Cognitive Computation (2024) 16:1–44
https://doi.org/10.1007/s12559-023-10192-x

Explainable Artificial Intelligence in Alzheimer's Disease Classification: A Systematic Review

Vimbi Viswan[1] · **Noushath Shaffi**[1] · **Mufti Mahmud**[2,3,4] · **Karthikeyan Subramanian**[1] · **Faizal Hajamohideen**[1]

Received: 29 March 2023 / Accepted: 12 August 2023 / Published online: 13 November 2023

Abstract

The unprecedented growth of computational capabilities in recent years has allowed Artificial Intelligence (AI) models to be developed for medical applications with remarkable results. However, a large number of Computer Aided Diagnosis (CAD) methods powered by AI have limited acceptance and adoption in the medical domain due to the typical blackbox nature of these AI models. Therefore, to facilitate the adoption of these AI models among the medical practitioners, the models' predictions must be explainable and interpretable. The emerging field of explainable AI (XAI) aims to justify the trustworthiness of these models' predictions. This work presents a systematic review of the literature reporting Alzheimer's disease (AD) detection using XAI that were communicated during the last decade. Research questions were carefully formulated to categorise AI models into different conceptual approaches (e.g., Post-hoc, Ante-hoc, Model-Agnostic, Model-Specific, Global, Local etc.) and frameworks (Local Interpretable Model-Agnostic Explanation or LIME, SHapley Additive exPlanations or SHAP, Gradient-weighted Class Activation Mapping or GradCAM, Layer-wise Relevance Propagation or LRP, etc.) of XAI. This categorisation provides broad coverage of the interpretation spectrum from intrinsic (e.g., Model-Specific, Ante-hoc models) to complex patterns (e.g., Model-Agnostic, Post-hoc models) and by taking local explanations to a global scope. Additionally, different forms of interpretations providing in-depth insight into the factors that support the clinical diagnosis of AD are also discussed. Finally, limitations, needs and open challenges of XAI research are outlined with possible prospects of their usage in AD detection.

Keywords Alzheimer's Disease Classification · Ante-hoc · Blackbox Models · Explainable Artificial Intelligence · Intrepretable Machine Learning · Model-Agnostic · Model-Specific · Post-hoc · XAI

Vimbi Viswan and Noushath Shaffi are joint first authors.

✉ Mufti Mahmud
muftimahmud@gmail.com

Vimbi Viswan
vismaya97@gmail.com

Noushath Shaffi
noushath.mys@gmail.com

Karthikeyan Subramanian
skaarthi@gmail.com

Faizal Hajamohideen
faizal.h@gmail.com

[1] College of Computing and Information Sciences, University of Technology and Applied Sciences, Suhar, Oman

[2] Department of Computer Science, Nottingham Trent University, Clifton, Nottingham NG11 8NS, UK

[3] Medical Technologies Innovation Facility, Nottingham Trent University, Clifton, Nottingham NG11 8NS , UK

[4] Computing and Informatics Research Centre, Nottingham Trent University, Clifton, Nottingham NG11 8NS, UK

Alzheimer's Disease (AD) is an untreatable, life-changing neurological sickness affecting the elderly population, leading to several hardships for the patients [1, 2]. According to the latest World Alzheimer Report [3], about 55 million clinically diagnosed AD patients live worldwide, with an estimated rise of cases to 139 million by 2050 [3]. The report also quotes a staggering 75% issues that go undiagnosed for several reasons. AD patients will have to endure many difficulties, such as memory loss, behavioural disturbances, vision, and mobility complications that interfere with daily routine tasks [1, 4, 5]. These sufferings will increase to the extent of interfering with one's ability to lead a self-reliant personal and social life and causing numerous tribulations for the caretaking family members [2, 6].

Of late, Artificial Intelligence (AI) techniques involving machine learning (ML) and deep learning (DL) algorithms have contributed to diverse application domains including: anomaly detection [7–9], biosignal and image analysis [10–21], neurodevelopmental disorder assessment and classification focusing on autism [22–32], neurological

disorder detection and management [2, 33–44], supporting the detection and management of the COVID-19 pandemic [45–52], elderly monitoring and care [53–56], cyber security and trust management [57–62], various disease diagnosis [63–69], smart healthcare service delivery [70–72], text and social media mining [73–76], personalised learning [77–80], earthquake detection [81, 82], etc.

Part of these methods has significantly boosted clinical diagnosis of AD in an incredibly accurate, fast, and efficient manner using compound medical data (2D or 3D MRI, PET, CT, etc) [83–85]. This success can be attributed to several factors, such as certain algorithmic advancements and the availability of powerful GPUs, which come pre-loaded with a spectrum of open-source computational tools [84, 85]. These have facilitated the accurate identification of AD in a remarkable manner. The AI-driven AD prediction is based on the concept that systems can identify stages of dementia by learning patterns through the input data so that optimal decisions can be made with minimal human intervention [86, 87]. The contemporary ML and DL algorithms for AD detection have achieved highly admirable results on various scales of metrics [34, 85].

However, these AI models are considered mainly as blackbox models by medical practitioners due to the inability to derive justifiable reasons (explainability) for the predictions delivered by them, leading to ambiguity [88]. The high opaqueness of these modern AI techniques often poses difficulty for even skilled medical experts to comprehend the solutions [89]. For this reason, it will lead to a trade-off of accuracy over trustworthiness by decision-makers [90]. Consequently, stakeholders and policymakers often prefer responsible and reliable decision-making instead of accurate decision-making. This lack of explainability keeps the medical field reluctant to deploy AI-driven computer-aided diagnosis (CAD), despite proven accuracy demonstrated in the recent literature [91].

In the last decade, several ML and DL algorithms have achieved breakthrough results in various AI-based decision-making, such as disease prognosis and prediction [92], drug discovery and development [93], solid-state material science [94] and machine fault analysis [95]. Furthermore, applications of DL are found in biomedicine [96], biology [96], and speech recognition, synthesising and audio processing in [97]. Sometimes these performances surpassed human-level accuracy.

Such blackbox models will often lead to unclear circumstances such as "Why did you predict/classify that as class x, why not class y?", "When will you succeed or fail ?", "How to correct the wrong feature selection?", "Which dominant feature are you looking to train the model?", "Can I rely on the prediction you gave?" and so on [89] (see Fig. 1). On the contrary, the Explainable AI (XAI) models can deliver reassuring outcomes to the user, such as "I know why you are classifying that as class x and why not as class y", "I

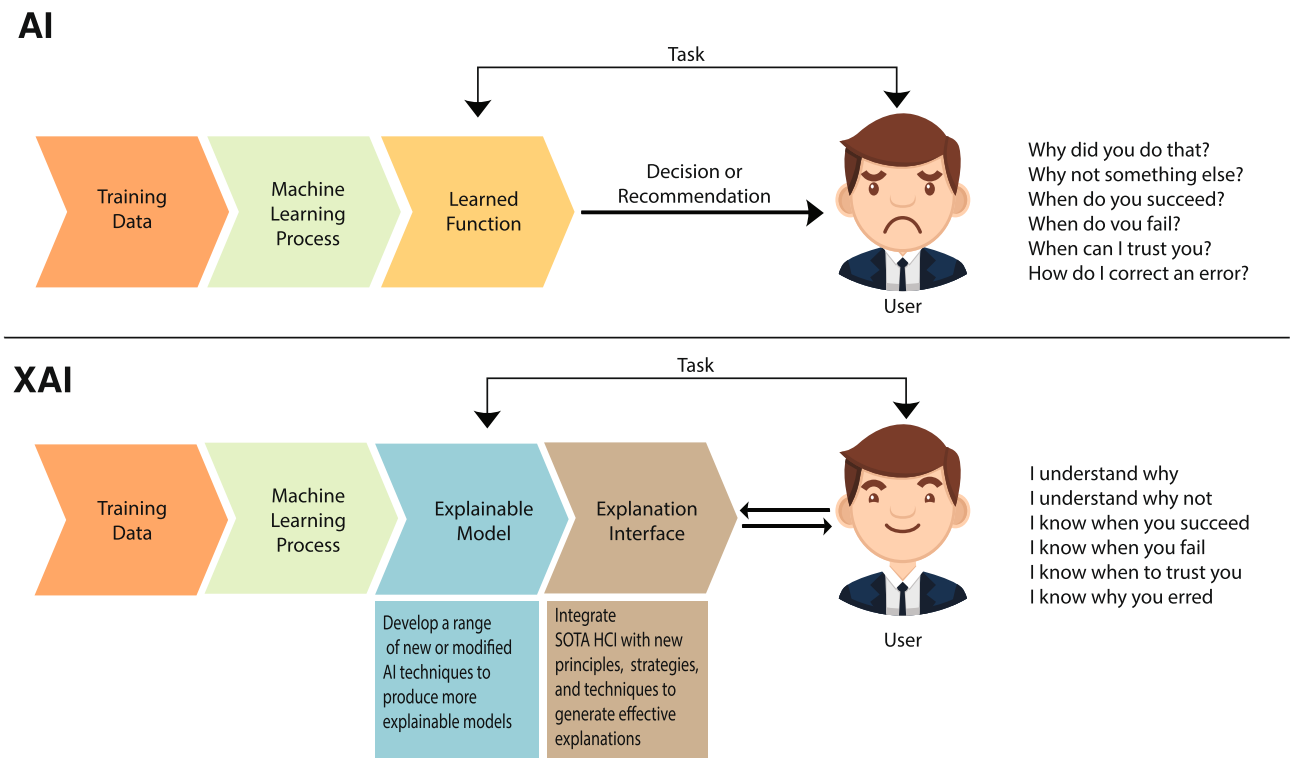

Fig. 1 A Typical AI systems (Top) and Explainable AI systems (Bottom)

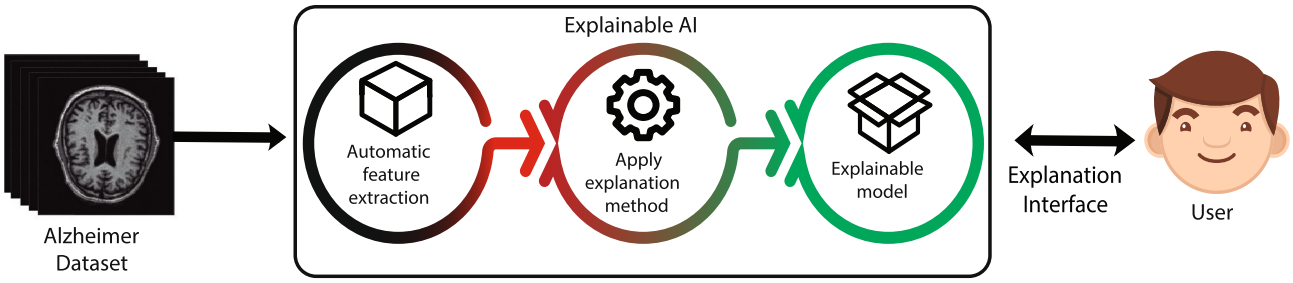

Fig. 2 A general outlook of explaining models

know to rectify the wrong feature selection", "I can rely on the prediction you gave" and alike. Hence, XAI is crucial if AI-based CAD are to be reliably deployed.

XAI or interpretable AI are convertible terms that refer to an emerging sub-field of AI [98]. The XAI is a series of features that interprets how a blackbox model is constructed, perform predictions and get humans to trust and use the system efficiently [99, 100]. The need for a model to be explainable is to justify the model output, make the functioning of the blackbox models transparent, gain new knowledge for smarter decision-making to improve model performances and increase trust by users in the model result [98]. It aims to produce methods and tools to make the AI systems' decisions, recommendations, or guidance understandable for decision-makers. For instance, in an AI system for classifying MRI images in early AD diagnosis, XAI can explain the model's working and synthesise influential factors considered for prediction [101]. Furthermore, if the model generates an adverse result, then with XAI's interpretability, the model will be able to identify and rectify the errors [102]. Explanation and interpretation of the model's output are therefore required to bring fidelity, trust and use in clinical applications [88, 100, 102]. The stakeholder's trust at every level is needed to maximally leverage these AI solutions, which is possible only through XAI. In addition to providing advanced insights into AI solutions, XAI can also deliver new opportunities. For instance, involving a human in the decision is a typical medical scenario, where AI solutions and human expertise go in tandem to tackle complex situations where neither can provide a satisfactory solution [103]. Figure 2 shows a general outlook of translating the blackbox model to an explainable model.

There is often a trade-off between model accuracy and associated explainability [104]. Linear regression models or decision tree(DT) models are intuitive, inherently interpretable, and easy to validate and understand by a novice in AI. This increases the trust in such models. However, to solve a complex problem, ML algorithms may derive a non-linear model which would yield good results but compromises on explainability. For instance, Convolutional Neural Network (CNN) often performs the best but is least explainable [105]. Figure 3 shows that ML models with high explainability are less accurate but more intuitive to humans. As the model complexity and performance increase, leading to more accuracy in results, explainability decreases. In healthcare systems, predictive accuracy is the most important measure of clinical validation. From the patient's perspective, there is more trust in the clinician and less tolerance for the machine, which naturally raises the importance of explainability, allowing more complex models and functions to be explainable.

In the last couple of years, XAI has gained paramount importance in the AI community not only because they are used in high-stake decisions but also because companies are held accountable by regulators for the decisions made by their AI models. It has grown exponentially in a short time span, potentially transforming how AI is seen and deployed in real-time in the coming days. Several diverse fields have embraced the explainable component of AI, prioritising trustworthiness over accuracy. XAI has been applied in drug discovery [106, 107], industrial applications [108, 109], gaming [110, 111], neurological disorders [112–114], neuroscience [115, 116] and recommender systems [117, 118]. This tremendous growth has led to several XAI-based review articles in the healthcare domain in the past years.

The interpretability of ML algorithms was the subject of a comprehensive survey by Tjoa and Guan [119]. The findings were further categorised into different groups by the author. These categorisations are studied from the perspective of application in the medical field. Authors in [89] have surveyed the progress of XAI in healthcare applications. They have also introduced solutions for XAI leveraging the fusion of data from multi-centric data with different modalities. The results of which were analysed and subsequently validated in two real clinical scenarios. In the review presented by Loh et al. [99], a detailed review of areas of healthcare deserving more attention of XAI was presented by considering three major healthcare datasets: clinical, textual, and

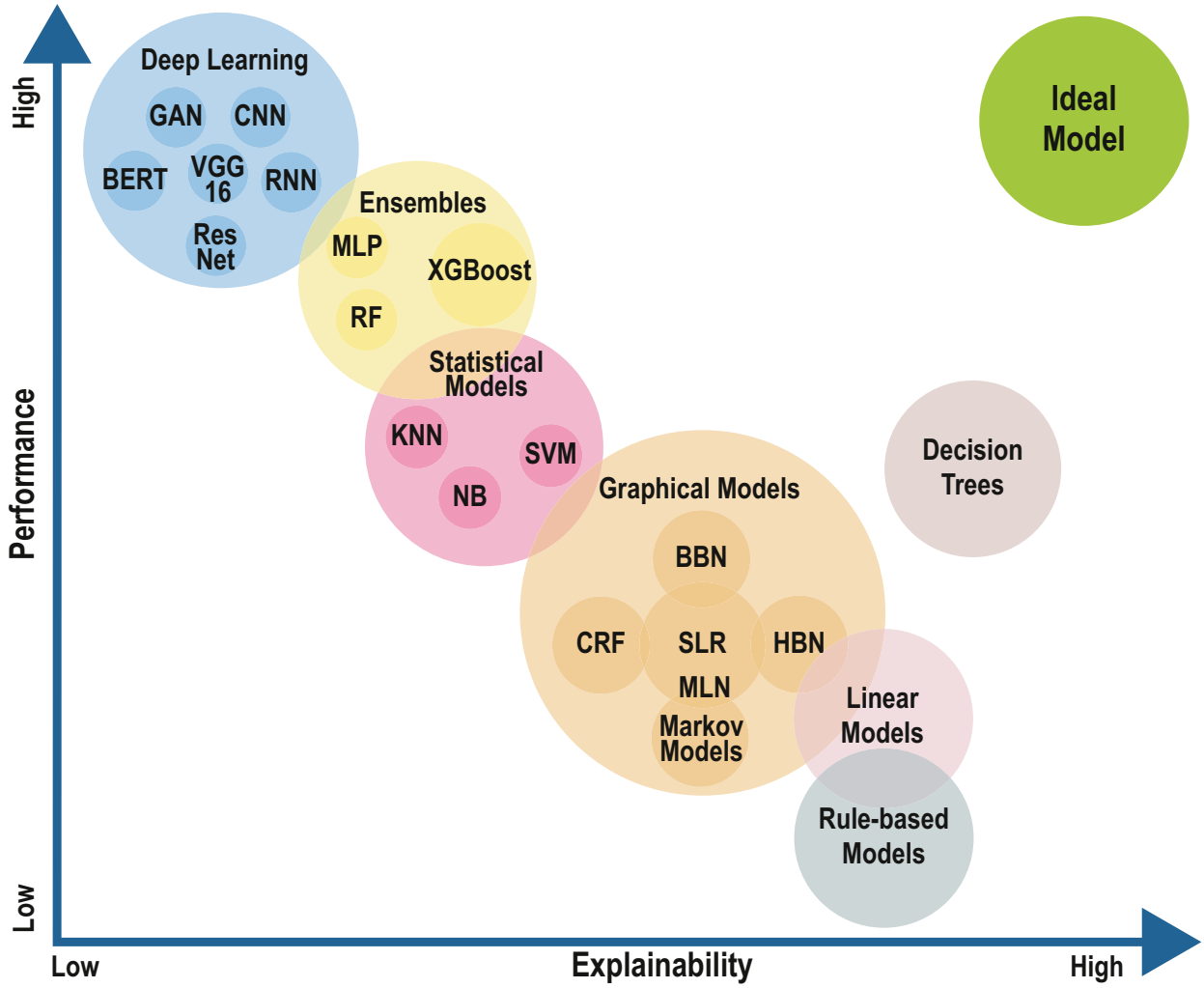

Fig. 3 Accuracy - Interpretability Trade-off

high-dimensional data. Nazar et al. [91] have discussed the XAI from the perspectives of Human-Computer Interaction (HCI) models. The authors focused on using AI, HCI, and XAI in healthcare. There have been several noteworthy applications of XAI in AD classification too [120–134] However, an exclusive systematic review article on XAI for AD classification that points to various XAI frameworks and blackbox models used inside these frameworks is yet to be proposed by the research community.

Also, a wide spectrum of XAI reviews involve crucial components such as various blackboxes considered for interpretation, XAI frameworks, XAI methods, and various output forms of interpretation. Other components include open-source XAI tools, their implementation platform, and associated datasets. Addressing blackboxes interpreted in AD studies alone will only provide partial coverage of this spectrum. Hence, it is essential to fully comprehend the complete XAI for the AD platform to do any worthwhile research in the future. The novelty of this review article is that it covers the entire XAI spectrum in interpretability of blackbox models used for AD detection. To the best of our knowledge, this is among the first attempts to review the XAI models in the context of AD diagnosis. The nomenclature used in this article is listed in Table 1.

The primary contribution of this work can be outlined as follows:

1. A systematic review adhering to the guidelines proposed by both Kitchenham [135] and PRISMA [136].
2. Formulation of essential research questions (RQ) covering the entire spectrum of XAI for AD classification.
3. Collection of different XAI techniques with their GitHub links used in interpreting blackbox models applied for AD detection.

Table 1 Nomenclature

Abbreviation	Description	Abbreviation	Description
18F-FDG	18F-Fluorodeoxyglucose	LR	Logistic Regression
3D CNN	Three-dimensional Convolutional Neural Network	LRP	Layer-wise Relevance Propagation
3D ResAttNet	3D Residual Attention Deep Neural Network	Ma	Model agnostic
3D VGG16	three-dimensional Visual Geometry Group 16	MCI	Mild Cognitive Impairment
AD	Alzheimer's Disease	mdDem	Mild Dementia
AdaBoost	Adaptive Boosting	ML	Machine Learning
ADNI	Alzheimer's Disease Neuroimaging Initiative	MLP	Multilayer Perceptron
ADReSS	Alzheimers Dementia Recognition through Spontaneous Speech	MMSE	Mini-Mental State Examination
Ah	Anti-hoc	MoCA	Montreal Cognitive Assessment
AI	Artificial intelligence	moDem	Moderate Dementia
AIBL	The Australian Imaging, Biomarker & Lifestyle Flagship Study of Ageing	MPRAGE	Magnetisation Prepared - Rapid Gradient Echo
aMCI	amnestic Mild Cognitive Impairment	MRI	Magnetic Resonance Imaging
ANN	Artificial Neural Network	Ms	Model specific
ApOE	Apolipoprotein E.	NGLY1	N-Glycanase 1
AUROC	Area Under the Receiver Operating Characteristic	NLP	Natural Language Processing
BERT	Bidirectional Encoder Representations from Transformers	noDem	No Dementia
bvFTD	Behavioural Fronto Temporal Dementia	nWBV	normalised Whole Brain Volume
CAD	Computer Aided Diagnosis	OASIS	Open Access Series of Imaging Studies
CDRSB	Clinical Dementia Rating Sum of Boxes	OCM	Occlusion Sensitivity Mapping
CDT	Clock Drawing Test	PCA	Principal Component Analysis
CNN	Convolutional Neural Network	PCR	Prediction basis Creation and Retrieval
CSF	Cerebrospinal Fluid	PET	Positron Emission Tomography
CT	Computerised Tomography	Ph	Post-hoc
D-BAC	DCGAN-based Augmentation and Classification	pMCI	progressive Mild Cognitive Impairment
DCGAN	Deep Convolutional Generative Adversarial Network	PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses
DL	Deep Learniing	Py	Python
DT	Decision Tree	R	R Language
EEG	Electroencephalogram	RAVLT-perc-for-getting	Rey's Auditory Verbal Learning Test
EMCI	Early Mild Cognitive Impairment	ResAttNET	Residual Attention Network for Image Classification
FDG	Fluorodeoxyglucose	ResNetGAP	Residual Network Global-Average-Pooling
FTD	Fronto Temporal Dementia	RF	Random Forest
GAN	Generative Adversarial Networks	RNN	Recurrent Neural Network
GAP	Global Average Pooling	RQ	Research Question
Gl	Global	SHAP	SHapley Additive exPlanation
GNNExplainer	Generative Adversarial Networks Explainer	SM	Saliency Map
GradCAM	Gradient-weighted Class Activation Mapping	sMCI	stable Mild Cognitive Impairment
HAM	High-Resolution Activation Mapping	SMILE	Statistical Machine Intelligence and Learning Engine
HC	Healthy Controls	sMRI	Structural Magnetic Resonance Imaging
HCI	Human-Computer Interaction	SVC	Support Vector Classifier
HTR1F	Hydroxytryptamine Receptor 1F	SVM	Support Vector Machine
ICE	Individual Conditional Expectation	SVM-SMOTE	Support Vector Machines -Synthetic Minority Oversampling Technique
IML	Interpretable Machine Learning	T-GNN	Transferable Graph Neural Network

Table 1 (continued)

Abbreviation	Description	Abbreviation	Description
IoT	Internet of Things	TADPOLE	The Alzheimer's Disease Prediction Of Longitudinal Evolution
		UBAC2	Ubiquitin-Associated Domain-Containing protein 2
kNN	k-Nearest Neighbours	VGG16	Visual Geometry Group 16
LDELTOTAL	Logical Memory Delayed Recall Total	vmDem	very mild Dementia
LGBM	Light Gradient-Boosting Machine	WMH	White Matter Hyperintensities
LIME	Local Interpretable Model Agnostic Explanation	XAI	Explainable Artificial Intelligence
Ll	Local	XGBoost	Extreme Gradient Boosting
LMCI	Late Mild Cognitive Impairment	XGNN	Explanations of Graph Neural Networks

4. A survey of XAI methods for AD classification reported in the last ten years, with critical analysis of findings, results, abilities, and limitations.
5. Identification of the XAI models' strengths for AD detection to ensure their reliability and trustworthiness for adoption by clinicians.
6. A focused discussion on current XAI research, it's benefits, limitations, and challenges along with future directions.

These significant findings will help the research community fill various research gaps, instigating new models that assist clinicians in elucidating the perception of an AI system.

The rest of this article is structured as follows: "Concepts and Background" provides necessary concepts and background in XAI. The data synthesis needed for the systematic review is detailed in "Search Strategy". The findings for the research questions are discussed in "Results and Discussions", and concluding remarks are drawn in "Conclusion".

Concepts and Background

This section provides a succinct background to different XAI methods in general (XAI Methods). It also provides a brief overview of popular XAI frameworks used in the AD prediction (XAI Frameworks). The primary aim of this section is to provide a comprehensive background helpful for discussions in later sections.

XAI Methods

The XAI methods can broadly be classified into four categories [98], as shown in Fig. 4, based on: i) scope of explanations, ii) stages of implementation, iii) applicability to models, and iv) forms of explanation.

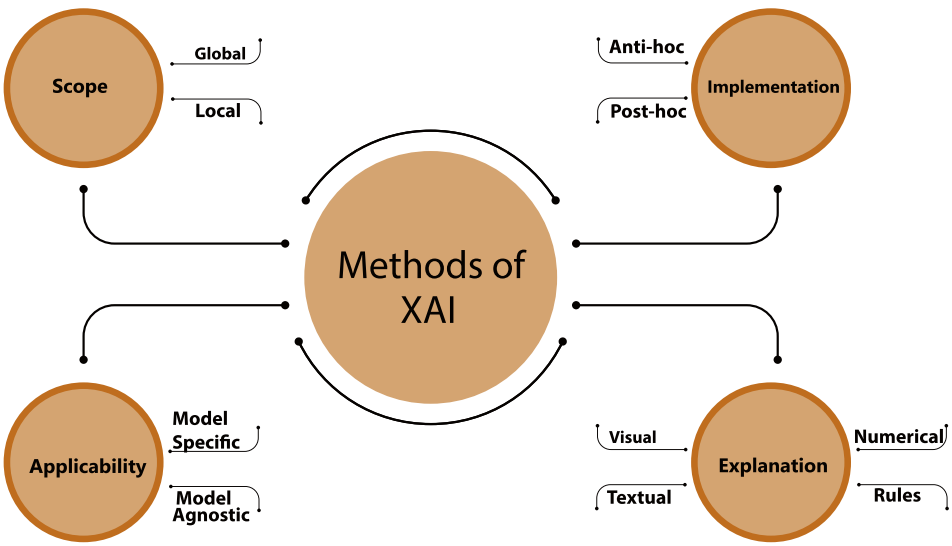

Fig. 4 Classification of XAI Methods

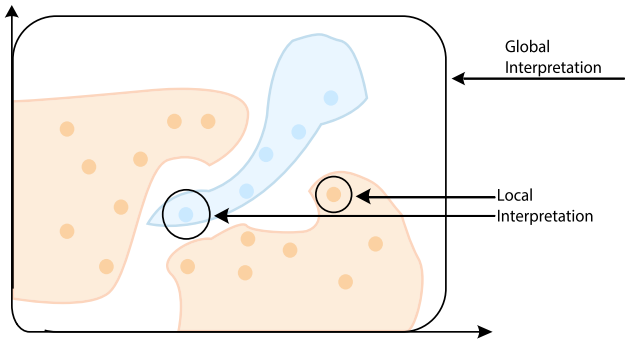

Fig. 5 XAI Explanation: Local vs. Global

XAI Methods Based On Scope

The explainability scope is the extent of explanations generated by the XAI method. It can interpret either an entire model or a specific instance of the model based on input test data. Accordingly, the explainability of a model can be either local or global. A global method explains the whole model by considering the entire inferential data set of the model. It gives a general perspective of the relationship of the model with all input instances. The popular DT algorithm can be intrinsically global in nature because the decision-making for all input data instances can be easily explained by visualising in a tree form.

On the other hand, the goal of a local method is to explain only a few instances of test data to the user. Local explanations help understand why certain decisions were made and can increase user confidence in specific examples. In the case of DTs, a local explanation can correspond to a single branch in the tree. It is worth noticing that, combining local inferences made through different input instances can yield global insights for the model. Pictorial representation of local vs global explanation can be represented as shown in Fig. 5.

XAI Methods based on Stages of Implementation

An XAI method can generate explanations for a model either during or after the training of the model (see Fig. 6). Based on these two ways, XAI methods are further classified as Ante-hoc and Post-hoc [98]. Ante-hoc is a Latin word that literally translates into ***before-this***. Hence, the goal of ante-hoc XAI methods is to provide explainability ***before*** the beginning of model training. Such Ante-hoc methods are transparent and self-explanatory and make the model explainable naturally while maintaining optimal accuracy [98]. ML algorithms that are ante-hoc in nature are linear regression, DT, and Bayesian models. These models are also referred to as white box or glass box models.

The Latin word Post-hoc translates to ***after-this***. Such methods aim to provide explainability ***after*** a model has been trained. An external explainable model, called a surrogate model, is augmented to provide explanations for a trained blackbox model. Generally, support vector machines (SVM), and CNN are the ones where the inference mechanisms remain completely unknown to users that necessitate post-hoc models. Gradient-weighted Class Activation Mapping (GradCAM) [137], Layer-wise Relevance Propagation (LRP) [138] and Local Interpretable Model Agnostic Explanations (LIME) [100] are some common examples of XAI frameworks that can be applied on surrogate models for generating explanations.

XAI Methods based on Applicability of Models

Applicability of models refers to a concept of XAI where explainable methods are restricted to particular models or applied to any model as a post-process. The former is called model-specific, and the latter is a post-hoc approach called model-agnostic. Model-specific is an intrinsic approach where explainability is integrated into the model architecture and is not transferable to any other model

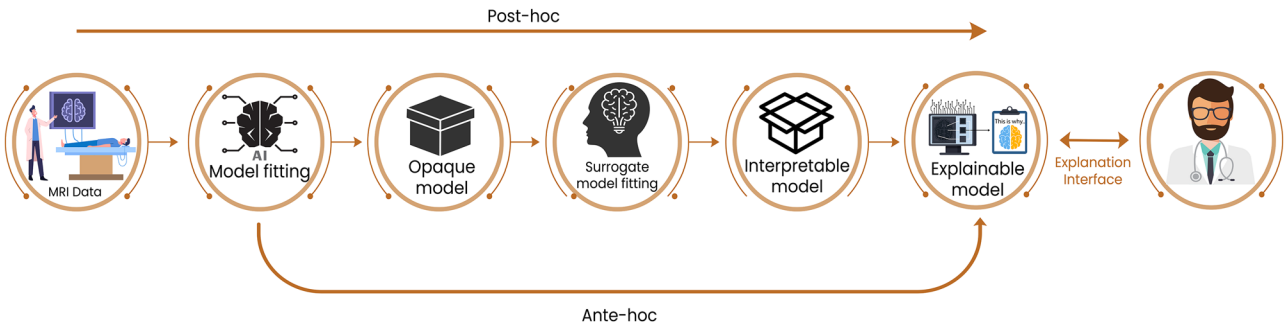

Fig. 6 XAI Explanation: Ante-hoc vs. Post-hoc

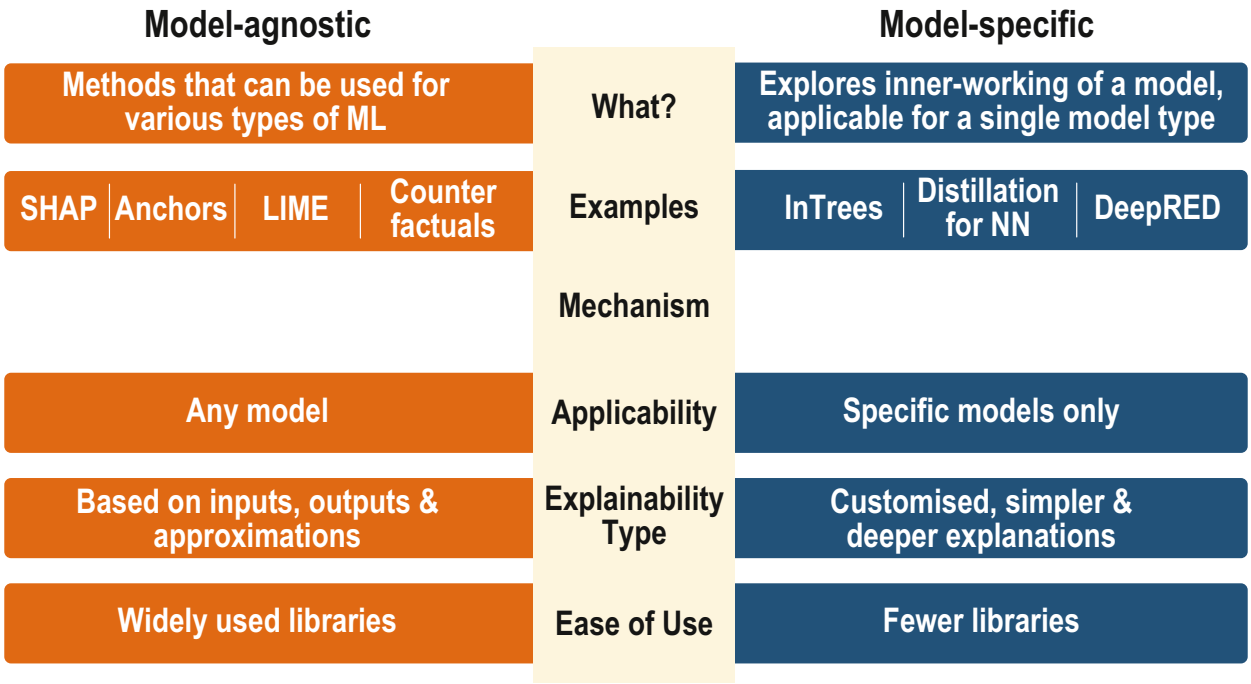

Fig. 7 XAI Explanation: Model Specific vs. Model Agnostic

architecture. For instance, interpreting a neural network model's weight or activation values is specific to that neural network learning approach. Model-specific approaches for deep neural networks work by traversing the path of CNN in reverse order highlighting specific regions of the input image that majorly contributed to the decision. The guided backpropagation [139] and LRP [138] are example model-specific approaches.

On the other hand, model agnostic methods do not consider any of the internal components like weight or activation values and can be used with any learning approach. They extract explanations by causing perturbation and mutation to the input data and subsequently observing the sensitivity of the performance compared to the original data. In other words, by perturbing the input or weights of important features we can measure how much it has influenced the model performance. This will provide valuable insights into a localised region of input data that underwent perturbation. Alternative methods used are Occlusion Sensitivy analysis [134], GradCAM [137] and Feature Importance [126]. Some popular model-agnostic approaches are LIME [100] and SHapley Additive exPlanations (SHAP) [140]. Figure 7 depicts a few key differences between these two types of XAI methods.

XAI Methods-based on Forms of Explanation

The classification model for images can differ substantially from those that classify using text, categorical, or temporal data such as speech. Therefore, the input formats (numerical, visual, textual or temporal) for a model can play a vital role in framing different forms of explanations for XAI methods. The interpretations of predictions can be in many forms and depend on the end users' needs and concerns. There are four forms of explanations commonly used to interpret a prediction: numerical, visual, rule-based, and textual [98].

The numerical explanations generated by models are usually a measure of the input variables that contribute to the model's outcome. They represent numerical formats like values, vectors of numbers, or matrices. A numerical explanation can also be a probability measure assigned to a neural network layer. Visual explanations are the most common way to explain the functioning of a model in a graphical way. For example, heatmaps can highlight important areas of an image that are influential for the decision. A visual explanation can be easily understood by novice end users of the AI model. Textual explanations are precise and specific and generally used for individual predictions. It is not a commonly used form

Table 2 Popular XAI Frameworks in the AD Classification

XAI	Ll	Gl	Ah	Ph	Ms	Ma	Forms	GitHub link	Lang Frameworks
LIME [100]	✓	✓	crossmark	✓	crossmark	✓	T,V	https://github.com/marcotcr/lime	Py, JS
GradCAM [137]	✓	✓	crossmark	✓	crossmark	✓	V	https://github.com/leftthomas/GradCAM	Py
LRP [141]	✓	✓	crossmark	✓	✓‘	crossmark	V	https://github.com/chr5tphr/zennit	Py
XGNN [142]	✓	✓	crossmark	✓	✓	crossmark	V	https://github.com/divelab/DIG	Py
SM [143]	✓	✓	crossmark	✓	crossmark	✓	V	https://github.com/raghakot/keras-vis	Py
Occlusion [144]	✓	✓	crossmark	✓	crossmark	✓	V	https://github.com/deel-ai/xplique	Py
Sensitivity ICE [145]	✓	✓	crossmark	✓	crossmark	✓	V	https://christophm.github.io/	R, Py
SHAP [140]	✓	✓	crossmark	✓	crossmark	✓	N,V	https://github.com/slundberg/shap	Py, C++, JS

Legends: Ll–Local; Gl–Global; Ah–Ante-hoc; Ph–Posthoc; Ms–Model Specific; Ma–Model Agnostic; T-Textual; V–Visual; N–Numeric; Py–Python; JS–Java Script; R – R programming; Lang – Programming Language

of explanation due to its high computational complexity requiring natural language processing (NLP). However, they are easily understandable to humans and are mostly generated for local scope. Rule-based explanations are simple and basic forms of explanation which are more structured than visual and textual explanations. They are intuitive to humans and can be used to explain the prediction of models using IF-THEN rules or trees with AND/OR operators [98]. This type of explanation is mainly utilised for ante-hoc methods and interprets models with global scope. A detailed discussion of these forms of explanations is dealt with in "XAI Frameworks for AD Detection" and "Benefits of using XAI Methods for AD Detection" in the context of AD.

XAI Frameworks

Local Interpretable Model-Agnostic Explanations

LIME is an open-source tool used to generate explanations for a single instance instead of the entire dataset, hence the term local. LIME provides explanations by perturbing the model's input data, creating a surrogate model, and observing the changes in prediction and selecting the top significant features [100]. Due to agnosticity of the LIME model, it is used after the model has been trained for prediction and can be used for any blackbox model. For blackbox explanations, LIME can interpret image classifications, explain text-based models and tabular datasets in either textual, numeric or visual form (for more details, see section XAI Frameworks for AD Detection).

SHapley Additive exPlanations

SHAP is an XAI technique that assigns a weight, called Shapley value, to each feature of a trained model [140]. These features with an assigned weight are observed for all possible weighted input combinations. The contribution of each of the Shapley value-added features, for all possible weighed input combinations, is observed based on its efficiency, symmetry, features with no zero contributions and cumulative contribution of a feature with sub-parts. SHAP shows performance consistency and provides good accuracy for predictions in the local scope. In this review, SHAP was commonly used in conjunction with numerical data to provide a visual analysis of blackbox models (see Figs. 15, 16, and 26).

Gradient-weighted Class Activation Mapping

GradCAM is a technique used to make CNN models more transparent by identifying the important regions of an input image for predictions [146]. GradCAM is applied using gradient information of the output layer of the CNN model to produce a localisation map representing crucial regions in an image. This is achieved by assigning important value for each neuron for making specific decisions. Therefore, the final output of GradCAM is a coarse heatmap that highlights important regions suitable for prediction and explanation (see Fig. 19).

Layer-wise Relevance Propagation

LRP is another tool like GradCAM that generates a heatmap with highlighted regions of an image [138]. LRP is used in CNN where inputs can be images or videos. LRP assigns relevance scores to all neurons of a specific output for the last layer of a CNN. LRP then propagates in reverse until the input layer by computing scores for every activation unit (neuron) in each layer. Using the final relevance score, a heat map is generated by LRP as an explanation that can be used to identify influential regions in the prediction (see Figs. 17 and 18).

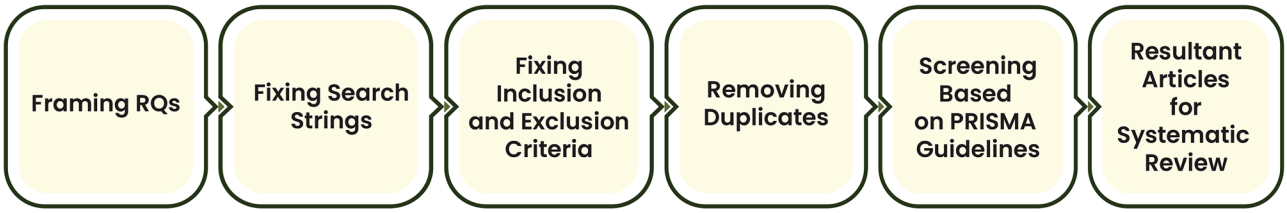

Fig. 8 Sequence of Steps in Search Strategy to Identify Relevant Papers

Individual Conditional Expectations

ICE is an extension of Partial Dependence Plot (PDP) that is used to produce visual explanations for blackbox models [145]. PDP is a visualisation obtained by plotting the average predicted outcome of a model by varying the value of one feature of interest and keeping the other feature values constant. ICE disaggregates the PDP by creating individual plots for each instance of blackbox model predictions by altering the value of a feature of interest and leaving other feature values unchanged [126]. The outcome can be visualised as a line plot which is a set of points for an instance with the altered feature value and the respective predictions (see Fig. 20).

Occlusion Sensitivity Analysis

In an image predicting AD, it is necessary to explain or identify areas in the image that contribute to AD classification. OSA is a technique initially proposed by Zeiler and Fergus [144]. In this technique, portions of the input image are occluded or hidden with a grey or black patch, creating a heatmap. The variations in the output probability of the occluded image are observed [147]. The most critical region, if occluded, will have the highest impact with low probability. An occlusion sensitivity map is therefore used to locate important patches of the image responsible for AD.

Saliency Map

The SM is another important concept of deep learning, which was first introduced by Simonyan et al. [148]. Unlike an occlusion map where portions of the input image are hidden with a black patch and creating a heatmap, in SM each pixel of an AD-classified image is removed and subsequently processed. The heatmap obtained is checked for probability variations where a low probability indicates that the removed pixel plays a vital role in AD classification [149]. Therefore the output heatmap that undergoes the Saliency technique has all important pixels in the image eligible for explaining the disease.

Table 2 provides XAI framework categorisation based on scope, applicability, implementation and interpretable forms for some popular XAI tools used in the literature. The table also provides links to GitHub repositories for rapid reproducibility.

Search Strategy

This section presents the overall steps involved in searching and identifying relevant papers needed for conducting a systematic review. Figure 8 depicts the total stages involved.

This review aims to investigate research articles that use XAI in diagnosing/early detection and thereby interpret the reasons for classifying. To locate contributions and summarise the results, published articles on the subject of artificial intelligence and its associated fields are examined.

The prime aim of this review is to identify research gaps that would instigate XAI-based research for AD detection.

We adopted concrete guidelines proposed by PRISMA [136] and Kitchenham [135] to retrieve relevant papers for this systematic review. The overall process can be outlined below:

- Formulating the research questions.
- Framing search strings.
- Identifying the digital libraries and conducting searches.
- Choosing the relevant inclusion/ exclusion criteria and filtering the papers based on their relevance to the study topics.
- Extracting necessary information from the selected articles.
- Investigating research questions allowing critical analyses to perform a thorough study of the existing methods and their benefits, future needs and limitations.

Research Questions

The main goal behind framing research questions is to lay out a well-defined plan to retrieve papers exclusively from the focused areas of consideration. This way, the reader can apprehend the knowledge disseminated more comprehensively. Table 3 lists the research questions that were addressed in this paper.

Table 3 Research Questions

Sr.No	Research Questions	Motivation
RQ1	What AI systems are available for AD research that incorporate XAI?	To know the blackbox models for AD detection that uses XAI for enhanced clinical fidelity.
RQ2	What different XAI methods are used for blackbox interpretability to detect AD?	To find the number and type of blackbox models that are interpretablity for AD, to know the preliminary steps taken to be post-hoc, ante-hoc, model agnostic, model specific, etc and whether it opens further avenues.
RQ3	What XAI frameworks are available in the literature which are used in AD detection in healthcare in general?	To discover different XAI frameworks and tools relevant to AD detection for the last decade.
RQ4	What are the proven benefits of using XAI in AD?	To know the distinction and applicability of using XAI tools/methods in explaining AD predictions and implications in the medical community.
RQ5	What are the limitations, challenges, needs, and prospects of XAI in AD detection and healthcare in general?	To grasp fundamental capabilities and limitations of existing XAI approaches in AD detection, to identify research gaps instigating further research.

Identification of Articles

One of the challenging tasks for a comprehensive and inclusive systematic review is to select the appropriate search strings. For this work, the search strings were carefully picked so that they are not too generic to avoid irrelevant papers and not too narrow to lead to missing relevant articles [136]. After several trials of combinations and permutations of relevant keywords, we arrived at the search strings as shown in Table 4.

Research papers were selected from widely accessed databases, as shown in Table 5. Apart from these, we have also considered some books and other online resources that satisfied our research questions.

Screening of Articles

A consolidated output of the individual searches produced a total of 1551 records of publications (ACM Digital Library=206, IEEEXplore=147, SpringerLink=158 PubMed=780, ScienceDirect = 260). We decided to include all research articles from the past decade until June 2023. The records were then pruned with duplicates and all those records published before 2012.

In the following task, we screened the identified articles using the inclusion-exclusion criteria shown in Table 6. To begin with, we examined and marked all duplicate records from each of the database search collections. The records after duplication from each collection are combined into a single collection, and duplicates are removed. The task resulted in 928 unique records. We then examined publication titles and abstracts and removed pilot studies, editorials, non-journal articles, conference proceedings, books, posters, and studies published before 2012. The process effectively reduced the number of articles to 73 eligible records.

Furthermore, we excluded inaccessible records, studies that presented only discussion without evidence of model performances and results, and studies that did not relate to the previously framed research questions. The resulting records were then screened for articles relevant to the research questions. Additionally, understanding the accuracy, specificity, sensitivity, and AUROC metrics of ML or DL models is crucial for this review study. We, therefore, excluded studies that did not provide model performances. Overall, this step resulted in 37 credible research articles from quality journals in accordance with our framed RQs. Figure 9 shows a proper understanding of the steps taken in the process. Figure 10 depicts the source and Fig. 11 shows the year-wise statistics of articles considered in our study. These numbers make it clear that XAI's scientific research for AD is limited and has only grown rapidly in the last few years. To our knowledge, this review can be considered as unique as there were no articles found exclusively on XAI with AD.

Results and Discussions

In this section, we present our findings by extensively reviewing the 37 articles through the RQs shown in Table 3.

XAI-based AI Systems for AD Research

This section aims addressing RQ1: What AI systems are available for AD research that incorporate XAI?

The focus on AI in disease diagnosis and treatment began in the early 1970 [84] and has achieved significant momentum over the years. Research in AI-based AD

Table 4 The Search Strings

Sl.No	Search Strings
1	"Alzheimer" explainable AI
2	"Alzheimer" interpretable AI
3	"Alzheimer" explainable ML
4	"Alzheimer" interpretable ML
5	"Alzheimer" explainable DL
6	"Alzheimer" interpretable DL
7	"Alzheimer" post hoc explainable AI
8	"Alzheimer" blackbox explainable AI
9	"Alzheimer" XAI

Table 5 The databases considered

Sr. No	Database
1	IEEEXplore (https://ieeexplore.ieee.org/)
2	ScienceDirect (https://www.sciencedirect.com/)
3	SpringerLink (https://link.springer.com/)
4	ACM Digital Library (https://dl.acm.org/)
5	PubMed (https://pubmed.ncbi.nlm.nih.gov/)

prediction did not involve XAI until the recent decade [150]. Although the time has not yet come for computers to replace doctors, XAI has recently been incorporated into AI-based AD prediction due to a growing demand for transparency and explainability in healthcare and medical practice.

Several studies for AI-based AD detections incorporating XAI have been identified [see Table 7]. Many studies have used datasets of numeric data type for training AI models and obtained explainable results. El Sappagh et al. [122] have developed and utilised a multi-layered multi-model system for an accurate and explainable AD diagnosis. Lombardi et al. [151] present a robust framework for classification between Healthy Control (HC), Mild Cognitive Impairment (MCI), and AD and interpret the predictions with XAI methods. Xu and Yan [152] propose a reliable multi-class classification model supported by XAI methods to explain the predictions accurately. A computer approach called Systems Metabolomics utilising Interpretable Learning and Evolution (SMILE) is proposed by Sha et al. [153]. This article involves a supervised metabolomics data analysis and uses the XAI method to learn and identify the most informative metabolites to understand and diagnose disease development and progression. Hammond et al. [154] analyse Beta-amyloid, tau, and the neuro-degenerative biomarkers, for AD classification. Additionally, the author uses XAI methods to identify the biomarker that is most influential in AD detection. The research article used a numeric data type dataset as input for subjects of different categories like HC, MCI, or AD selected from the Alzheimer's Disease Neuroimaging Initiative dataset (ADNI). Bloch and Friedrich [123] state that the diverse causes of AD can lead to inconsistencies in disease patterns, protocols used for acquiring scans, and preprocessing errors of MRI scans resulting in improper ML classification. This study investigates whether selecting the most informative participants from the ADNI and Australian Imaging Biomarker and Lifestyle (AIBL) cohorts can enhance ML classification using an automatic and fair data valuation method based on XAI techniques.

Hernandez et al. [155] compare the performances of the best three models from 'The Alzheimer's Disease Prediction of Longitudinal Evolution' (TADPOLE) challenge concerning prediction and interpretability within a common XAI framework. Based on interpretable machine learning, Lai et al. [156] investigate the Endoplasmic Reticulum (ER) stress-related gene function in AD patients and identify six feature-rich genes (RNF5, UBA C2, DNAJC10, RNF103, DDX3X, and NGLY1) that enable accurate prediction of AD progression. An XAI method can now illustrate which feature-rich gene will influence the prediction output for an ML model. The datasets are chosen from an indigenous Gene expression dataset having numeric measures for genes. Chun et al. [126] try to improve the predictive power of progression from amnestic MCI to AD using an interpretable ML algorithm. This study uses datasets of numeric input values of neuropsychological and apolipoprotein test data. Sidulova et al. [157] propose a novel approach for classifying Electroencephalogram (EEG) signals to provide early AD diagnosis. The XAI method used in the study provides quantitative features that help arrive at the prediction using EEG recordings obtained from individuals with probable AD, MCI, and HC.

Many of the research articles have utilised datasets that include ADNI, OASIS, and Kaggle data for training AI-based

Table 6 Inclusion-Exclusion Criteria

Inclusion Criteria	Exclusion Criteria
Studies related to AD diagnosis using AI techniques. Studies related to Explainable AI for AD prediction.	Pilot papers, Editorials, proceedings, magazines. Articles not related to AI-based AD and AD disease diagnosis
Studies related to performance results of ML/DL models for AD	Article on AD but not on detecting it (Eg: supportive care).

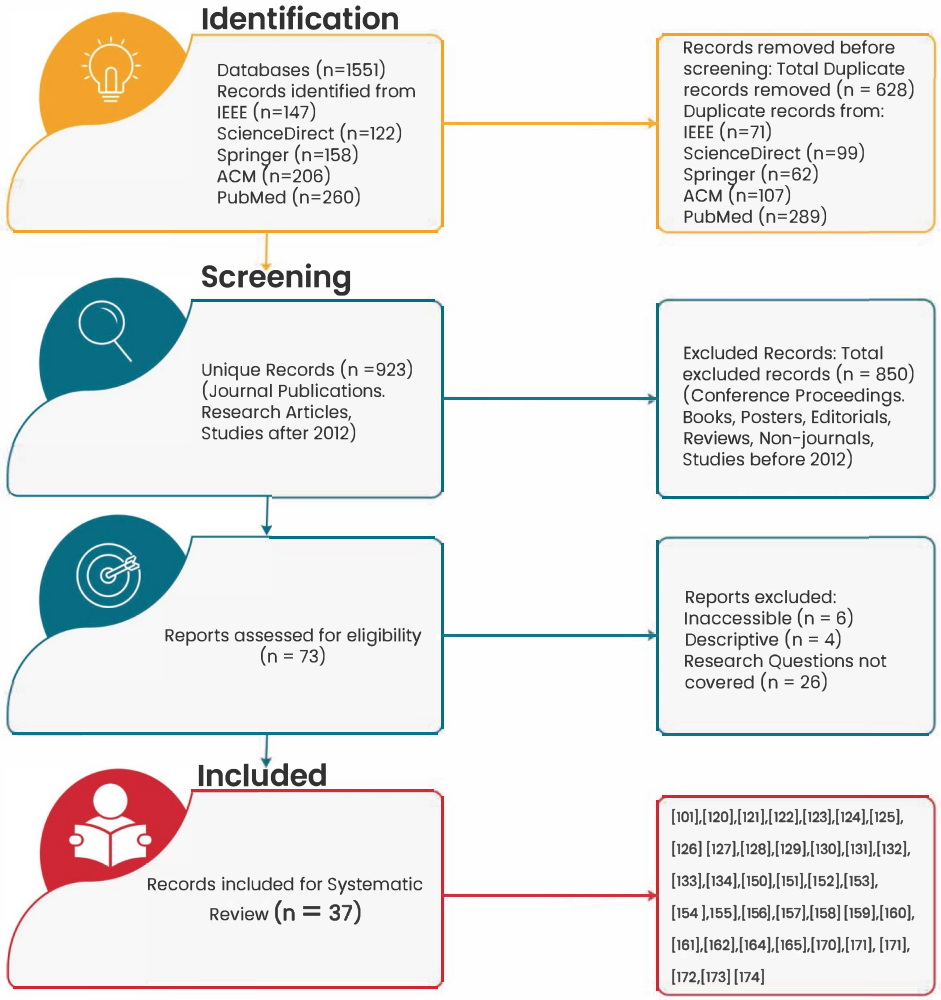

Fig. 9 PRISMA chart showing the identification, screening and inclusion of articles

AD detection models with MRI as input data. From Table 7, the articles [101, 121, 130, 158, 159], and [132] propose classifiers of deep neural networks for prediction and classification between HC, MCI and AD. All these articles use datasets from ADNI and choose MRI as input. According to the prominence and severity of dementia in the available MRI, Jain et al. [160] offer a DCGAN-based Augmentation and Classification (D-BAC) model strategy to identify and categorise dementia into several categories. The MRI scan datasets for the purpose are collected from Kaggle. Shad et al. [128] experimented with neural network models for early AD detection by employing classification approaches utilising a hybrid dataset from Kaggle and OASIS. Bloch and Friedrich [161] propose a machine learning workflow to train and interpret different blackbox models and to compare its performance. All models were trained and evaluated on ADNI, AIBL and OASIS datasets. Deep learning models have been created by Ruengchaijatuporn et al. [124] to classify MCI and AD utilising tasks like the traditional clock drawing, cube-copying, and trail-making test. Multiple drawing task images are used as input and have proved to have significantly improved the classification performance between HC and AD. By combining an interpretable graph neural network with the dataset collected from ADNI, Kim et al. [129] bridge the gap between efficiently integrating longitudinal neuroimaging data and biologically meaningful structure and knowledge to develop precise and understandable systems. García-Gutierrez et al. [162] present a Python-based computational tool to deal with the data obtained during clinical diagnosis. The tool integrates data processing, designing predictive models and features of XAI for explainability. Yang et al. [150] have developed three approaches for generating visual explanations from 3D CNN for AD classification and all the approaches identify important brain parts for AD diagnosis. For all the approaches the models were trained with brain MRI scans from the ADNI database.

Some of the articles have used hybrid datasets as input for the training of AI models. Kamal et al. [127] have used images and gene expression to classify AD and also explained the results. Another article by Ilias et al. [131] has used speech

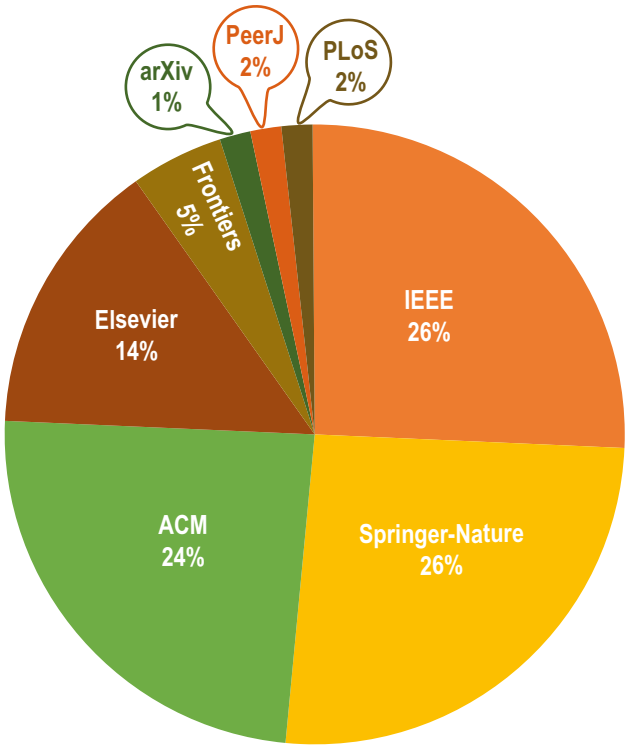

Fig. 10 Sources of Articles Considered in the Systematic Review

recordings and associated transcripts from the ADReSS Challenge dataset to detect AD.

The sunburst chart in Fig. 12 reveals a significant involvement of XAI methods for AD detection with ML techniques. A vast number of DL classifiers, such as CNN, VGG16 etc. are frequently utilised for classifying AD with subsequent explanations. Intuitively, Figs. 11 and 12 also establish that more research is dealt with within the area of ML, which utilises RF, XGBoost, SVM and many other classifiers. It is also to be noted that each research article under ML uses multiple classifiers, whereas articles under DL use very few classifiers. Therefore Fig. 12 shows a comprehensive coverage for ML-based studies.

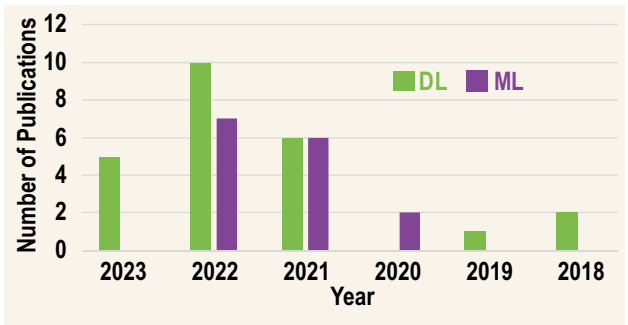

Fig. 11 Year-wise statistics of XAI papers for AD Classsification

XAI Methods for Interpreting Blackbox Models to Detect AD

This section addresses the RQ2: What different XAI methods are used for blackbox interpretability to detect AD?

This research question is devised to find the number and type of XAI methods currently available for the blackbox interpretability in AD detection. It provides essential details such as understanding primary steps taken to be local/global, posthoc/ante-hoc, and model agnostic/model-specific in the XAI context of AD detection studies. While finding answers to this RQ, below additional questions were raised:

1. Why are the scope of some explanations local AND global, local only, global only?
2. Why are some blackbox models, Random Forest (RF) for instance, present in different XAI method categories?
3. Why is CNN considered both in Model Agnostic and Model Specific contexts?
4. Why would an XAI method be considered Posthoc and Model-Agnostic simultaneously?
5. Why would an XAI method be considered Antehoc and Model-Specific simultaneously?

We answer these questions to provide enhanced clarity for addressing research question 02.

(a) Why are some explanations local AND global, local only, global only?

Some of the XAI methods behave either in a local interpretable format or global. However, it is the prerogative of the researcher to use those methods to interpret globally by aggregating the local explanations. Therefore, there is no fixed distinction between saying that model can only be local or global. For instance, the XAI framework SHAP is predominantly used for local interpretation. However, SHAP can also be used to interpret a global population. Similarly, the LIME which is a local explainer method can also be used for global understanding by aggregating local explanations.

(b) Why are some blackbox models, RF for instance, present in different XAI method categories?

The blackbox models LGBM and XGBoost are tree-based models where the classification in LGBM is done branch-wise, while in the XGBoost, it is done level-wise. Therefore, the scope of explainability for LGBM can be local as a branch-wise classifier. Subsequently, local results can be aggregated to establish global explanations. Since XGBoost is classified level-wise, it can achieve a local description at each tree level, and the final path to the last level can be

Table 7 Summary of AI for AD research that incorporate XAI

Ref.	Contribution	Dataset	Input Type	Output
[120]	Creating Explanation Driven HCI Model	Dementia dataset	Numeric	Numeric
[121]	Layer-Wise Relevance Propagation to explain deep neural network using MRI	ADNI	Image	Visual
[122]	Explaining the predictions of a multi-layer model	ADNI	Numeric	Numeric, Rule-Based
[123]	Examining an automatic and fair data valuation	ADNI	Numeric	Numeric
[151]	Systematically classifying and explaining MCI and AD	ADNI	Numeric	Numeric
[124]	Classify healthy control and MCI using clock drawing and other drawing tasks as input	Clock Drawing, Cube Drawing and Trail Making	Image	Image
[155]	Prediction of clinical AD status using TADPOLE challenge	TADPOLE	Numeric	Numeric Rule-Based,
[156]	Prediction of AD using stress associated gene based on interpretable machine learning	Gene Expression Data	Numeric	Numeric, Rule-Based
[125]	Prediction of AD with XGBoost model and making it explainable	Personal Information	Numeric, Categoric	Numeric, Rule-Based
[126]	Prediction of MCI to AD using interpretable ML	Apolipoprotein-E (APOE) Genotype Data	Numeric	Numeric, Rule-Based
[163]	HealthXAI system for early diagnosis of cognitive decline using IoT	CASAS dataset	Numeric, Categoric	Numeric, Rule-Based
[158]	Diagnosis using form of CNN explaining and HAM (High Resolution activation Mapping) and PCR (Prediction basis Creation and Retrieval Module)	MRI Scans (ADNI)	Image	Visual Textual

Table 7 (continued)

Ref.	Contribution	Dataset	Input Type	Output
[160]	CNN based technique to identify and classify dementia into various categories	MRI scans from Kaggle	Image	Visual
[127]	Classifying AD and explaining using image and Gene expression data.	MRI Gene Expression	Image Numeric	Numeric, Visual
[128]	XAI for AD prediction in Deep Learning Models.	MRI Scans	Image	Numeric, Visual
[129]	Propose interpretable graphical neural network model for AD prediction based on longitudinal neuroimaging data	ADNI	Image	Rule-Based
[157]	XAI-Image Analysis for MCI diagnosis	EEG Recordings of AD, MCI probables	Numeric	Numeric, Textual, Visual
[152]	Propose a solution RN-SSAS for reliable multi-class classification	MRI, Clinical Demographic	Numeric	Numeric, Visual
[164]	LRP to identify the stages of AD	MRI Data	Image	Visual
[130]	GradCAM based AD Diagnosis Using Structural MRI	MRI Scans from ADNI	Image	Visual
[165]	Occlusion Sensitivity method to reveal the White Matter Hyperintensivity regions of AD	MRI Scans from OASIS	Image	Numeric, Visual
[131]	Using transformer network,BERT, to use transcripts to detect AD with LIME	Speech recordings and associated transcripts ADRess dataset	Categoric	Textual
[159]	Exploring ML and MRI-based features with XAI (obtained numeric measures)	T1 weighted images (ADNI-3)	Image	Numeric

Table 7 (continued)

Ref.	Contribution	Dataset	Input Type	Output
[153]	Introducing a interpretable and computational framework SMILE, with evolutionary algorithm	Metabolomic dataset	Numeric	Numeric
[101]	Using 18F-FDG PET images and CNN to explain predictions for early diagnosis of AD	18F-FDG PET	Image	Numeric
[162]	Python-based framework for early and automated diagnosis using neuroimages and neurocognitive assessments and validation using Evolutionary Algorithm	Cognitive and and PET images	Image	Visual
[161]	Evaluation and Classification from Multiple datasets for different blackbox models to explain prediction using Shapley values	MRI data (ADNI/AIBL/ OASIS)	Image	Numeric
[132]	Interpreting DNN for AD diagnosis using LRP with new propagation rules	MRI scans from ADNI and TADPOLE	Image	Visual
[133]	Predicting and explaining the risk of Dementia with Ensemble Learning algorithms and explaining with Rule-based and SHAP	Dataset from Survey of Health Ageing and Retirement (PREVENT Program)	Numeric, Categoric	Numeric, Rule-based
[150]	Give visual explanations for a deep 3D CNN for AD using GradCAM	ADNI	Image	Visual
[134]	Understanding visualising XAI methods for MRI based diagnosis of AD from CNN	MRI, PET, Biological markers, Clinical and	Image, Numeric, Categorical	Visual, Numeric

Table 7 (continued)

Ref.	Contribution	Dataset	Input Type	Output
		neuropsychological assessments (ADNI)		
[154]	Using amyloid, tau and neurodegenerative biomarkers with RF to analyse and predict AD using DT and SHAP	Four biomarkers -phosphorylated tau -amyloid beta -glucose uptake -volumetric measures (ADNI)	Numeric	Numeric

considered a global explanation. Hence, LGBM and XGBoost, though blackbox models, can be aggregated locally and globally. The case is similar to that of RF as it is a collection of DTs where the above description that we have provided fits into each tree either locally or globally in a similar manner. Therefore, the same blackbox model is aggregated in different XAI method categories based on the nature of explanations provided in the respective study.

(c) Why is CNN considered both in Model-agnostic and Model-specific contexts?

Contrary to the widespread understanding of CNNs being considered for model-agnostic interpretation, some of the CNNs could be interpreted in a model-specific manner. For instance, a model-agnostic approach can explain the prediction of a CNN model without affecting the internal layers (e.g., kernel SHAP). On the other hand, a model-specific system can give perturbations to each layer of a CNN and back-propagate to the input to achieve feature-rich values for better explainability (e.g., LRP) [105].

(d) Why would an XAI method be considered Post-hoc and Model-Agnostic simultaneously?

Post-hoc models are primarily applied to such blackbox models where the inner workings of these models remain untouched. The prediction thus obtained must undergo an XAI method for producing explainability-this concept can be termed both post-hoc and model-agnostic. Hence, some blackbox model is both Post-hoc and Model-Agnostic.

(e) Why would an XAI method be considered Ante-hoc and Model-specific simultaneously?

The Ante-hoc model is where the essential details of a training model are inherently available. To derive explainability out of an ante-hoc model, a model-specific or a model-agnostic XAI method can be employed. However, a model-specific XAI method necessitates the inner working details to integrate explainability during the training of an ante-hoc model.

With this backdrop, we now answer the main RQ. The XAI-based AD papers in our study are broadly classified under three main categories:

1. Local, Global, Post-hoc, Model-agnostic
2. Local, Post-hoc, Model-agnostic
3. Global, Post-hoc, Model-agnostic

Local, Global, Post-hoc, Model-agnostic

El-Sappagh et al. [122] have proposed a multimodal prediction and detection of AD in two stages. In the first stage, the model

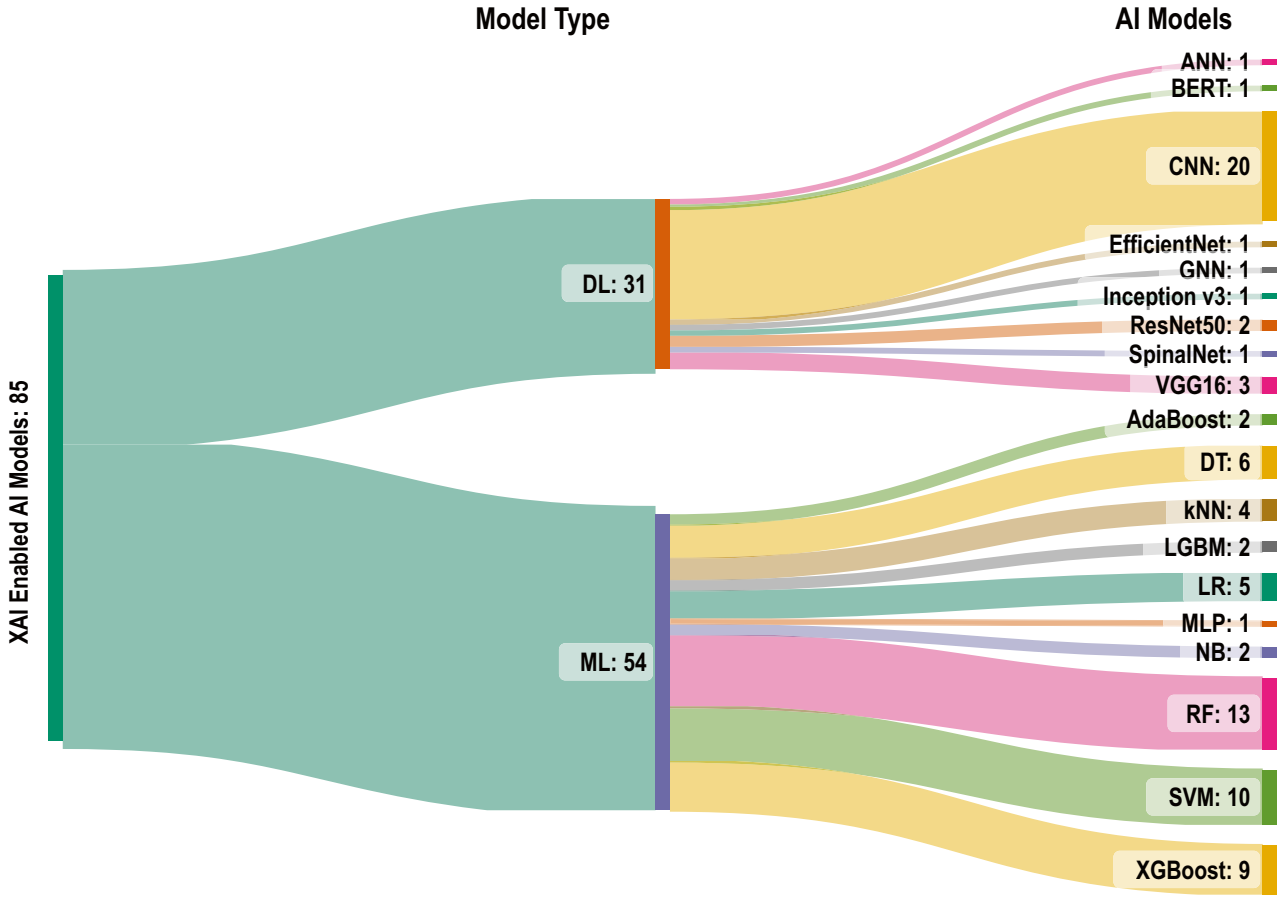

Fig. 12 Sankey diagram of various AI models that incorporate XAI for AD detection

performs a multi-class classification for the early diagnosis of AD. A binary classification model will follow in the second stage to detect possible MCI to AD progression. The authors have utilised 11 different modalities: PET, MRI, cognitive scores, genetic data, demographic, and other clinical data. The classification was done using the RF algorithm resulting in an overall F1-score of 93.94% and 87.09% in two classification stages. The authors have used explainers based on the DT and fuzzy rule, providing complementary justifications for every prediction made in each stage.

Authors in [125] provided local and global interpretation in the conversion of Dementia to MCI using the XGBoost algorithm. They have utilised multimodal data (personal information, gene expression, PET and MRI measures, cognitive score) in the four-way classification of stages of AD progression. The model achieved an accuracy of 84.0%. The interpretation methods provided insights about data modality influential in each stage of AD progression.

Yang et al. [150] provided visual explanations for the deep 3D CNN in AD classification. The authors have utilised the brain MRI scans from the ADNI dataset for AD vs. HC classification using ResNet and VGG16 architectures. The authors also proposed a variant of ResNet architecture called ResNetGAP, where the Global Average Pooling(GAP) layer was introduced in the original ResNet architecture instead of the conventional Max-pool layer. The approach yielded an overall accuracy of 76.6% and an AUC of 0.863. Regarding interpretability, the authors could produce visual explanations for AD prediction using the network weight map from three different network architectures. The visual interpretation highlighted the cerebral cortex, lateral ventricle, and hippocampus regions in the 2D slices of the brain MRI.

In another study [133], authors used RF and XGBoost algorithms in the HC vs. AD classification using socio-demographic data, medical history, and lifestyle parameters (daily activity and smoking). The study developed an ensemble-based ML model to predict AD and explained the prediction in local and global contexts. The study also includes feature importance analysis and ranked the dominant features influential in AD. The top 7 features considered by both classifiers (RF and XGBoost) in AD prediction were the same. The feature importance analysis also found the least suspected risk factors driving the risk of AD.

Local, Post-hoc, Model-agnostic

Ilias and Askounis [131] have proposed transformer-based models for the identification of Dementia using voice transcripts as data modality. This is the only work found in the review that uses voice with associated text as data modality in Dementia identification and subsequent interpretation. The study involves the identification of Dementia in the first stage (HC vs. AD), followed by identifying the severity of Dementia in the second stage. The authors have employed Bidirectional Encoder Representations from Transformers (BERT). To distinguish between the languages used by AD patients and non-AD patients, word symbols were colour-coded for interpretation.

Another study [157] identifies MCI and AD patients (3-way classification HC vs. MCI vs. AD) using 90 seconds recording of resting state EEG. The study compares the performance of classification using three classifiers: SVM, ANN, and CNN. The explainable component in this study aimed to highlight the brain region most indicative of the onset or progression of MCI/AD.

Lombardi et al. [151] used multimodal data for AD vs. MCI vs. HC classification using the RF classifier. The clinical and neurophysiological indices were used to train the RF classifier in the AD classification. The authors explored various neurophysiological data's capabilities in predicting different degrees of cognitive impairment. The dominant features used by the classifier for prediction have been enriched with explainable values.

Authors in [153] have used a metabolomic dataset to identify the key metabolites and their interaction associated with AD. The authors have used SVM and RF as classifiers. The model interpretation was provided by ranking significant metabolite features in the prediction based on the Gene importance. The authors claimed that the study provided explanations that could give additional background for the metabolomic backdrop of AD.

Rieke et al. [134] have used 3D CNN in the binary classification task (AD vs. HC) using structural MRI scans of the brain. The authors emphasised the importance of applying different visualisation methods for identifying various brain regions. For instance, a particular visualisation method could highlight the temporal lobe, whereas other techniques could focus on cortical areas. Such details obtained from different visualisation methods help find the distribution of relevant patterns which could vary across patients.

Global, Post-hoc, Model-agnostic

Bloch and Friedrich [161] used Shapley values to interpret XGBoost, SVM, and RF blackBox models using the ADNI dataset. The study considered MRI volume features, cognitive test results, sociodemographic data, and Apolipoprotein alleles. The Shapley values were employed to visualise the feature association in the blackbox classification. The models were trained individually using separate data modalities. The examination found a biological correlation and enhanced results when these models included cognitive test results.

Ruengchaijatuporn et al. [124] proposed a two-way classification for differentiating MCI vs. HC patients using the VGG16 and custom CNN architecture incorporating a self-attention mechanism (Conv-Att). The authors considered digital drawings (clock, cube, and trail making) collected from HC and MCI patients to train the models. The VGG16 model interpretability was provided using the GradCAM, and custom CNN has a built-in self-attention mechanism to offer visual cues. Clinical experts validated the visualisation produced by the GradCAM and the Conv-Att model using a simple rating mechanism. The authors concluded that the heatmap produced by the Conv-Att model was better aligned with the expert's clinical experience. However, a serious limitation of this study is the non-consideration of a biomarker.

Another study [101] utilised the CNN architecture using 18F-FDG PET images to classify AD vs. MCI vs. HC. The model achieved AUC scores of 0.81, 0.63, and 0.77 for HC, MCI, and AD cases. For explanations, heatmaps were generated and registered with the Talairach atlas (3-dimensional coordinate system of the human brain), indicating each voxel's importance for the final classification decision.

Table 8 provides a complete summary of XAI methods used in the blackbox interpretability for AD detection. The chart in Fig. 13 provides a clear perspective of research using different XAI methods. In particular, we find that studies conforming to the concepts of Local (Ll), Post-hoc (Ph), and Model-agnostic (Ma) make up the totality of the volume. In this context, most of the studies have been concentrating on classifiers under DL, of which a majority study deals with CNN. Furthermore, it is clear that a subsidiary part of the research focuses on Global, Post-hoc, and Model-agnostic, where classification techniques are widely used within the framework of ML approaches. Cumulatively, it can be understood that the Model agnostic approach covers 31 out of 37 studies considered in our review.

XAI Frameworks for AD Detection

This section addresses RQ3: What XAI frameworks are available in the literature which are used in AD detection?

This RQ aims to identify the XAI frameworks and techniques used in the studies to interpret AI-based AD classification. The discussions in this area will encourage researchers, developers, and subject matter experts to comprehend the inner workings of a machine-learning model. Explainable embedded machines, especially in healthcare, can significantly reduce the time medical professionals spend on recurrent patient studies and spend time concentrating on interpreting disease diagnoses. Many XAI frameworks exist to help tackle the problem of blackbox models where

Table 8 Summary of XAI methods used for blackbox interpretability to detect AD

XAI Method	Ref	Classifier	Blackbox	Classification Task	Input Data Modality
Ll, Gl, Ph, Ma	[122]	ML	RF	HC vs MCI vs AD	Numeric, Image
	[156]	ML	AdaBoost	AD vs HC	Gene Expression Data
	[125]	ML	XGBoost	HC vs EMCI vs LMCI vs AD	Sociodemographic, Gene Expression, PET, MRI Features, Neurophysiological
	[130]	DL	3D VGG16	AD vs HC	sMRI
	[133]	ML	RF	HC vs AD	Sociodemographic, Clinical, Life Style
	[150]	DL	3D CNN	HC vs AD	Normalised, Masked, N3-Corrected T1 MRI
Ll, Ph, Ma					
	[121]	DL	CNN	AD vs HC	sMRI
	[151]	ML	RF	HC vs MCI vs AD	Clinical, Neurophysiological Data
	[158]	DL	CNN	AD vs HC	T1-Weighted, Preprocessed, Baseline MRI
	[127]	DL	SpinalNet	ND vs VMD vs MiD vs MoD	T1-Weighted MRI, Gene Expression Data
	[128]	DL	VGG16	ND vs VMD vs MiD vs MoD	T1-Weighted MRI
	[157]	DL	ANN	HC vs MCI vs AD	Resting State EEG
	[164]	DL	VGG16	ND vs VMD vs MiD vs MoD	MRI Scan
	[165]	DL	EfficientNet	ND vs VMD vs MiD vs MoD	MRI Scan
	[131]	DL	BERT	AD vs HC	Voice Transcripts
	[153]	ML	EA	HC vs MCI vs AD	Metabolomic Dataset
	[162]	ML	BC	HC vs AD	PET, Neurophysiological Data
	[132]	DL	CNN	HC vs AD	T1-Weighted MRI
	[134]	DL	3DCNN	HC vs MCI vs AD	MRI, PET, Biological Marker, Clinical and Neurophysiological Assessment
Gl, Ph, Ma	[123]	ML	RF	sMCI vs pMCI	Sociodemographic, Neurophysiological Data
	[124]	DL	VGG16	HC vs MCI	Digital Drawings of Clock, Cube, Trail Making
	[155]	ML	XGBoost	HC vs MCI vs AD	Neurophysiological Data
	[160]	DL	CNN	ND vs V vs MiD vs MoD	MRI Scan
	[129]	DL	TGNN	HC vs MCI vs AD	T1-Weighted MRI
	[152]	ML	SVM-SMOTE	HC vs MCI vs AD	Clinical Data
	[101]	DL	3DCNN	HC vs MCI vs AD	18F-FDG PET Scans
	[161]	ML	XGBoost	HC vs MCI vs sMCI vs pMCI vs AD	MRI Features, Sociodemographic, Clinical, Neurophysiological Data
	[154]	ML	RF	HC vs LMCI vs AD	Sociodemographic, Neurophysiological Data

Ll Local, *Gl* Global, *Ph* Post-hoc. *Ah* Ante-hoc, *Ma* Model-Agnostic, *Ms* Model-Specific, *RF* Random Forest, *LR* Logistic Regression, *EA* Evolutionary Algorithm, *BC* Bayesian Classifier, *DT* Decision Trees, *TGNN* Temporal Graph Neural Network, *D* Demented, *ND* Non Dememted, *VDM* Very Mild Demented, *MiD* Mild Demented, *MoD* Moderate Demented

predictions are highly accurate, and the inner workings are hidden. We have identified popular XAI frameworks like LIME, SHAP, and GradCAM, among many others, that are extensively used in AD and of interest to RQ3. The methods are classified as follows: Tables 9, 10, 11, and 12 are presented as a list of the studies that have attempted to use the methods LIME, SHAP, LRP, and GradCAM, respectively. Table 13 lists studies that have used a combination of explainable methods, for instance, LIME and SHAP, where one algorithm gives a local explanation and global for the other.

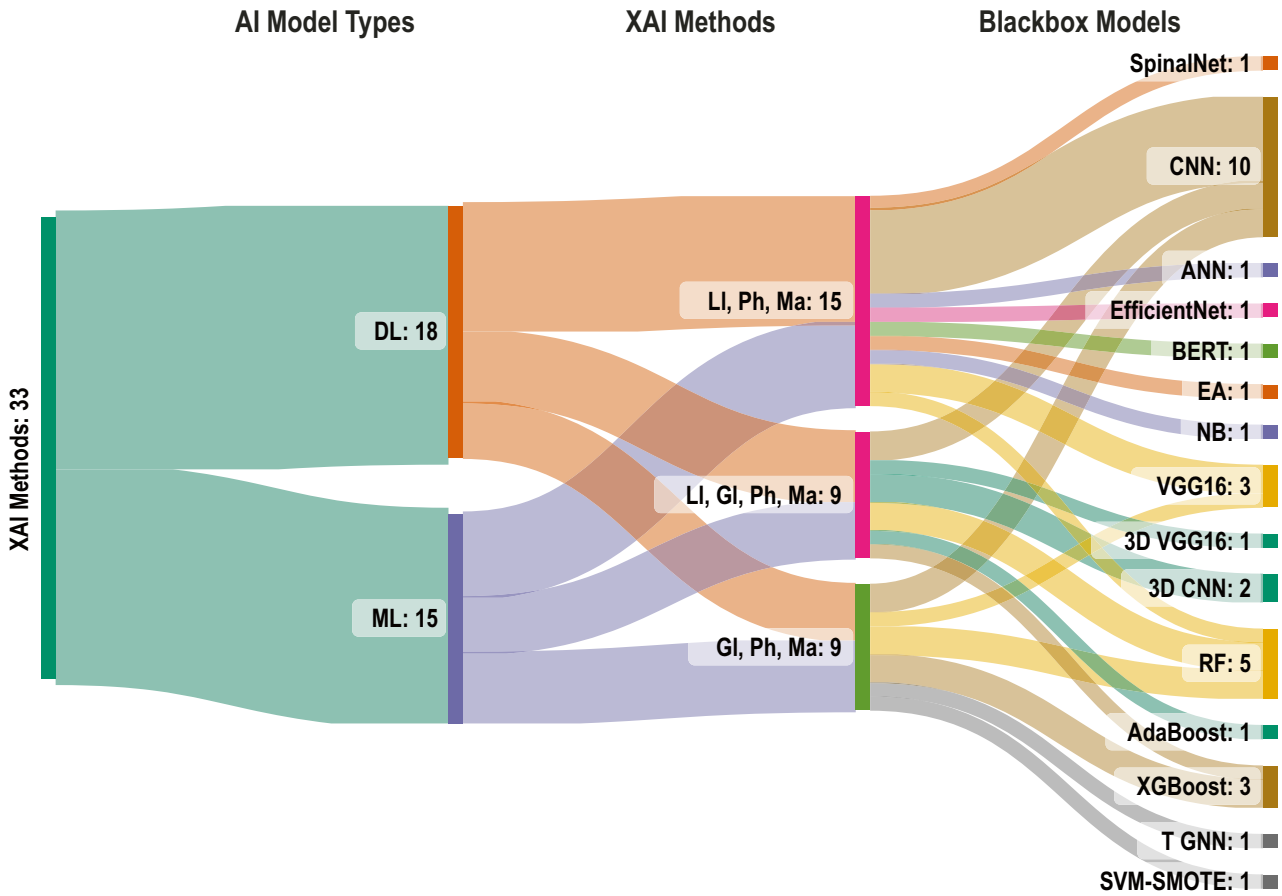

Fig. 13 Sankey diagram of XAI methods for different classifiers used in AD Detection. **Legends** Ll–Local; Gl–Global; Ph–Post-hoc; Ah–Ante-hoc; Ma – Model-Agnostic; Ms – Model-Specefic

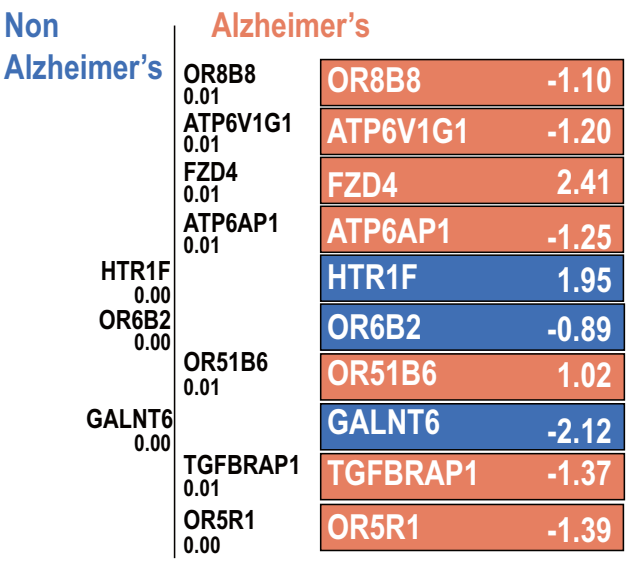

Fig. 14 LIME Explanation. Reproduced with permission from [127]

The LIME is a popular method for simple human interpretations of predictive models. The studies in Table 9 use LIME to interpret the predictions from a wide range of ML/DL classifiers, including CNN, SpinalNet, kNN, XGBoost, SVM, and transfer-based model BERT. In the studies, the classifiers have used datasets of the type, including MRI, gene expressions, EEG signals, and linguistic or textual data. Kamal et al. [127] propose a study of four-way classification between mild dementia, moderate dementia, no dementia, and very mild dementia using MRI scans and gene expression. The author uses LIME to obtain local explanations of AD classifications from MRI with CNN and gene expressions with kNN and XGBoost. LIME proved instrumental in identifying and ranking the significant sets of features based on probability values responsible for an AD patient. Figure 14 illustrates how LIME selects the most critical genes from the gene expression data and interprets the predicted genes that are critically responsible for an AD patient. In Fig. 14, a ranking of the genes is shown based on probability values of prediction and separated into AD and non-AD categories. LIME allows users to understand which features contribute positively and negatively to the prediction. Though

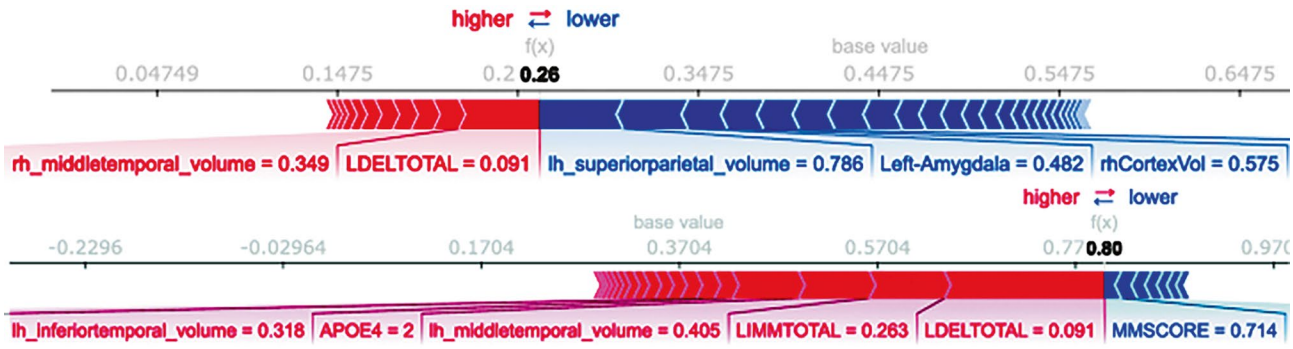

Fig. 15 SHAP Explanation – Force Plot. Reproduced with permission from [122]

trust is not inherently quantified the trust in the explanation can be explained based on the probability values. For instance, LIME interprets OR8B8 and ATPV1G1 as the most significant genes for AD and HTR1F and OR6B2 of a lower significance.

Illias and Askounis [131] undertake a thorough linguistic analysis from a medical transcript dataset using the transfer learning model, BERT, with the co-attention mechanism to classify between control and dementia patients. Subsequently, personal pronouns, interjections, adverbs, and past tense verbs are all used by AD patients, according to LIME. Healthy controls, on the other hand, employ present-tense verbs, nouns, and determiners. The studies in Table 9 show how LIME creates local explanations for any machine learning classifier by constructing a trainable interpretable model on data that recognises differences in classification performances.

SHAP is model-agnostic and utilises an approach of game theory for explaining the output of any machine learning model. In this review, it was found that SHAP is another XAI framework that is being used frequently. The papers in Table 10 use SHAP to explain classifications from machine learning models, including RF, XGBoost, SVM, and Logistic Regression(LR). Most studies use datasets of type, including demographic data, Apolipoprotein measures, Mini-Mental state examinations, Clinical Dementia ratings, and other volumetric measurements of MRI and PET scans. The fundamental principle of SHAP is to determine the Shapley values for each sample feature that needs to be understood. Each Shapley value reflects the influence of the corresponding

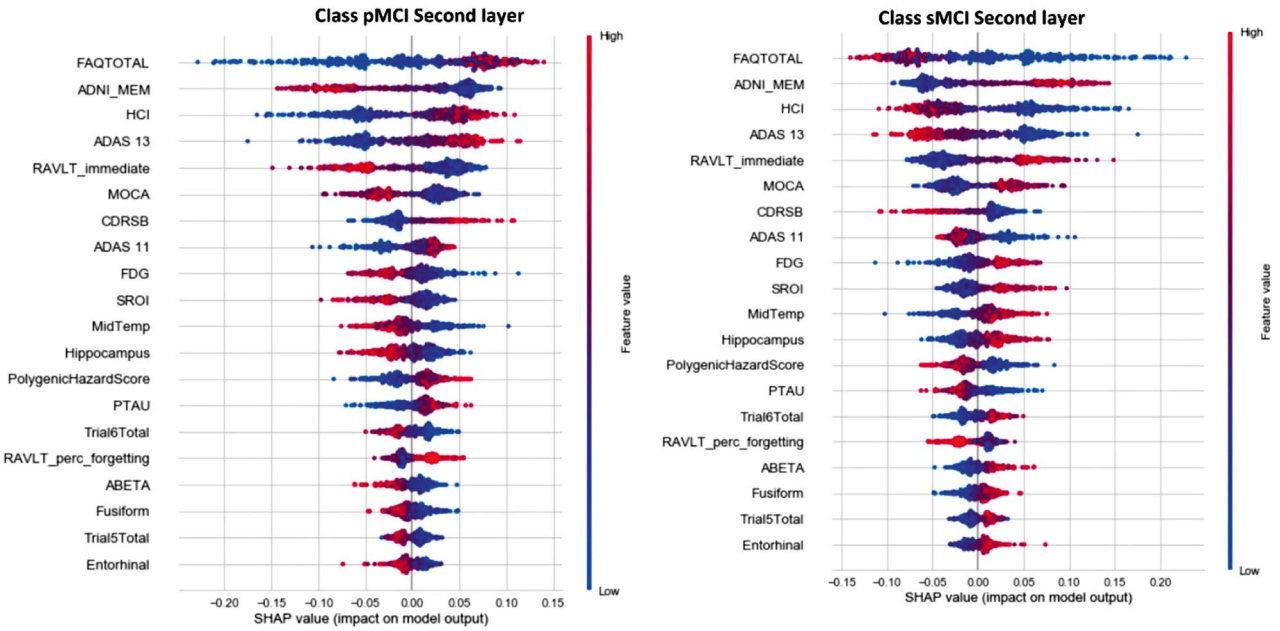

Fig. 16 SHAP Explanation – Summary Plot. Reproduced with permission from [122]

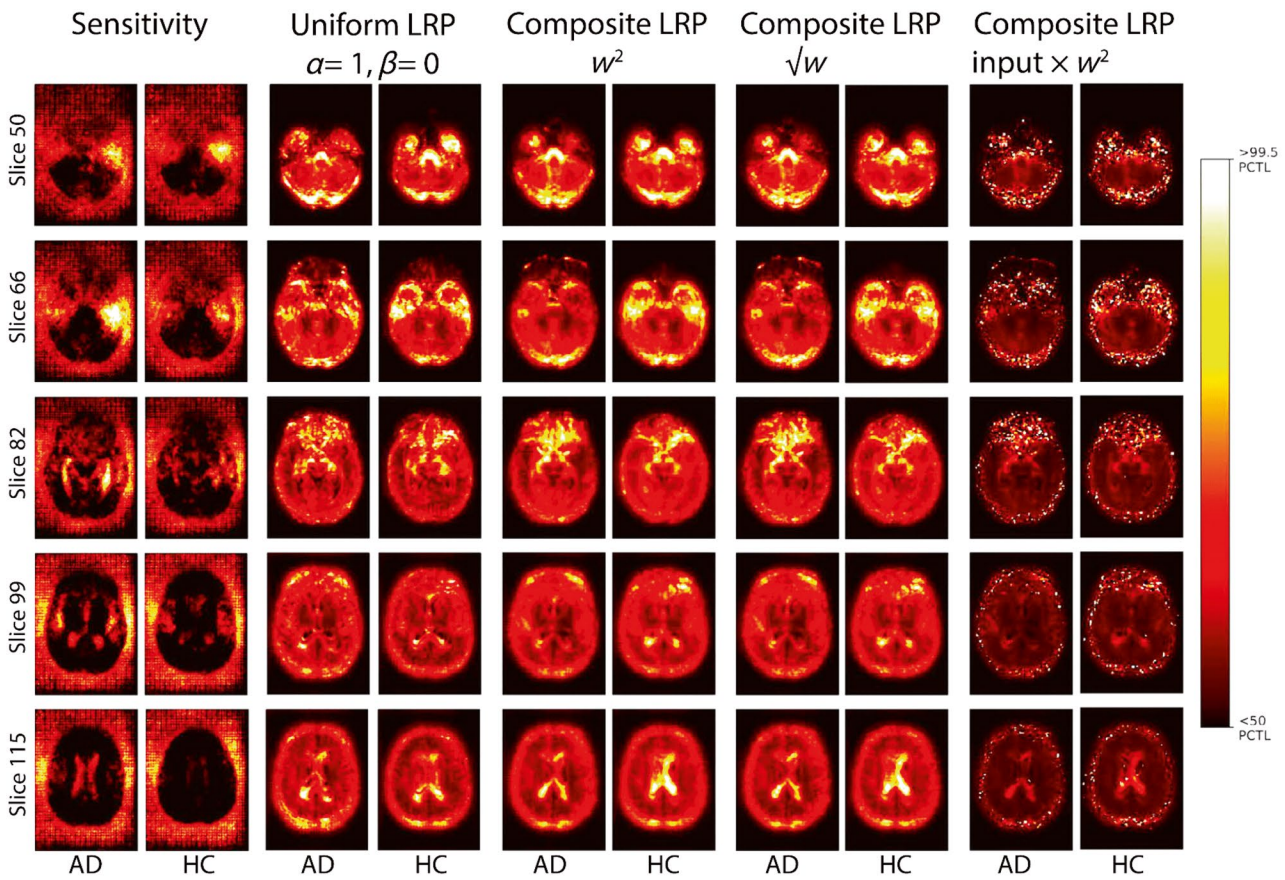

Fig. 17 LRP Explanation. Reproduced with permission from [132]

feature on the prediction. It is possible to obtain sample features with high and low Shapley values. Both sets of features are visualised with Shapley plots to understand the impact of the features on a specific sample for a given prediction. Bloch and Friedrich [123] propose a study that compares the classification using RF and XGBoost for volumetric measurements of MRI scans of control and dementia patients. Shapley values are obtained for features of both the classifiers and ranked accordingly. The effect of the attributes on AD prediction is then displayed using Shapley plots, namely force plots and summary plots. Figure 15 is an example of a force plot that shows features that had the most influence on the model's prediction for a single observation. Figure 16 is an example of a SHAP summary plot used to show the contribution of all features for every instance. Similar approaches are handled in their studies by Bogdanovic et al. [125] and Danso et al. [133]. SHAP force and summary plots are not explicitly quantified to show trustworthiness. However, the visual representations show quantifiable insights based on the magnitude of feature importance.

Table 11 lists XAI studies that use the LRP model-specific interpretation tool. Complex deep neural networks with video, or picture inputs can now be explained using LRP [138]. The prediction is transmitted back through the neural network using local propagation rules. The decisions made by CNN using AD-based MRI data are visualised using LRP by Böhle et al. [121]. LRP creates a heatmap that explains the significance of each voxel that contributes to a specific classification. The study also computes a sum of all layerwise relevance metrics of the MRI that helps to identify critical areas of the image. Based on trained CNN, the author's individual categorisation choices for AD and HC are explained using LRP.

Pohl et al. [132] propose LRP with multiple rules, also known as composite LRP. On the contrary, LRP with a single rule, also known as uniform LRP, uses a single rule for interpretation. LRP of both uniform and composite forms are used in this study to compare the evaluation measures quantitatively. Figure 17 shows a comparison of interpretations for AD classification, termed positive evidence, between uniform LRP and composite LRP. The study proves that the composite LRP rule, compared with the uniform rule, gives a more focused visualisation of only the relevant regions of the brain for positive AD by filtering out the least relevant ones. The advantage of composite LRP is visualised from the last column in Fig. 17 where a predominant relevance

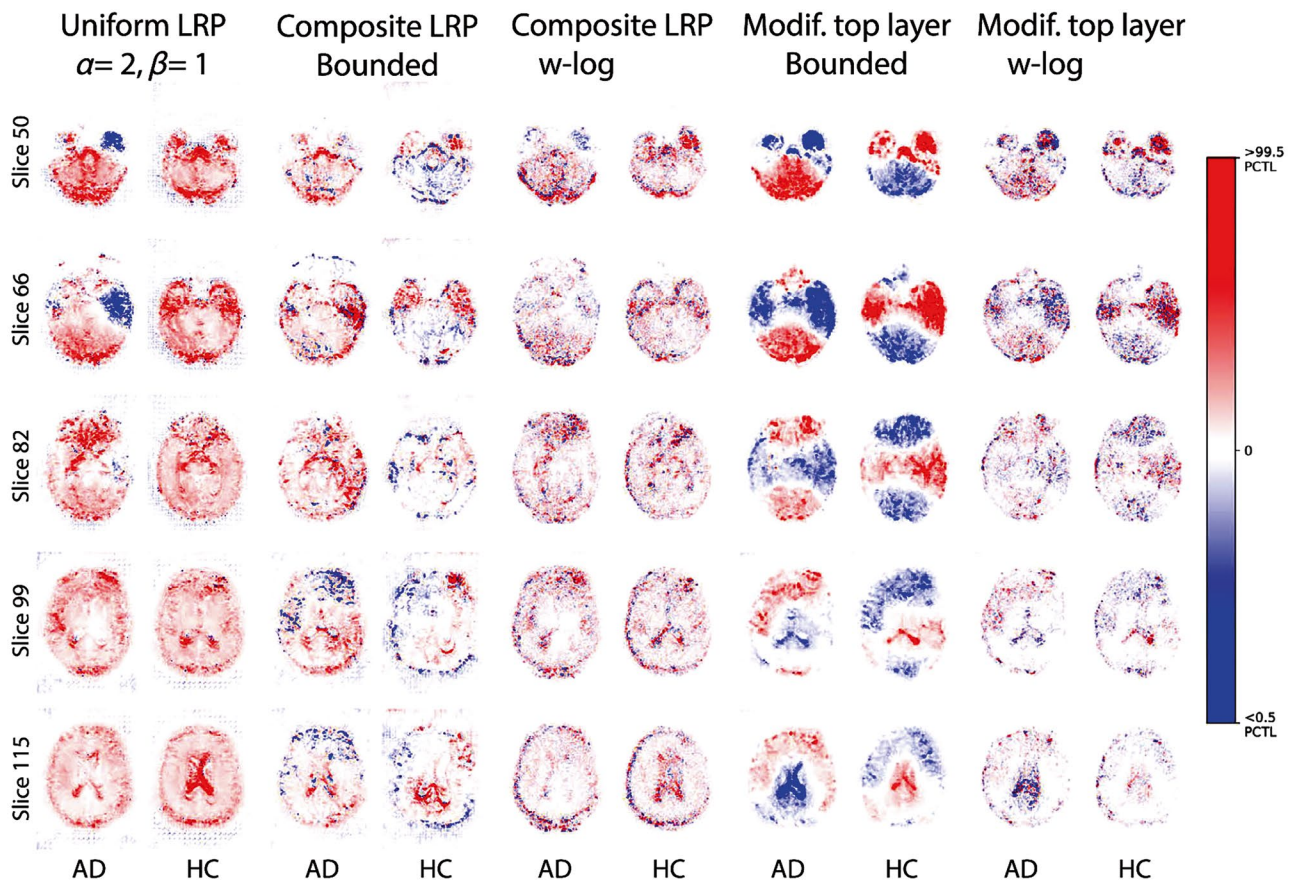

Fig. 18 LRP Explanation. Reproduced with permission from [132]

can be observed from the heatmap. Additionally, Fig. 18 shows that in the visualisations of non-AD outcomes (negative evidence), composite LRP proves beneficial. The figure shows negative visualisation for both classes - HC and AD. In Fig. 18 the last column visualises the positive contribution to the HC class (shown in red) and the negative contribution to the AD class (shown in blue). As a result, the LRP studies in Table 11 have a good chance of helping doctors by outlining the neural network decisions used to diagnose AD and other disorders using structural MRI data.

The GradCAM is a model-agnostic XAI tool used by studies in Table 12. GradCAM is typically used to produce visual explanations of the key input regions for predictions, increasing the transparency of CNN-based models. Using a gradient of the localised classification score for the features selected by the network, this technique can identify the areas of the image that are most crucial for prediction [166, 167]. Combining the localised scores creates a high-resolution and class-discriminative visualisation. Ruengchaijatuporn et al. [124] use images of bedside tasks like clock drawing tests, cube-copying and tail-making tests to classify between HC and AD patients in a deep neural network. For improving interpretation, convolutional self-attention and output of class probability as a soft label are applied with the GradCAM tool to visualise the model for essential input regions. The author also compares the CNN outputs with VGG16 with the explanation of the visuals using GradCAM. Figure 19 is an example of the visual explanation obtained from the multi-input VGG16 model with GradCAM and the author-proposed (Conv-self-attention, soft label) model for an AD test sample. The last column in Fig. 19 depicts the crucial regions of interest to be classified as AD compared to the HC column when used with GradCAM. Jain et al. [160] construct a heatmap emphasising the characteristics discovered from the input MRI scan for each layer of the CNN model using GradCAM. Then they combine it for a final interpretation. Though GradCAM provides qualitative visualisations and does not offer quantitative metrics for trust, it indirectly supports the quantitative analysis by assessing a model's attention and localisation. GradCAM can complement trust assessment by offering visual insights into the model's attention. Additionally, Zhang et al. [130] employ GradCAM to produce heatmaps or visual explanations

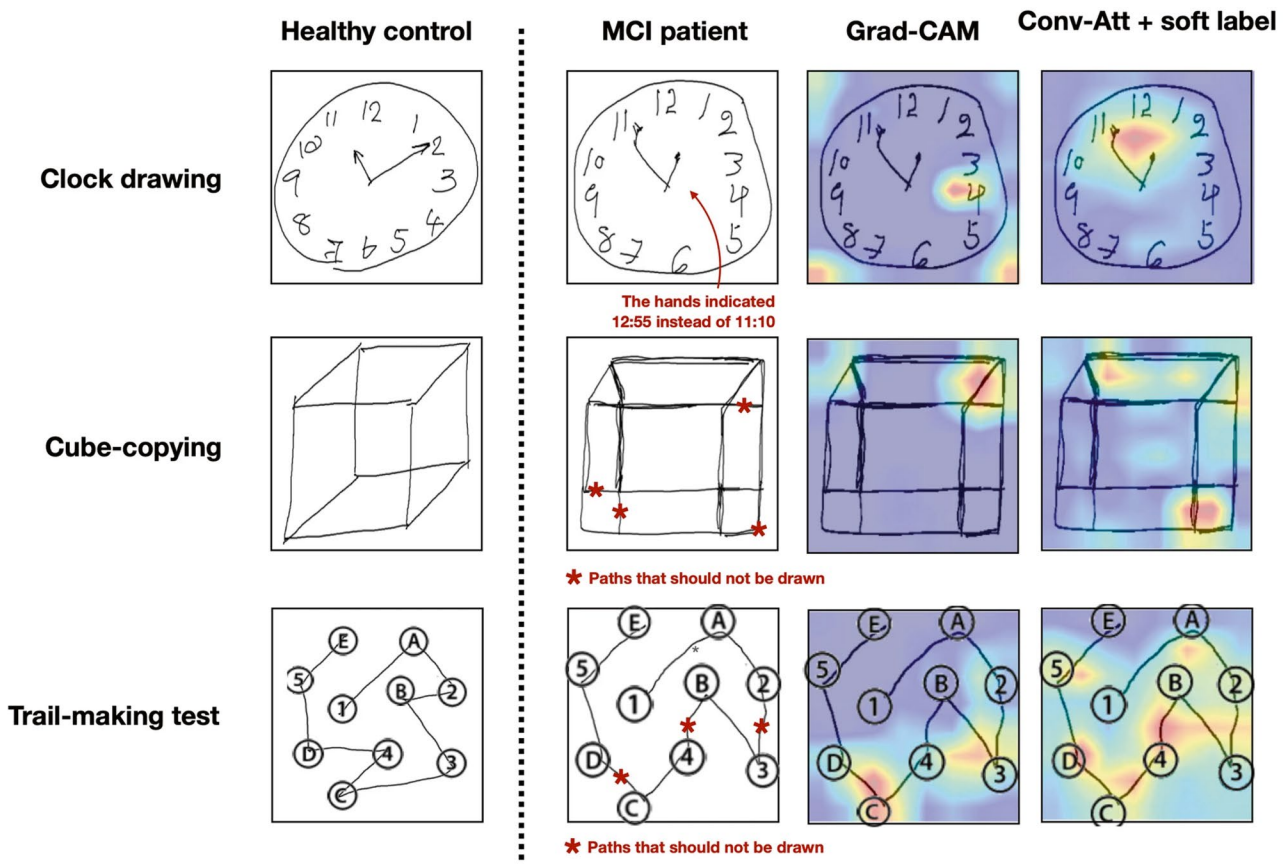

Fig. 19 GradCam Explanation. Reproduced with permission from [124]

from a 3D ResAttNet (Residual Attention network) that will emphasise the features discovered from the input MRI scans for each layer. GradCAM can be used with multimodal inputs without architectural changes or re-training and provides visual explanations by measuring their ability to discriminate between classes. The tool inspires trust in humans, particularly in the healthcare domain.

The review also identified several other XAI frameworks such as GNNExplainer (GNNE), ICE, OCA, and SM (see Table 13). Without needing to change the underlying GNN architecture, GNNE is a model-agnostic method that is used to deliver trustworthy justifications for predictions made by any Graph Neural Network (GNN) based ML task [129]. The explanation pinpoints a subgraph structure and a selection of node attributes for a specific instance that are essential for accurately forecasting the GNN in a local scope [168]. The GNNE can also produce global explanations for a whole group of instances. By successfully combining longitudinal neuroimaging and biologically significant data, Kim et al. [129] offer an interpretable GNN model for AD prediction. GNNE is used to find significant nodes that contribute to the prediction. This tool creates a subgraph structure and a subset of node attributes crucial to the prediction. The ability to display syntactically relevant structures and interpretations and the capacity to get insight into faulty GNNs are two features that make GNNE useful.

Chun et al. [126] in their paper provides a local explanation for the prediction of conversion from amnestic MCI (aMCI) to dementia or AD using ICE and SHAP for each patient. The XGBoost has shown the best performance for prediction in the paper. ICE show plots for each individual instance with a variation of values for a feature of interest and keeping values of other features constant. Figure 20 shows ICE plots of eight important features - Age, Controlled Oral Word Association (COWAT), Education, Minimental State Examination (MMSE), Rey-Osterrieth Complex Figure Test (RCFT) with delayed recall, RCFT with copy time, Clinical Dementia Rating- Sum of Boxes (CDR-SOB) and Seoul Verbal Learning Test (SVLT) for six patients numbered 1 to 6 in different colours. For instance, for the feature Age, line plots for each patient are drawn by varying the Age feature values and keeping values of other features constant [126].

Bordin et al. [165] use the Occlusion Sensitivity method to reveal the relevant measure of white matter hyperintensities lesions with healthy lesions. Understanding which elements

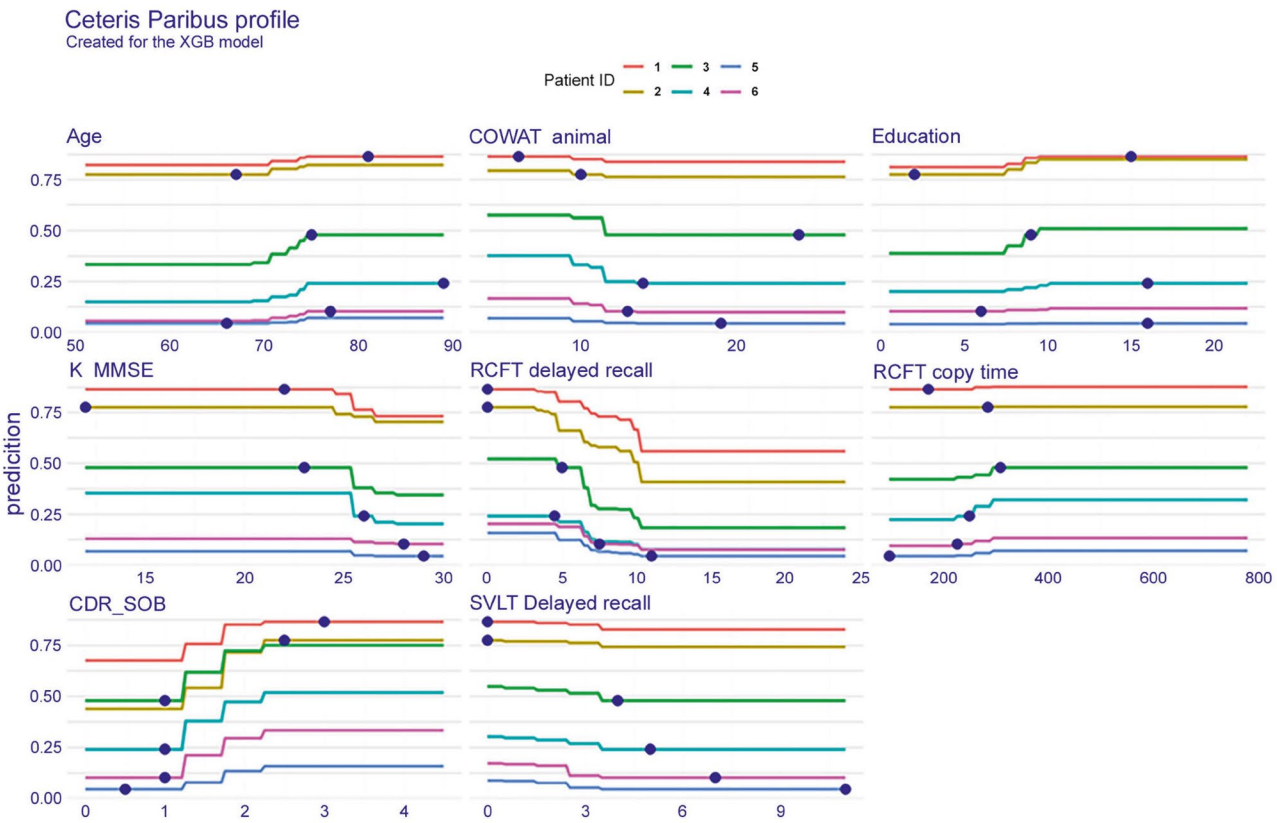

Fig. 20 ICE Explanation. Reproduced with permission from [126]

of a picture are most crucial for a deep network's classification can be done simply using occlusion sensitivity analysis. By eliminating a patch from an image's input dimension and comparing the output, the study determines an image's susceptibility to occlusion in various image regions. The removed patch is important for classification if the variation is significant. The authors have successfully classified the brain areas that mainly contribute to the classification using the Occlusion sensitivity technique. As a result, occlusion sensitivity aids in gaining a high-level knowledge of the image attributes that a network employs to produce a specific classification and sheds light on why a network could misclassify an image. Rieke et al. [134] also use the Occlusion sensitivity method to visualise heatmaps that classify HC and AD. Figure 21 shows the brain area occlusion for AD and HC where the red area indicates the importance of the classification decision.

Saliency Map (SM) is another XAI tool in which an image voxel brightness represents the voxel's saliency. SMs are also called heat maps; they refer to those regions of the image that significantly impact predicting the class to which the object belongs [169]. Volumetric 18F-Fluorodeoxyglucose (FDG) PET scans were used by De Santi et al. [101] to train a CNN that conducts a multiclass classification task (HC, MCI, AD) and explains using two different post-hoc explanation strategies, SM and LRP. While maintaining a constant overall relevance across all layers, the authors used LRP to break down the output of the network into individual contributions of input neurons. The authors then created unique heat maps for each input image using SM to show the significance of each voxel for the categorisation process. Figure 22 is an example of an SM plot showing the evaluation of the averages in each brain region. SM measures the influence of the output on changes in the input image.

We understand from this RQ that a wide range of input parameters, like visual features and volumetric measurements of CT, MRI, and PET scan images and clinical data have been used to train ML and DL models. From Fig. 23, it is evident that SHAP has occupied a predominant position in interpreting AD diagnosis. It is also to be noted that SHAP is employed only on ML techniques. As can be seen from the figure, LIME, DT, GradCAM, and other XAI tools have been used in many other research studies. Furthermore, several XAI frameworks identified in the review prove to reduce model biasing, increase the system's confidence, and try to bridge the gap with the healthcare domain. The RQ also reveals many limitations, including a lack of ground rules for explanations, data imbalance, non-availability of a

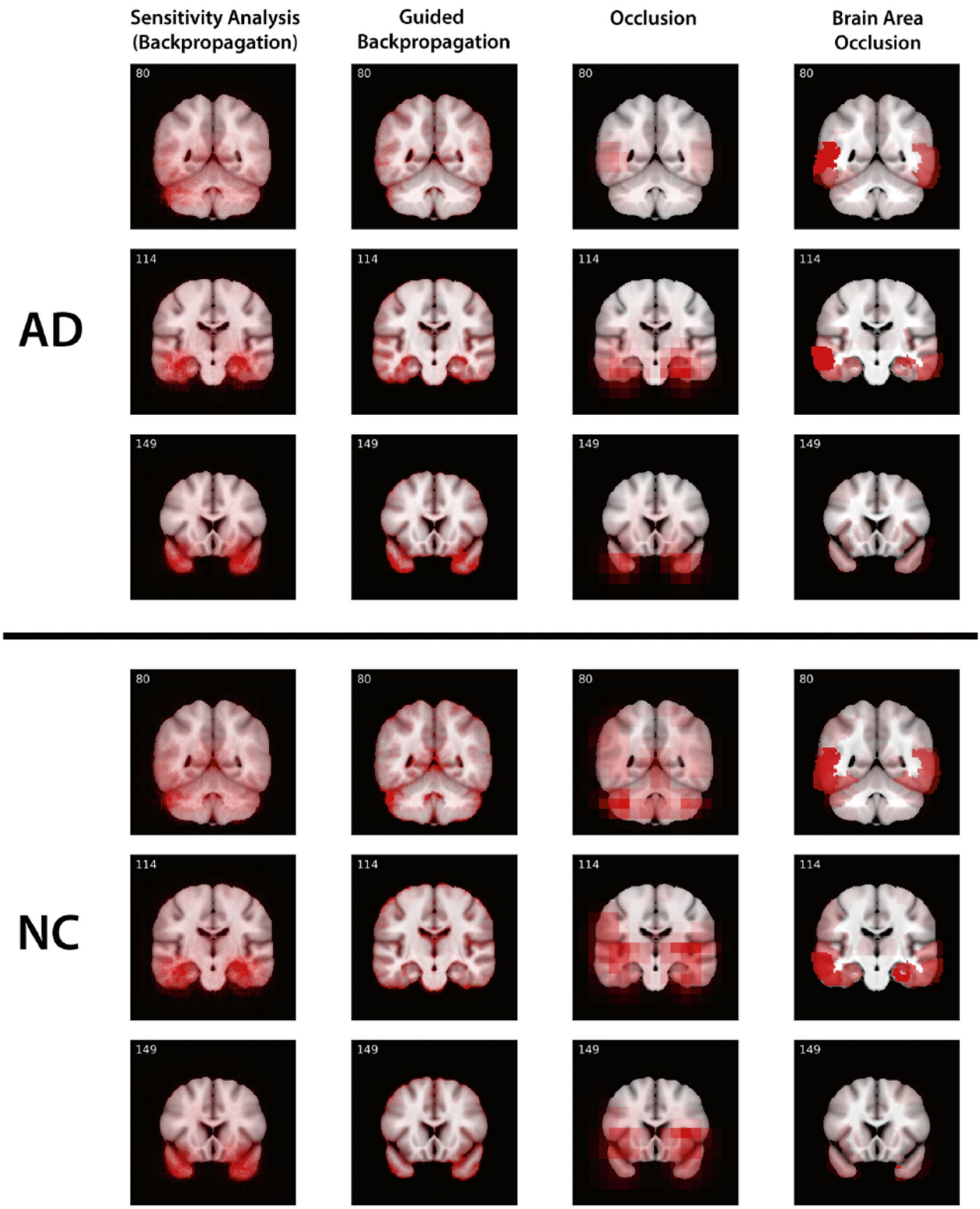

Fig. 21 Occlusion Sensitivity Mapping. Reproduced with permission from [134]

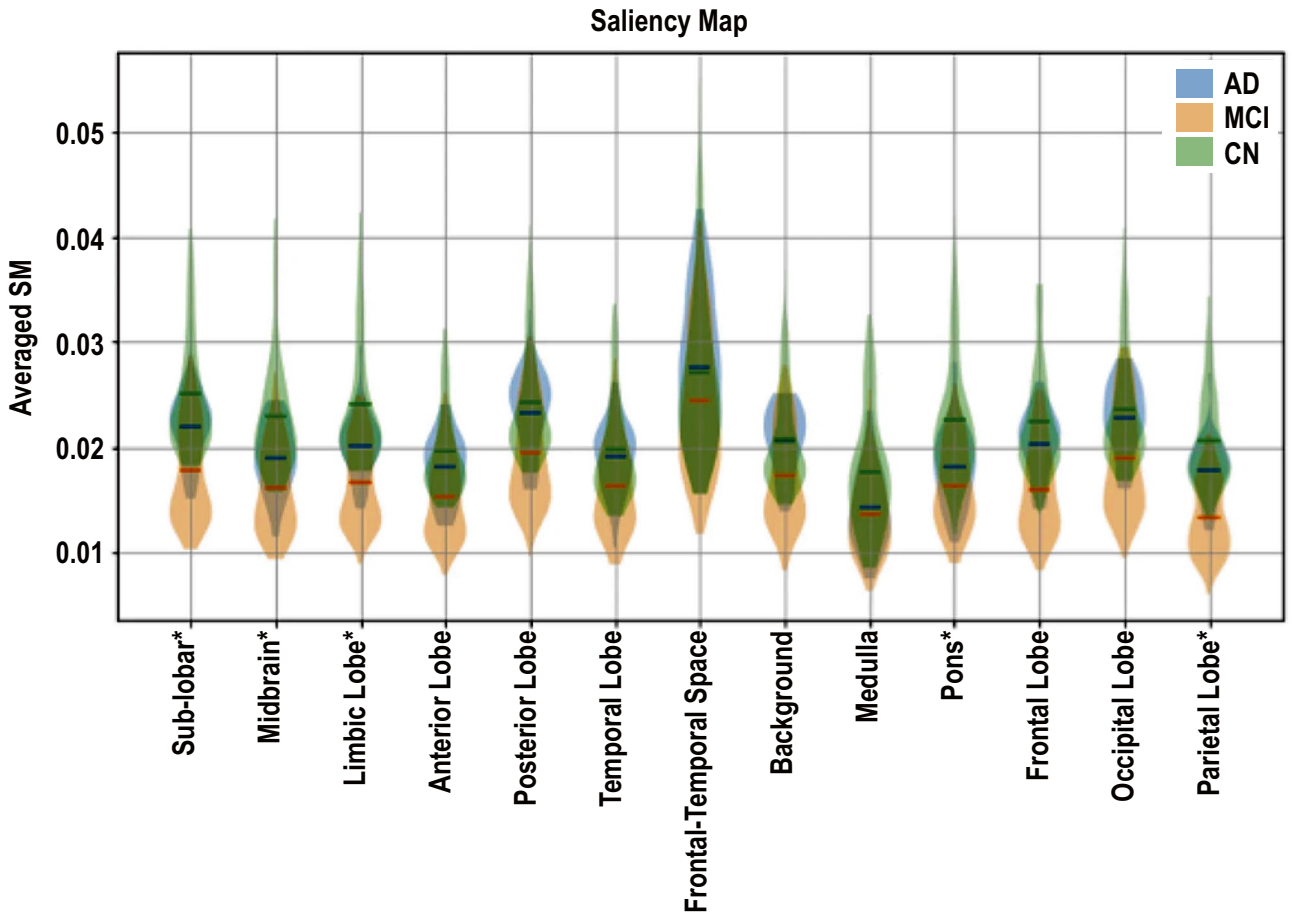

Fig. 22 Saliency Mapping. Reproduced with permission from [101]

comprehensive dataset, and non-inclusiveness of a professional from the healthcare domain. Section "Limitations, Challenges, Needs, and Prospects of XAI in AD Detection" for RQ5 elaborates on the future needs and limitations of AI-based AD detection with XAI.

Benefits of using XAI Methods for AD Detection

This section addresses the RQ4: What are the proven benefits of using XAI in AD detection and healthcare in general?

In this review, studies have reported several benefits of using the concept of XAI in AI-based AD detection. Most studies have tried to report model accuracy, fairness, and transparency. They have highlighted the importance of XAI in fostering confidence and trust when using AI models for prediction, particularly in the medical industry. Independent studies have shown benefits, demonstrating a responsible approach to the development of AI with XAI. In this section, we categorise the benefits from the selected studies based on the four forms of explanation - Numeric, Rule-based, Textual, and Visual. This classification will help researchers to decide appropriate explanations to be sought based on available data modality. While most studies using XAI tools produced explanations in visual form, nominal studies have interpretations in textual, rule-based, and numeric outcomes.

Textual

The field of dementia detection using transcripts with the transformer-based network - BERT by Ilias et al. [131] produces promising classification results. The authors illustrate how transcripts using LIME explain the classification between dementia and non-dementia patients. Figure 24 shows texts in different colours to identify between the labels AD and HC. The tokens or textual forms in transcripts are assigned different colours, indicating which tokens indicate a control group. The intensity of colours for the tokens indicates the importance of these markers for the final transcript classification.

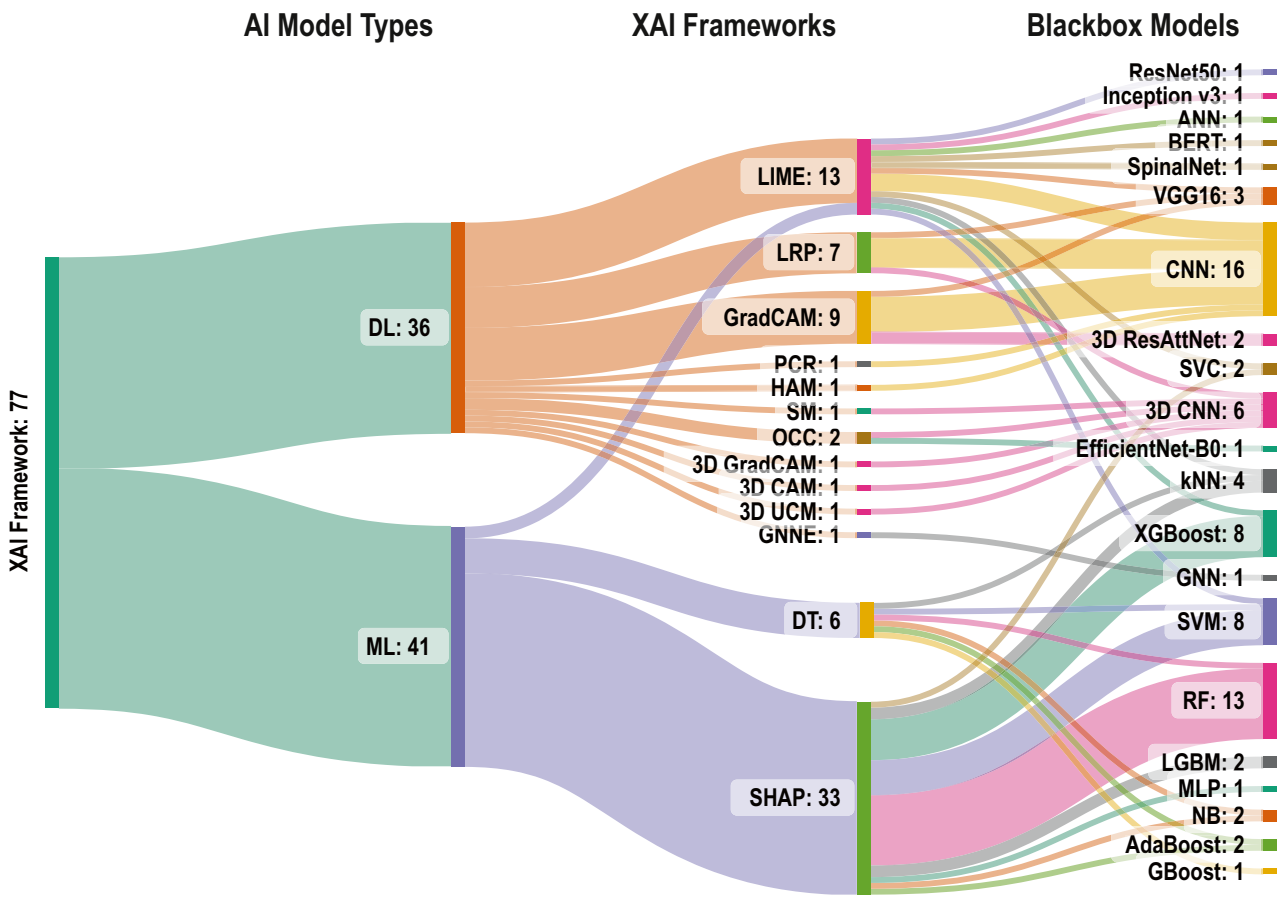

Fig. 23 Sankey diagram showing the various popular XAI frameworks applied to blackbox models

Numeric

In another study, Salih et al. [159] try to develop a proxy that will check for stability in the explanation by choosing the proper XAI method, classifier, and available data. The authors have used Principal Component Analysis (PCA) to verify the stability of the identified predictors with the chosen explainer by quantifying (in numeric form) the informative predictors. In this study, the measure associated with predictors using SHAP and the proxy PCA produces uncorrelated variables that give stable ranking for most classifiers. Figure 25 shows a correlation score of different models for identified features. Due to the widespread use of XAI in delicate fields, including the prognosis of long-term mortality, admission to critical care units, and extubation failure, the results are beneficial to the medical community.

Rule-based

We found two articles in our review that obtain explanations in the form of rules. One study integrates the Internet of Things (IoT) and AI agents to remotely monitor seniors' health status. Khodabandehloo et al. [163] offer a novel HealthXAI system that employs a DT regression method to aid in the early identification of cognitive decline and give caregivers high-level numerical scores reporting inappropriate behaviours and explanations of the forecasts in natural language. The decision rule predicts the value of the target variable and interprets it with a natural language description as either HC or AD. The suggested strategy addresses the problem of ongoing remote monitoring of elderly individuals to aid in the early identification of cognitive decline and to better assist clinicians in reaching a diagnosis. In another study, García-Gutierrez et al. [162] proposed a diagnostic

Table 9 Studies incorporating LIME framework for explaining model predictions

Ref.	Classification Task	Data Type	Significant Features	Classifier	Blackbox
[127]	mdDem vs moDem vs noDem vs vmDem	Image, Numeric	OR8B8, ATP6V1G1, FZD4, HTR1F, OR68B2, GALNT6, ATP6AP1, TGFBRAP1, ORGR1	ML DL	SpinalNet CNN, SVC, KNN, XGBoost Accuracy: CNN - 97.3% SpinalNet - 87.5%
[128]	noDem vs vmDem	Image	Super pixel generation	DL	VGG16, CNN, ResNet50, Inception v3 Accuracy: ResNet - 86.86% VGG16 - 87.44%
[157]	MCI vs AD	Numeric	Headplot Spectogram	ML DL	SVM ANN, CNN Accuracy: ANN - 79-96% CNN - 83-96%
[131]	HC vs AD	Categoric	Text, Vocubulary, Word, Linguistic	DL	BERT, BioBERT, BioClinicalBERT, ConvBERT RoBERTa, ALBERT, XLNet, MTL-BERT, MTL-BERT-DE Accuracy: 86.25%

Table 10 Studies incorporating SHAP framework for explaining model predictions

Ref.	Classification Task	Data Type	Significant Features	Classifier	Blackbox
[122]	HC vs sMCI vs pMCI vs AD	Numeric	Cognitive, PET, MRI, Neuropsychological battery, Genetics, Medical history, CSF data, Other Individual modalities	ML	RF Accuracy: Cross Validation First Layer - 93.95% Second Layer - 87.08%
[123]	HC vs MCI vs AD	Numeric	Volumetric measurements, Demographic features, Cognitive tests, ApoE allele	ML	RF XGBoost
[151]	HC vs MCI vs AD	Numeric	Demographic, Clinical, Neuropsychological	ML	RF Accuracy: 75%
[155]	HC vs MCI vs AD	Numeric	Clinical history, Cognitive features, ApoE4, Summary anatomical Metabolic features, Cerebrospinal fluid biomarkers	ML	XGBoost RF SVM
[156]	HC vs AD	Numeric	Endoplasmic Reticulum stress related differentially expressed genes measures	ML	LGBM SVM Accuracy: SVM - 80.8%
[125]	HC vs erMCI vs ltMCI vs AD	Numeric Categorical	CDRSB, Age, MMSE, RAVLT, FDG, MRI whole brain measure, MRI hippocampus measure, MRI middle temporal artery measure, MRI entorhinal measure, Gender, ApoE	ML	XGBoost, RF Accuracy: 84.2%
[126]	aMCI vs AD	Numeric	Clinical, Demographic, ApoE genotype, Neuropsychological features	ML	RF, SVM, XGBoost Accuracy: 80.7%
[152]	HC vs MCI vs	Numeric	CDRSB, MMSE, EcogSPTotal, RAVLT-perc-for-getting, LDELTOTAL, FAQ, ADAS11, MOCA	ML	Random Seeds and Nested Cross-Validation, SVM, SVM-SMOTE, RF
[159]	HC vs erMCI vs ltMCI vs AD	Numeric Categorical	MRI Volumetric measures, Age, Gender, Education, ApoE	ML	LGBM RF, SVC
[161]	HC vs sMCI vs pMCI vs AD	Numeric	MRI Volumetric measures, Socio-demographic data, ApoE4 alleles, Cognitive test results	ML	XGBoost, RF, SVM Accuracy: 92.6%
[133]	HC vs AD	Numeric	Socio-demographic data, Self-reported medical history, Life Style measures	ML	RF XGBoost
[154]	HC vs ltMCI vs AD	Numeric	Amyloid beta features, glucose uptake features, MRI Volumetric measures, Phosphor tau	ML	RF

tool that uses a DT that provides a simple and unambiguous set of decision rules to provide capabilities to clinicians to give insights into the pathophysiology of AD and behavioural Fronto Temporal Dementia (bvFTD). This paper is beneficial for early detection and diagnosis in the medical field because it outlines all the processes needed to evaluate the datasets, including data preparation, selection of features using an evolutionary approach, and in the creation of a model for the disease discussed in the paper.

Visual

The data models in the studies that use LRP as an AI explainer include CNN and 3D CNN. LRP provides visual explanations as heat maps of significant areas of the brain in identifying brain atrophy. The studies' significant features recognised for interpretation by LRP include the hippocampus, entorhinal cortex, and amygdala. Böhle et al. [121] discuss using LRP with guided backpropagation in discovering

Table 11 Studies incorporating LRP framework for explaining model predictions

Ref.	Classification Task	Data Type	Significnt Features	Classifier	Blackbox
[121]	HC vs AD	Image	Structural MRI data (T1-weighted MPRAGE) Neurobiological data	DL	CNN Specificity - 94%
[164]	HC vs vmDem vs MCI vs moDEM	Image	CNN's feature map	DL	VGG-16, CNN Accuracy - 78.12%
[170]	HC vs AD	Image	CNN's heatmaps	DL	CNN
[132]	HC vs AD	Image	Back propagating the network architecture in the input feature map	DL	3D CNN
[171]	HC vs AD	Image	Clinically-guided prototype learning	DL	XADLiME

the heat maps with relevant significant features. Pohl et al. [132] state that they have discovered similar significant features by using composite LRP, using many propagation rules. The author also identifies that damage to the left and right temporal lobes causes problems with verbal semantic memory and visual memory, respectively. Both authors contribute these findings to the benefit of clinicians and radiologists in diagnosis and building trust in the system.

Several studies use the GradCAM XAI tool for a visual explanation of the predictions of a DL model. Ruengchaijatuporn et al. [124] use GradCAM to visually explain predictions from a VGG16 deep learning model. The DL model has three types of neuropsychological test inputs: clock score prediction, cube-copying drawing, and trail-making inputs. However, the authors prove the benefit of using a CNN model with self-attention work more efficiently than VGG16 with GradCAM. The heat maps proved beneficial to experts with clinical experience and are rated far superior to the baseline model. The authors also claim the model yields better classification performance and interpretability and benefits the domain community. In another study by Jain et al. [160], GradCAM was materialised to show heat maps of a four-way classification of AD predicted using a Generative Adversarial Network (GAN) model. Differently coloured heatmaps obtained from the system help inform predictions of the early onset and severity of dementia. The system has proved beneficial in accurately distinguishing between different classes and making appropriate early predictions. The research community benefits from the authors' use of GAN to create a newly balanced dataset and their awareness of the serious issue with unbalanced datasets. The coloured heat map in the article, which showed the advanced characteristics of various stages of dementia, would aid medical professionals in making judgments. By proposing a 3D Residual Attention Deep Neural Network (3D ResAttNet) that is easy to understand, Zhang et al. [130] have developed an innovative computer-aided technique for the early diagnosis of AD. The authors assert

Table 12 Studies incorporating GradCAM framework for explaining model predictions

Ref.	Classification Task	Data Type	Significant Features	Classifier	Blackbox
[124]	HC vs MCI	Image	Clock drawings, Cube copying, Tail-making	DL	CNN with self attention mechanism VGG-16, CNN Accuracy: 81%
[160]	HC vs aMCI vs pMCI vs AD	Image	CNN models are visualised for their features	DL	CNN Accuracy: VGG16 - 87% CNN - 82%
[130]	HC vs AD and pMCI vs aMCI	Image	CNN's feature map	DL	3D ResAttNet CNN Accuracy: VGG16 - 80.7% ResNet - 85.1% ResAttNet - 86.0%
[172]	HC vs MCI vs AD	Image	LEAR - learn-explain-reinforce	DL	CNN
[173]	HC vs AD	Image	CNN intra-slice features	DL	3DCNN and BRNN
[174]	HC vs AD	Image	Occlusion maps for feature extraction	DL	CNN

Table 13 Studies incorporating a combination of XAI frameworks for explaining model predictions

Ref.	Classification Task	Data Type	XAI Framework	Significant Features	Classifier	Blackbox
[120]	HC vs AD	Numeric Categoric	LIME, SHAP	Normal Whole brain volume, Years of Education, Socio-economic status, Age, MMSE, Gender, Total intracranial volume, Atlas Scaling factor	ML	SVM with radial basis kernel, kNN, MLP Accuracy: SVM - 85.9% KNN - 87.27% MLP - 91.94%
[158]	HC vs AD	Image	HAM, PCR	Salient features related to AD (e.g., atrophy of cerebral cortex and hippocampus	DL	CNN Accuracy - 95.4%
[129]	HC vs MCI vs AD	Image	GNNExplainer	Volume, area of cortical region, average and standard deviation of vertex-based thickness measures of cortical region	DL	Graph Neural Network (GNN) Accuracy: 53.5 ± 4.5%
[165]	HC vs AD	Image	Occlusion Sensitivity Mapping,	White Matter Hyperintensities (WMH)	DL	EfficientNet-B0 Accuracy - 80.0%
[101]	HC vs MCI vs AD	Image	Saliency Map, LRP	MRI, 3D PET, Biological markers, clinical and neuro-psychological assessments	DL	3D CNN
[162]	AD vs FTD (Frontotemporal Dementia)	Image	DT	Demographic data, Cognitive and Brain metabolism data	ML	Bernoulli NB, SVM, kNN, RF, AdaBoost, Gradient Boosting (GBoost) Accuracy - 91.0%
[150]	HC vs AD	Image	3D Ultrametric Contour Map, 3D Class Activation Map, 3D GradCAM	Features of 3D MRI	DL	3D CNN Accuracy - 76.6%
[134]	HC vs MCI vs AD	Image	Sensitivity Analysis Occlusion	Features of 3D Image	DL	3D CNN Accuracy - 77.0%

that the 3D ResAttNet enhances the diagnostic performance and interpretability of MRI with GradCAM by capturing local, global, and spatial information. The study offers an entire end-to-end learning system for automated disease diagnosis. Furthermore, the suggested approach's explainable process can identify and emphasise the role that crucial brain regions like the hippocampus, lateral ventricle, and most of the cortex play in transparent decision-making. Another study by Yang et al. [150] used different 3D-CNNs for classification and AI explainers, including 3D GradCAM. Experts in medicine can gain from the heat maps because they demonstrate how vital the lateral ventricle and most cortical regions are in classifying AD.

Most visualisation techniques consider only the last convolutional layer that extracts global features of pathological abnormalities but do not consider the small subjects and discrepancies. The research by Yu et al. [158] used the High-Resolution Activation Mapping (HAM) approach, which created high-resolution visual explanations that take into account values from the last convolutional layer and intermediate features. Compared to the previous efforts, high-quality heatmaps that display discriminative localisation of brain anomalies perform better. The authors validated the model's effectiveness with good diagnostic accuracy and insightful explanations, which affirm fidelity in clinical applications.

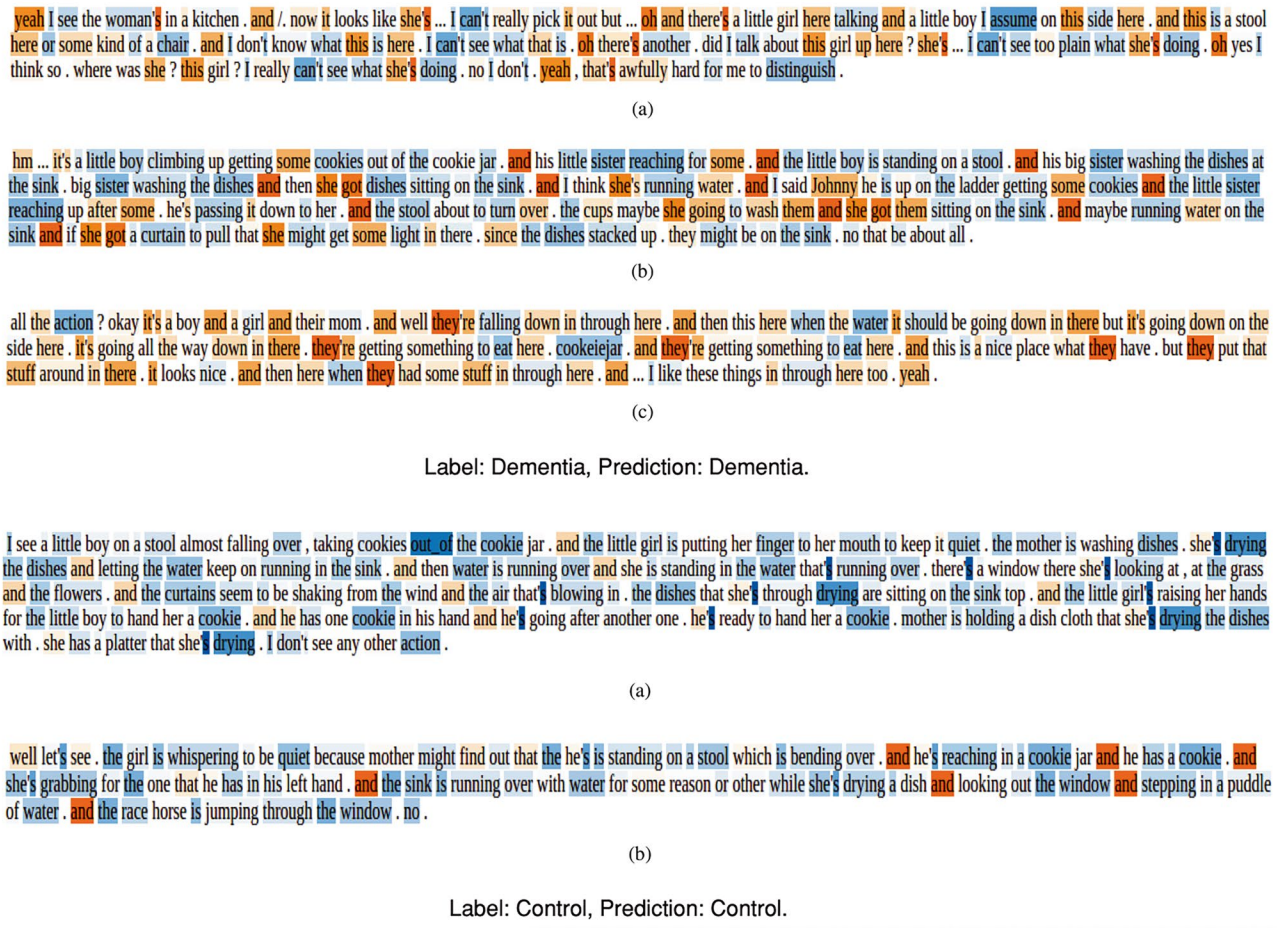

Fig. 24 LIME Explanation – Textual Explanation. Reproduced with permission from [131]

Bordin et al. [165] created heatmaps using the occlusion sensitivity method by occluding a section of the input image with a black patch. The model's brain regions contributing to the classification decision were easily discernible from fluctuations in the output probability predictions. The authors identified and reinforced the relevance of white matter hyperintensity as a neuroimaging biomarker for dementia. One of the studies used LRP to decompose the output score of the network with input 18 FDG PET images into individual contributions while maintaining the conservation principle and heat map produced. Using a Saliency map, the study also generates a voxel-wise heat map for each contribution. In their work, De Santi et al. [101] establish that the colour distribution, as opposed to LRP, emphasises a higher variation among the classes in the saliency map. This study demonstrates that the saliency map regarded the frontal-temporal space of the brain as a vital region for classifying all the classes. The occipital lobe, however, was the area that mattered most in LRP. In both studies, their finding proves significant clinical relevance and, in the long run, leads to increased trust and use of AI models.

Model	DT	LGBM	LR	RF	SVC
DT	1	0.31	0.09	0.02	0.64
LGBM	0.84	1	0.44	0.74	0.01
LR	0.62	0.62	1	0.48	-0.08
RF	0.78	0.90	0.66	1	-0.19
SVC	0.50	0.30	0.43	0.45	1

Fig. 25 Correlation Score – Numeric Explanation. Reproduced with permission from [159]

The study proposed by Kim et al. [129] uses an interpretable Graph Neural Network (GNN) for classifying AD and MCI. Using GNNExplainer, the proposed model's predictions are explained in light of the actual and predicted labels for HC, MCI, and AD. GNNExplainer visualises nodes of importance with a high region of interests representing a high contribution to the classification. The authors found that

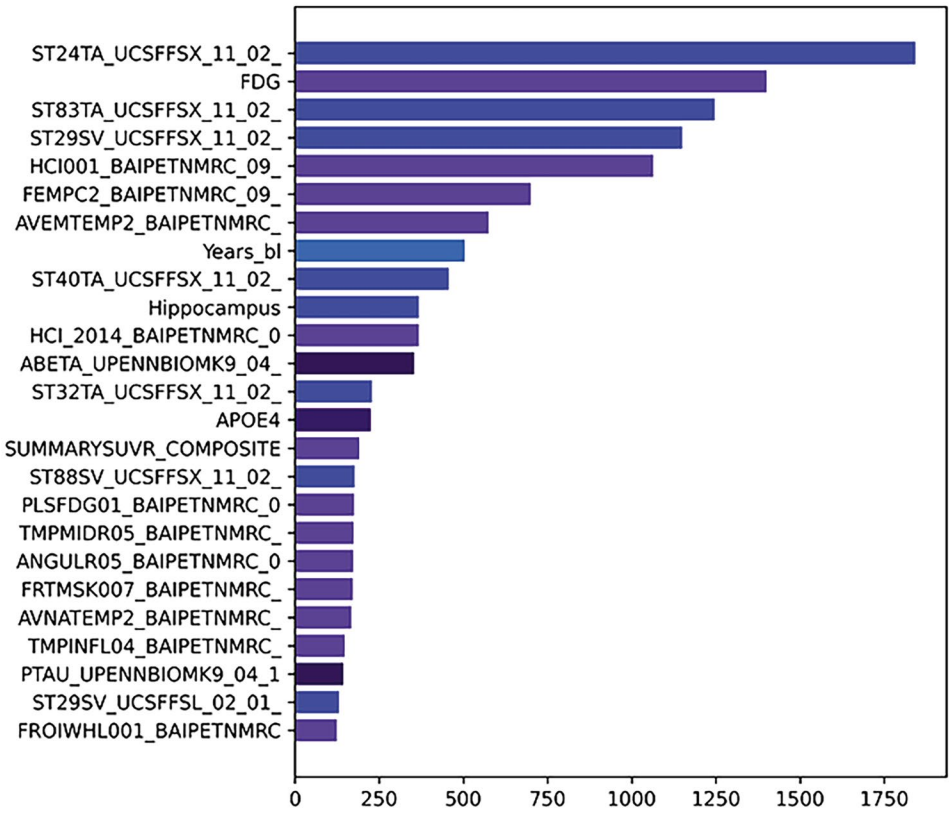

Fig. 26 SHAP Explanation – Violin Plot. Reproduced with permission from [155]

GNNExplainer gives encouraging interpretable results. Also, the explainer can capture the predictor's neuro-anatomical contribution, giving more biological interpretations to better understand AD progression. The authors find the GNNExplainer beneficial as it outperforms other competing models (i.e., DNN, SVM) concerning prediction accuracy.

Several articles considered in the review have used LIME to visualise the explanations of AD predictions. Kamal et al. [127] have used LIME to discover the critical genes responsible for AD. The genes OR8B8 and ATP6V1G1 are found to be very important for AD by the authors. HTR1F and OR6B2 are therefore discovered to be important characteristics of HCs. The predictions about the likely outcome of the generated data are visualised using LIME by Shad et al. [128] and Sidulova et al. [157]. Coloured areas are used to denote the places that prompt models to classify images to make appropriate predictions.

The RF model has been used in numerous research to classify AD ([122, 123, 151, 154]) using SHAP to depict the explanations using force, summary and violin plots (see Figs. 15, 16 and 26 respectively). According to the study in [122], the output decision is supported by several complimentary, credible, and visible justifications. Additionally, the model displays a significant accuracy-interpretability tradeoff due to the accurate outcomes and great interpretability it produced. The proposed model is accurate and understandable, according to the authors. In [123], the author displays SHAP force plots that can explain specific model predictions widely used in clinical practice. The model displays the most significant features that are learned and show an acceptable relationship. The absolute value of each SHAP score reflects how much each attribute contributes to the final prognosis, as shown in [151] by the authors. The internal workings of the RF classifier that are trained with cognitive and clinical data are explained by SHAP, demonstrating a potential connection between feature relevancy patterns and diagnosis. In a different study [154], the authors demonstrate models with great prediction accuracy because they merge many DTs to create a single global forecast. Additionally, it repeated the study using the SHAP method and returned feature ranking results that agreed with those from RF. The study used AD biomarkers strong enough to predict HC, LMCI, and AD correctly and ranked biomarkers according to their significance. The paper also shows that the Amyloid beta (A), tau (T), and neurogenerative biomarkers(N) have different importance in predicting dementia. The study also establishes that the amyloid beta and tau status throughout disease progression plays a more significant role in predicting early cognitive impairment. The study also demonstrates that glucose consumption is more significant in predicting future cognitive impairment. The study incorporates biomarkers from all A, T, and N framework arms into a single integrated analysis, utilising RF to categorise

dementia status and rank biomarker characteristics in order of relative importance.

XGBoost and RF are used in the study by Bogdanovic et al. [125] for AD classification and interpreted with SHAP. The classification model proves beneficial in obtaining exactness and validity in prediction results. The SHAP force plot in either model indicates that the feature clinical dementia rating scale has the highest impact. The features of gender and apolipoprotein, as seen from the SHAP force plot, have the most negligible impact and are not decisive factors for having an AD diagnosis [125]. On the other hand, the study also reveals that mini-mental state examination values impact mainly healthy subjects, and age influences the LMCI class. Danso et al. [133] also go through the benefits of the tree-based approach and how it can give details on decisions made concerning forecasts. The research created a machine-learning model with multiple classifiers to predict AD at both global and local levels. Traits such as education, hypertension, hearing loss, smoking, obesity, depression, physical inactivity, diabetes, and infrequent social interaction were highlighted as potential modifiable risk factors in the report and were among the best-ranked predictive model.

In their study, Hernandez et al. [155] compare XGBoost, RF, and SVM models to understand how to quantify each feature's contribution and achieve the best accuracy. With the help of SHAP violin plots, the study identifies the best models that use information coherent with clinical knowledge. Figure 26 illustrates a violin plot that shows the important features based on the XGBoost classifier for the complete test samples. In Fig. 26 the feature values for various test samples are shown by a colour code, which helps to relate whether a specific feature value favours the high or low probabilities predicted by the model. Blue hues indicate low values, while red hues indicate high values on the colour scale. The author employs a similar justification to demonstrate the utility of the features in distinguishing between the classes.

Lai et al. [156] make use of learning models AdaBoost, LR, LGBM, DT, XGBoost, RF, kNN, Naive Bayes, and SVM along with SHAP to generate force plots to illustrate profiles of the afflicted patient and normal subjects. In this study, the authors found six genes that could accurately predict AD progression and used SHAP to explain the decision-making process of the model used. The study offers fresh perspectives on the function of ER stress-related genes in AD heterogeneity and the creation of brand-new immunotherapy targets for AD patients. The work of Chun et al. [126] uses learning models RF, SVM, and XGBoost. The study is significant because it shows that the Interpretable Machine Learning (IML) method can calculate the individual probability of dementia conversion for each MCI patient. This study's fundamental discovery is that the IML, consisting of ICE, SHAP, and BreakDown plots, enabled the interpretation of variables crucial in each patient's conversion to dementia. The authors affirm that a model using any IML techniques enables predicting patients' conversion from amnestic MCI to dementia.

The study of Xu et al. [152] involves deep learning models that include Random Seeds and Nested Cross-validation, SVM-SMOTE, and RF for a three-way AD classification. Using SHAP, the paper identifies the feature RAVLT-perc-forgetting, and an explanation force-plot for every instance is obtained. The explanations of each instance of the test set can be rotated ninety degrees. Subsequently, the rotated instances are finally stacked horizontally, producing a SHAP summary plot [152]. They consequently provide the doctors with an understanding of how and why the model makes judgements. SHAP is used by Salih et al. [159] and Bloch et al. [161] to determine the order of informative predictors in test data. ML models and their relationships were also visualised and analysed using SHAP summary plots. SHAP force plots examined the individual forecasts of chosen individuals, and the summary plots of those models primarily displayed biologically conceivable outcomes. Moderate to significant correlations were found when comparing the relevance of natural and permutational features to SHAP values.

To summarise, the LIME explainer interprets transcripts predicted by BERT to predict textual tokens. SHAP was used to produce probabilistic prediction in numeric format, DTs produced rule-based ante-hoc interpretations, and other explainers like LRP and GradCAM supported the AD diagnosis by visualising heatmaps showing significant features. A total of 28 research articles out of 37 resorted to visual form representation, one article each for numeric and textual form, and the remaining two explained using the rule-based technique. This research question helped us bring to light the different forms of explanation for AD prediction and will be of significant use for future research.

Limitations, Challenges, Needs, and Prospects of XAI in AD Detection

This section addresses the RQ5: What are the limitations, challenges, needs, and prospects of XAI in AD detection in general?

In the last few years, several studies have been proposed using the XAI concept to better explain AI systems' decisions. Easy access to several XAI frameworks with readily available source code and the availability of high-performance computers has enabled effortless integration of these explainers into standalone AI systems. Unsurprisingly, these efforts have several limitations despite the promising results demonstrated by independent studies. Here, we list several limitations and research gaps in XAI-based AD detection intending to instigate further research in this field.

1. XAI researchers often resort to self-intuition to determine what establishes a good explanation without vali-

dating with a professional from the medical domain [175]. Also, the derived AI explanations are mainly data-driven without domain experts' input. Delivering maximum benefit to stakeholders necessitates a concurrent involvement of medical and AI experts in ascertaining the interpretability evolved by the XAI framework. None of the papers considered in our study has this distinct aspect.

2. One of the significant drawbacks of XAI-based AD diagnosis is the absence of ground truth data [176]. Several neuroimaging and clinical biomarker datasets exist for AD, but none provide ground truth to validate the explainability elicited by XAI models. For instance, in the case of visual explainers (GradCAM, LRP, SM, etc.), heatmaps are often visually assessed. Heatmaps highlight voxels based on classifier decisions without stating underlying atrophy or shape differences in brain regions. This dilutes the heatmap interpretation to a mere indication of where the trained model sees the evidence. Sometimes, heatmaps and the presence of actual biomarkers may be uncorrelated in the case of a poorly trained classifier. Hence, there is a need for rationalising visual assessments in the case of explainers with visual outputs through appropriate ground truth [121].
3. Furthermore, the influences of XAI explanations dramatically vary when delivered to people with varying levels of domain expertise [121]. When people observe explanations contradicting their own intuition, a confusing situation arises, questioning the counterintuitive relationship delivered by the XAI systems. Such situations lead to further doubting the correctness of the model even though the model delivers a valid explanation [121]. The only way to circumvent such a situation is to have ground truth where one can objectively validate the explanation against the ground truth data without challenging the decisions made by the XAI systems.
4. Confidence measures are crucial in computer-aided diagnosis, where a wrong prediction is almost always life-threatening. When the system cannot deliver a confident prediction, it must warrant a manual intervention to arrive at an appropriate decision. Hence, XAI methods must also incorporate a confidence score to identify situations when the classifier is incorrect before providing explanations. Otherwise, the end user may create false trust in the system [177]. Therefore it is vital to evaluate not only whether an explanation is intuitive to the user but also to arrive at an optimal decision[177].
5. Some papers used multiple XAI frameworks for enhanced explainability. It may be good from an academic standpoint but contributes to added opaqueness in real-time. For instance, LIME and SHAP frameworks were used jointly in one study [120]. The feature rankings derived by these individual frameworks did not correlate with each other. The Mini-Mental State Examination (MMSE) significantly contributes to SHAP, whereas normalised Whole Brain Value (nWBV) dominates the LIME features [120]. In yet another study [161], SHAP was used with other methods to validate the interpretability. Again, a weaker correlation was found between feature rankings of the SHAP values and other models. Such scenarios lead to ambiguity in the explanations delivered by the models resulting in a complete loss of clinicians' trust in the models.
6. Another significant lapse in almost all studies we considered is the limited use of medical datasets or the non-availability of a comprehensive benchmark dataset that exhibits variations representing real-world scenarios [178]. It impedes testing of the model on an extensive dataset which is crucial in determining the actual robustness [131, 155, 156, 179]. Hence, most of the studies in the literature ended up with subjective claims but exhibited subpar performance due to generalisability issues when tested on a different dataset [178]. Another closely related issue with the dataset is the issue of class imbalance [123, 165]. The ML or DL learning algorithms predict dominant classes more accurately than classes with inadequate samples. Most studies had limited AD samples compared to HC or MCI cohorts [123, 125, 133, 151, 157, 161, 165]. Only a balanced dataset can draw meaningful insights. Applying XAI-based AI techniques in AD diagnosis will become genuinely influential if only research efforts can be diverted into creating such a comprehensive, balanced, and benchmark dataset.
7. Although some studies utilised multimodal data (clinical, sociodemographic, MRI features, neuroimages, etc.) to predict AD [122, 127], explanations were derived for a single modality only. This may be due to the absence of correlation among the interpretation obtained from different modalities (see point 5 above). Hence, having medical experts in the loop and deriving interpretations for every modality used is the way forward.
8. Even though some studies used XAI tools in AD prediction, they did not consider disease biomarkers such as MRI volumetry, cortical thickness, etc., which correlate well with dementia [124, 126].
9. Most studies have not indicated factors (hyperparameter values, the split proportion of train-test data, data preprocessing, etc.) affecting model accuracy and subsequent explainability derived [125].
10. Another huge concern that adds to the reluctance of medical experts to use AI solutions reliably is the inability of AI to consider the history of anomalies that contributed to cognitive decline. The lack of real-world labelled data sets of individuals collected over a long

period of time is a genuine limitation for any medical field, not just in AI-based AD diagnosis [163].

11. Even though researchers applied either single or multiple XAI frameworks in the AD prediction, sometimes there was no specific correlation between the AI prediction and the associated brain region. [157].

We have seen numerous studies proposed to explain the AD prognosis and diagnosis using several XAI frameworks. Although these studies have greatly facilitated clinical fidelity in the associated predictions, this RQ made us realise that we are far from making use of the XAI-based AD systems in real medical eventualities due to the aforementioned limitations and challenges. In future, AI technocrats must thoroughly investigate these needs by involving medical experts in the loop to deliver profound trustworthiness to the medical community for AI-driven AD diagnosis.

Conclusion

Explainable Artificial Intelligence has gained tremendous importance over the last several years due to scientific demands and regulatory compliance. Researchers are exploring different XAI frameworks that characterise the accuracy of the model, rationality and clarity in AI-assisted decision-making, which is impeccable in healthcare. XAI aids in creating synergistic environments where it can efficiently address the solution to predictions such as long-term mortality and extubation failures. Hence, promoting wider dissemination of XAI concepts, backgrounds, and techniques to the research community is crucial.

Towards this aim and to serve as a reference source, this article presents a systematic review of XAI models and frameworks' application on multimodal AD data. We have reviewed articles based on XAI for AD diagnosis for the last decade. The study included 37 research articles thoroughly reviewed through carefully framed RQs. The RQs highlighted different XAI-based studies adopted for AD diagnosis and unveiled various ML and DL models that have embraced XAI frameworks to imbibe transparency and fidelity in AI predictions. The study also reveals several benefits, limitations, and future avenues for clinical diagnosis. We understand it is too early to comment on reducing the gap between medical and AI domains to a minimal zero. Nevertheless, such reviews will reveal the benefits and limitations to the research community so that the trade-off between accuracy in AI solutions and explainability can be sorted out to an acceptable level of fidelity. This review will help explore many healthcare domains to leverage the true capabilities of AI in fostering fidelity in the clinical decision support system.

Acknowledgements N.S., F.H., K.S and V.V were funded by the Ministry of Higher Education, Research and Innovation (MoHERI) of the sultanate of Oman under the Block Funding Program (Grant number-MoHERI/BFP/UoTAS/01/2021). M.M. was funded by Nottingham Trent University, UK, through the Quality of Research grant 2023 from the Computing and Informatics Research Centre. V.V. is funded by the internal research grant 2023 of the UTAS-Sohar, Oman.

Author Contributions All authors worked closely together to complete this project. M.M. conceived the presented concept. All authors were involved in the identification of the relevant articles for this study. M.M. supervised the findings of this work. V.V. and N.S. wrote the manuscript. All authors have edited the manuscript. All authors have contributed to and approved the final version of the manuscript.

Data Availability There were no datasets generated during this study.

Declarations

Ethics Approval In this study, no human or animal participants were involved; hence ethical approval was not needed.

Informed Consent Informed consent was obtained from all individual participants included in the study.

Conflicts of Interest The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. McDade EM. Alzheimer Disease. CONTINUUM: Lifelong Learning in Neurology. 2022;28(3):648–75.
2. Shaffi N, Hajamohideen F, Mahmud M, Abdesselam A, Subramanian K, Sariri AA. Triplet-Loss Based Siamese Convolutional Neural Network for 4-Way Classification of Alzheimer's Disease. In: International Conference on Brain Informatics. Springer 2022; 277–87.
3. Gauthier S, Webster C, Sarvaes S, Morais J, Rosa-Neto P. World Alzheimer Report. Life After Diagnosis - Navigating Treatment. Alzheimer's Disease International: Care and Support; 2022. p. 2022.
4. Dubois B, Picard G, Sarazin M. Early detection of Alzheimer's disease: new diagnostic criteria. Dialogues in clinical neuroscience. 2022.

5. Tatulian SA. Challenges and hopes for Alzheimer's disease. Drug Discovery Today. 2022
6. Knopman DS, Amieva H, Petersen RC, Chételat G, Holtzman DM, Hyman BT, et al. Alzheimer disease. Nature reviews Disease primers. 2021;7(1):1–21.
7. Yahaya SW, Lotfi A, Mahmud M. Towards the Development of an Adaptive System for Detecting Anomaly in Human Activities. In: Proc SSCI 2020;534–41.
8. Yahaya SW, Lotfi A, Mahmud M. Towards a data-driven adaptive anomaly detection system for human activity. Pattern Recognit Lett. 2021;145:200–7.
9. Lalotra GS, Kumar V, Bhatt A, Chen T, Mahmud M. iReTADS: An Intelligent Real-Time Anomaly Detection System for Cloud Communications Using Temporal Data Summarization and Neural Network. Secur Commun Netw. 2022;1–15. ArticleID: 9149164.
10. Fabietti M, et al. Neural network-based artifact detection in local field potentials recorded from chronically implanted neural probes. In: Proc IJCNN 2020;1–8.
11. Fabietti M, et al. Artifact detection in chronically recorded local field potentials using long-short term memory neural network. In: Proc AICT 2020;1–6.
12. Fabietti M, et al. Adaptation of convolutional neural networks for multi-channel artifact detection in chronically recorded local field potentials. In: Proc SSCI 2020;1607–13.
13. Fabietti M, Mahmud M, Lotfi A. Machine learning in analysing invasively recorded neuronal signals: available open access data sources. In: Proc. Brain Inform 2020;151–62.
14. Fabietti M, Mahmud M, Lotfi A. Artefact Detection in Chronically Recorded Local Field Potentials: An Explainable Machine Learning-based Approach. In: Proc. IJCNN 2022; 1–7.
15. Rahman S, Sharma T, Mahmud M. Improving alcoholism diagnosis: comparing instance-based classifiers against neural networks for classifying EEG signal. In: Proc Brain Inform 2020;239–50.
16. Tahura S, Hasnat Samiul S, ShamimKaiser M, Mahmud M. Anomaly detection in electroencephalography signal using deep learning model. In: Proc TCCE 2021;205–17.
17. Wadhera T, Mahmud M. Computing Hierarchical Complexity of the Brain from Electroencephalogram Signals: A Graph Convolutional Network-based Approach. In: Proc IJCNN 2022;1–6.
18. Fabietti MI, et al. Detection of Healthy and Unhealthy Brain States from Local Field Potentials Using Machine Learning. In: Proc Brain Inform 2022;27–39
19. Das S, Obaidullah SM, Mahmud M, Kaiser MS, Roy K, Saha CK, et al. A machine learning pipeline to classify foetal heart rate deceleration with optimal feature set. Sci Rep. 2023;13(1):2495.
20. Singh R, Mahmud M, Yovera L. Classification of First Trimester Ultrasound Images Using Deep Convolutional Neural Network. In: Proc AII 2021;92–105.
21. Sutton S, Mahmud M, Singh R, Yovera L. Identification of Crown and Rump in First-Trimester Ultrasound Images Using Deep Convolutional Neural Network. In: Proc. AII; 2023;231–47.
22. Sumi AI, et al. fASSERT: A fuzzy assistive system for children with autism using internet of things. In: Proc Brain Inform 2018;403–12.
23. Al Banna M, et al. A monitoring system for patients of autism spectrum disorder using artificial intelligence. In: Proc Brain Inform 2020;251–62.
24. Akter T, et al. Towards autism subtype detection through identification of discriminatory factors using machine learning. In: Proc Brain Inform 2021;401–10.
25. Biswas M, Kaiser MS, Mahmud M, AlMamun S, Hossain M, Rahman MA, et al. An xai based autism detection: The context behind the detection. In: Proc Brain Inform 2021;448–59.
26. Ghosh T, et al. Artificial intelligence and internet of things in screening and management of autism spectrum disorder. Sustain Cities Soc. 2021;74:103189.
27. Ahmed S, et al. Toward Machine Learning-Based Psychological Assessment of Autism Spectrum Disorders in School and Community. In: Proc TEHI 2022;139–49.
28. Mahmud M, et al. Towards explainable and privacy-preserving artificial intelligence for personalisation in autism spectrum disorder. In: Proc HCII; 2022;356–70.
29. Wadhera T, Mahmud M. Brain Networks in Autism Spectrum Disorder, Epilepsy and Their Relationship: A Machine Learning Approach. In: Artificial Intelligence in Healthcare; 2022;125–42.
30. Wadhera T, Mahmud M. Influences of Social Learning in Individual Perception and Decision Making in People with Autism: A Computational Approach. In: Proc. Brain Inform 2022;50–61.
31. Wadhera T, Mahmud M. Brain Functional Network Topology in Autism Spectrum Disorder: A Novel Weighted Hierarchical Complexity Metric for Electroencephalogram. IEEE J Biomed Health Inform. 2022;1–8.
32. Wadhera T, Mahmud M. A Deep Concatenated Convolutional Neural Network-based Method to Classify Autism. In: Proc ICONIP 2022;1–10.
33. Akhund NU, et al. ADEPTNESS: Alzheimer's disease patient management system using pervasive sensors-early prototype and preliminary results. In: Proc. Brain Inform 2018;413–22.
34. Noor MBT, Zenia NZ, Kaiser MS, Mamun SA, Mahmud M. Application of deep learning in detecting neurological disorders from magnetic resonance images: a survey on the detection of Alzheimer's disease. Parkinson's disease and schizophrenia Brain informatics. 2020;7(1):1–21.
35. Jesmin S, Kaiser MS, Mahmud M. Towards artificial intelligence driven stress monitoring for mental wellbeing tracking during COVID-19. In: Proc. WI-IAT; 2020; 845–51.
36. Biswas M, Rahman A, Kaiser MS, Al Mamun S, Ebne Mizan KS, Islam MS, et al. Indoor navigation support system for patients with neurodegenerative diseases. In: Proc. Brain Inform.; 2021;411–22.
37. AlMamun S, Kaiser MS, Mahmud M. An artificial intelligence based approach towards inclusive healthcare provisioning in society 5.0: A perspective on brain disorder. In Proc Brain Inform 2021;157–69
38. Motin MA, Mahmud M, Brown DJ. Detecting Parkinson's Disease from Electroencephalogram Signals: An Explainable Machine Learning Approach. In: Proc AICT 2022;1–6.
39. Shaffi N, Mahmud M, Hajamohideen F, Subramanian K, Kaiser MS. Machine Learning and Deep Learning Methods for the Detection of Schizophrenia using Magnetic Resonance Images and EEG Signals: An Overview of the Recent Advancements. In: Proc. ICTCS 2022;1-18.
40. Shaffi N, Hajamohideen F, Abdesselam A, Mahmud M, Subramanian K. Ensemble Classifiers for a 4-Way Classification of Alzheimer's Disease. In: Proc. AII; 2023;219–30.
41. Hajamohideen F, Shaffi N, Mahmud M, Subramanian K, Al Sariri A, Vimbi V, et al. Four-way classification of Alzheimer's disease using deep Siamese convolutional neural network with triplet-loss function. Brain Informatics. 2023;10(1):1–13.
42. Shafiq S, Ahmed S, Kaiser MS, Mahmud M, Hossain MS, Andersson K. Comprehensive Analysis of Nature-Inspired Algorithms for Parkinson's Disease Diagnosis. IEEE Access. 2023;11:1629–53.
43. Javed AR, Saadia A, Mughal H, Gadekallu TR, Rizwan M, Maddikunta PKR, et al. Artificial Intelligence for Cognitive Health Assessment: State-of-the-Art, Open Challenges and Future Directions. Cognitive Computation 2023;1–46.
44. Fabietti M, Mahmud M, Lotfi A, Leparulo A, Fontana R, Vassanelli S, et al. Early Detection of Alzheimer's Disease from Cortical and Hippocampal Local Field Potentials using an Ensembled Machine Learning Model. IEEE Transactions on Neural Systems and Rehabilitation Engineering. 2023;31(2839):2848.
45. Jesmin S, Kaiser MS, Mahmud M. Artificial and internet of healthcare things based Alzheimer care during COVID 19. In: Proc Brain Inform 2020;263–74.

46. Satu MS, et al. Short-term prediction of COVID-19 cases using machine learning models. Applied Sciences. 2021;11(9):4266.
47. Bhapkar HR, Mahalle PN, Shinde GR, Mahmud M. Rough sets in COVID-19 to predict symptomatic cases. In: COVID-19: Prediction, Decision-Making, and its Impacts 2021;57-68.
48. Kumar S, Viral R, Deep V, Sharma P, Kumar M, Mahmud M, et al. Forecasting major impacts of COVID-19 pandemic on country-driven sectors: challenges, lessons, and future roadmap. Pers Ubiquitous Comput 2021;1–24.
49. Mahmud M, Kaiser MS. Machine learning in fighting pandemics: a COVID-19 case study. In: COVID-19: prediction, decision-making, and its impacts; 2021;77–81.
50. Prakash N, Murugappan M, Hemalakshmi G, Jayalakshmi M, Mahmud M. Deep transfer learning for COVID-19 detection and infection localization with superpixel based segmentation. Sustain Cities Soc. 2021;75.
51. Paul A, Basu A, Mahmud M, Kaiser MS, Sarkar R. Inverted bell-curve-based ensemble of deep learning models for detection of COVID-19 from chest X-rays. Neural Comput Appl 2022;1–15.
52. Banerjee JS, Chakraborty A, Mahmud M, Kar U, Lahby M, Saha G. Explainable Artificial Intelligence (XAI) Based Analysis of Stress Among Tech Workers Amidst COVID-19 Pandemic. In: Advanced AI and Internet of Health Things for Combating Pandemics 2023;151–74.
53. Nahian MJA, Ghosh T, Uddin MN, Islam MM, Mahmud M, Kaiser MS. Towards artificial intelligence driven emotion aware fall monitoring framework suitable for elderly people with neurological disorder. In: Proc Brain Inform 2020;275–86.
54. Nahiduzzaman M, Tasnim M, Newaz NT, Kaiser MS, Mahmud M. Machine learning based early fall detection for elderly people with neurological disorder using multimodal data fusion. In: Proc Brain Inform 2020;204–14.
55. Nahian MJA, Ghosh T, Al Banna MH, Aseeri MA, Uddin MN, Ahmed MR, et al. Towards an accelerometer-based elderly fall detection system using cross-disciplinary time series features. IEEE Access. 2021;9:39413–31.
56. Nahian MJA, Raju MH, Tasnim Z, Mahmud M, Ahad MAR, Kaiser MS. Contactless fall detection for the elderly. In: Contactless Human Activity Analysis. Springer 2021;203–35.
57. Farhin F, Kaiser MS, Mahmud M. Towards secured service provisioning for the internet of healthcare things. In: Proc AICT 2020;1–6.
58. Farhin F, Sultana I, Islam N, Kaiser MS, Rahman MS, Mahmud M. Attack detection in internet of things using software defined network and fuzzy neural network. In: Proc. ICIEV and ICIVPR 2020;1–6.
59. Ahmed S, et al. Artificial intelligence and machine learning for ensuring security in smart cities. In: Data-driven mining, learning and analytics for secured smart cities. Springer 2021;23–47.
60. Islam N, et al. Towards machine learning based intrusion detection in IoT networks. Comput Mater Contin. 2021;69(2):1801–21.
61. Esha NH, et al. Trust IoHT: A trust management model for internet of healthcare things. In: Proc. ICDSA 2021;47–57
62. Zaman S, et al. Security threats and artificial intelligence based countermeasures for internet of things networks: a comprehensive survey. Ieee Access. 2021;9:94668–90.
63. Zohora MF, Tania MH, Kaiser MS, Mahmud M. Forecasting the risk of type ii diabetes using reinforcement learning. In: Proc ICIEV and ICIVPR 2020;1–6.
64. Mukherjee H, et al. Automatic lung health screening using respiratory sounds. J Med Syst. 2021;45(2):1–9.
65. Deepa B, Murugappan M, Sumithra M, Mahmud M, Al-Rakhami MS. Pattern Descriptors Orientation and MAP Firefly Algorithm Based Brain Pathology Classification Using Hybridized Machine Learning Algorithm. IEEE Access. 2021;10:3848–63.
66. Mammoottil MJ, et al. Detection of Breast Cancer from Five-View Thermal Images Using Convolutional Neural Networks. J Healthc Eng 2022.
67. Chen T, et al. A dominant set-informed interpretable fuzzy system for automated diagnosis of dementia. Front Neurosci. 2022;16:86766.
68. Mukherjee P, et al. iConDet: An Intelligent Portable Healthcare App for the Detection of Conjunctivitis. In: Proc. AII 2021;29–42.
69. Tasnim N, Al Mamun S, Shahidul Islam M, Kaiser MS, Mahmud M. Explainable Mortality Prediction Model for Congestive Heart Failure with Nature-Based Feature Selection Method. Applied Sciences. 2023;13(10):6138.
70. Farhin F, Kaiser MS, Mahmud M. Secured smart healthcare system: blockchain and bayesian inference based approach. In: Proc TCCE 2021;455–65.
71. Kaiser MS, et al. 6G access network for intelligent internet of healthcare things: opportunity, challenges, and research directions. In: Proc. TCCE 2021;317–28.
72. Biswas M, et al. ACCU3RATE: A mobile health application rating scale based on user reviews. PloS one. 2021;16(12).
73. Adiba FI, Islam T, Kaiser MS, Mahmud M, Rahman MA. Effect of corpora on classification of fake news using naive Bayes classifier. Int J Autom Artif Intell Mach Learn. 2020;1(1):80–92.
74. Rabby G, et al. A flexible keyphrase extraction technique for academic literature. Procedia Comput Sci. 2018;135:553–63.
75. Ghosh T, et al. An Attention-Based Mood Controlling Framework for Social Media Users. In: Proc Brain Inform 2021;245–56.
76. Ghosh T, Al Banna MH, Al Nahian MJ, Uddin MN, Kaiser MS, Mahmud M. An attention-based hybrid architecture with explainability for depressive social media text detection in Bangla. Expert Systems with Applications. 2023;213.
77. Ahuja NJ, et al. An Investigative Study on the Effects of Pedagogical Agents on Intrinsic, Extraneous and Germane Cognitive Load: Experimental Findings With Dyscalculia and Non-Dyscalculia Learners. IEEE Access. 2021;10:3904–22.
78. Rahman MA, et al. Explainable multimodal machine learning for engagement analysis by continuous performance test. In: Proc. HCII; 2022;386–99.
79. Rahman MA, Brown DJ, Shopland N, Harris MC, Turabee ZB, Heym N, et al. Towards machine learning driven self-guided virtual reality exposure therapy based on arousal state detection from multimodal data. In: International Conference on Brain Informatics. Springer; 2022;195–209.
80. Rahman MA, Brown DJ, Mahmud M, Harris M, Shopland N, Heym N, et al. Enhancing biofeedback-driven self-guided virtual reality exposure therapy through arousal detection from multimodal data using machine learning. Brain Informatics. 2023;10(1):1–18.
81. Al Banna MH, Taher KA, Kaiser MS, Mahmud M, Rahman MS, Hosen AS, et al. Application of artificial intelligence in predicting earthquakes: state-of-the-art and future challenges. IEEE Access. 2020;8:192880–923.
82. Al Banna MH, et al. Attention-based bi-directional long-short term memory network for earthquake prediction. IEEE Access. 2021;9:56589–603.
83. Ahmed Z, Mohamed K, Zeeshan S, Dong X. Artificial intelligence with multi-functional machine learning platform development for better healthcare and precision medicine. Database. 2020.
84. Davenport T, Kalakota R. The potential for artificial intelligence in healthcare. Future healthcare journal. 2019;6(2):94.
85. Mahmud M, Kaiser MS, McGinnity TM, Hussain A. Deep learning in mining biological data. Cognitive computation. 2021;13(1):1–33.
86. Fabrizio C, Termine A, Caltagirone C, Sancesario G. Artificial Intelligence for Alzheimer's Disease: Promise or Challenge? Diagnostics. 2021;11(8):1473.

87. Mahmud M, Kaiser MS, Hussain A, Vassanelli S. Applications of Deep Learning and Reinforcement Learning to Biological Data. IEEE Transactions on Neural Networks and Learning Systems. 2018;29(6):2063–79.
88. Rai A. Explainable AI: From black box to glass box. Journal of the Academy of Marketing Science. 2020;48(1):137–41.
89. Yang G, Ye Q, Xia J. Unbox the black-box for the medical explainable AI via multi-modal and multi-centre data fusion: A mini-review, two showcases and beyond. Information Fusion. 2022;77:29–52.
90. Kaur D, Uslu S, Rittichier KJ, Durresi A. Trustworthy artificial intelligence: a review. ACM Computing Surveys (CSUR). 2022;55(2):1–38.
91. Nazar M, Alam MM, Yafi E, Su'ud MM. A systematic review of human-computer interaction and explainable artificial intelligence in healthcare with artificial intelligence techniques. IEEE Access. 2021;9:153316–48.
92. Kourou K, Exarchos TP, Exarchos KP, Karamouzis MV, Fotiadis DI. Machine learning applications in cancer prognosis and prediction. Computational and structural biotechnology journal. 2015;13:8–17.
93. Vamathevan J, Clark D, Czodrowski P, Dunham I, Ferran E, Lee G, et al. Applications of machine learning in drug discovery and development. Nature reviews Drug discovery. 2019;18(6):463–77.
94. Schmidt J, Marques MR, Botti S, Marques MA. Recent advances and applications of machine learning in solid-state materials science. NPJ Computational Materials. 2019;5(1):1-36.
95. Lei Y, Yang B, Jiang X, Jia F, Li N, Nandi AK. Applications of machine learning to machine fault diagnosis: A review and roadmap. Mechanical Systems and Signal Processing. 2020;138.
96. Mamoshina P, Vieira A, Putin E, Zhavoronkov A. Applications of deep learning in biomedicine. Molecular pharmaceutics. 2016;13(5):1445–54.
97. Deng L, Yu D, et al. Deep learning: methods and applications. Foundations and trends® in signal processing. 2014;7(3–4):197-387.
98. Kumar D, Mehta MA. 3. In: An Overview of Explainable AI Methods, Forms and Frameworks. Cham: Springer International Publishing; 2023;43–59. Available from: https://doi.org/10.1007/978-3-031-12807-3_3.
99. Loh HW, Ooi CP, Seoni S, Barua PD, Molinari F, Acharya UR. Application of Explainable Artificial Intelligence for Healthcare: A Systematic Review of the Last Decade (2011–2022). Computer Methods and Programs in Biomedicine. 2022:107161.
100. Ribeiro MT, Singh S, Guestrin C. " Why should i trust you?" Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining; 2016;1135–44.
101. DeSanti LA, Pasini E, Santarelli MF, Genovesi D, Positano V. An Explainable Convolutional Neural Network for the Early Diagnosis of Alzheimer's Disease from 18F-FDG PET. J Digit Imaging. 2022;1–15.
102. Ghassemi M, Oakden-Rayner L, Beam AL. The false hope of current approaches to explainable artificial intelligence in health care. The Lancet Digital Health. 2021;3(11):e745-50.
103. Pawar U, O'Shea D, Rea S, O'Reilly R. Explainable ai in healthcare. In: 2020 International Conference on Cyber Situational Awareness, Data Analytics and Assessment (CyberSA). IEEE 2020;1–2
104. Wanner J, Herm LV, Heinrich K, Janiesch C. Stop ordering machine learning algorithms by their explainability! An empirical investigation of the tradeoff between performance and explainability. In: Conference on e-Business, e-Services and e-Society. Springer; 2021;245–58.
105. Jung YJ, Han SH, Choi HJ. Explaining CNN and RNN using selective layer-wise relevance propagation. IEEE Access. 2021;9:18670–81.
106. Jiménez-Luna J, Grisoni F, Schneider G. Drug discovery with explainable artificial intelligence. Nature Machine Intelligence. 2020;2(10):573–84.
107. Preuer K, Klambauer G, Rippmann F, Hochreiter S, Unterthiner T. Interpretable deep learning in drug discovery. In: Explainable AI: Interpreting, Explaining and Visualizing Deep Learning. Springer 2019;331–45.
108. Gade K, Geyik S, Kenthapadi K, Mithal V, Taly A. Explainable AI in industry: Practical challenges and lessons learned. In: Companion Proceedings of the Web Conference 2020;303–4.
109. Ahmed I, Jeon G, Piccialli F. From artificial intelligence to explainable artificial intelligence in industry 4.0: a survey on what, how, and where. IEEE Transactions on Industrial Informatics. 2022;18(8):5031–42.
110. Tao J, Xiong Y, Zhao S, Wu R, Shen X, Lyu T, et al. Explainable AI for Cheating Detection and Churn Prediction in Online Games. IEEE Transactions on Games. 2022.
111. Fulton LB, Lee JY, Wang Q, Yuan Z, Hammer J, Perer A. Getting playful with explainable AI: games with a purpose to improve human understanding of AI. In: Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems; 2020;1–8.
112. Mellem MS, Kollada M, Tiller J, Lauritzen T. Explainable AI enables clinical trial patient selection to retrospectively improve treatment effects in schizophrenia. BMC medical informatics and decision making. 2021;21(1):1–10.
113. Korda AI, Andreou C, Rogg HV, Avram M, Ruef A, Davatzikos C, et al. Identification of texture MRI brain abnormalities on first-episode psychosis and clinical high-risk subjects using explainable artificial intelligence. Translational Psychiatry. 2022;12(1):1–12.
114. Galazzo IB, Cruciani F, Brusini L, Salih A, Radeva P, Storti SF, et al. Explainable Artificial Intelligence for Magnetic Resonance Imaging Aging Brainprints: Grounds and challenges. IEEE Signal Processing Magazine. 2022;39(2):99–116.
115. Fellous JM, Sapiro G, Rossi A, Mayberg H, Ferrante M. Explainable artificial intelligence for neuroscience: behavioral neurostimulation. Frontiers in neuroscience. 2019;13:1346.
116. Chen K, Hwu T, Kashyap HJ, Krichmar JL, Stewart K, Xing J, et al. Neurorobots as a means toward neuroethology and explainable AI. Frontiers in Neurorobotics. 2020;14.
117. Ravi M, Negi A, Comparative Chitnis S. A Review of Expert Systems, Recommender Systems, and Explainable AI. In,. IEEE 7th International conference for Convergence in Technology (I2CT). IEEE. 2022;1–8.
118. Vultureanu-Albişi A, Bădică C. Recommender systems: an explainable AI perspective. In: 2021 International Conference on INnovations in Intelligent SysTems and Applications (INISTA). IEEE; 2021. p. 1-6.
119. Tjoa E, Guan C. A survey on explainable artificial intelligence (xai): Toward medical xai. IEEE transactions on neural networks and learning systems. 2020;32(11):4793–813.
120. Loveleen G, Mohan B, Shikhar BS, Nz J, Shorfuzzaman M, Masud M. Explanation-driven HCI Model to Examine the Mini-Mental State for Alzheimer's Disease. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM). 2022.
121. Böhle M, Eitel F, Weygandt M, Ritter K. Layer-wise relevance propagation for explaining deep neural network decisions in MRI-based Alzheimer's disease classification. Frontiers in aging neuroscience. 2019;11:194.
122. El-Sappagh S, Alonso JM, Islam SMR, Sultan AM, Kwak KS. A multilayer multimodal detection and prediction model based on explainable artificial intelligence for Alzheimer's disease. Scientific Reports. 2021;11(1):2660.
123. Bloch L, Friedrich CM. Data analysis with Shapley values for automatic subject selection in Alzheimer's disease data sets

using interpretable machine learning. Alzheimer's Research & Therapy. 2021;13(1):1–30.
124. Ruengchaijatuporn N, Chatnuntawech I, Teerapittayanon S, Sriswasdi S, Itthipuripat S, Hemrungrojn S, et al. An explainable self-attention deep neural network for detecting mild cognitive impairment using multi-input digital drawing tasks. Alzheimer's Research & Therapy. 2022;14(1):1–11.
125. Bogdanovic B, Eftimov T, Simjanoska M. In-depth insights into Alzheimer's disease by using explainable machine learning approach. Scientific Reports. 2022;12(1):1–26.
126. Chun MY, Park CJ, Kim J, Jeong JH, Jang H, Kim K, et al. Prediction of conversion to dementia using interpretable machine learning in patients with amnestic mild cognitive impairment. Frontiers in Aging Neuroscience. 2022;14.
127. Kamal MS, Northcote A, Chowdhury L, Dey N, Crespo RG, Herrera-Viedma E. Alzheimer's patient analysis using image and gene expression data and explainable-AI to present associated genes. IEEE Transactions on Instrumentation and Measurement. 2021;70:1–7.
128. Shad HA, Rahman QA, Asad NB, Bakshi AZ, Mursalin SF, Reza MT, et al. Exploring Alzheimer's Disease Prediction with XAI in various Neural Network Models. In: TENCON 2021-2021 IEEE Region 10 Conference (TENCON). IEEE 2021;720–5.
129. Kim M, Kim J, Qu J, Huang H, Long Q, Sohn KA, et al. Interpretable temporal graph neural network for prognostic prediction of Alzheimer's disease using longitudinal neuroimaging data. In: 2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE; 2021. p. 1381–4.
130. Zhang X, Han L, Zhu W, Sun L, Zhang D. An explainable 3D residual self-attention deep neural network for joint atrophy localization and Alzheimer's disease diagnosis using structural MRI. IEEE journal of biomedical and health informatics. 2021.
131. Ilias L, Askounis D. Explainable identification of dementia from transcripts using transformer networks. IEEE Journal of Biomedical and Health Informatics. 2022;26(8):4153–64.
132. Pohl T, Jakab M, Benesova W. Interpretability of deep neural networks used for the diagnosis of Alzheimer's disease. International Journal of Imaging Systems and Technology. 2022;32(2):673–86.
133. Danso SO, Zeng Z, Muniz-Terrera G, Ritchie CW. Developing an explainable machine learning-based personalised dementia risk prediction model: A transfer learning approach with ensemble learning algorithms. Frontiers in big Data. 2021;4:21.
134. Rieke J, Eitel F, Weygandt M, Haynes JD, Ritter K. Visualizing convolutional networks for MRI-based diagnosis of Alzheimer's disease. In: Understanding and Interpreting Machine Learning in Medical Image Computing Applications. Springer; 2018. p. 24-31.
135. Kitchenham B, Charters S. Guidelines for performing Systematic Literature Reviews in Software Engineering. Keele University and Durham University Joint Report; 2007. EBSE 2007-001. Available from: http://www.dur.ac.uk/ebse/resources/Systematic-reviews-5-8.pdf.
136. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. Systematic reviews. 2021;10(1):1–11.
137. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE International Conference on Computer Vision; 2017;618;26.
138. Montavon G, Binder A, Lapuschkin S, Samek W, Müller KR. Layer-wise relevance propagation: an overview. Explainable AI: interpreting, explaining and visualizing deep learning. 2019:193-209.
139. Springenberg JT, Dosovitskiy A, Brox T, Riedmiller M. Striving for simplicity: The all convolutional net. arXiv preprint http://arxiv.org/abs/1412.6806. 2014.
140. Lundberg SM, Lee SI. A Unified Approach to Interpreting Model Predictions. In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, et al. editors. Advances in Neural Information Processing Systems 30. Curran Associates, Inc. 2017;4765–74. Available from: http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf.
141. Anders CJ, Neumann D, Samek W, Müller KR, Lapuschkin S. Software for Dataset-wide XAI: From Local Explanations to Global Insights with Zennit, CoRelAy, and ViRelAy. CoRR. 2021;abs/2106.13200.
142. Liu M, Luo Y, Wang L, Xie Y, Yuan H, Gui S, et al. DIG: A Turnkey Library for Diving into Graph Deep Learning Research. J Mach Learn Res. 2021;22(240):1–9. Available from: http://jmlr.org/papers/v22/21-0343.html.
143. Alqaraawi A, Schuessler M, Weiß P, Costanza E, Berthouze N. Evaluating saliency map explanations for convolutional neural networks: a user study. In: Proceedings of the 25th International Conference on Intelligent User Interfaces; 2020;275–85.
144. Zeiler MD, Fergus R. Visualizing and understanding convolutional networks. In: European conference on computer vision. Springer; 2014;818–33.
145. Goldstein A, Kapelner A, Bleich J, Pitkin E. Peeking Inside the Black Box: Visualizing Statistical Learning With Plots of Individual Conditional Expectation. Journal of Computational and Graphical Statistics. 2015.
146. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. International Journal of Computer Vision. 2019.
147. Folego G, Weiler M, Casseb RF, Pires R, Rocha A. Alzheimer's disease detection through whole-brain 3D-CNN MRI. Frontiers in bioengineering and biotechnology. 2020;8.
148. Simonyan K, Vedaldi A, Zisserman A. Deep inside convolutional networks: Visualising image classification models and saliency maps. arXiv preprint http://arxiv.org/abs/1312.6034. 2013.
149. Petsiuk V, Jain R, Manjunatha V, Morariu VI, Mehra A, Ordonez V, et al. Black-box explanation of object detectors via saliency maps. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2021;11443–52.
150. Yang C, Rangarajan A, Ranka S. Visual explanations from deep 3D convolutional neural networks for Alzheimer's disease classification. In: AMIA annual symposium proceedings. vol. 2018. American Medical Informatics Association. 2018;1571.
151. Lombardi A, Diacono D, Amoroso N, Biecek P, Monaco A, Bellantuono L, 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. 2022;9(1):1–17.
152. Xu X, Yan X. A Convenient and Reliable Multi-Class Classification Model based on Explainable Artificial Intelligence for Alzheimer's Disease. In: 2022 IEEE International Conference on Advances in Electrical Engineering and Computer Applications (AEECA). IEEE; 2022;671–5.
153. Sha C, Cuperlovic-Culf M, Hu T. SMILE: systems metabolomics using interpretable learning and evolution. BMC bioinformatics. 2021;22(1):1–17.
154. Hammond TC, Xing X, Wang C, Ma D, Nho K, Crane PK, et al. β-amyloid and tau drive early Alzheimer's disease decline while glucose hypometabolism drives late decline. Communications biology. 2020;3(1):1–13.
155. Hernandez M, Ramon-Julvez U, Ferraz F. Withthe ADNI-Consortium. Explainable AI toward understanding the

performance of the top three TADPOLE Challenge methods in the forecast of Alzheimer's disease diagnosis. PloS one. 2022;17(5):e0264695.
156. Lai Y, Lin X, Lin C, Lin X, Chen Z, Zhang L. Identification of endoplasmic reticulum stress-associated genes and subtypes for prediction of Alzheimer's disease based on interpretable machine learning. Frontiers in Pharmacology. 2022;13.
157. Sidulova M, Nehme N, Towards Park CH. Analysis Explainable Image, for Alzheimer's Disease and Mild Cognitive Impairment Diagnosis. In,. IEEE Applied Imagery Pattern Recognition Workshop (AIPR). IEEE. 2021;2021:1–6.
158. Yu L, Xiang W, Fang J, Chen YPP, Zhu R. A novel explainable neural network for Alzheimer's disease diagnosis. Pattern Recognition. 2022;131.
159. Salih A, Galazzo IB, Cruciani F, Brusini L, Radeva P. Investigating Explainable Artificial Intelligence for MRI-based Classification of Dementia: a New Stability Criterion for Explainable Methods. In: 2022 IEEE International Conference on Image Processing (ICIP). IEEE; 2022;4003–7.
160. Jain V, Nankar O, Jerrish DJ, Gite S, Patil S, Kotecha K. A novel AI-based system for detection and severity prediction of dementia using MRI. IEEE Access. 2021;9:154324–46.
161. Bloch L, Friedrich CM. Machine Learning Workflow to Explain Black-box Models for Early Alzheimer's Disease Classification Evaluated for Multiple Datasets. arXiv preprint http://arxiv.org/abs/2205.05907. 2022.
162. García-Gutierrez F, Díaz-Álvarez J, Matias-Guiu JA, Pytel V, Matías-Guiu J, Cabrera-Martín MN, et al. GA-MADRID: Design and validation of a machine learning tool for the diagnosis of Alzheimer's disease and frontotemporal dementia using genetic algorithms. Medical & Biological Engineering & Computing. 2022;60(9):2737–56.
163. Khodabandehloo E, Riboni D, Alimohammadi A. HealthXAI: Collaborative and explainable AI for supporting early diagnosis of cognitive decline. Future Generation Computer Systems. 2021;116:168–89.
164. Sudar KM, Nagaraj P, Nithisaa S, Aishwarya R, Aakash M, Lakshmi SI. Alzheimer's Disease Analysis using Explainable Artificial Intelligence (XAI). In: 2022 International Conference on Sustainable Computing and Data Communication Systems (ICSCDS). IEEE; 2022;419–23.
165. Bordin V, Coluzzi D, Rivolta MW, Baselli G. Explainable AI Points to White Matter Hyperintensities for Alzheimer's Disease Identification: a Preliminary Study. In. 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC). IEEE. 2022;2022:484–7.
166. Selvaraju R, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: visual explanations from deep networks via gradient-based localization. 2016. arXiv preprint http://arxiv.org/abs/1610.02391. 2016.
167. Selvaraju RR, Das A, Vedantam R, Cogswell M, Parikh D, Batra D. Grad-CAM: Why did you say that? arXiv preprint http://arxiv.org/abs/1611.07450. 2016.
168. Ying Z, Bourgeois D, You J, Zitnik M, Leskovec J. Gnnexplainer: Generating explanations for graph neural networks. Advances in neural information processing systems. 2019;32.
169. Mundhenk TN, Chen BY, Friedland G. Efficient saliency maps for explainable AI. arXiv preprint http://arxiv.org/abs/1911.11293. 2019.
170. Wang D, Honnorat N, Fox PT, Ritter K, Eickhoff SB, Seshadri S, et al. Deep neural network heatmaps capture Alzheimer's disease patterns reported in a large meta-analysis of neuroimaging studies. Neuroimage. 2023;269.
171. Mulyadi AW, Jung W, Oh K, Yoon JS, Lee KH, Suk HI. Estimating explainable Alzheimer's disease likelihood map via clinically-guided prototype learning. NeuroImage. 2023;273.
172. Oh K, Yoon JS, Suk HI. Learn-explain-reinforce: counterfactual reasoning and its guidance to reinforce an Alzheimer's Disease diagnosis model. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2023;45(4):4843–57.
173. Rahim N, El-Sappagh S, Ali S, Muhammad K, Del Ser J, Abuhmed T. Prediction of Alzheimer's progression based on multimodal Deep-Learning-based fusion and visual Explainability of time-series data. Information Fusion. 2023;92:363–88.
174. Shojaei S, Abadeh MS, Momeni Z. An evolutionary explainable deep learning approach for Alzheimer's MRI classification. Expert Systems with Applications. 2023;220.
175. Kou Y, Gui X. Mediating community-AI interaction through situated explanation: the case of AI-Led moderation. Proceedings of the ACM on Human-Computer Interaction. 2020;4(CSCW2):1–27.
176. Slijepcevic D, Horst F, Lapuschkin S, Horsak B, Raberger AM, Kranzl A, et al. Explaining machine learning models for clinical gait analysis. ACM Transactions on Computing for Healthcare. 2021;3(2):1–27.
177. Arrotta L, Civitarese G, Bettini C. DeXAR: Deep Explainable Sensor-Based Activity Recognition in Smart-Home Environments. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies. 2022;6(1):1–30.
178. Winterburn JL, Voineskos AN, Devenyi GA, Plitman E, de la Fuente-Sandoval C, Bhagwat N, et al. Can we accurately classify schizophrenia patients from healthy controls using magnetic resonance imaging and machine learning? A multimethod and multi-dataset study. Schizophrenia Research. 2019;214:3–10.
179. Bhandari M, Shahi TB, Siku B, Neupane A. Explanatory classification of CXR images into COVID-19, Pneumonia and Tuberculosis using deep learning and XAI. Computers in Biology and Medicine. 2022;150.