explainability strategies in AI models, such as LIME or SHAP, to make predictions clear and understandable. This helps healthcare professionals make informed decisions and promotes self-assurance. However, the opacity of AI models can lead to biases and discrimination, affecting healthcare access and understanding of AD. XAI offers a promising approach for creating reliable and honest AI systems, potentially empowering scientists, improving patient care, and accelerating improvements in AD prevention, diagnosis, and treatment [9, 10]. This issue of interpretability needs to be reduced by employing XAI, in which case you could optimally deploy the assets from AI to prevent AD. The comparison of this review with the previous study for a good understanding is shown in Table 1.

The primary aim of this systematic review is to identify, evaluate, and synthesize available research on XAI techniques to improve the diagnosis of AD. This review integrates technical, ethical, and clinical perspectives on XAI for the diagnosis of AD, distinguishing it from the current literature. Unlike the majority of reviews of AI interpretability in general, this article is the first to have a comparative overview of prominent XAI techniques (LIME, SHAP, Grad-CAM, LRP) used in real clinical applications and an evaluation of how they affect clinician trust, patient-centered explanation, and health care adoption. Moreover, the paper discusses practical implementation issues, multimodal dataset application with diversity, and ethical issues, which are usually neglected in the current literature. With both algorithmic transparency and clinical usability in focus, this paper is a complete guide to developing XAI-based Alzheimer's diagnosis. The nomenclature used in this article is listed in Table 2.

1.1. Problem Statement

A major obstacle to the clinical integration of AI is the "black-box issue"—where deep algorithms generate diagnoses without transparent reasoning. This interpretability deficiency limits clinicians' confidence in AI-driven diagnostics for AD [2]. Clinicians may hesitate to adopt AI-recommended interventions due to a lack of justifications, affecting trust and safety in real-world use [4, 5].

1.2. Research Gap

Although several studies have explored AI in healthcare, there is a lack of comprehensive reviews that focus specifically on XAI techniques applied to AD. Existing works often limit their scope to general AI interpretability or a narrow set of models (e.g., only SHAP and LIME), and many overlook clinical implementation aspects, dataset diversity, and ethical implications.

1.3. Objectives and Methodology

This work's main contribution can be summarized as a systematic review that deepens the understanding of XAI approaches used in AD in multiple unprecedented ways. Firstly, it delineates the boundaries of current research. It provides a comparative analysis of LIME, SHAP, Grad-CAM, and other emerging XAI frameworks towards an integration of clinical use, a feat no prior review has accomplished. The paper also highlights the advantages of employing XAI methods specifically for

AD. To achieve this, the study applied a systematic search methodology to identify relevant articles published in the last five years. This process yielded 37 papers focusing on the use of XAI in AD, employing various ML and DL methods. Through a comprehensive analysis of these papers, the review compares the different XAI methods utilized. It also identifies the limitations of existing works and pinpoints research gaps, which are notably summarized as an absence of standardization and limited generalizations. The paper further includes a detailed discussion of the challenges associated with limited datasets and the interpretation of findings within the current research landscape. Ultimately, this work is aimed at providing clear future directions for research into XAI models in AD. We can summarize the main aims of this paper as follows:

1. Provide a comprehensive review of recent XAI methods used in AD diagnosis.
2. Compare and evaluate popular XAI frameworks based on clinical usability, interpretability, and model performance.
3. Identify limitations in the current literature, including challenges in standardization, generalization, and dataset usage.
4. Recommend future research directions for more transparent and effective AI adoption in AD.

Table 1.
Advancements of this review compared to previous studies.

Aspect	Previous Reviews	This Review
AI Models	Focused on traditional ML (e.g., SVM, Random Forest) with minimal DL coverage Band, et al. [4] and Segato, et al. [5].	Includes DL models like CNNs, Effi-***cientNet alongside classical ML***
XAI Techniques	Limited to SHAP & LIME Kamal, et al. [11] and Vimbi, et al. [12].	Expands to Grad-CAM, GNNExplainer, Occlusion Sensitivity, etc.
Clinical Focus	General XAI surveys with no specific focus on AD Kumar, et al. [2] and Payrovnaziri, et al. [3]	Specific focus on Alzheimer's diagnosis and explainability in clinical context
Dataset Scope	Small, curated datasets (early ADNI subsets Chun, et al. [13].	Uses larger, real-world clinical datasets (e.g., ADNI, NACC, multi- modal EHR)
Practical Implementation	Mostly conceptual/theoretical, limited clinical deployment analysis Payrovnaziri, et al. [3] and Loh, et al. [14].	Analyzes real-world deployment challenges in hospitals including usability and ethics

Table 2.
The nomenclature used in this paper.

Abbreviation	Description	Abbreviation	Description
AD	Alzheimer's Disease	DL	Deep Learning
AI	Artificial Intelligence	ML	Machine Learning
XAI	Explainable Artificial Intelligence	RF	Random Forest
MCI	Mild Cognitive Impairment	LR	Logistic Regression
MRI	Magnetic Resonance Imaging	VGG16	Visual Geometry Group 16
PET	Positron Emission Tomography	SVM	Support Vector Machine
CNN	Convolutional Neural Network	GNN	Graph Neural Network
CAD	Computer Aided Diagnosis	RQ	Research Questions
Grad-CAM	Gradient-Weighted Class Activation Mapping	DT	Decision Tree
LIME	Local Interpretable Model-Agnostic Explanations	AUC	Area Under the Curve
SHAP	SHapley Additive exPlanations	GNNE	GNNExplainer
ADNI	Alzheimer's Disease Neuroimaging Initiative	ApOE	Apolipoprotein E
EEG	Electroencephalogram	SM	Saliency Map
IML	Interpretable Machine Learning	Ma	Model Agnostic
LIME	Local Interpretable Model-Agnostic Explanation	NLP	Natural Language Processing
LRP	Layer-wise Relevance Propagation	TADPOLE	The Alzheimer's Disease Prediction of Longitudinal Evolution
OSA	Occlusion Sensitivity Analysis	sMCI	stable Mild Cognitive Impairment

This paper is organized as follows: We begin by discussing the advantages of using XAI methods to detect AD in Section 2. Section 3 details our methodology, followed by a review of existing literature in Section 4. The discussion, including ethical considerations and patient-centric XAI explanation in Subsection 5, ethical consideration is presented in Section 6. Finally, Section 7 outlines future directions, and Section 8 concludes the paper.

2. The Important Role of XAI in AD Detection

XAI approaches alleviate some of the drawbacks of conventional black-box AI models and provide advantages for AD's. The main benefits include enhanced transparency, as XAI techniques increase the model's transparency and illuminate the tactics and propensities utilized by a model in generating its predictions. In the medical enterprise, consumers and healthcare carriers believe best on evidentiary grounds when the reason is apparent. This transparency also facilitates widespread trust and adoption among medical professionals and patients by increasing the interpretability of AI systems. The more likely medical practitioners are to accept AI technology, the more likely it is that the technology can be explained by being able to understand its basis, and that the basis is valid according to scientific knowledge.

Another key advantage lies in identifying relevant biomarkers. XAI techniques can help pinpoint characteristics that play a large role in some of the most relevant brain areas and test a brain's performance on cognitive tests that carry the most predictive energy for AD. This helps to improve the records of the contamination and directly guide future investigation and restoration exercises. Furthermore, XAI may help enable custom diagnosis and treatment by being custom designed to consider the effects of all factors that impact each person's analysis. In addition, physicians could tailor the treatment regimens accordingly, based on the distinct traits that the AI model finds most important for each patient.

Finally, XAI offers Contributions to enhancing model robustness and safety. Through XAI, developers can gain insights to test how different inputs result in an impact on the target output. This function is crucial for enhancing the model's robustness and ensuring that AI generation produces equitable and strong consequences for many patient companies. Moreover, XAI aids in regulatory compliance. Regulatory organizations in several locations want transparency in clinical AI architecture. XAI tactics help to gather such requirements and make it less difficult for AI improvement to be covered as against the law and used in scientific settings by supplying clear, succinct explanations for model picks.

To sum up, XAI strategies are critical to enhancing AI-driven AD detection systems' transparency, reliability, and medical applicability. These advantages enhance AI's applicability and adoption within the healthcare industry, as well as individual consequences, by allowing. greater individualized and knowledgeable remedies. Figure 2 shows the advantages of using XAI methods to detect AD [15].

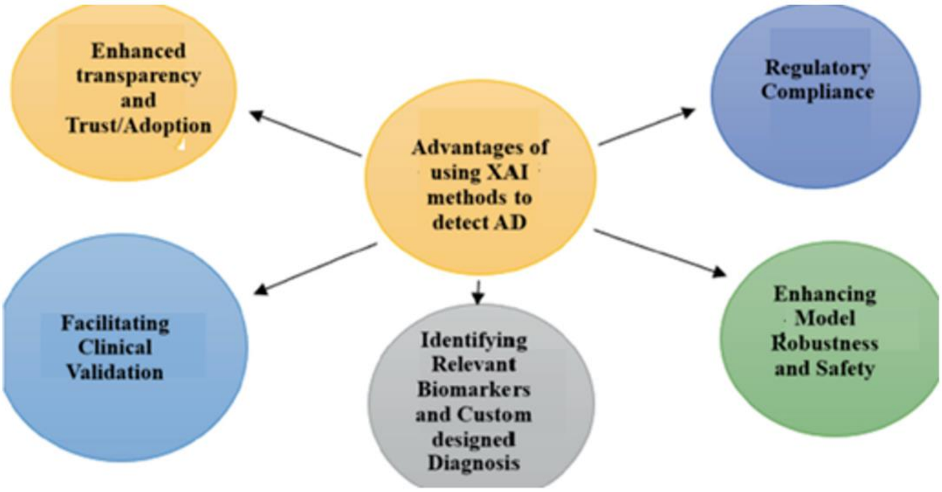

Enhanced transparency and Trust/Adoption
Regulatory Compliance
Advantages of using XAI methods to detect AD
Facilitating Clinical Validation
Identifying Relevant Biomarkers and Custom designed Diagnosis
Enhancing Model Robustness and Safety

Figure 2.
Advantages of using XAI methods to detect AD.

Table 3.
Overview of the systematic literature search strategy used in this study, detailing the databases, keywords, Boolean operators, inclusion criteria, and exclusion criteria.

Component	Description
Databases Searched	PubMed, IEEE Xplore, Scopus, ScienceDirect, SpringerLink
Search Period	Last five years.
Keywords and Controlled Terms	"Explainable AI," "XAI," "Alzheimer's Diagnosis," "Machine Learning in AD," "Interpretable AI"
Boolean Operators	("XAI" OR "Explainable AI") AND ("Alzheimer's" OR "AD Diagnosis")
Inclusion Criteria	Published in peer-reviewed journals
Exclusion Criteria	Studies unrelated to AD or XAI Non-English papers

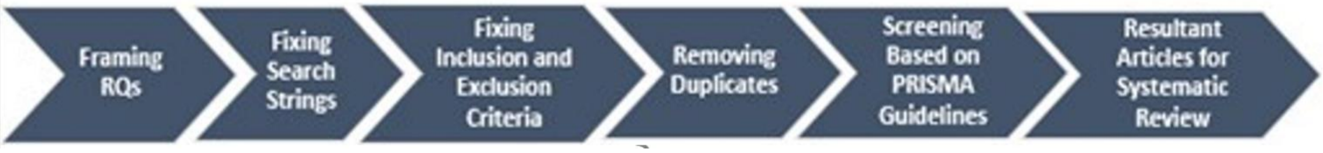

Framing RQs
Fixing Search Strings
Fixing Inclusion and Exclusion Criteria
Removing Duplicates
Screening Based on PRISMA Guidelines
Resultant Articles for Systematic Review

Figure 3.
Sequence of Steps in Search Strategy to identify Relevant Paper Viswan, et al. [1].

3. Methodology

This methodology explained the difficult and structural procedures involved in the collection, assessment, and integration of related literature about the application of XAI to AD diagnosis. It also describes the approach used in the literature search, the criteria for paper inclusion, and the method of extraction that ensures a thorough and impartial assessment.

3.1. Search Strategy

This section describes the major stages involved in locating and identifying suitable publications for a systematic review. The following are the steps shown in Figures 3 and 4. To consider only the relevant papers, we discussed guidelines based on Kumar, et al. [2] and Payrovnaziri, et al. [3]. Notably, we used Boolean operators to establish search strings that simultaneously improved memorization and accuracy as can be seen in Table 3.

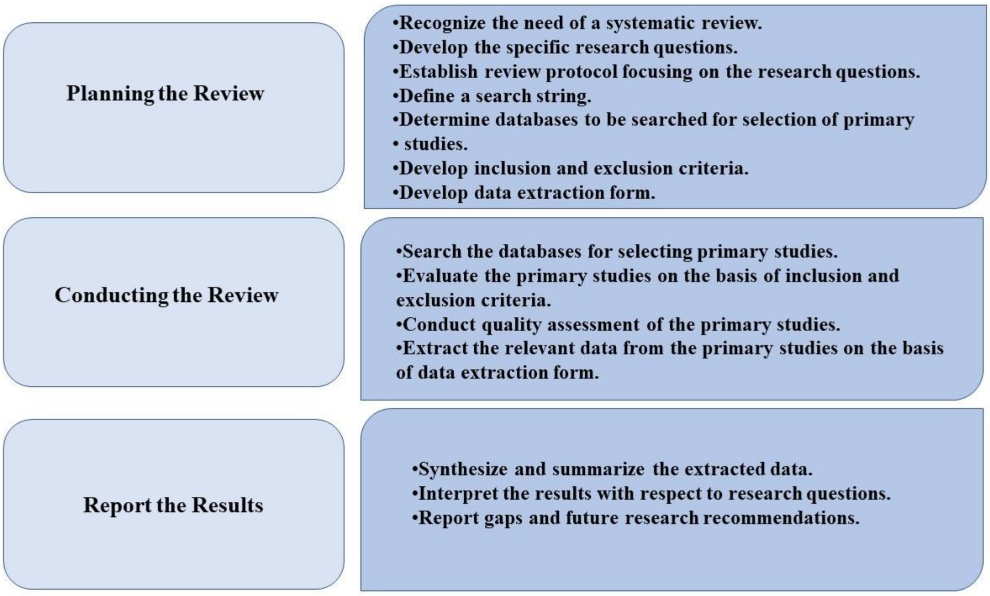
3.2. Research Questions

The study questions aim to identify a proven strategy for retrieving publications that are specific to the topics under investigation. In this paper, we determined the research questions listed in Table 4. The aim is to fill existing research gaps by providing a balance between diagnostic accuracy and interpretability and applying XAI in clinical setups. These questions also pose the structure of XAI evolving beyond just focusing on algorithms. It centers on the patient care providers because they want to know how they can depend on AI to provide useful insights and why it has made certain conclusions.

3.3. Articles Selection Criteria

In this study, we precisely selected our search terms to strike a balance: just specific enough to exclude papers unrelated to the topic, but not so specific that potential articles related to the topic might not include the proper keywords. Furthermore, we added some BOOS resources that are useful for answering research questions and other online resources related to the subject. However, to conduct a specific and selective examination of XAI applications in diagnosing AD, specific search terms need to be selected, and proper acceptance and rejection criteria must be applied.

To this end, we conducted several experiments to compare different sets and placements of keywords related to XAI and AD diagnosis. Both keywords and documents were preprocessed using standard analysis tools as we experimented iteratively toward discovering effective search terms for identifying relevant documentation. The search terms included, among others, *"explainable AI," "XAI," "Alzheimer's Disease," "AD," "diagnosis," "machine learning," "deep learning," "interpretability,"* and *"classification."*



Planning the Review

•Recognize the need of a systematic review.
•Develop the specific research questions.
•Establish review protocol focusing on the research questions.
•Define a search string.
•Determine databases to be searched for selection of primary
• studies.
•Develop inclusion and exclusion criteria.
•Develop data extraction form.

Conducting the Review

•Search the databases for selecting primary studies.
•Evaluate the primary studies on the basis of inclusion and exclusion criteria.
•Conduct quality assessment of the primary studies.
•Extract the relevant data from the primary studies on the basis of data extraction form.

Report the Results

•Synthesize and summarize the extracted data.
•Interpret the results with respect to research questions.
•Report gaps and future research recommendations.

Figure 4.
Steps of Systematic Review.

Table 4.
Research Questions.

No.	Research Question	Motivation
RQ1	Which XAI tools and techniques demonstrate the greatest efficacy in improving clinician trust and decision-making for AD patients?	This research focuses on the various XAI techniques that have been used in the diagnosis of AD. It tries to assess their effectiveness and how they balance between the supply of precise forecasts and simple mechanisms to explain them to healthcare professionals and patients.
RQ2	What patient-centric XAI interfaces facilitate greater usability and adoption in real- world clinical settings for AD diagnosis?	This subject concerns explanation interfaces emphasizing patient- and caregiver-centered approaches. The purpose is to make it possible to understand and quickly apply the information that XAI systems offer to patients so that patients become more involved in decision-making processes.

Data for this study were obtained from peer-reviewed journals using sources such as PubMed, Scopus, IEEE Xplore, Science Direct, Springer Link, Google Scholar, Wiley Online Library, ACM Digital Library, and Taylor and Francis. This approach accessed all records from the various fields and sources where the investigation was carried out. All the selected studies meet the inclusion criteria. Only the articles published in the last five years have been incorporated into this paper to include the most recent developments in AI and AD.

Meta-analyses and systematic reviews that did not address the application of AI-based AD diagnostic models in clinical practice were not included due to the research focus and practical use of the findings. Search strings and Boolean operators are shown in Table 5.

Table 5.
Search Strings and Boolean Operators.

No. of Databases	Searching String
IEEE Xplore	"explainable AI" AND "Alzheimer's Disease" OR "AD" AND "machine learning"
ScienceDirect	"XAI" AND "diagnosis" AND ("machine learning" OR "deep learning" OR "interpretability")
SpringerLink	"explainable artificial intelligence" AND "Alzheimer's" AND "clinical diagnosis"
Google Scholar	("interpretability" OR "classification") AND "Alzheimer's Disease" AND "AI models"
ACM Digital Library	"explainable AI" AND "diagnosis" AND "clinical applications" AND "Alzheimer's research"
Taylor and Francis	"AI models" AND "Alzheimer's disease diagnosis" AND ("frameworks" OR "methodologies")
PubMed	"explainable AI" AND "AD" AND "cognitive function" AND "patient care" AND "diagnostic tools"

3.3.1. Criteria for Exclusion

The subsequent studies were excluded from the review:

1. Exclusion of Superfluous Emphasis: To keep the research analysis manageable and the identified papers relevant, we excluded papers that did not report the identification or diagnosis of AD.
2. Exclusion from Research Article: To filter out irrelevant studies, we excluded the articles that did not present XAI frameworks or methods for understanding the AI-based categorization of AD.
3. Date Limitations: We excluded papers with a publish date of more than five years ago to focus on recent developments. Review papers focusing solely on a specific class of AI models (e.g., DL) were excluded to avoid bias.
4. Exclusion of Particular Data Types:

- Studies relying solely on a single data type (e.g., MRI) were excluded.
- Studies that focused only on one XAI framework or interpretability method were excluded to ensure a broader analysis.
- Studies not detailing the practical application of AI in Alzheimer's care were excluded for lacking clinical relevance.

5. Exclusion of Non-Empirical Work: Non-empirical and qualitative articles with unclear or poorly described methodologies were excluded to ensure quality.
6. Exclusion of Secondary Publications: Only the first, most authoritative, and reliable reports were considered.
7. Exclusion of Studies Without Baseline Comparison:

- Studies that did not compare AI-based diagnosis to traditional methods were excluded.
- Articles lacking interpretability insights for AD patterns were excluded.

8. Exclusion of Unvalidated Studies: Papers without scientific confirmation of AI models were not included.

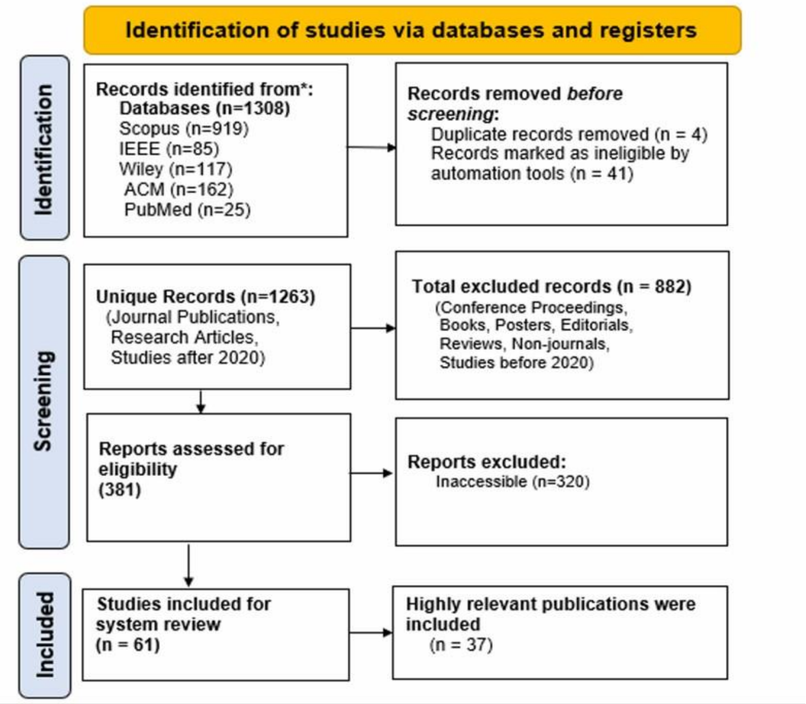

Identification of studies via databases and registers
Identification
Records identified from*:
Databases (n=1308)
Scopus (n=919)
IEEE (n=85)
Wiley (n=117)
ACM (n=162)
PubMed (n=25)
Records removed *before screening*:
Duplicate records removed (n = 4)
Records marked as ineligible by automation tools (n = 41)
Screening
Unique Records (n=1263)
(Journal Publications, Research Articles, Studies after 2020)
Total excluded records (n = 882)
(Conference Proceedings, Books, Posters, Editorials, Reviews, Non-journals, Studies before 2020)
Reports assessed for eligibility (381)
Reports excluded:
Inaccessible (n=320)
Included
Studies included for system review (n = 61)
Highly relevant publications were included (n = 37)

Figure 5.
PRISMA flowchart illustrating the identification, filtering, and inclusion of articles.

9. Exclusion Due to Lack of Novelty: Research without novel ideas or modifications to existing XAI methods for AD detection was excluded.
10. Lack of Collaboration: Papers that did not show cooperation between clinicians and AI experts were excluded.
11. Insufficient Outcome Measures: Trials without clear evaluation metrics (e.g., AUC, accuracy) were excluded from the analysis.

3.3.2. Screening of Articles, Data Extraction, and Analysis

Figure 5 illustrates the PRISMA 2020 flow diagram used to guide the study selection process. An initial search of both databases provided 1,308 reaches in total. Using a screening process consistent with our inclusion-exclusion criteria, four duplicate papers were also excluded, as were 41 papers identified by automation tools as ineligible. This gave 1,263 record reaches, including journal articles, research articles, and studies published after 2020. Nevertheless, 882 studies were excluded as conference proceedings, books, posters, editorials, reviews, non-journal articles, or articles published before January 2020. When we screened the remaining reports for inclusion, an additional 381 reports were excluded, which included 61 reports that could not be obtained.

Finally, our systematic review included 320 articles that met the criteria for sufficient reporting quality and aligned with our proposed research questions. These studies were then meticulously reviewed based on their relevance and the quality of available data. This allowed us to provide comprehensive and accurate model performance metrics for the ML and DL models employed, specifically including accuracy, F1 score, specificity, sensitivity, and the area under the curve (AUC). Among these, only 61 papers were deemed highly relevant.

Further filters were applied to refine our focus specifically on XAI for AD diagnosis. This more thorough evaluation led to the inclusion of 37 papers in our final review. We eliminated articles that did not meet our stringent quality standards, such as those lacking accessible data, papers that discussed the topic without presenting empirical data or model performance, and those unrelated to our core research areas.

3.3.3. Data Extraction

After the last set of primary studies that satisfied the quality rating criteria was identified, relevant data were gathered to answer the research questions. Table 6 lists the data extraction techniques used here.

3.4. Results

This systematic review encompassed 37 studies that reported methodologies utilizing XAI for diagnosing AD through various ML approaches. The studies employed various types of ML, with DL methodologies being the most frequently investigated.

Table 6.
Data Extraction Form.

Data Item Extracted	*Description*
Bibliographic Information	Title, Author, Year, Source
Type of Article	Journal, Workshop Paper, Conference Article
Field of Study	The specific domain of code clone research explored in the study
Characteristics of Source Code	The attributes of source code considered for implementing a ML technique
Utilized ML Algorithm	The category of ML algorithm applied in the research
Training Methodology	The approach used for training the model
Validation Methodology	The validation technique employed to evaluate model accuracy
Assessment Criteria	The metrics used to measure the performance of the ML algorithm
Dataset	The dataset utilized in the study for evaluating ML methodologies
Clone Detection Instrument	The existing clone detection technology employed for clone detection or for benchmarking the proposed approach
Comparative Analysis	The ML methods compared within the study

These models demonstrated utility in handling large data inputs, such as neuroimaging and clinical profiles. Generative clustering and conventional methods, including SVM and RF, also featured prominently. Less common were exploratory techniques like contingency analysis, though they were employed for identifying relationships within datasets. A majority of the studies were published between 2019 and 2024, reflecting a burgeoning interest in XAI's role within healthcare. This trend underscores the escalating importance of interpretability in AI systems applied to clinical decision-making.

Explainability in XAI is categorized as either *global* (explaining the entire model) or *local* (explaining specific predictions). For example, Decision Trees (DTs) inherently provide global explanations, while local explanations examine specific prediction branches. Combining local insights allows for a deeper understanding of overall model behavior [16]. In AD diagnosis, DTs help illustrate the influence of age, biomarkers, and cognitive scores on patient outcomes.

4. Review of Existing Literature

This review compares the most prominent XAI frameworks applied to AD diagnosis. The intention here is to examine the strengths, weaknesses, and use cases of these frameworks to determine how these frameworks could be improved to achieve more interpretability in using AI to design diagnostic tools. We explore these frameworks to argue for their use, e.g., LIME, SHAP, Grad-CAM, LRP, and occlusion sensitivity analysis, in solving some of the problems with AD diagnosis, e.g., explaining predictions to clinicians and patients. The comparative analysis of the AD systems, as shown in Table 7, summarizes the comparison between various frameworks regarding their applicability and their restrictions on AD detection.

Table 7.
Comparison of XAI Frameworks: Strengths, Weaknesses, and Use Cases in Medical Diagnosis

XAI Frame-work	Strengths	Weaknesses	Use Case Example
LIME	Model-agnostic, pro- vides local explanations, easy to implement.	Computationally intensive, inconsistent explanations for similar predictions.	Interpreting black-box models predicting dementia progression, highlighting hippocampal atrophy in AD diagnosis.
SHAP	Consistent, theoretically grounded, provides global and local explanations.	Computationally expensive, complexity increases with features.	Assessing feature importance in AD risk models during clinical trials.
Grad-CAM	Provides visual explanations for CNNs, highlighting important image regions.	Limited to CNNs, subjective interpretation.	Identifying important regions in brain MRI scans for AD diagnosis.
LRP	Detailed explanations for neural networks, applicable to various networks	Complexity increases with model depth, challenging implementation.	Unraveling DL decisions in AD classification tasks.
Occlusion Sensitivity Analysis	Simple to understand and implement, highlights important features.	Slow, computationally demanding, less precise explanations.	Identifying influential brain regions by occluding image segments in AD progression analysis.

SHAP offers strong feature attribution but is too computationally expensive for real-time use. In contrast, LIME is computationally light but provides inconsistent explanations, which lowers clinician confidence. Grad-CAM is useful for MRI interpretation but cannot be applied to tabular data, while LRP offers deep insights but can be challenging for non-experts to interpret. Therefore, hybrid models that combine SHAP's attribution with Grad-CAM's visualization are recommended for improved clinical usability. Tables 8, 9, 11 and 10 provide a structured overview of key studies applying

XAI techniques to the diagnosis of AD. These tables capture the diversity of AI approaches and emphasize the increasing emphasis on both precision and interpretability in the diagnosis of AD.

The main taxonomy of XAI methods in AD prediction is proposed as shown in Figure 6. The taxonomy categorizes a set of key elements and highlights the techniques used to ensure interpretability, ethical transparency, trustworthiness, and decision support. This proposed taxonomy represents an initial research step toward guiding future directions in the application of XAI for AD diagnosis.

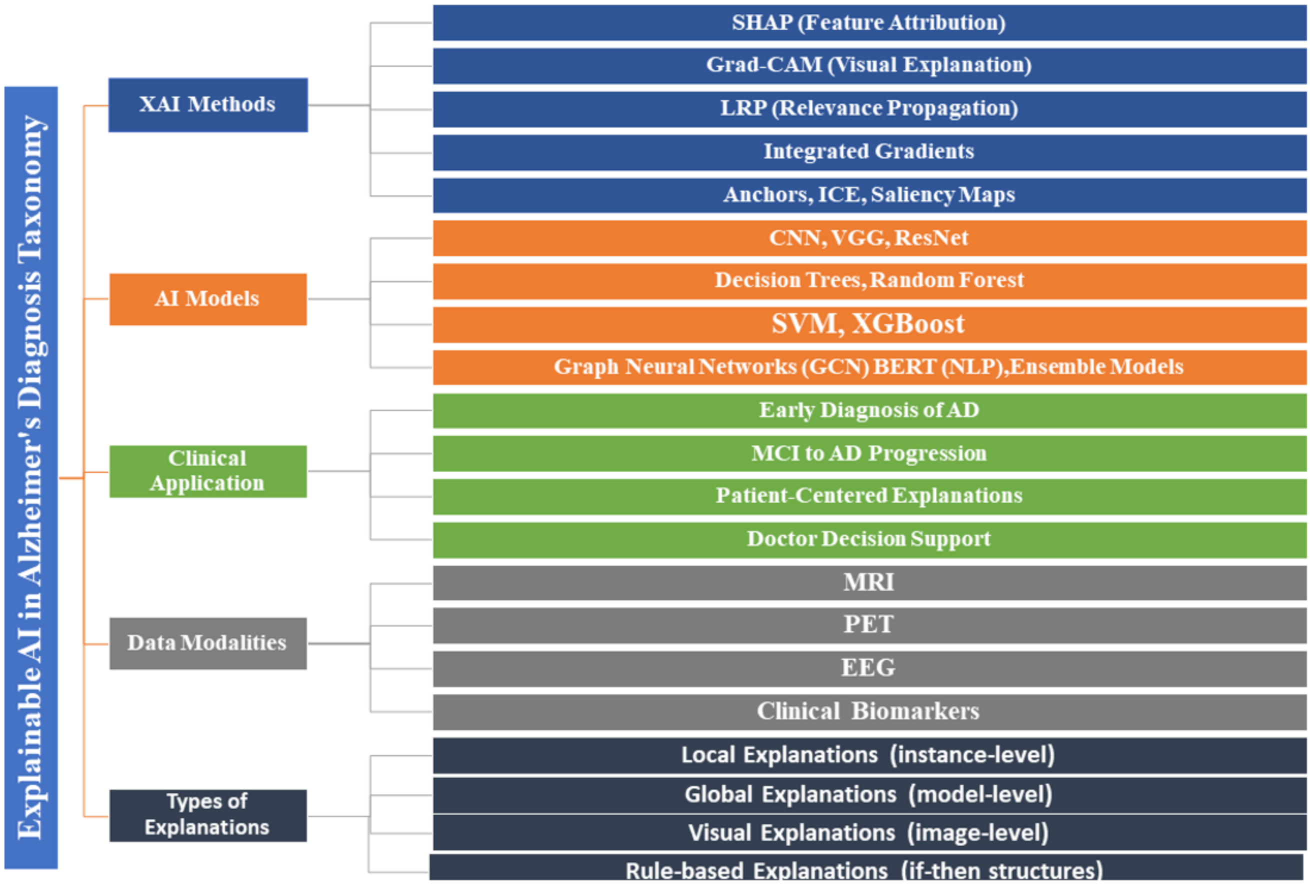

Explainable AI in Alzheimer's Diagnosis Taxonomy
XAI Methods
SHAP (Feature Attribution)
Grad-CAM (Visual Explanation)
LRP (Relevance Propagation)
Integrated Gradients
Anchors, ICE, Saliency Maps
AI Models
CNN, VGG, ResNet
Decision Trees, Random Forest
SVM, XGBoost
Graph Neural Networks (GCN) BERT (NLP),Ensemble Models
Clinical Application
Early Diagnosis of AD
MCI to AD Progression
Patient-Centered Explanations
Doctor Decision Support
Data Modalities
MRI
PET
EEG
Clinical Biomarkers
Types of Explanations
Local Explanations (instance-level)
Global Explanations (model-level)
Visual Explanations (image-level)
Rule-based Explanations (if-then structures)

Figure 6.
Taxonomy of XAI in AD.

Table 8.
Comprehensive Overview of XAI-Based AD Studies: Combined Performance Metrics, Frameworks, Algorithms, Datasets, and Novel Contributions (Part 1/4).

Ref	XAI Framework	ML Algorithm	Dataset	Accuracy	Novel Contributions
Viswan, et al. [1]	SHAP, LIME, Grad- CAM, LRP	CNN, RNN, SVM, RF, XGBoost	ADNI, OASIS, Kaggle (various)	Varies by study (Systematic Review)	Comprehensive categorization of AI models & XAI techniques; identified strengths/weaknesses; emphasized patient- centric explanations.
Danso, et al. [10]	SHAP	Random Forest, XG- Boost	SHARE, PREVENT	AUROC=96% (SHARE), +16.9% GA (PRE- VENT)	SHAP-based ensemble for dementia pre- diction; used transfer learning; identified modifiable risk factors visually.
Kamal, et al. [11]	SHAP, ICE, Break- down	RF, SVM, XGBoost	Clinical, ApoE, Neuropsychological	80.7%	Combined multiple XAI techniques
Vimbi, et al. [12]	LIME, SHAP	CNN, SVM, RF, XGBoost (various ML/DL models)	ADNI, OASIS, Kaggle, ADReSS Challenge	Not reported as numeric accuracy	Systematic review of 23 studies applying LIME and SHAP to AD diagnosis; discusses capabilities, input modalities, strengths, limitations, and research gaps; emphasizes future need for patient-centric explanations and large-scale

					validation
Chun, et al. [13]	SHAP, ICE	XGBoost, SVM, RF, LR	Samsung Medical Center (760 aMCI patients)	AUC=0.83, Accuracy=0.76, F1=0.65	Used XGBoost + SHAP/ICE to predict dementia in aMCI; highlighted key neuropsychological/genetic features; offered personalized explanation insights.
Xu and Yan [15]	SHAP (local and global)	RN-SSAS (custom robust classifier)	Cognitive Scores dataset (4 features, multi-class, imbalanced, small sample)	F-measure = 0.878	Proposed RN-SSAS model for imbalanced, multi-class AD screening using only cognitive scores; integrated SHAP to provide interpretable, instance-level explanations for early-stage AD diagnosis
Yu, et al. [16]	Custom XAI Tool with High-Resolution Heatmaps	CNN with Attention Mechanisms	Brain MRI scans	Higher than state-of-the-art (Exact not specified)	Developed an explainable neural net- work combining multi-scale attention- based CNN and novel XAI tool; offered clear visualizations and prediction-basis retrieval for neurologist validation.
Ilias and Askounis [17]	LIME	BERT	ADReSS Challenge, voice transcripts	86.25%	First use of BERT with LIME for AD detection
Alsubaie, et al. [18]	SHAP, Grad-CAM, LIME (surveyed)	CNN, RNN, GAN, Autoencoder,3D CNN, Transfer Learning	ADNI, AIBL, OASIS, MIRIAD, Kaggle	70%–99.95% (varies by study)	Comprehensive review of deep learning models for AD detection via neuroimag- ing; highlighted interpretability challenges; emphasized role of multimodal data and 3D CNNs; called for benchmark datasets and model transparency.
Zhang, et al. [19]	Grad-CAM	3D ResAttNet (Residual Attention DNN)	sMRI (AD vs NC, pMCI vs sMCI)	Competitive with SOTA (exact % not given)	Introduced explainable 3D ResAttNet with self-attention; Grad-CAM for interpretability; end-to-end diagnosis framework high- lighting key brain regions (hippocampus, cortex)
Shad, et al. [20]	LIME	ResNet50, VGG16, Inception v3	Kaggle (T1-weighted MRI)	82.04%–86.82%	Applied CNNs for early AD classification; used LIME to identify contributing brain regions in MRI scans

Table 9.
Comprehensive Overview of XAI-Based AD Studies: Combined Performance Metrics, Frameworks, Algorithms, Datasets, and Novel Contributions (Part 2/4).

Ref	XAI Framework	ML Algorithm	Dataset	Accuracy	Novel Contributions
Lombardi, et al. [21]	SHAP	Not specified	Cognitive test indices (e.g., MMSE)	Not reported	Proposed a robust XAI framework to classify healthy, MCI, and AD subjects using cognitive scores. Introduced longitudinal SHAP analysis to explain feature contributions and track.
Bogdanovic, et al. [22]	SHAP	XGBoost	Medical, cognitive, and lifestyle features (12,000+ subjects)	F1-score: 0.84(PREVENT)	Applied SHAP for both global and local interpretability of XGBoost model trained on large-scale tabular data; introduced a feature influence scheme to validate or refute diagnostic hypotheses; emphasized explain- able ML in supporting early AD diagnosis.

Lai, et al. [23]	SHAP	SVM, XGBoost, RF, LightGBM, KNN, NB, AdaBoost, DT, LR	GEO (Microarray gene expression data)	SVM: AUC =0.879, Accuracy= 0.808, Recall = 0.773, Precision= 0.809	Identified six ER stress-related genes predictive of AD progression; used SHAP to interpret model outputs; performed clustering to define AD subtypes; conducted immune infiltration and enrichment analysis; pro- posed gene-based nomogram and subtype- specific small-molecule compounds.
Jain, et al. [24]	Grad-CAM (Visual XAI)	CNN, VGG-16, VGG-19, DCGAN (augmentation)	MRI scans (original, geometrically transformed, and GAN-augmented)	MCI prediction: 74%, CNN: 82%, VGG-16: 84%, VGG-19: 87%	Introduced D-BAC framework using DC- GAN for data augmentation; classified dementia by severity; used Grad-CAM for visual interpretability; demonstrated improved performance using progressive resizing and augmentation.
Bloch, et al. [25]	SHAP, Data Shapley (TMC)	RF, XGBoost	ADNI, AIBL	65.36 % (AIBL), +3.61% over base (ADNI)	Used Data Shapley for subject selection; evaluated SHAP visualizations; compared with LOO and random exclusion.
Yue, et al. [26]	SHAP	Ensemble Models	China Longitudinal Aging Study (CLAS)	99.2% (AD/NC), 89.2% (MCI/NC)	High-accuracy AD/MCI prediction using ReliefF & SHAP; clinical insights; explained model with SHAP force/summary plots.
El-Sappagh, et al. [27]	SHAP, DT, Fuzzy Rule-Based Systems	RF, RFE,SVM, KNN, NB	ADNI (baseline)	93.95% (AD), 87.08% (MCI→AD)	Multilayer fusion model with 11 data types; interpretable AI via SHAP+FRBS; natural language explanations.
Ruengchaijatuporn, et al. [28]	Self-Attention, Attention Rollout, Grad-CAM	CNN-based Multi- Input Attention Network	Digital Drawing Tasks (918 elderly)	High AUC/F1 (unspecified)	MCI detection via drawing tasks; interpretable heatmaps; improved over Grad- CAM and baseline CNNs.
Tekkesinoglu and Pudas [29]	Decomposition-Based Explanation, SHAP	Graph Convolutional Network (GCN)	ADNI; validated by 11 clinical experts	Accuracy ~80%, 71% expert agreement	Proposed graph-specific XAI method; explanations were stable and interpretable; outperformed SHAP in efficiency; validated via clinician survey.
Mahmud, et al. [30]	Saliency Maps, Grad- CAM	VGG16, V G G 19, DenseNet169, DenseNet201, Ensemble CNNs	OASIS-2 (MRI; 6,400 images; 4 AD stages)	Accuracy: 96% (Proposed), 95% (Ensembles)	Applied deep transfer learning for AD diagnosis; integrated Grad-CAM for visual interpretability; addressed class imbalance; highlighted potential for clinical deployment.
Kim, et al. [31]	GNNExplainer	GNN (GCN, GAT, GraphSAGE, GIN), RNN (LSTM, Attention)	ADNI (806), HCP brain connectome	Not explicitly stated; outperformed DNN, SVM	Combined temporal GNN with GNNEx- plainer for AD prediction; used longitudinal data; identified neuroanatomical biomarkers with clinical relevance.

Table 10.
Comprehensive Overview of XAI-Based AD Studies: Combined Performance Metrics, Frameworks, Algorithms, Datasets, and Novel Contributions (Part 3/4)

Ref	XAI Framework	ML Algorithm	Dataset	Accuracy	Novel Contributions
Chethana, et al. [32]	Grad-CAM	Custom CNN and RNN	Kaggle AD Dataset (4-class MRI images)	CNN: 95.06%, RNN: 79.71%	Developed a custom CNN-RNN hybrid; CNN outperformed baseline models; applied Grad-CAM for brain region localization; deployed a Streamlit GUI for interpretability.
Pohl, et al. [33]	LRP (custom propagation rules)	Deep Neural Networks (CNN)	MRI scans (AD vs. healthy controls)	92%	Proposed a DNN for AD diagnosis; introduced three new LRP rule configurations; enhanced interpretability and evidence visualization to support clinical decision-making.
Xu and Yan [15]	SHAP (instance-level and global explanations)	Multi-class ML classifier (RN-SSAS)	Cognitive Scores (4 types)	F-measure = 0.878	Developed a robust model for small, imbalanced datasets using cognitive scores; enhanced model trust with SHAP-based local and global interpretability; aimed to increase AD screening rates.
Rahim, et al. [34]	Visual Explainability, Custom XAI	3D CNN, BRNN	Longitudinal 3D MRI, Demographic and Cog-native Scores	Accuracy=96%, Precision=99%, Recall=92%, AUC=96%	Hybrid multimodal deep learning model combining 3D CNN and BRNN; fused MRI with demographic and cognitive features; introduced a novel visual explainability module highlighting relevant brain regions for AD progression.
De Santi, et al. [35]	Saliency Map, LRP	3D CNN	ADNI (18F-FDG PET,2552 scans)	AUC: 0.81 (CN), 0.63 (MCI), 0.77 (AD)	Developed a 3D CNN for multiclass classification of PET scans (CN, MCI, AD); evaluated LRP vs. Saliency for explanation quality; LRP better mapped relevance to anatomical regions using Talairach Atlas.
Kamal, et al. [11]	LIME	CNN, SpinalNet, KNN, SVC, XG-Boost	MRI images, microarray gene expression data	CNN: 97.6%, SVC: Highest for genes	Multimodal AD classification using MRI and gene expression; introduced LIME for gene-level interpretability; first to integrate both modalities with trustworthy explanations for gene contribution analysis.
Jahan, et al. [36]	SHAP	RF, LR, DT, MLP, KNN, GB, AdaB, SVM, NB	OASIS-3	RF: 98.81% (10-fold CV)	Proposed a multimodal five-class AD classification model using clinical, MRI segmentation, and psychological data; applied SHAP for model interpretability; introduced a novel patient management architecture.
Hernandez, et al. [37]	SHAP	Random Forest, Gradient Boosting, Tree-based Ensembles	TADPOLE (ADNI)	Not explicitly reported (focus on feature importance, not accuracy)	Investigated the top 3 TADPOLE challenge models using a unified framework; applied SHAP for feature attribution; assessed alignment between model decisions and clinical knowledge; emphasized the need for interpretability in AD progression models.

Alatrany, et al. [38]	SHAP, LIME, Rule-based Explanations	Support Vector Machine (SVM)	National Alzheimer's Coordinating Center (NACC)	F1 Score: 98.9% (binary), 90.7% (multiclass), 88% (4-year binary prediction), 72.8% (4-year multiclass prediction)	Developed interpretable SVM-based model for AD diagnosis and progression prediction; introduced rule-extraction techniques for human-understandable explanations; validated clinical relevance of key cognitive features using SHAP and LIME; emphasized importance of Clinical Dementia Rating metrics.

Table 11.
Comprehensive Overview of XAI-Based AD Studies: Combined Performance Metrics, Frameworks, Algorithms, Datasets, and Novel Contributions (Part 4/4)

Ref	XAI Framework	ML Algorithm	Dataset	Accuracy	Novel Contributions
Ekuma [39]	SHAP (SHapley Addative exPlanations)	Deep CNNs (DenseNet121, DenseNet169, Inception-ResNet-v2)	MRI dataset with varying AD severity, APOE genotype, and neurocognitive scores	High sensitivity and specificity (exact values not specified)	Developed an explainable CNN framework for predicting AD severity; used SHAP to localize brain regions and correlated misclassifications with APOE and cognitive scores to improve interpretability
Yang, et al. [40]	SHAP, LIME, Grad-CAM, LRP, Saliency Maps	CNNs, RNNs, GNNs, SVM, RF, XGBoost (various)	ADNI, OASIS, AIBL, Kaggle MRI datasets (various)	Not applicable (Survey)	Provided a comprehensive taxonomy of XAI techniques in AD diagnosis; analyzed model interpretability, datasets, and clinical applicability; identified challenges and future directions for explainable AD prediction systems
Vimbi V. et al [41]	SHAP (SHapley Additive exPlanations), LIME (Local Interpretable Model-agnostic Explanations)	XGBoost, RF, LR, SVM, KNN	ADNI clinical and neuropsychological data	Accuracy up to 91% (LR)	Compared performance and interpretability of five ML models on ADNI data using SHAP and LIME; highlighted model-specific feature attribution for Alzheimer's diagnosis; emphasized importance of transparent AI in healthcare decision-making
Cruciani [42]	Model-agnostic XAI; Stability, Consistency, Understandability, Plausibility validation	AI models for neurodegeneration analysis (unspecified ensemble of morphological, microstructural, and functional models)	Imaging and genetic datasets in neuroscience (specific dataset not named; possibly AD-related)	Not explicitly stated	Proposed neuroscience-specific XAI guidelines and defined four novel XAI validation attributes; applied XAI to decode brain modulations in neurodegeneration and evaluated explanation alignment with literature
Eitel [43]	Heatmap-based XAI (e.g., Saliency Maps) for visual explanation	CNN architectures for neurodegenerative disease diagnosis	Brain MRI datasets (specific datasets not named; focus on neurodegenerative conditions, likely including AD)	Not explicitly stated	Conducted five experimental studies validating CNN use in neurological MRI; introduced a custom CNN architecture optimized for brain imaging; used XAI heatmaps to align model outputs with clinical findings and improve interpretability

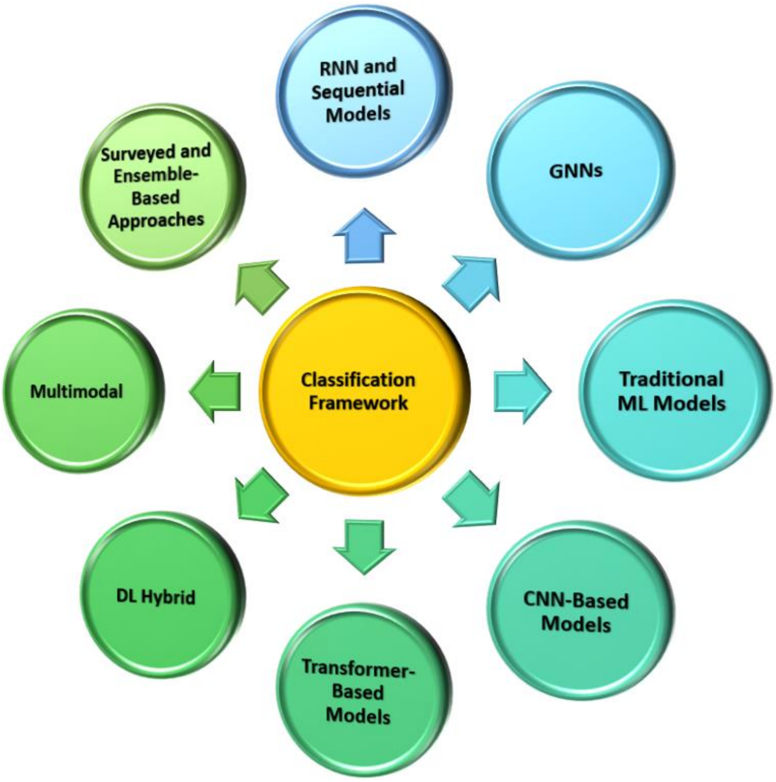


Figure 7.
Classification Framework for ML Models in AD Research with XAI.

4.1. Classification Framework for ML Models in AD Research with XAI

To get a clearer picture of the patterns in the studies we reviewed, we classified the included ML models into eight different categories according to their architectural characteristics and methodological design. This classification enables us to have a clearer understanding of how various ML methods contribute to diagnostic accuracy and explainability in AD research. The following is a breakdown of each group, along with representative studies and Figure 7 showing the XAI classifications of AD research.

4.1.1. Convolutional Neural Network (CNN)-Based Models

CNN architectures dominated the reviewed studies, particularly in neuroimaging tasks involving MRI or PET scans. Models such as VGG16, ResNet, DenseNet, and custom CNNs were frequently used in conjunction with visual explanation tools such as Grad-CAM and LRP. Representative studies include [19, 20, 24, 28, 30, 35, 43].

To begin with, Zhang, et al. [19] presented an explainable 3D residual self-attention deep neural network for AD using structural MRI scans. With regard to the sample size, the study does not provide enough information, which in turn limits the assessment of how easily the results can be generalized to large populations. Moreover, the absence of a detailed description of statistical preprocessing methods raises concerns about the integrity and reliability of the dataset. In addition, the authors' failure to discuss validation techniques undermines confidence in the model's strength and stability. Second, the sole use of Grad-CAM for interpretability can miss showing the full range of the decision-making process of the model. Third, the study does not discuss biases in data or model construction, which is troubling for the fairness and transparency of the research. In contrast, Shad, et al. [20] developed an explainable deep learning approach for early AD using T1-weighted MRI scans. They applied Convolutional Neural Network (CNN) models (ResNet50, VGG16, Inception v3) on a Kaggle-based hybrid dataset, achieving up to 86.82% accuracy. To interpret predictions, they integrated Local Interpretable Model-Agnostic Explanations (LIME), enabling the localization of brain regions that contribute to classification decisions, thus improving the transparency of the model for clinical relevance.

Similarly, Jain, et al. [24] implemented a DL framework for dementia classification using MRI scans and integrated data augmentation based on DCGAN (D-BAC) with CNNs VGG-16 and VGG-19. The authors evaluated three datasets: original, geometric transformation-augmented, and GAN-augmented MRI images. Their results indicated that augmentation yielded better predictive performance. The model demonstrated testing accuracies of 82%, 84%, and 87% utilizing CNN, VGG-16, and VGG-19 in that order. For interpretability, visual feature relevance was provided using Grad-CAM. This research illustrates the impact of applying generative models with explainable AI to enhance the early detection of dementia and MCI.

Building on CNN-based diagnostics, Ruengchaijatuporn, et al. [28] incorporated multiple drawing features into existing CNN-based diagnostic modeling (e.g., clock drawing, cube-copying, trail-making), which led to improvements in classification outcomes alongside interpretability of the resulting models. Although useful, their claims lack external dataset validation, which greatly limits their generalizability, alongside the limited use of Grad-CAM for visually driven explanatory narration reduces the robustness of their claims pertaining to the overall strength of the model's interpretability.

On that note, Mahmud, et al. [30] conducted a work that only used Grad-CAM for visual interpretability of AD. However, the study lacks information on participant number and dataset origin, which hinders understanding of sample representativeness and sufficiency. Furthermore, the absence of information on data preprocessing procedures raises concerns about data quality and the reliability of the conclusions. Methodologically, the reliance on a single XAI technique (Grad-CAM) limits the comprehensiveness of the interpretability framework. Finally, the study does not address potential biases in the dataset or model development process, which are essential considerations for clinical applicability and trustworthiness.

In a more comprehensive endeavor, De Santi, et al. [35] developed a 3D Convolutional Neural Network (CNN) to classify 18F-FDG PET scans from the ADNI dataset into three diagnostic categories: Cognitively Normal (CN), MCI, and AD. The model achieved Area Under the Curve (AUC) results of 0.81 for CN, 0.63 for MCI, and 0.77 for AD. In order to address the black-box issue associated with AI-driven frameworks, two post hoc explainability methods, namely, separation maps and layerwise relevance propagation (LRP), were utilized. A quantitative evaluation demonstrated that LRP better localized relevance to anatomically meaningful brain regions, as validated using the Talairach Atlas.

Finally, Eitel [43] evaluated the use of neural networks (CNNs) to diagnose neurodegenerative diseases using MRI data in five experimental studies. The work emphasized explainability by generating heat maps using XAI methods to visualize important regions of the brain. The results showed high classification performance and clinical relevance, supported by alignment between the model explanations and the medical literature. A novel CNN architecture optimized for spatial MRI features was also developed.

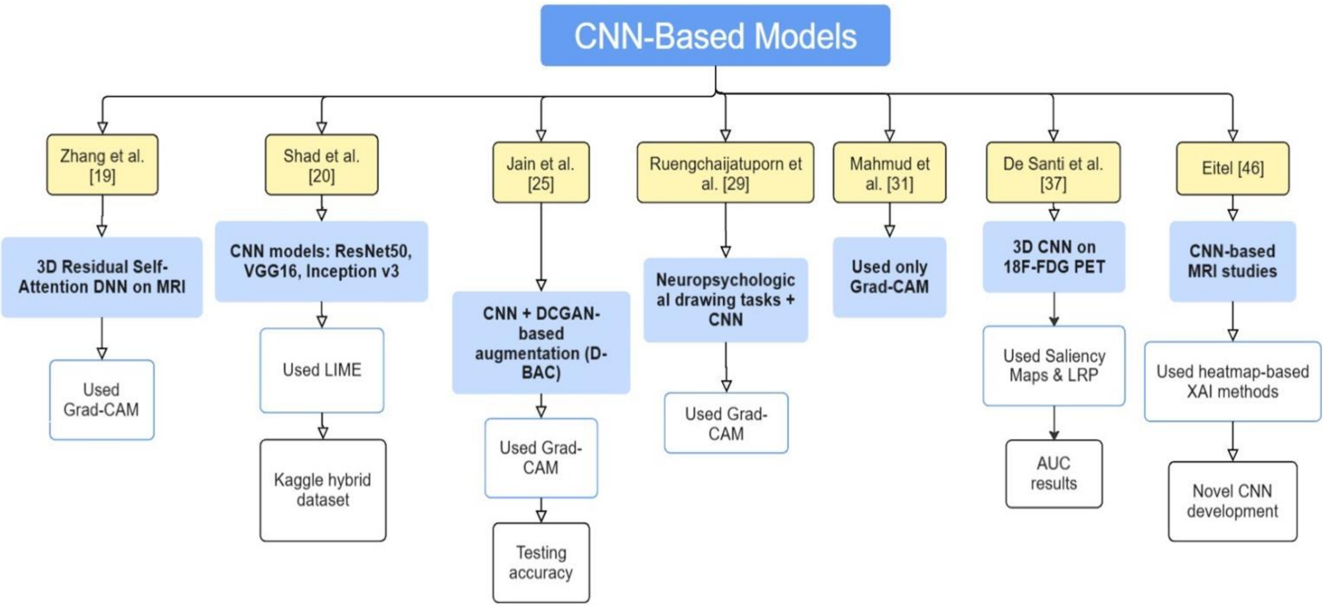


Figure 8.
Overview of CNN-based models applied in AD studies, highlighting architectures (e.g., ResNet, VGG16, 3D CNN), datasets, and explainability methods (Grad-CAM, LIME, Saliency Maps, LRP) used across selected works.

4.1.2. Recurrent Neural Network (RNN) and Sequential Models

RNNs and their variants (LSTM, BRNN) were applied to handle sequential or longitudinal data, often in multimodal pipelines. These models were particularly effective in modeling disease progression, and were typically paired with CNNs and inter- pretability techniques like Grad-CAM. Rahim, et al. [34] developed a DL model that combined a 3D-CNN with a BRNN to capture both spatial and temporal dependencies in the sequence of longitudinal MRI scans. Along with demographic, cognitive, and other relevant features, they fused these biomarkers methodically, resulting in strong performance (96% accuracy). Moreover, they added a visual XAI component that highlighted brain regions in images corresponding to those marked by specialists and corroborated with their interpretations.

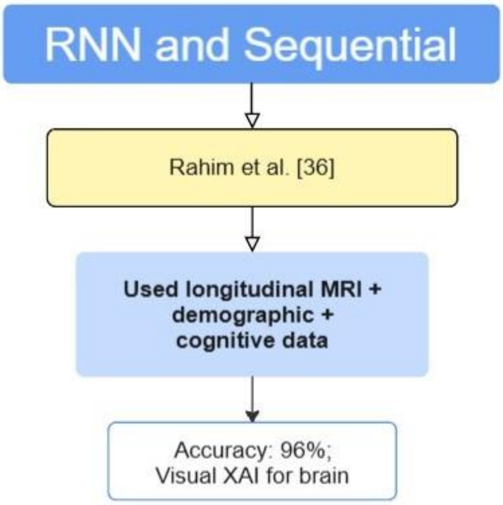

RNN and Sequential
Rahim et al. [36]
Used longitudinal MRI + demographic + cognitive data
Accuracy: 96%; Visual XAI for brain

Figure 9.
Overview of RNN-based models for AD diagnosis. Rahim, et al. [34] used a hybrid 3D CNN and bidirectional RNN model incorporating longitudinal MRI, demographic, and cognitive data. A visual XAI module was employed to localize relevant brain regions, achieving 96% accuracy.

4.1.3. Traditional ML Models

Tree-based models such as RF, XGBoost, and kernel-based models like SVM were widely used for their interpretability and performance on tabular clinical or cognitive data. These were commonly explained using SHAP, LIME, or ICE. Studies in this category include [6, 15, 21-23, 25, 41].

In a 2022 study, Xu and Yan [15] presented RN-SSAS, an explainable ML model designed to provide predictions for AD based solely on cognitive test scores. The model achieved an F-measure of 0.878 and was tailored to address the challenges associated with small, imbalanced, multi-class datasets. Moreover, to ensure interpretability, the authors employed SHAP to generate both global and instance-level explanations, offering insights into how specific features influenced the model's predictions.

Similarly, Loveleen, et al. [44] proposed an explainable HCI model for AD diagnosis, integrating deep learning classifiers with SHAP and LIME to provide interpretability and insight into the model's decision-making processes. Their framework demonstrated the importance of transparency in medical AI applications while also achieving high classification accuracy, including a 100% success rate using a decision tree model.

Further, Lombardi, et al. [21] also created a robust XAI framework with SHAP for the classification of AD, MCI, and healthy controls from cognitive test data. Their framework emphasized the temporal stability and reliability of SHAP values, providing interpretable explanations of typical contributions and allowing longitudinal monitoring of cognitive decline.

In their study, Bogdanovic, et al. [22] used an XGBoost model trained on a large-scale dataset of more than 12,000 patients with medical, cognitive, and lifestyle characteristics to help early detection of AD. Although achieving a competitive F1 value of 0.84, predictive accuracy was not the focus of the study, but instead model explainability. The authors applied SHAP for both global and local interpretability, developing a feature influence scheme that enabled them to validate or challenge initial diagnostic hypotheses. Their findings highlight the importance of explainable machine learning in uncovering meaningful feature diagnosis relationships, providing clinicians with an additional layer of understanding for the assessment of AD in early stages.

In another contribution to the field, Lai, et al. [23] used interpretable machine learning to investigate the role of stress-related genes in the endoplasmic reticulum (ER) in AD. Using gene expression data from the GEO microarray repository, the authors assessed the predictive performance of nine traditional classifiers, including SVM, XGBoost, LightGBM, and RF, for their ability to model AD progression. The SVM model demonstrated the highest performance with an AUC of 0.879 and an accuracy of 0.808. In addition, SHAP was employed to interpret model outputs, highlighting six key genes (*RNF5*, *UBAC2*, *DNAJC10*, *RNF103*, *DDX3X*, and *NGLY1*) as significant predictors. Moreover, the study identified two AD subtypes through consensus clustering based on ER stress gene expression, each showing distinct immune infiltration characteristics. Connectivity Map (CMap) analysis further suggested potential therapeutic compounds tailored to each subtype. These findings emphasize

Furthermore, Bloch, et al. [25] introduced the integration of Data Shapley (TMC) along with SHAP to guide subject selection and interpret the outputs of RF and XGBoost models. This approach resulted in an accuracy improvement of 3.61% over baseline models. Generally, these studies emphasize the value of traditional ML techniques when combined with XAI methods, highlighting their potential to enhance diagnostic accuracy and improve the clinical interpretability of model decision-making in AD detection.

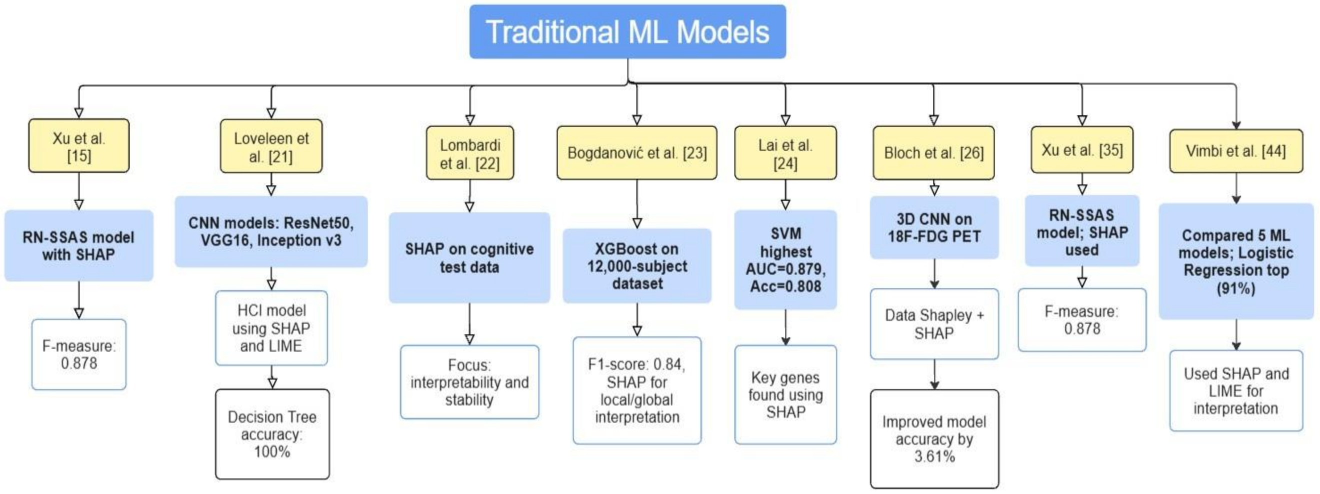

Traditional ML Models
Xu et al. [15]
RN-SSAS model with SHAP
F-measure: 0.878
Loveleen et al. [21]
CNN models: ResNet50, VGG16, Inception v3
HCI model using SHAP and LIME
Decision Tree accuracy: 100%
Lombardi et al. [22]
SHAP on cognitive test data
Focus: interpretability and stability
Bogdanović et al. [23]
XGBoost on 12,000-subject dataset
F1-score: 0.84, SHAP for local/global interpretation
Lai et al. [24]
SVM highest AUC=0.879, Acc=0.808
Key genes found using SHAP
Bloch et al. [26]
3D CNN on 18F-FDG PET
Data Shapley + SHAP
Improved model accuracy by 3.61%
Xu et al. [35]
RN-SSAS model; SHAP used
F-measure: 0.878
Vimbi et al. [44]
Compared 5 ML models; Logistic Regression top (91%)
Used SHAP and LIME for interpretation

Figure 10.
Overview of traditional ML models used in AD diagnosis and their corresponding explainability tools (e.g., SHAP, LIME, Data Shapley). Models include SVM, XGBoost,decision trees, and hybrid CNN-ML approaches applied across varied datasets.

On another front, Xu and Yan [15] addressed the limitations of small-sample, multi-class classification tasks using cognitive scores by developing a robust RN-SSAS framework. Their model achieved an F-measure of 0.878 and incorporated SHAP to provide both global and local interpretability, thereby enhancing clinical trust in the predictions.

Finally, Vimbi V. et al [41] evaluated five machine learning models—LR, RF, SVM, KNN, and XGBoost—using clinical and neuropsychological data from the ADNI dataset to support AD classification. The study prioritized both predictive accuracy and model interpretability. To this end, the authors employed SHAP and LIME to explain model outputs and uncover influential diagnostic features, thereby promoting transparency and clinical trust. LR achieved the highest classification accuracy at 91%. The study emphasizes the importance of balancing interpretability and predictive performance when. implementing AI models for clinical decision-making.

4.1.4. Graph Neural Networks (GNNs)

GNNs such as GCN, GAT, and GraphSAGE were employed to model relationships between brain regions or across time, using connectomic or longitudinal data. XAI tools like GNNExplainer and SHAP were used to interpret model predictions. Relevant studies include [29, 31].

. GNNs are gaining traction in Alzheimer's research due to their ability to model complex relationships—such as brain connectivity or patient similarities—using graph structures. In their study, Tekkesinoglu and Pudas [29] applied a GCN framework to the ADNI dataset to categorize individuals into Normal Cognition (NC), MCI, and AD. To further enhance model interpretability, the authors introduced a decomposition-based explanation method that identifies the contribution of both node and edge influences. Therefore, the model not only achieved notable performance and robustness to input perturbations, but also received favorable validation from clinicians, with 71% rating the explanations as understandable and clinically relevant.

Similarly, Kim, et al. [31] investigated disease progression prediction using longitudinal neuroimaging data by leveraging a GNN model. Their approach demonstrated superior performance compared to conventional models such as Deep Neural Net- works (DNNs) and SVMs. Notably, the study emphasized interpretability by establishing biologically meaningful connections between model predictions and structural brain regions. This highlights the capacity of GNN-based frameworks to extend beyond predictive accuracy, offering clinically relevant insights that enhance transparency and foster trust in AI-driven medical decision-making.

4.1.5. Transformer-Based Models

Transformer architectures, particularly BERT, were used in NLP tasks such as analyzing speech transcripts. LIME was employed to explain predictions in this context. A key study is Ilias and Askounis [17]. Recent advances in transformer architectures have enabled significant progress in applying NLP methods to AD detection. Ilias and Askounis [17] proposed a multi-task framework.

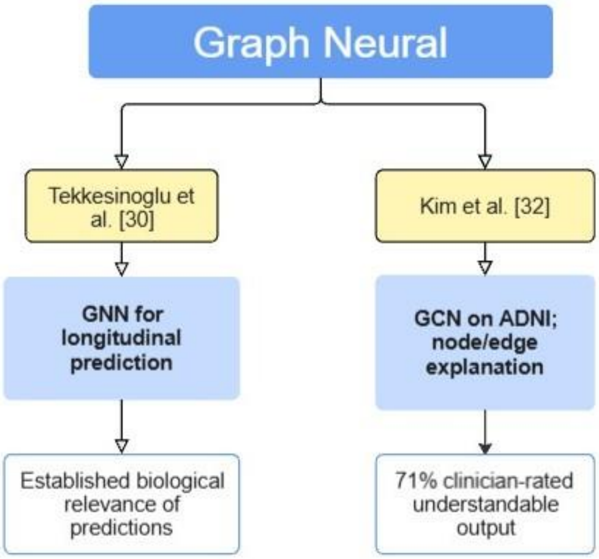

Graph Neural
Tekkesinoglu et al. [30]
Kim et al. [32]
GNN for longitudinal prediction
GCN on ADNI; node/edge explanation
Established biological relevance of predictions
71% clinician-rated understandable output

Figure 11.
Overview of GNN-based approaches for AD diagnosis.

Tekkesinoglu and Pudas [29] applied GNNs for longitudinal prediction, validating biological relevance. Kim, et al. [31] implemented GCNs on ADNI data, achieving 71% clinician-rated understandable explanations through node/edge analysis

leveraging BERT to analyze voice transcripts from the ADReSS Challenge, aiming to identify dementia and predict MMSE scores concurrently. Their work addresses two critical gaps: the limited use of transformer-based networks in AD research and the lack of interpretability in existing NLP models. By achieving an accuracy of 86.25% in the binary classification of AD patients and employing LIME to interpret the linguistic features influencing BERT's predictions, the study contributes to both diagnostic performance and model transparency. Additionally, they conducted a detailed linguistic analysis to uncover significant language use differences between AD and non-AD subjects. As one of the first applications of BERT integrated with LIME for explainable AD detection, this study represents a pivotal shift toward interpretable, transformer-driven diagnostic tools in neurocognitive research.

4.1.6. DL Hybrid and Unspecified Architectures

Some studies proposed hybrid or novel deep learning architectures combining elements of CNNs, autoencoders, or attention mechanisms. These models leveraged high-capacity learning for complex data but varied in explainability integration. Examples include [16, 32, 33].

For instance, Yu, et al. [16] proposed an explainable deep learning framework for brain MRI classification that integrates attention-based CNNs with a custom-designed XAI module. The model applies attention mechanisms to multi-scale feature representations to enhance class-discriminative learning, and incorporates a novel high-resolution visualization method as well as a prediction-based retrieval tool to support clinical interpretability. However, the study does not clearly report the dataset size, patient demographics, or diagnostic subcategories used, which limits the transparency and generalizability of the reported performance. Moreover, while the proposed visualization module claims superiority over existing Grad-CAM-based techniques, the study lacks comparative quantitative evaluation against established XAI methods such as LRP or SHAP. Additionally, there is no discussion on the potential biases in the training data or model predictions, nor any validation involving clinical experts to confirm the usefulness of the retrieved prediction basis. These omissions reduce the robustness of the interpretability claims despite the model's architectural innovation.

Notable examples include Chethana, et al. [32] who proposed a hybrid deep learning framework combining CNNs and RNNs for stage classification of AD using active MRI scans. Their model leverages the spatial strengths of CNNs and the temporal sequence learning capabilities of RNNs. To enhance transparency, the study incorporated Gradient-weighted Class Activation Mapping (Grad-CAM), which provided localized visual explanations of the brain regions associated with each classification outcome. This integration of hybrid architectures with explainable AI mechanisms demonstrates the potential of deep learning models to improve diagnostic accuracy while supporting clinical trust and interpretability. For instance, Pohl, et al. [33] proposed a deep neural network model augmented with customized Layer-wise Relevance Propagation (LRP) rules, achieving a classification accuracy of 92% in distinguishing AD patients from healthy controls. This approach also advanced

evidence visualization, enhancing clinical transparency.

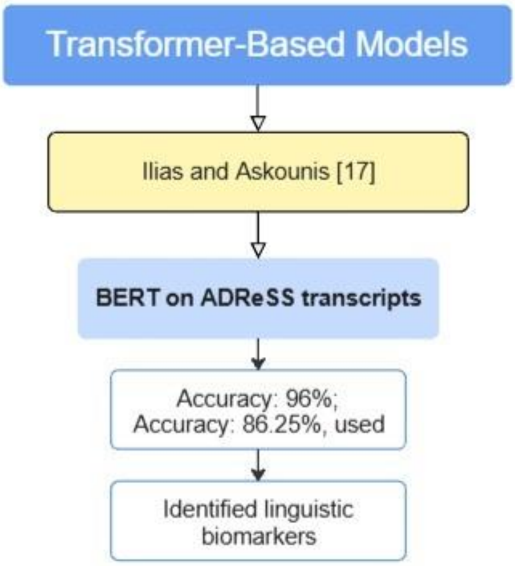


Figure 12.
Overview of transformer-based approaches in AD diagnosis.
Source: Ilias and Askounis [17].

Ilias and Askounis [17] utilized BERT on ADReSS speech transcripts to identify linguistic biomarkers. Their model achieved 96% and 86.25% accuracy in different evaluation settings, highlighting the role of language data in AD detection.

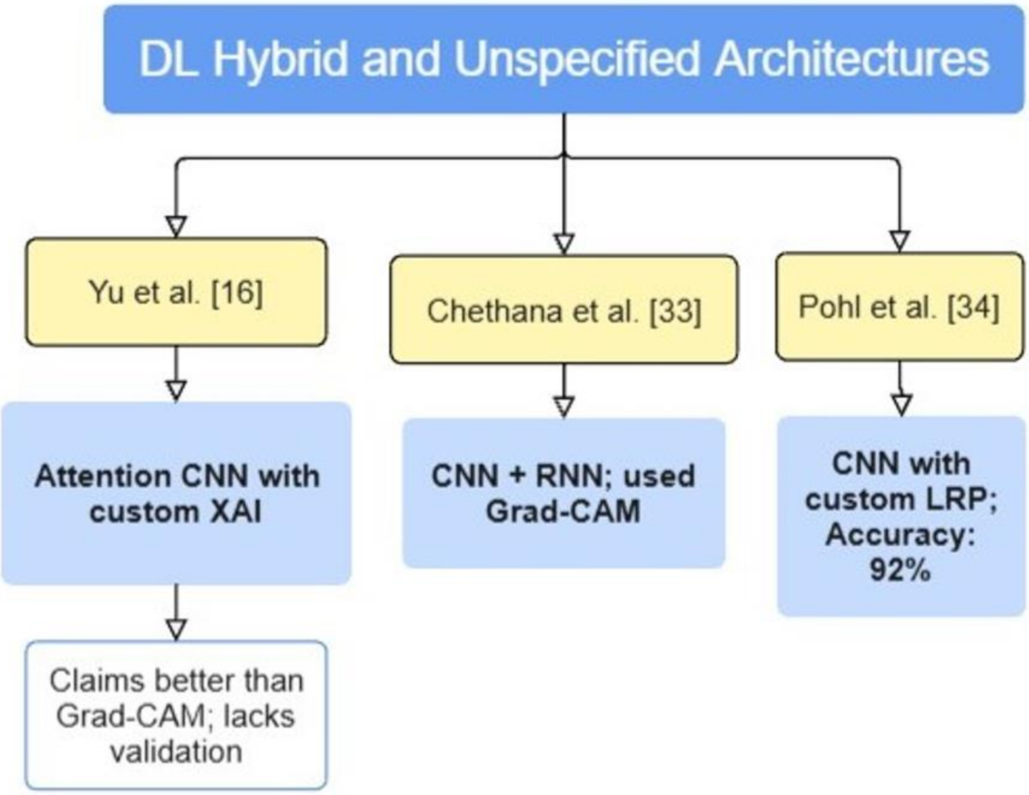

DL Hybrid and Unspecified Architectures
Yu et al. [16]
Chethana et al. [33]
Pohl et al. [34]
Attention CNN with custom XAI
CNN + RNN; used Grad-CAM
CNN with custom LRP; Accuracy: 92%
Claims better than Grad-CAM; lacks validation

Figure 13.
Summary of DL hybrid and unspecified architectures in Alzheimer's Disease diagnosis.
Source: Yu, et al. [16]; Chethana, et al. [32]; Pohl, et al. [33]

Yu, et al. [16] introduced an attention-based CNN model with a custom XAI mechanism claiming superiority over Grad-CAM, though lacking extensive validation. Chethana, et al. [32] combined CNN and RNN architectures and applied Grad-CAM to interpret AD stages from MRI data. Pohl, et al. [33] developed a CNN model with custom LRP, reporting a classification accuracy of 92%.

4.1.7. Multimodal and Fusion-Based Architectures

Several studies integrated multiple data types (e.g., MRI with gene expression, cognitive scores, or demographic information) using model fusion approaches such as CNN+RF or joint feature spaces. These architectures aimed to improve diagnostic robustness and explanation richness. Examples include [11, 27, 36, 39].

Several studies demonstrated the advantages of integrating multiple data modalities to enhance diagnostic performance and interpretability in AD research. For example, Kamal, et al. [11] employed SVM, XGBoost, and KNN to classify AD using gene expression and MRI data, integrating SHAP, ICE, Breakdown, and LIME to identify predictive biomarkers.

El-Sappagh, et al. [27] introduced a comprehensive two-layer fusion model based on random forest classifiers that integrated [11] modalities—including clinical, biological, and cognitive data—from the ADNI dataset. Their model not only achieved high performance (93.95% accuracy in AD diagnosis and 87.08% in MCI-to-AD progression), but also incorporated SHAP and fuzzy rule-based systems to deliver natural-language explanations, making it clinically interpretable and trustworthy.

In another multimodal approach, Kamal, et al. [11] combined MRI images with microarray gene expression data, applying a variety of classifiers including CNN, SpinalNet, SVC, and XGBoost. They employed the LIME framework to offer gene-level explanations and successfully identified key genetic contributors for individual patients. Collectively, these studies illustrate the growing utility of multimodal architectures in producing diagnostic models that are both accurate and meaningfully interpretable, meeting clinical demands for transparency and trustworthiness. Jahan, et al. [36] proposed a novel explainable machine learning framework for the five-class classification of AD using a multimodal dataset comprising clinical, MRI segmentation, and psychological data. The study evaluated nine traditional ML models, with Random Forest achieving the highest performance (98.81% accuracy via 10-fold cross-validation). To enhance transparency, the authors employed SHAP for interpreting model outputs and identifying key predictive features. This work represents one of the first attempts to integrate these three modalities within a trustworthy, interpretable architecture for AD diagnosis and patient management.

Finally, Ekuma [39] proposed an explainable deep learning framework for predicting the severity of Alzheimer's disease (AD) using MRI neuroimaging data. The study evaluated three CNN architectures—DenseNet121, DenseNet169, and Inception-ResNet-v2—for classifying AD severity levels. To enhance interpretability, SHAP were employed to identify the contribution of specific brain regions to model predictions. Furthermore, the authors examined the relationship between model misclassifications and five neurocognitive assessment scores, as well as the APOE genotype biomarker, to understand the clinical factors affecting model performance. The framework demonstrated high sensitivity and specificity, highlighting its potential for reliable, interpretable AD diagnosis.

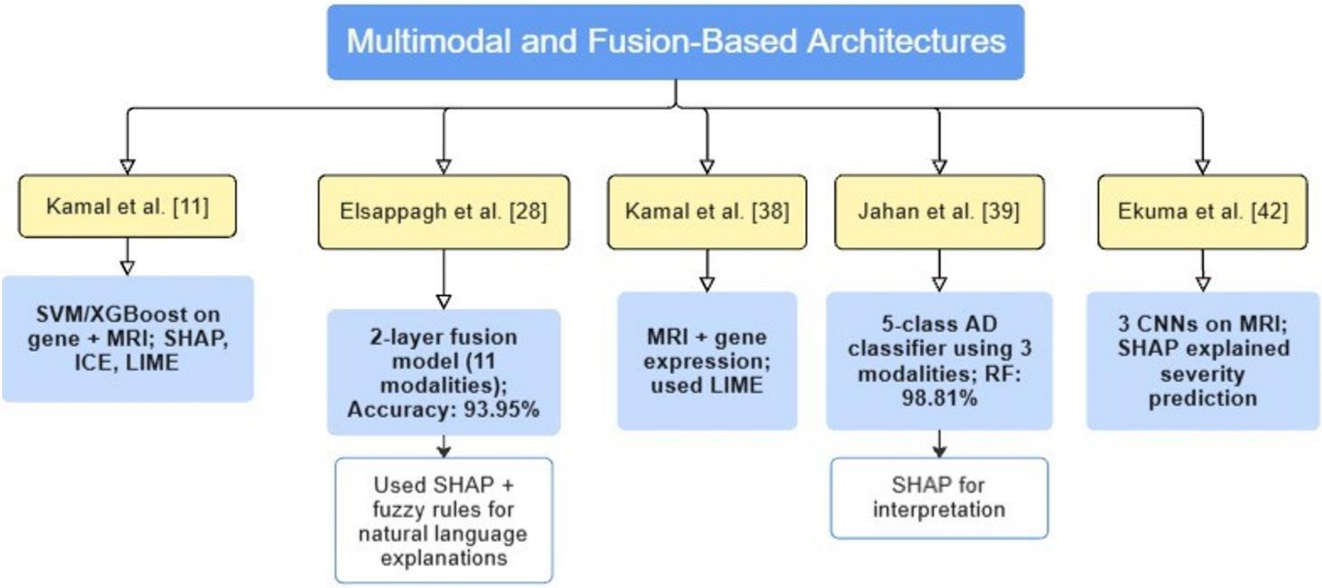

Multimodal and Fusion-Based Architectures
Kamal et al. [11]
Elsappagh et al. [28]
Kamal et al. [38]
Jahan et al. [39]
Ekuma et al. [42]
SVM/XGBoost on gene + MRI; SHAP, ICE, LIME
2-layer fusion model (11 modalities); Accuracy: 93.95%
MRI + gene expression; used LIME
5-class AD classifier using 3 modalities; RF: 98.81%
3 CNNs on MRI; SHAP explained severity prediction
Used SHAP + fuzzy rules for natural language explanations
SHAP for interpretation

Figure 14.
Overview of multimodal and fusion-based architectures in AD diagnosis. Kamal, et al. [11] utilized gene expression and MRI data with SHAP, ICE, and LIME to enhance diagnostic interpretation. El-Sappagh, et al. [27] proposed a 2-layer fusion model combining 11 modalities, integrating SHAP with fuzzy rule-based explanations. Kamal, et al. [11] applied XGBoost with multimodal inputs, focusing on natural language justification. Jahan, et al. [36] developed a five-class AD classifier using random forest with SHAP for interpretation. Ekuma [39] employed CNNs on MRI scans with SHAP to explain severity prediction.

4.1.8. Surveyed and Ensemble-Based Approaches

Review and ensemble-based studies provided meta-level analysis or combined multiple algorithms to boost performance. These works often highlighted strengths and trade-offs between different ML and XAI techniques. Notable references are Yang, et al. [40] and Cruciani [42] diagnostics.

Viswan, et al. [1] presented a systematic review of XAI applications in AD diagnosis, emphasizing the need for model inter- pretability to improve clinical adoption. The study categorized AI models—such as CNN, RNN, SVM, RF, and XGBoost—and interpretability techniques like SHAP, LIME, Grad-CAM, and LRP across datasets including ADNI, OASIS, and Kaggle. It classified XAI approaches into model-specific vs. model-agnostic, and local vs. global methods, providing a broad interpretive spectrum. The authors highlighted strengths, limitations, and the importance of patient-centric explanations, offering key insights for enhancing trust in AI-driven AD diagnostics.

Danso, et al. [10] developed a transfer learning-based ensemble framework for dementia risk prediction, leveraging large-scale longitudinal data and adapting it to younger populations. Using SHAP for model interpretability, their approach achieved a geometric accuracy of 87% and significantly improved prediction metrics in the target domain. The study emphasized clinical utility by enabling personalized and explainable early risk detection across diverse cohorts.

In their systematic review, Vimbi V. et al [41] examined the application of XAI techniques—specifically LIME and SHAP—in the context of AD diagnosis. Following PRISMA and Kitchenham's systematic review guidelines, the authors analyzed 23 peer-reviewed studies that employed these interpretability frameworks with various ML and DL models. The review highlighted the critical role of XAI in improving the transparency and trustworthiness of AD predictions. It also explored the advantages and limitations of LIME and SHAP across diverse datasets and modeling approaches. The authors emphasized the need for patient-centric explanations and robust validation strategies to support the integration of XAI into clinical decision support systems for AD prognosis.

Chun, et al. [13] proposed an interpretable machine learning framework to forecast the conversion from amnestic mild cognitive impairment (aMCI) to Alzheimer's dementia using longitudinal clinical data. Evaluating LR, random forest, SVM, and XGBoost, the latter demonstrated superior performance (AUC = 0.852, Accuracy = 0.807). The study applied SHAP and ICE to elucidate both global and individual model behavior, revealing key features such as age, education level, neuropsychological scores, and APOE genotype. The framework highlights the importance of explainability in clinical AI applications for personalized dementia risk assessment.

Alsubaie, et al. [18] conducted a systematic review of deep learning models for AD detection using neuroimaging. The study surveyed CNNs, RNNs, GANs, autoencoders, and 3D CNNs applied across datasets like ADNI, AIBL, OASIS, MIRIAD, and Kaggle. Reported accuracies ranged from 70% to 99.95%. The review highlighted the role of explainable AI (SHAP, Grad-CAM, LIME), discussed interpretability limitations, and emphasized the need for benchmark datasets and transparent, multimodal models.

Hernandez, et al. [37] conducted a post-hoc interpretability analysis of the top three predictive models from the TADPOLE Challenge, which aimed to forecast the longitudinal progression of AD. The study emphasized the need for model transparency by applying SHAP to ensemble-based algorithms such as Random Forest and Gradient Boosting. This framework enabled the identification and quantification of key features contributing to AD prognosis and examined the coherence between model decisions and established clinical knowledge. The authors highlighted that while the original challenge focused on accuracy, their analysis addressed the critical gap in interpretability, advocating for more explainable and trustworthy AI systems in clinical AD prediction tasks.

Alatrany, et al. [38] developed an interpretable machine learning approach for early (AD) diagnosis using structured clinical data from the NACC dataset. Utilizing a SVM on over 169,000 records with 1,024 features, their model achieved high F1 scores of 98.9% (binary) and 90.7% (multiclass), along with robust performance in four-year progression prediction. To enhance transparency, they employed rule-extraction techniques alongside SHAP and LIME, identifying clinically relevant features such as MEMORY, JUDGMENT, and ORIENT. The study contributes a clinically-aligned, interpretable framework for trustworthy AD prediction systems.

Yang, et al. [40] presented a comprehensive survey of explainable AI (XAI) techniques in AD (AD) diagnosis, reviewing ML and DL models such as CNNs, RNNs, GNNs, SVMs, and RFs alongside XAI methods including SHAP, LIME, Grad-CAM, and LRP. The study introduced a taxonomy based on interpretability scope, data modality, and clinical relevance, and identified key challenges including lack of standardization and clinical validation. The survey offers critical insights and future directions for deploying domain-specific, interpretable AI in neurodegenerative disease diagnostics.

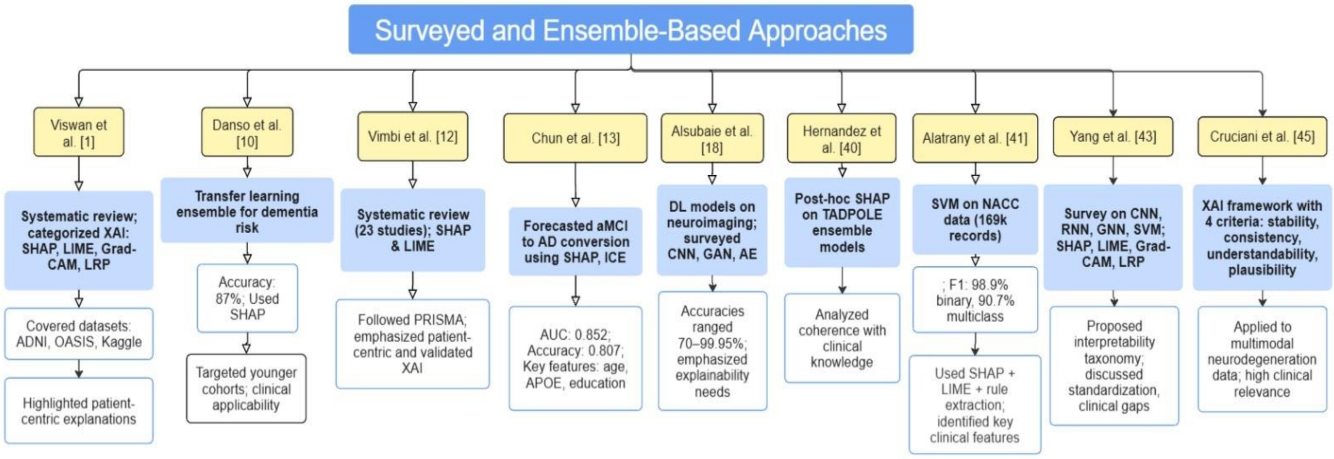

Surveyed and Ensemble-Based Approaches
Viswan et al. [1]
Systematic review; categorized XAI: SHAP, LIME, Grad-CAM, LRP
Covered datasets: ADNI, OASIS, Kaggle
Highlighted patient-centric explanations
Danso et al. [10]
Transfer learning ensemble for dementia risk
Accuracy: 87%; Used SHAP
Targeted younger cohorts; clinical applicability
Vimbi et al. [12]
Systematic review (23 studies); SHAP & LIME
Followed PRISMA; emphasized patient-centric and validated XAI
Chun et al. [13]
Forecasted aMCI to AD conversion using SHAP, ICE
AUC: 0.852; Accuracy: 0.807; Key features: age, APOE, education
Alsubaie et al. [18]
DL models on neuroimaging; surveyed CNN, GAN, AE
Accuracies ranged 70–99.95%; emphasized explainability needs
Hernandez et al. [40]
Post-hoc SHAP on TADPOLE ensemble models
Analyzed coherence with clinical knowledge
Alatrany et al. [41]
SVM on NACC data (169k records)
; F1: 98.9% binary, 90.7% multiclass
Used SHAP + LIME + rule extraction; identified key clinical features
Yang et al. [43]
Survey on CNN, RNN, GNN, SVM; SHAP, LIME, Grad-CAM, LRP
Proposed interpretability taxonomy; discussed standardization, clinical gaps
Cruciani et al. [45]
XAI framework with 4 criteria: stability, consistency, understandability, plausibility
Applied to multimodal neurodegeneration data; high clinical relevance

Figure 15.
Surveyed and ensemble-based XAI approaches in AD diagnosis. Studies include systematic reviews and applied frameworks integrating SHAP, LIME, Grad-CAM, and ICE across diverse models (e.g., CNN, GAN, SVM) and datasets. Key contributions involve patient-centric explanations, transfer learning, interpretability validation, and multimodal Analysis Viswan, et al. [1]; Danso, et al. [10]; Vimbi V. et al [41]; Chun, et al. [13]; Alsubaie, et al. [18]; Yu, et al. [16]; Hernandez, et al. [37]; Alatrany, et al. [38]; Yang, et al. [40] and Cruciani [42].

Cruciani [42] introduced a domain-specific XAI framework for neuroscience, focusing on neurodegenerative disease interpretation. The framework includes methodological guidelines and a validation scheme grounded in four key attributes: stability, consistency, understandability, and plausibility. Applied to brain modulations using multimodal data, the approach enhances explanation robustness across AI models. While not explicitly AD-specific, the study aligns closely with Alzheimer's research and advances the clinical reliability of interpretable decision support systems.

Across multiple research efforts, several common limitations are evident:

- Small Sample Sizes: Many studies are restricted by small sample sizes, limiting broader applicability.
- Insufficient Data Preprocessing Details: Lack of transparency in preprocessing methods impedes the ability to assess data quality.
- Limited Validation Techniques: The predominant use of simple validation strategies (e.g., k-fold cross-validation) may not adequately capture model reliability in clinical settings.

4.2. Research Gaps

Many research papers have explored the application of XAI in AD diagnosis. However, researchers in this area face several significant challenges, including limited generalizability, lack of standardization, and other technical and clinical barriers. This section highlights these challenges in detail. Figure 16 presents a taxonomy of XAI in AD diagnosis, categorizing the various aspects and approaches used in this field. It serves as a conceptual map of the current landscape of XAI in AD diagnosis. The main challenges in this area can be summarized as follows:

1. Limited Generalizability: Several prior works are constrained in generating generalizable results since they use datasets with low demographic heterogeneity and small sample sizes. To achieve strong and clinically relevant performance, explainable AI (XAI) models should incorporate information from diverse groups of populations. Inclusivity makes the models more reliable and generalizable across various clinical and demographic environments.
2. Absence of Standardization: Standardization methods remain nonexistent in the XAI diagnosis technology implementation process for AD. Many evaluation metrics and available methods, and tools hinder scientists from comparing study results with one another to identify ideal approaches.
3. Inadequate Clinical Validation: Research evidence shows XAI tools function well under controlled conditions, but insufficient extensive clinical results continue to validate their use. Further investigation is required to evaluate these tools' practical effectiveness and dependability in clinical environments.
4. User Acceptance and Trust: Healthcare professionals' acceptance of XAI solutions continues to be difficult. Many studies fail to recognize the importance of understanding clinicians' viewpoints, which is critical for effectively using XAI systems in real-world settings.
5. Ethical and Regulatory Concerns: With the increasing integration of AI systems into clinical decision-making, it is crucial to address ethical concerns related to patient data protection, accountability for AI system judgments, and regulatory compliance comprehensively.

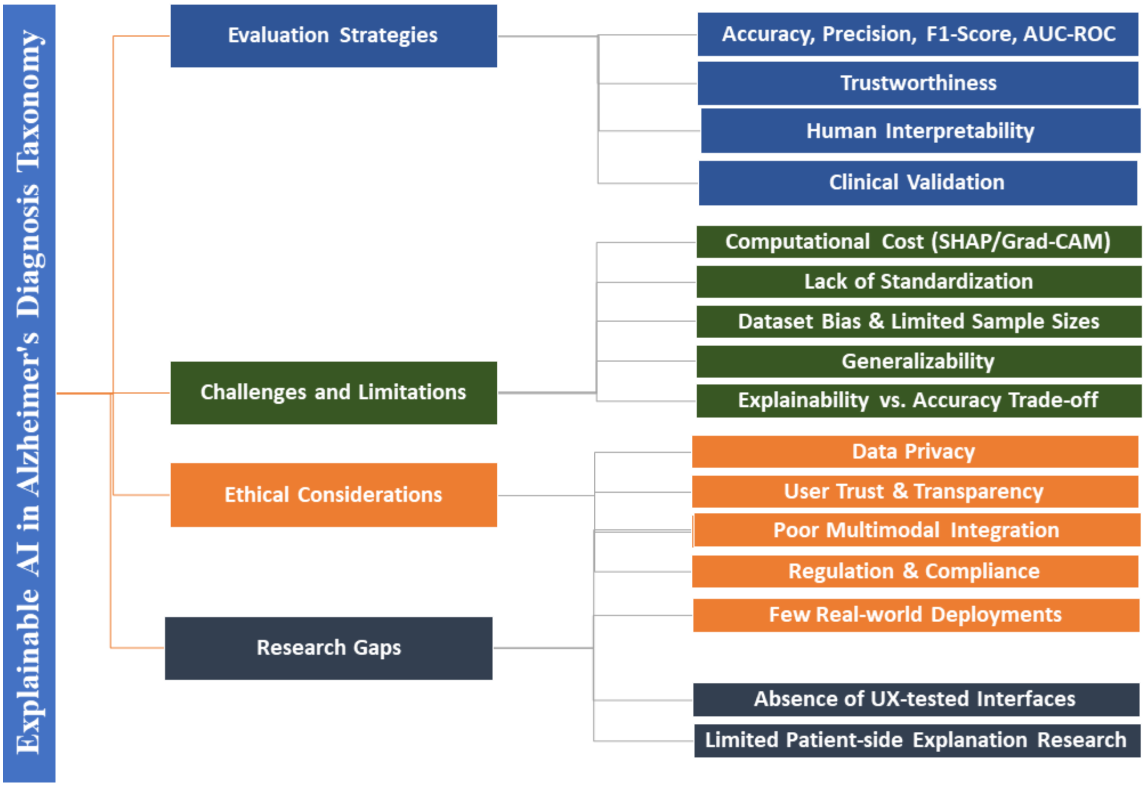

Explainable AI in Alzheimer's Diagnosis Taxonomy
Evaluation Strategies
Accuracy, Precision, F1-Score, AUC-ROC
Trustworthiness
Human Interpretability
Clinical Validation
Challenges and Limitations
Computational Cost (SHAP/Grad-CAM)
Lack of Standardization
Dataset Bias & Limited Sample Sizes
Generalizability
Explainability vs. Accuracy Trade-off
Ethical Considerations
Data Privacy
User Trust & Transparency
Poor Multimodal Integration
Regulation & Compliance
Few Real-world Deployments
Research Gaps
Absence of UX-tested Interfaces
Limited Patient-side Explanation Research

Figure 16.
Research Taxonomy of xAI in AD, this figure outlines key evaluation strategies, challenges, ethical considerations, and research gaps in XAI for AD.

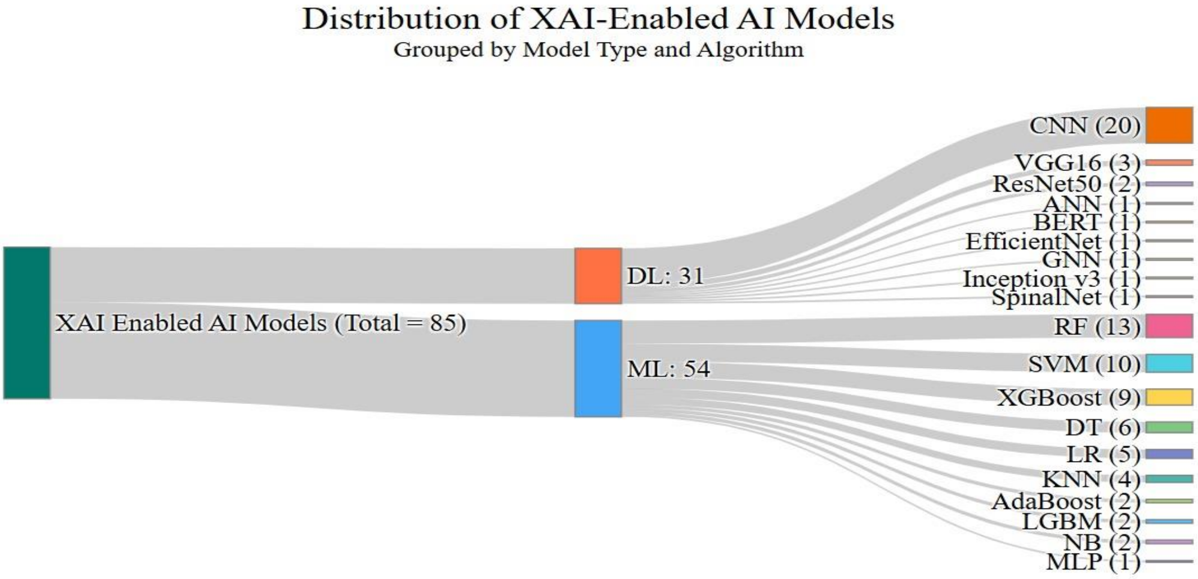


Figure 17.
Sankey diagram showing the distribution of XAI frameworks in AD. SHAP and Grad-CAM are commonly used for image studies, while LIME is applied to structured data Loh, et al. [14].

The Sankey diagram is shown in Figures 17, 18, and 19. XAI frameworks in both ML and DL models are shown in Figure 20. Figure 17 shows the distribution of XAI techniques in AD models. As we can see, XAI techniques are applied more frequently to ML models than to DL models, which indicates a higher maturity or prevalence of XAI methods for ML models. Additionally, within the DL category, CNNs are the dominant model type that utilizes XAI techniques. Figure 18 shows the frequency of utilization of various XAI techniques across ML and DL models in AD. Consistent with Figure 17, this figure emphasizes that XAI techniques are more frequently applied to ML models than to DL models. Figure 19 shows the XAI framework usage in ML and DL models. Within the DL category, CNNs (including 3D CNNs) are overwhelmingly the most common models for which XAI is applied. Furthermore, SHAP is used largely compared to other XAI techniques and is applied to both DL and ML models.

Figure 20 shows the XAI framework usage in ML and DL models. We can see that SHAP and LIME are frequently used to explain both ML and DL models. Also, it represents the use of XAI frameworks in ML and DL models used to diagnose Alzheimer's, with CNNs being the most common. DL models are common, whereas traditional ML models such as DT and SVM are rarely used. CNNs occur most frequently because they can offer better image processing of high-dimensional imaging data (MRI and PET scans, widely used in AD). Their structure could search spatial hierarchies and local features through convolutional layers for very detailed pattern recognition for neuroimaging. CNNs can be coupled with some explainable AI techniques, such as Grad-CAM, making them excellent interpretable DL tools in medical imaging.

Figure 21 shows the use of XAI frameworks in ML and DL models for AD diagnosis, with CNNs being the most commonly applied. DL models dominate the landscape, whereas traditional ML models such as decision trees and SVMs are less frequently utilized. The review investigated the combination of XAI approaches into numerous AI models for the prediction of AD as a part of addressing RQ1. An assessment of methods exposed an extensive range of tactics in numerous studies, from implementing DL systems to conventional device-gaining algorithms. To enhance interpretability, a CNN was applied, combined with SHAP. They demonstrated a thorough technique and extensively validated it on several datasets. Nevertheless, it was observed that complex models have limited interpretability. From these figures, we can summarize three main points, concluded as follows:

- Dominance of XAI in ML Models, but CNNs Lead in DL.
- SHAP and LIME are the Preferred XAI frameworks.

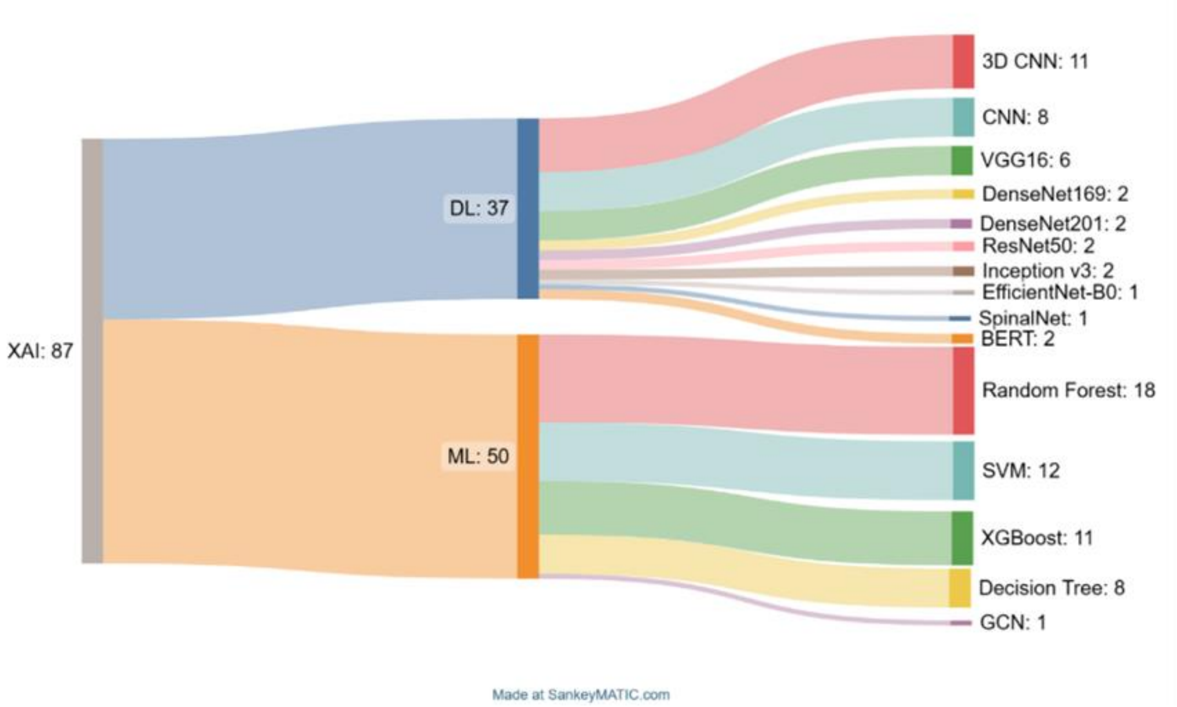


Figure 18.
A bar chart depicting the frequency of utilization of various XAI frameworks in ML and DL models to diagnose AD.

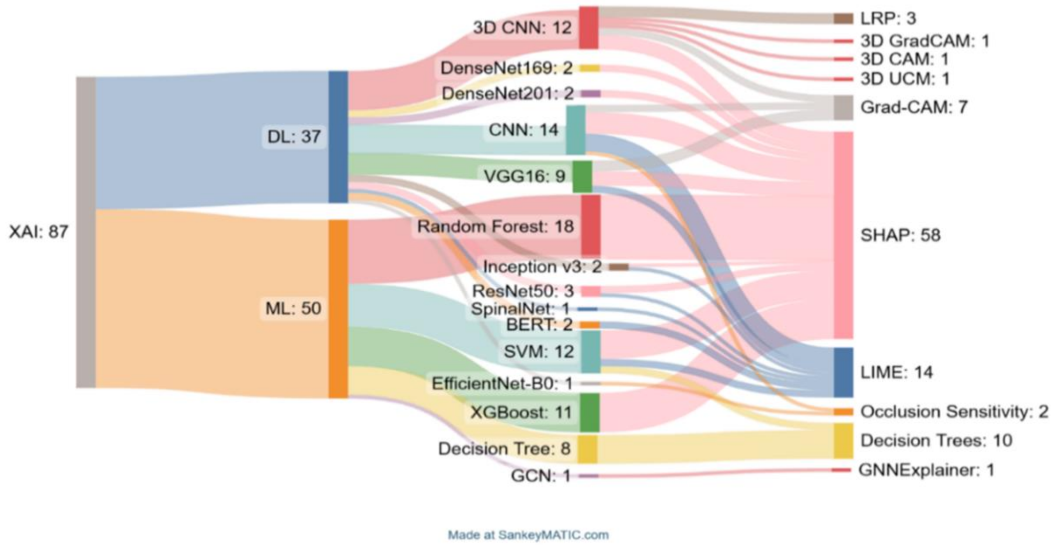


Figure 19.
Sankey diagram representing the application of XAI frameworks to ML and DL models for diagnosing AD.

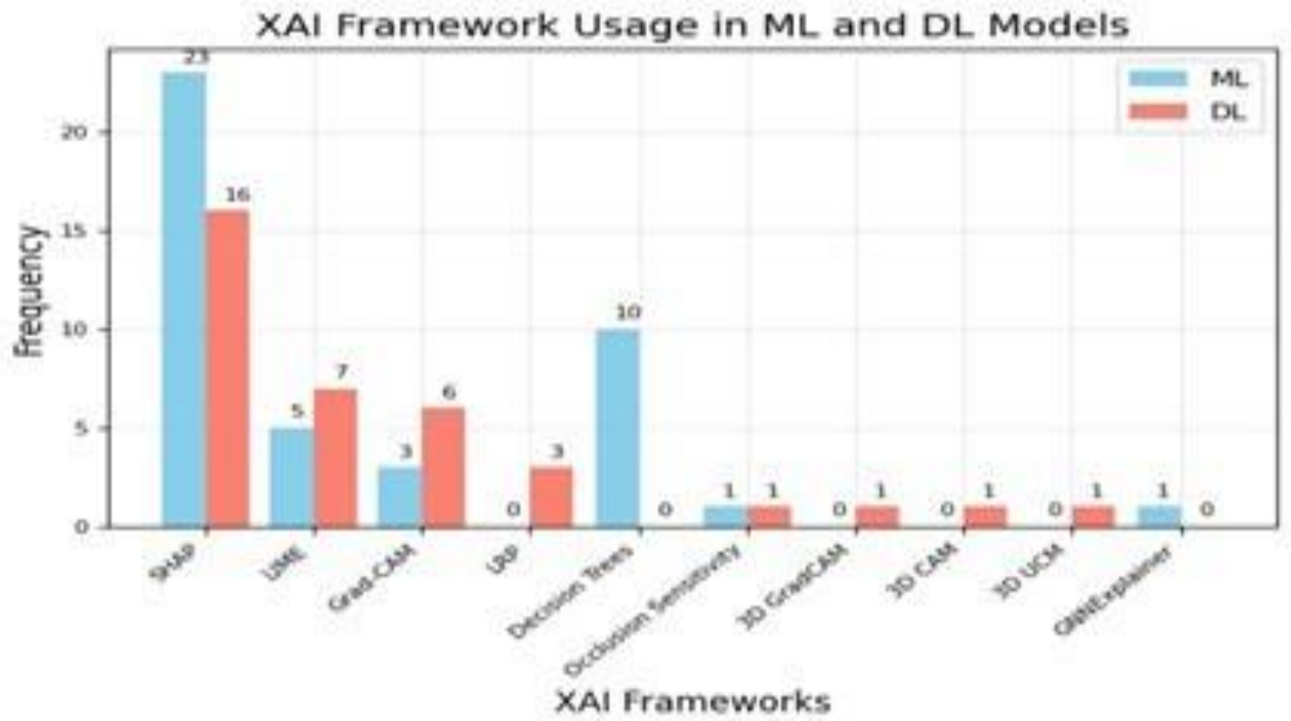


Figure 20.
XAI frameworks in ML and DL models.

- Emphasis on Interpretability for Complex Models in AD Diagnosis
-

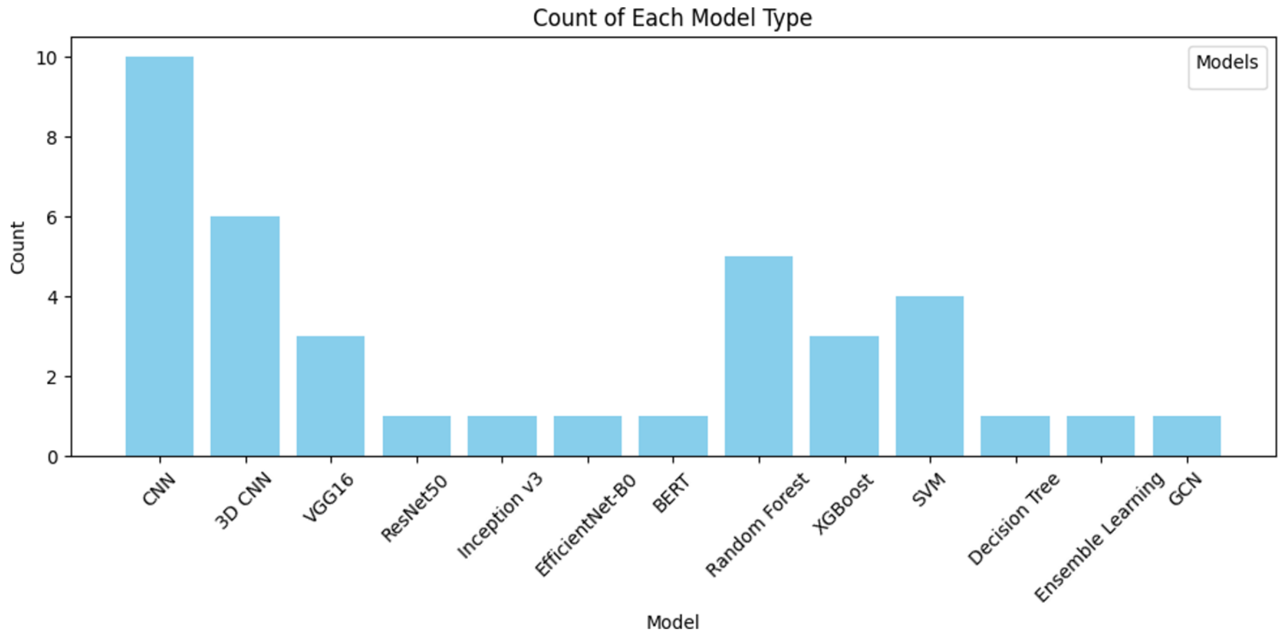


Figure 21.
Frequency of utilizing various XAI frameworks in ML and DL models to diagnose AD.

Regarding **RQ2**, the review examined how XAI predictions are shared with patients and caregivers in an effective way. The findings showed that it is important to design simple and easy-to-use interfaces that also match the needs and preferences of users. Key factors such as health literacy, cultural background, and personal preferences were considered. The studies revealed that the quality of explanations varied—some were clear, while others were hard to understand. However, some research introduced creative solutions to make the explanations easier. These included using simple language and visual aids. Such user-friendly designs helped patients and caregivers better understand the AI's predictions and work together when making.

5. Discussion

The reviewed studies illustrate the flexibility and utility of the XAI frameworks around healthcare, especially when identifying AD. A major advantage is that these models enhance the explanation of other opaque algorithms in ML, such as

DL models and black-box models, such as SVM and CNN. These XAI frameworks, such as Grad-CAM and LIME, are helpful because they explain the model's predictions to clinicians, making them accept AI systems' decisions and build trust in AI systems.

5.1. Limitations in Datasets

However, the analysis also pointed to some weaknesses, primarily regarding the datasets employed for XAI applications. Most of the datasets were relatively small, even though more than half of the included studies recruited over 100 participants; the variability of the datasets was sometimes limited. This limitation could limit the transportability of the findings across various patients [40].

The reviewed studies also showed limitations in the datasets used for analysis. Furthermore, these limitations in the diversity and quality of training datasets can lead to algorithmic bias. Studies Hernandez, et al. [37] and Yang, et al. [40] suggest that algorithmic bias resulting from imbalanced training data can lead to health disparities; for example, certain ML models may perform worse with patient demographics that may not have been represented fairly in the training data.

To avoid such consequences, AI algorithms must use data that fairly represent different populations to avoid exacerbating disparities. This risk is particularly concerning in the context of AD, where factors such as age, race, socioeconomic status, and genetic background can all interact to influence disease progression and clinical manifestations, requiring greater care when selecting data samples.

5.2. Interpretation of Findings

This study focuses on multimodal AD prediction using XAI. Various works in this area used genetic data, medical histories, neuropsychological questionnaires, and cognitive scores data. LIME and SHAP are useful explanation models because local and additive explanations make complex models interpretable.

To enhance model interpretability, this research suggests an XAI-based solution that includes DL with ensemble modeling. The approach integrates tools such as Grad-CAM and saliency maps for interpreting model predictions. Many experiments conducted on large datasets showed that ensemble models like Ensemble-1 (VGG-16 and VGG-19) and Ensemble-2 (DenseNet- 169 and DenseNet-201) outperformed individual models, achieving 95% accuracy. A new model with 96% accuracy presented an exemplary visualization of brain regions relevant to AD diagnosis, assisting clinicians in understanding predictions [30]. A PRISMA-based scoping review of XAI in healthcare from 2019 to 2024 identified key benefits and limitations. Key factors discussed include safety, performance evaluation, standardized language, and cross-care adaptability. Challenges include organizational barriers, policies, legal risks, and communication hurdles. Recommendations include deploying AI strategically, enhancing provider-patient communication, and increasing patient engagement [38].

Deep CNNs remain standard for medical image classification. Despite their predictive strength, issues such as limited datasets, skewed class distributions, and lack of explainability hinder precision medicine application. This thesis addresses these limitations using Synthetic Minority Over-sampling Technique (SMOTE) to rebalance data and compares three CNNs: DenseNet121, DenseNet169, and Inception-ResNet-v2. The model showed high sensitivity and specificity. Misclassifications were linked to neurocognitive outcomes and the APOE biomarker using SHAP [39].

Using a Kaggle hybrid MRI dataset, one study experimented with ResNet50, VGG16, and Inception v3, achieving accuracies of 82.56%, 86.82%, and 82.04% respectively. LIME was used to explain prediction regions for each patient based on T1-weighted scans [20]. A model based on the China Longitudinal Aging Study (CLAS), with 3,514 individuals, used feature selection and ensemble learning, achieving 89.2% accuracy for MCI and 99.2% for AD. SHAP visualizations improved clinicians' understanding of individual and global prediction features [26].

Using data from the ADNI database, decomposition-based explanation methods were proposed to clarify how features from multiple modalities impact diagnosis. Compared with SHAP, this method saved computation time and yielded consistent results. A survey of 11 domain experts confirmed interpretability, with 71% rating explanations as true and understandable [35]. The TADPOLE Challenge further evaluated model interpretability and alignment with clinical expectations using SHAP [37]. Another study used a multimodal dataset (clinical, MRI, cognitive) and tested nine ML models (e.g., RF, LR, Decision Tree), improving explainability with SHAP. GCNs were employed to classify from NC through MCI to AD, using neurocognitive, genetic, and imaging data [36].Using a dataset from the National Alzheimer's Coordinating Center (169,408 records, 1024 features), SVM models achieved F1 scores of 98.9% (binary) and 90.7% (multiclass). Two interpretable rule extraction methods were used, supported by SHAP and LIME explanations, identifying memory, judgment, and orientation as critical predictors [40].

Despite early Computer-Aided Diagnosis (CAD) models being unreliable, XAI has improved their reliability. A systematic PRISMA-based review categorized XAI methods as post-hoc, ante-hoc, model-agnostic, or model-specific using LIME, SHAP, GradCAM, and LRP to answer research questions [41].

Biomedical image analysis seeks to model biological states using AI. CNNs applied to MRI for AD diagnosis have achieved 95.70% training accuracy and 99.71% validation accuracy. Studies focusing on detection, prediction, or rehabilitation using AI in neurological disorders were included; others were excluded [45].

Research on AD diagnosis and progression often lacks clinical relevance due to overemphasis on single modalities, separate problem framing, and poor explainability. A dual-layer model using 11 modalities and 1,048 subjects from ADNI applied Random Forest classification. The first layer did multiclass prediction; the second predicted MCI-to-AD conversion within three years. SHAP and fuzzy rules explained model predictions with an F1 of 93.94% in layer one and 87.09% in layer two [45].XAI-based, multilayer, multimodal models show potential for improving AD prediction. Still, reluctance persists due to concerns about generalizability, evidence, and bias. Only studies related to diagnosis, prediction, or interpretability were retained in this review [35].

Table 12 provides a summary of XAI approaches in multimodal AD prediction. We can conclude some key points: LIME and SHAP are commonly and frequently used in XAI techniques. Grad-CAM is also used but focuses more on neuroimaging contexts. There is also more focus on using diverse multimodal datasets to improve the accuracy of AD prediction. Ensemble learning models demonstrate better accuracy compared to individual models for AD prediction. Some researchers addressed challenges related to limited datasets to enhance robustness and generalizability. Finally, XAI models are used to enhance clinicians' understanding

5.3. Implications for Practice

Several obstacles ought to be addressed even though this work offers insightful data on the integration of XAI strategies through diverse AI methods and the effective transmission of AI-generated predictions to patients and caregivers:

1. Many XAI techniques, including SHAP and Grad-CAM, require significant computational resources. This limitation hampers their real-time use in healthcare environments with limited resources. Optimizing these techniques to balance interpretability and computational efficiency should be a principal aim for future improvements [46].
2. Large-scale datasets pose challenges for scalable and consistent application of methods like LIME and Anchors, despite their simplicity. The large volumes of data common in clinical contexts restrict their practical deployment [21].
3. Patient autonomy, algorithmic bias, and data privacy are ethical implications often underexplored in studies. The ethical and responsible use of XAI in healthcare depends on thoroughly examining these issues.
4. Most AD studies focus on diagnosis at a single time point. Future research should explore XAI models capable of analyzing longitudinal data to forecast disease trajectory and detect early biomarkers.
5. Static datasets dominate AD prediction studies. Real-time data streams (e.g., from wearable sensors or EHRs) are largely unexamined. Integrating such data is crucial for improving the value and precision of AD diagnosis [47].

By addressing these constraints, XAI strategies can become more reliable, scalable, and applicable in real-world clinical settings. These advancements will be key to realizing XAI's full potential in enhancing AD diagnosis and patient care. LIME is a widely adopted technique for explaining model predictions by approximating model behavior around specific instances. In AD detection, studies have applied LIME to interpret classifiers trained on MRI scans and gene expression data, identifying key genes involved in AD progression [47, 48]. Another study used LIME to analyze clinical transcripts and detect linguistic patterns associated with dementia.

Researchers have used SHAP to interpret the influence of MRI volumetric data on AD classification and to examine model decisions based on clinical and demographic records [27]. Grad-CAM has been used to highlight important image regions in AD diagnosis, including task-specific photos [48], MRI scans [19] and saliency maps for visualizing neural activity linked to AD [19]. Beyond LIME, SHAP, and Grad-CAM, additional frameworks such as GNNExplainer have been applied to GNNs to uncover the structural roles of brain networks in AD prediction [31]. ICE plots, occlusion sensitivity, and saliency maps have also been employed to explain classifier decisions at the individual level, improving transparency in AD model predictions. The review acknowledges the transformative role of XAI frameworks like LIME, SHAP, and Grad-CAM in advancing AD diagnosis [42]. However, a deeper analysis of their mechanisms and applications is necessary. Understanding how each method handles different AI architectures and the types of insights each provides would improve their clinical applicability. Including case studies from real-world scenarios, such as how SHAP helped interpret demographic influences or how Grad-CAM visualized MRI regions, would provide critical evidence of their clinical utility.

Table 12.
Comparison of XAI Approaches in Multimodal AD Prediction.

Study	Data Modalities	Key AI Models	XAI Methods Used	Key Contribution/Performance/Focusties
Xu and Yan [15]; Yu, et al. [16] and Mahmud, et al. [30]	Ensemble-1	Neuroimaging (VGG-16, VGG-19), Ensemble-2 (DenseNet-169, DenseNet-201), New model	Grad-CAM, Saliency Maps	DL with ensemble modeling; New model 96% accuracy; visualization of brain regions.
Alatrany, et al. [38]	Not specified (Scoping Re-view)	Not applicable (Review)	Not applicable (Re- view)	PRISMA-based scoping review of XAI in healthcare; benefits, limitations, challenges, recommendations.
Rahim, et al. [34] and Ekuma [39]	MedicalImage	DenseNet121, DenseNet169, Inception- ResNet-v2	SHAP	Addresses limited datasets/skewed distributions (SMOTE); links misclassifications to neurocognitive outcomes/APOE.
Shad, et al. [20]	MRI (T1-weighted)	ResNet50, VGG16, Inception v3	LIME	Kaggle hybrid MRI dataset; Accuracies: 82.56% (ResNet50), 86.82% (VGG16), 82.04% (Inception v3).
Yue, et al. [26]	Clinical	Ensemble learning	SHAP	CLAS dataset (3,514 individuals); 89.2% accuracy for MCI, 99.2% for AD; improved clinician understanding.

Anjomshoae and Pudas [49]	Multimodal (ADNI)	Decomposition- based methods	Compared to SHAP	Faster computation than SHAP, consistent results; 71% expert rating for interpretability.
Tekkesinoglu and Pudas [29] and Hernandez, et al. [37]	TADPOLE Challenge data	Various(evaluated)	SHAP	Evaluated model interpretability and clinical alignment.
Jahan, et al. [36]	Clinical, MRI, Cognitive; Neurocognitive, Genetic, Imaging	9 ML models (RF, LR, Decision Tree); GCNs	SHAP	Multimodal dataset; improved explainability; GCNs for NC to MCI to AD classification.
Hernandez, et al. [37]; Alatrany, et al. [38] and Yang, et al. [40]	Clinical (NACC: 169,408 records, 1024 features)	SVM	SHAP, LIME (and 2 rule extraction methods)	F1 scores: 98.9% (binary), 90.7% (multi- class); identified critical predictors (memory, judgment, orientation).
Vimbi V. et al [41]	Not specified (Systematic Review)	Not applicable (Review)	LIME, SHAP, Grad-CAM, LRP	PRISMA-based review categorizing XAI methods (post-hoc, ante-hoc, model- agnostic, model-specific) for CAD.
Subasi, et al. [45]	MRI	CNN	Not explicitly stated for XAI in this text	CNNs for AD diagnosis: 95.70% training accuracy, 99.71% validation accuracy.
Subasi, et al. [45]	11 modalities (ADNI, 1,048 subjects)	Random Forest	SHAP, Fuzzy Rules	Dual-layer model (multiclass prediction, MCI-to-AD conversion); F1: 93.94% (L1), 87.09% (L2).
Zhang, et al. [19]; Kim, et al. [31] and De Santi, et al. [35]	Not specified	Multilayer, Multi-modal	Not explicitly stated for XAI in this text	XAI-based models show potential, but concerns about generalizability, evidence, and bias persist.

Although the assessment recognizes the ethical implications of XAI in AD analysis, the paper's contribution might be strengthened by including a dedicated section that thoroughly explores these issues [32, 42, 43, 50, 51]. Examining topics such as data privacy, algorithmic bias, patient autonomy, and the potential societal consequences of XAI emphasizes the significance of responsible AI development and integration in the healthcare sector.

6. Ethical Considerations & Patient-Centric XAI Explanation

A dedicated section has been added to comprehensively address critical ethical concerns that arise in the integration of XAI into AD diagnosis. These include:

- Ensuring that AI models comply with regulatory frameworks such as HIPAA and GDPR to safeguard patient confidentiality.
- Addressing imbalances and under-representation in training data to prevent systematic misdiagnosis among certain demographic groups.
- Delivering model explanations in an interpretable and patient-friendly manner, empowering patients to make informed decisions.
- Promoting transparency in AI-based decision-making to foster trust among clinicians, patients, and stakeholders.
- Ensuring that AI-generated recommendations are cross-validated by medical professionals to reduce the risk of diagnostic errors and unintended consequences.

These considerations underscore the necessity of developing ethical, human-centered AI systems that align with clinical standards and patient well-being.

7. Future Direction

The field of XAI for AD diagnosis holds significant potential for improving clinical decision-making by offering interpretable, robust models. However, several key areas remain for future exploration. These include dataset sensitivity, where current XAI models like SHAP, LIME, and GradCAM are often evaluated on limited, homogenous datasets. Future models must be validated on diverse, large-scale, and multicenter datasets to ensure generalizability across populations and geographical regions. Furthermore, multimodal integration, incorporating data from neuroimaging, genetics, and cognitive assessments into XAI systems, can yield more holistic insights into AD pathology, which could enhance diagnostic performance and support precision medicine. Additionally, scalability and efficiency are crucial areas, with future research needing to focus on optimizing XAI methods to reduce computational overhead, making them suitable for real-time use in clinical settings.

Beyond these technical advancements, the development of interactive explanation tools that allow clinicians to dynamically interact with AI outputs, for example, visualizing decision paths or adjusting input features, can improve usability and trust. Crucially, ethical and social implications require additional research to explore how AI might inadvertently infringe on patient rights or propagate institutional biases. Addressing these issues is essential for safe and equitable implementation in healthcare. These directions collectively suggest a roadmap for advancing XAI methodologies toward clinically deployable, transparent, and ethical diagnostic tools.

8. Conclusions

This review underscores a significant advancement in the application of XAI for the diagnosis of AD. Through the use of techniques such as SHAP, LIME, and GradCAM, researchers and practitioners are now better equipped to interpret and validate the internal decision-making processes of ML models. These tools not only improve the transparency and accountability of diagnostic models, but also contribute to clinical acceptance and trust in AI-driven healthcare. XAI facilitates the communication of model predictions to both healthcare providers and patients, supporting informed and individualized treatment decisions. By bridging the gap between AI innovation and medical practice, XAI methodologies are poised to play a transformative role in early detection, prognosis, and personalized intervention for ADs.

References

[1] V. Viswan, N. Shaffi, M. Mahmud, K. Subramanian, and F. Hajamohideen, "Explainable artificial intelligence in alzheimer’s disease classification: A systematic reviewa," *Cognitive Computation*, vol. 16, no. 1, pp. 1-44, 2024. https://doi.org/10.1007/s12559-023-10192-x

[2] Y. Kumar, A. Koul, R. Singla, and M. F. Ijaz, "Artificial intelligence in disease diagnosis: A systematic literature review, synthesizing framework and future research agenda," *Journal of Ambient Intelligence and Humanized Computing*, vol. 14, no. 7, pp. 8459-8486, 2023. https://doi.org/10.1007/s12652-021-03612-z

[3] S. N. Payrovnaziri *et al.*, "Explainable artificial intelligence models using real-world electronic health record data: A systematic scoping review," *Journal of the American Medical Informatics Association*, vol. 27, no. 7, pp. 1173-1185, 2020. https://doi.org/10.1093/jamia/ocaa053

[4] S. S. Band *et al.*, "Application of explainable artificial intelligence in medical health: A systematic review of interpretability methods," *Informatics in Medicine Unlocked*, vol. 40, p. 101286, 2023. https://doi.org/10.1016/j.imu.2023.101286

[5] A. Segato, A. Marzullo, F. Calimeri, and E. De Momi, "Artificial intelligence for brain diseases: A systematic review," *APL Bioengineering*, vol. 4, no. 4, p. 041503, 2020. https://doi.org/10.1063/5.0011697

[6] F. Giuste *et al.*, "Explainable artificial intelligence methods in combating pandemics: A systematic review," *IEEE Reviews in Biomedical Engineering*, vol. 16, pp. 5-21, 2022.

[7] L. J. Quek, M. R. Heikkonen, and Y. Lau, "Use of artificial intelligence techniques for detection of mild cognitive impairment: A systematic scoping review," *Journal of Clinical Nursing*, vol. 32, no. 17-18, pp. 5752-5762, 2023. https://doi.org/10.1111/jocn.16699

[8] M. A. Ebrahimighahnavieh, S. Luo, and R. Chiong, "Deep learning to detect Alzheimer's disease from neuroimaging: A systematic literature review," *Computer Methods and Programs in Biomedicine*, vol. 187, p. 105242, 2020. https://doi.org/10.1016/j.cmpb.2019.105242

[9] D. Kumar and M. A. Mehta, "An overview of explainable ai methods, forms and frameworks." Cham: Springer International Publishing, 2023, pp. 43-59. https://doi.org/10.1007/978-3-031-12807-3_3

[10] S. O. Danso, Z. Zeng, G. Muniz-Terrera, and C. W. Ritchie, "Developing an explainable machine learning-based personalised dementia risk prediction model: A transfer learning approach with ensemble learning algorithms," *Frontiers in Big Data*, vol. 4, p. 613047, 2021.

[11] M. S. Kamal, A. Northcote, L. Chowdhury, N. Dey, R. G. Crespo, and E. Herrera-Viedma, "Alzheimer’s patient analysis using image and gene expression data and explainable-AI to present associated genes," *IEEE Transactions on Instrumentation and Measurement*, vol. 70, pp. 1-7, 2021. https://doi.org/10.1109/TIM.2021.3107056

[12] V. Vimbi, N. Shaffi, and M. Mahmud, "Interpreting artificial intelligence models: a systematic review on the application of LIME and SHAP in Alzheimer’s disease detection," *Brain Informatics*, vol. 11, no. 1, p. 10, 2024. https://doi.org/10.1186/s40708-024-00222-1

[13] M. Y. Chun *et al.*, "Prediction of conversion to dementia using interpretable machine learning in patients with amnestic mild cognitive impairment," *Frontiers in Aging Neuroscience*, vol. 14, p. 898940, 2022.

[14] H. W. Loh, C. P. Ooi, S. Seoni, P. D. Barua, F. Molinari, and U. R. Acharya, "Application of explainable artificial intelligence for healthcare: A systematic review of the last decade (2011–2022)," *Computer Methods and Programs in Biomedicine*, vol. 226, p. 107161, 2022. https://doi.org/10.1016/j.cmpb.2022.107161

[15] X. Xu and X. Yan, "A convenient and reliable multi-class classification model based on explainable artificial intelligence for Alzheimer’s disease," presented at the IEEE International Conference on Advances in Electrical Engineering and Computer Applications (AEECA), 2022.

[16] L. Yu, W. Xiang, J. Fang, Y.-P. Phoebe Chen, and R. Zhu, "A novel explainable neural network for Alzheimer’s disease diagnosis," *Pattern Recognition*, vol. 131, p. 108876, 2022. https://doi.org/10.1016/j.patcog.2022.108876

[17] L. Ilias and D. Askounis, "Explainable identification of dementia from transcripts using transformer networks," *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 8, pp. 4153-4164, 2022.

[18] M. G. Alsubaie, S. Luo, and K. Shaukat, "Alzheimer’s disease detection using deep learning on neuroimaging: A systematic review," *Machine Learning and Knowledge Extraction*, vol. 6, no. 1, pp. 464-505, 2024. https://doi.org/10.3390/make6010024

[19] X. Zhang, L. Han, W. Zhu, L. Sun, and D. Zhang, "An explainable 3D residual self-attention deep neural network for joint atrophy localization and Alzheimer's disease diagnosis using structural MRI," *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 11, pp. 5289-5297, 2021.

[20] H. A. Shad *et al.*, "Exploring Alzheimer's disease prediction with XAI in various neural network models," in *TENCON 2021-2021 IEEE Region 10 Conference (TENCON)*, 2021, pp. 720-725, doi: https://doi.org/10.1109/TENCON54134.2021.9707468.

[21] A. Lombardi *et al.*, "A robust framework to investigate the reliability and stability of explainable artificial intelligence markers of Mild Cognitive Impairment and Alzheimer's Disease," *Brain Informatics*, vol. 9, no. 1, p. 17, 2022. https://doi.org/10.1186/s40708-022-00165-5

[22] B. Bogdanovic, T. Eftimov, and M. Simjanoska, "In-depth insights into Alzheimer's disease by using explainable machine learning approach," *Scientific Reports*, vol. 12, no. 1, p. 6508, 2022. https://doi.org/10.1038/s41598-022-10202-2

[23] Y. Lai, X. Lin, C. Lin, X. Lin, Z. Chen, and L. Zhang, "Identification of endoplasmic reticulum stress-associated genes and subtypes for prediction of Alzheimer's disease based on interpretable machine learning," *Frontiers in Pharmacology*, vol. 13, p. 975774, 2022.

[24] V. Jain, O. Nankar, D. J. Jerrish, S. Gite, S. Patil, and K. Kotecha, "A novel AI-based system for detection and severity prediction of dementia using MRI," *IEEE Access*, vol. 9, pp. 154324-154346, 2021.

[25] L. Bloch, C. M. Friedrich, and A. s. D. N. Initiative, "Data analysis with Shapley values for automatic subject selection in Alzheimer's disease data sets using interpretable machine learning," *Alzheimer's Research & Therapy*, vol. 13, no. 1, p. 155, 2021.

[26] L. Yue, W.-g. Chen, S.-c. Liu, S.-b. Chen, and S.-f. Xiao, "An explainable machine learning based prediction model for Alzheimer's disease in China longitudinal aging study," *Frontiers in Aging Neuroscience*, vol. 15, p. 1267020, 2023.

[27] S. El-Sappagh, J. M. Alonso, S. M. R. Islam, A. M. Sultan, and K. S. Kwak, "A multilayer multimodal detection and prediction model based on explainable artificial intelligence for Alzheimer's disease," *Scientific Reports*, vol. 11, no. 1, p. 2660, 2021. https://doi.org/10.1038/s41598-021-82098-3

[28] N. Ruengchaijatuporn *et al.*, "An explainable self-attention deep neural network for detecting mild cognitive impairment using multi-input digital drawing tasks," *Alzheimer's Research & Therapy*, vol. 14, no. 1, p. 111, 2022.

[29] S. Tekkesinoglu and S. Pudas, "Explaining graph convolutional network predictions for clinicians—An explainable AI approach to Alzheimer's disease classification," *Frontiers in Artificial Intelligence*, vol. 6, p. 1334613, 2024.

[30] T. Mahmud, K. Barua, S. U. Habiba, N. Sharmen, M. S. Hossain, and K. Andersson, "An explainable ai paradigm for alzheimer's diagnosis using deep transfer learning," *Diagnostics*, vol. 14, no. 3, p. 345, 2024. [Online]. Available: https://www.mdpi.com/2075-4418/14/3/345

[31] M. Kim *et al.*, "Interpretable temporal graph neural network for prognostic prediction of Alzheimer's disease using longitudinal neuroimaging data," in *IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2021: IEEE, pp. 1381-1384.

[32] S. Chethana, S. S. Charan, V. Srihitha, S. Palaniswamy, and P. B. Pati, "A novel approach for alzheimer's disease detection using XAI and grad-CAM," in *IEEE Global Conference for Advancement in Technology (GCAT)*, 2023, pp. 1-6, doi: https://doi.org/10.1109/GCAT59970.2023.10353475.

[33] T. Pohl, M. Jakab, and W. Benesova, "Interpretability of deep neural networks used for the diagnosis of Alzheimer's disease," *International Journal of Imaging Systems and Technology*, vol. 32, no. 2, pp. 673-686, 2022.

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

[35] L. A. De Santi, E. Pasini, M. F. Santarelli, D. Genovesi, and V. Positano, "An explainable convolutional neural network for the early diagnosis of Alzheimer's disease from 18F-FDG PET," *Journal of Digital Imaging*, vol. 36, no. 1, pp. 189-203, 2023/02/01 2023. 10.1007/s10278-022-00719-3

[36] S. Jahan *et al.*, "Explainable AI-based Alzheimer's prediction and management using multimodal data," *Plos one*, vol. 18, no. 11, p. e0294253, 2023. https://doi.org/10.1371/journal.pone.0294253

[37] M. Hernandez, U. Ramon-Julvez, F. Ferraz, and w. t. A. Consortium, "Explainable AI toward understanding the performance of the top three TADPOLE Challenge methods in the forecast of Alzheimer's disease diagnosis," *PloS one*, vol. 17, no. 5, p. e0264695, 2022. https://doi.org/10.1371/journal.pone.0264695

[38] A. S. Alatrany, W. Khan, A. Hussain, H. Kolivand, and D. Al-Jumeily, "An explainable machine learning approach for Alzheimer's disease classification," *Scientific Reports*, vol. 14, no. 1, p. 2637, 2024. https://doi.org/10.1038/s41598-024-51985-w

[39] G. O. Ekuma, "An explainable deep learning prediction model for severity of Alzheimer's disease from brain images," Unpublished Manuscript, 2023.

[40] W. Yang *et al.*, "Survey on explainable AI: From approaches, limitations and applications aspects," *Human-Centric Intelligent Systems*, vol. 3, no. 3, pp. 161-188, 2023/09/01 2023. 10.1007/s44230-023-00038-y

[41] Vimbi V. et al, "Application of explainable artificial intelligence in Alzheimer's disease classification: A systematic review," *arXiv preprint arXiv:2304.12345*, 2023.

[42] F. Cruciani, "Explainable artificial intelligence: Enabling AI in neurosciences and beyond," White Paper, ExplainableAI.org, 2023.

[43] F. Eitel, "Explainable deep learning classifiers for disease detection based on structural brain MRI data," Master's Thesis, University of Tübingen, German, 2022.

[44] G. Loveleen, B. Mohan, B. S. Shikhar, J. Nz, M. Shorfuzzaman, and M. Masud, "Explanation-driven HCI model to examine the mini-mental state for Alzheimer's disease," *ACM Transactions on Multimedia Computing, Communications and Applications*, vol. 20, no. 2, pp. 1-16, 2023.

[45] A. Subasi, M. N. Kapadnis, and A. Kosal Bulbul, *4 - Alzheimer's disease detection using artificial intelligence*. Academic Press, 2022, pp. 53-74. https://doi.org/10.1016/B978-0-323-90037-9.00011-4

[46] J. Zacharias, M. von Zahn, J. Chen, and O. Hinz, "Designing a feature selection method based on explainable artificial intelligence," *Electronic Markets*, vol. 32, no. 4, pp. 2159-2184, 2022. https://doi.org/10.1007/s12525-022-00608-1

[47] O. Arreche, T. R. Guntur, J. W. Roberts, and M. Abdallah, "E-xai: Evaluating black-box explainable ai frameworks for network intrusion detection," *IEEE Access*, vol. 12, pp. 23954-23988, 2024.

[48] M. S. Awadallah, F. de Arriba-Pérez, E. Costa-Montenegro, M. Kholief, and N. El-Bendary, "Investigation of local interpretable model-agnostic explanations (LIME) framework with multi-dialect Arabic text sentiment classification," in *International Conference on Computer Theory and Applications (ICCTA)*, 2022, pp. 116-121.

[49] S. Anjomshoae and S. Pudas, "Explaining graph convolutional network predictions for clinicians—An explainable AI approach to Alzheimer’s disease classification," SSRN Preprint No. 4194675, 2022. https://ssrn.com/abstract=4194675

[50] A. Ali, T. Schnake, O. Eberle, G. Montavon, K.-R. Müller, and L. Wolf, "XAI for transformers: Better explanations through conservative propagation," in *International Conference on Machine Learning*, 2022, pp. 435-451.

[51] E. Tjoa and C. Guan, "Quantifying explainability of saliency methods in deep neural networks with a synthetic dataset," *IEEE Transactions on Artificial Intelligence,* vol. 4, no. 4, pp. 858-870, 2022.