applied sciences

MDPI

Review

Explainable AI for Alzheimer Detection: A Review of Current Methods and Applications

Fatima Hasan Saif [1], Mohamed Nasser Al-Andoli [2,*] and Wan Mohd Yaakob Wan Bejuri [1]

[1] Fakulti Teknologi Maklumat dan Komunikasi, Universiti Teknikal Malaysia Melaka, Melaka 76100, Malaysia; forignn@gmail.com (F.H.S.); yaakob@utem.edu.my (W.M.Y.W.B.)
[2] Faculty of Computing & Informatics, Multimedia University, Cyberjaya 63100, Malaysia
* Correspondence: nasser.alandoli@mmu.edu.my

Abstract: Alzheimer's disease (AD) is the most common cause of dementia, marked by cognitive decline and memory loss. Recently, machine learning and deep learning techniques have introduced promising solutions for improving AD detection through MRI, especially in settings where specialists may not be readily available. These techniques offer the potential to assist general practitioners and non-specialists in busy clinical environments. However, the 'black box' nature of many AI techniques makes it challenging for non-expert physicians to fully trust their diagnostic accuracy. In this review, we critically evaluate current explainable AI (XAI) methods applied to AD detection and highlight their limitations. In addition, a new interpretability framework, called "Feature-Augmented", was theoretically designed to improve model interpretability. This approach remains underexplored, primarily due to the scarcity of explainable AD-specific datasets. Furthermore, we underscore the importance of AI models being accurate and explainable, which enhance diagnostic confidence and patient care outcomes.

Keywords: Alzheimer's disease; artificial intelligence; machine learning; deep learning; explainable AI; convolutional neural networks; MRI; clinical decision

check for updates

Citation: Hasan Saif, F.; Al-Andoli, M.N.; Bejuri, W.M.Y.W. Explainable AI for Alzheimer Detection: A Review of Current Methods and Applications. *Appl. Sci.* **2024**, *14*, 10121. https://doi.org/10.3390/app142210121

Academic Editor: Qi-Huang Zheng

Received: 6 October 2024
Revised: 24 October 2024
Accepted: 2 November 2024
Published: 5 November 2024

Correction Statement: This article has been republished with a minor change. The change does not affect the scientific content of the article and further details are available within the backmatter of the website version of this article.

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Alzheimer's disease (AD) is one of the leading causes of dementia, which is characterized by a gradual decline in cognitive functions, including memory, reasoning, and communication. It is more prevalent among older people, and its impact is profound, not only on those affected but also on their families and caregivers [1,2]. AD is a major public health challenge, and the number of cases is expected to rise significantly as the world's population ages [3]. One of the oldest and most recognizable signs of AD is hippocampal atrophy [4], Which involves shrinkage of the hippocampus, an area of the brain important for memory formation. As the disease progresses, AD is often accompanied by cortical thinning and white matter disturbances, indicating more severe AD stages. These structural brain changes highlight the importance of early detection in managing and mitigating the effects of the disease [5]. Early detection of AD is critical because it allows timely intervention, which can slow the progression of symptoms, improve quality of life, and provide patients and their families more time to plan for the future [6]. Furthermore, early diagnosis enables therapeutic strategies to be applied when most effective, which may delay the appearance of more disabling symptoms [7,8]. Figure 1 shows MRI images of a normal brain, a brain with mild cognitive impairment (MCI), and a brain affected by AD [9]. The normal brain on the left appears structurally intact, with well-preserved gray matter (GM) volume. In contrast, the Alzheimer's brain on the right demonstrates significantly reduced GM volume, particularly in regions critical for memory and cognition, with certain areas becoming enlarged due to atrophy and brain tissue loss. In the case of MCI, an intermediate decline in GM volume can be observed between the two.

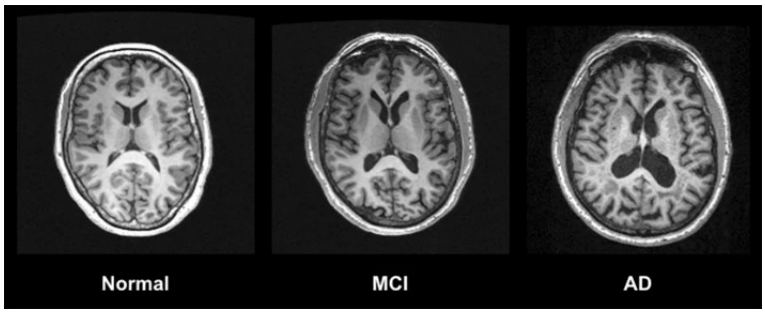


Figure 1. T1-weighted MRI image sequences compare a normal brain and a brain affected by AD [9].

1.1. Artificial Intelligence in Alzheimer's Disease Detection

Recent advances in artificial intelligence (AI), especially in machine learning (ML) and deep learning (DL) in [10–14], have introduced promising approaches to enhance the accuracy and efficiency of AD detection.

AI-based tools, especially those using convolutional neural networks (CNN), like [14–16] have demonstrated exceptional performance in image classification tasks, - such as MRI [17], or CT [18], making them particularly relevant for medical imaging applications. These AI models can detect subtle patterns in brain scans that human experts might ignore [19], thus improving diagnostic accuracy and providing valuable insights into disease progression [20,21]. However, despite technological advances, integrating AI into clinical practice remains a challenge. The "black box" nature of many AI models poses a major barrier, limiting their interpretability and reducing trust among healthcare providers. In clinical settings, where rapid and reliable decision-making is essential, the ability to understand and trust AI-driven recommendations is critical. The explainability of AI models is not just a technical concern; Rather, it is also a prerequisite for its acceptance and adoption in clinical settings [22,23].

1.2. The Need for XAI in Clinical Settings

AI tools are meant to assist both general physicians and non-specialists in diagnosing in the absence of specialists. AI might augment the ability of non-expert doctors to make accurate diagnoses, especially in crowded clinical settings. The need for explainable artificial intelligence (XAI) in detecting AD is crucial. Explainability refers to the ability of AI models to provide transparent and understandable reasons for their decisions. Without this, even the most accurate AI models as [10,13–16,24,25] may be met with skepticism and hesitation from general physicians and non-specialists. Current efforts in interpretable AI, including technologies such as Shapley additive explanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME), in [14,15,25–30] are steps in the right direction. However, these methods often fail to provide the kind of intuitive understanding necessary for clinical application. These models may provide insights into which features made the AI make the decisions, but these insights cannot always be easily interpreted by human users, especially in a high-stakes medical context, and presented without explaining the outputs making AI tools inapplicable in clinical decision-making. The lack of large datasets and truly interpretable AI data has hampered the implementation of such contributions in clinical settings [31,32]. It can be difficult to trust traditional AI in critical applications. If a model makes an error, it is often difficult to diagnose why, leading to potential risks and a lack of accountability.

1.3. Advancing XAI for Enhanced Clinical Decision-Making

This review aims to explore studies that focus on developing XAI tools that excel in detecting AD and provide clear, understandable, and actionable explanations to general physicians and non-specialists. This review focuses on examining the existence of XAI systems that can be trusted and utilized effectively in clinical settings by prioritizing interpretability alongside accuracy.

The DSM-5-TR criteria for AD focus on memory loss, learning difficulties, and problems in at least one other cognitive domain, all of which interfere with daily life, making memory impairment the primary manifestation of AD [33]. The studies referenced in this review consider the different stages of AD, as defined in the DSM-5 and other relevant clinical guidelines. In this review, references discussing AI models distinguish between these stages, AD and MCI are distinguished such as in the article [34,35], and the distinction between three categories: AD, MCI, and HC as [36] to ensure that the performance and interpretability of AI tools are evaluated across the spectrum of disease progression. Moreover, in XAI models, explanations are made based on two types: the first is model-agnostic—such as SHAP and LIME—as in articles in[14,15,25–30], and the second is model-specific—such as decision trees (DT) as in articles [16,24,25,37]. Table 1 lists the abbreviations and their corresponding full forms used in this study.

Table 1. List of shortcuts.

AD	Alzheimer's Disease
AI	Artificial Intelligence
XAI	Explainable Artificial Intelligence
SVM	Support Vector Machine
RF	Random Forest
CNN	Convolutional Neural Network
ML	Machine Learning
DL	Deep Learning
SHAP	SHapley Additive exPlanations
LIME	Local Interpretable Model-Agnostic Explanations
Grad-CAM	Gradient-weighted Class Activation Mapping
LAVA	Granular Neuron-level Explainer
PDA	Pixel Density Analysis
PGM	Probabilistic Graphical Model
MRI	Magnetic Resonance Imaging
CT	Cortical Thickness
DT	Decision Tree
XGB	Extreme Gradient Boosting
ADNI	Alzheimer's Disease Neuroimaging Initiative
OASIS	Open Access Series of Imaging Studies
MCI	Mild Cognitive Impairment
DPE	Deep Parallel Ensembles
KNN	K-Nearest Neighbors
HC	Healthy Control
DSM-5	Diagnostic and Statistical Manual of Mental Disorders, 5th Edition

2. Techniques in Alzheimer's Disease Detection

AI includes techniques that mimic human functions such as thinking, learning, planning, and forecasting. This area includes ML, computer vision, and natural language processing [38]. AI plays a crucial role in the development of AD diagnosis, as the field of diagnosis is considered one of the most prominent applications of AI in medicine. AI algorithms can analyze complex patterns in MRI scans with accuracy beyond human capabilities. These models can detect subtle changes in the hippocampus, which have been shown to play an important role in detecting AD in animal models [39]. AI can also analyze large sets of patient information, including clinical data, lifestyle factors, and treatment outcomes, and can identify patterns and predict the transition from MCI to AD [5,40]. AI can also provide relatively high performance in detecting and monitoring AD using natural language processing and speech technology [41]. The effectiveness of AI (DL models) in recognizing subtle facial signals associated with AD has been demonstrated in [16]. The article [30] analyzed fundus photographs for the diagnosis of AD using neural network-based methods by developing a framework that combines neuroimaging in CNN models and hierarchical clustering techniques. Based on AI, a patient management framework based on wearable sensing technology and cloud systems was also proposed in [25]. However,

no real translation into clinical practice was achieved in most of the papers reviewed [41] which highlighted the potential of AI and ML methods in detecting AD.

2.1. Conventional Machine Learning

Conventional machine learning methods have been widely used in different fields for classification, regression, and other prediction tasks. These methods, although not as sophisticated as modern DL models, have proven highly effective in many practical applications due to their power and relatively low computational requirements. In the context of medical diagnosis and other critical fields, these models provide valuable insights and are often easier to implement and interpret than more complex models. Figure 2 compares Conventional Machine Learning and DL workflows. In conventional machine learning, data preprocessing involves handcrafted feature extraction followed by feature selection or reduction. The processed data is then fed into a classifier, which generates outputs. In contrast, deep learning automatically processes data through an input layer, which is passed through multiple hidden layers to learn features, with the final output layer producing results. Unlike conventional methods, deep learning eliminates the need for manual feature extraction by learning patterns directly from the data.

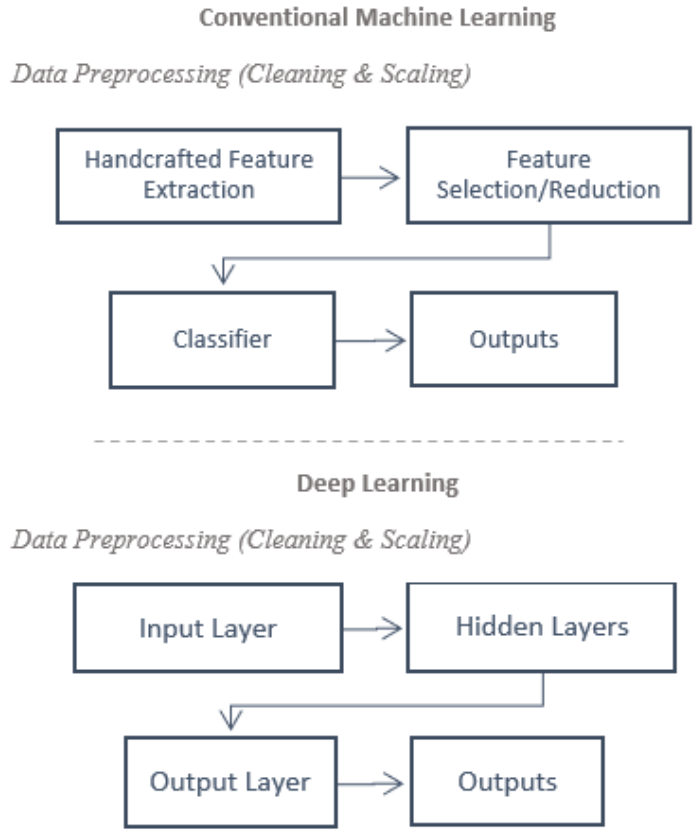


Figure 2. Conventional Machine Learning vs. DL. Conventional machine learning requires manual feature extraction before making predictions, while DL automatically extracts features from raw data.

2.1.1. Support Vector Machine (SVM)

Support vector machines (SVMs) are a common choice for binary classification tasks, including AD classification from MRI images. SVM works by finding the hyperplane that best separates two classes of data in a high-dimensional space [5]. The equation of an SVM can be represented as [42]:

$$F(x) = w^t x + b \quad (1)$$

where w is the weight vector, x is the input feature vector, and b is the bias term. The goal is to find the optimal values for w and b such that the distance between the hyperplane and the closest data points (support vectors) is maximized. SVM is particularly effective when the feature space is clear and well-defined, making it useful for early studies in detecting AD. An SVM classifier has been used to diagnose early stages of dementia in [34]. The power spectral density (PSD) and temporal data were combined for feature extraction. The work proposed a new method to differentiate between three categories—AD, MCI, and HC based on electroencephalographic (EEG) signals. The classification accuracy achieved after data augmentation using the VAE model increased by 3% compared to before augmentation. In [36] six basic models were used in addition to the SVM. The proposed models achieved the highest performance with selected features against regular ML classifiers and stacking models using full modality. In the study [43] the potential of RTOP/RTAP/ was evaluated and SVMs based on different discriminant metrics for early detection of AD. An ML model was trained to classify patients with AD and healthy controls based on their genetic data in article [44], the best-performing model was SVM, achieving 89% accuracy.

2.1.2. Random Forest (RF)

Random forests (RF) were used in the article [45] along with 5 basic classification models. In [36] RF was used among the six basic models. It is an ensemble learning method that combines multiple DTs to improve classification accuracy and robustness. The equation of an RF can be represented [46] as:

$$I_G(n) = 1 - \sum_{i=1}^{J} (p_i)^2 \quad (2)$$

where the node n is equal to 1 minus the total of the fraction of cases in each class pi squared over all classes J. The values of the Gini impurity index for the two split nodes are lower than the value for the parent node for a particular node split. The equivalent Gini significance value for each variable can be obtained by adding the Gini impurity decreases for three variables in a dataset across all trees in an RF model. This value can then be utilized for feature selection.

Each tree in the forest is trained on a random subset of the data, and a final prediction is made by averaging the predictions of all the trees. Although RF is more robust than individual DT, it is still susceptible to overfitting, especially with high-dimensional data. The paper [47] proposes an ML model as a first-level decision model using RF, BalancedRF, Bagging, and Extra Tree with smoothed clustering decisions followed by voting classifications. It explains the results using SHAP and LIME interpretations. Step-by-step data processing involves balancing the data using SMOTEENN, then transforming the data using Quantile Transformer, PCA dimensionality reduction technique for six features, and a Meta ML model to obtain 97.6% performance, 95.8% accuracy, 97.9% recall, and 96.8% F1 score. RF and DT algorithms [24] were applied to brain MRI images acquired from normal controls (NC) and AD subjects. The KNIME analytics platform was used to calculate the DT, and the R project was used for the RF.

Table 2 provides a detailed comparison of the different methods used in studies of AD detection. Various models, such as SVM, CNN, and random forest, have been applied to datasets such as ADNI [48] and OASIS [49] using MRI, DTI, and fundus images. Model-agnostic methods like SHAP and LIME provide general explanations, while model-specific methods like Grad-CAM and Score-CAM provide detailed insights. The table highlights that combining different models and explanation methods leads to different accuracies, with some studies achieving high accuracies (more than 90%), depending on the approach used.

Table 2. Detailed Comparison of Previous Methods.

Author	Dataset	Data Type	Models Used	Explanation (Model-Agnostic)	Explanation (Model-Specific)	Highest Accuracy
[16]	Participants	Facial images	Xception, SENet50, ResNet50, VGG16, simple CNN	-	Keras, TensorFlow	92.56%
[30]	UK Biobank	Fundus image	CNN	LAVA	-	71.4%
[45]	ADNI	MRI	SVM, RF, ETC, XGB, and MLP	-	-	86.57%
[13]	ADNI	MRI	VGG16	non	non	98.17%
[25]	OASIS	Clinical data, MRI	RF, LR, DT, MLP, KNN, GB, AdaB, SVM, and NB	SHAP	DT	98.81%
[50]	Participants	DTI	SVM, logistic regression, CNN, and XGB	non	non	82.35%
[24]	ADNI	MRI	DT, CNNs, and RF	-	DT	91%
[26]	ADNI	sMRI	ResNet-based 3D, CNN	Score-CAM	-	89.02%
[14]	Kaggle, OASIS	MRI	KNN, SVM, and CNN	SHAP	-	99.9%
[15]	Kaggle	MRI	CNN	Grad-CAM	-	93.82%
[28]	Kaggle	MRI	Resnet50, VGG16 and Inception v3	LIME	-	86.82%
[29]	Kaggle, UNBC	MRI	DenseNet, GoogLeNet, ResNet18, EfficientNet, and RegNet	Grad-CAM	-	88.4%
[27]	OASIS-3	MRI	CNN	SHAP	-	89%
[10]	ADNI	fMRI	CNN, DT and a KNN	non	DT	98%
[37]	OASIS	MRI	DT, RF, and AdaBoost	-	DT	86.84%

2.2. Deep Learning Models

Deep learning (DL) is the process of training a computer to apply its experience to solve a specific problem given to it [22]. Current models in DL often treat large networks as a single object, which requires huge trainable parameters. To understand the structure of complex networks, the paper [51] proposed a method for improving DL that achieves greater efficiency as the level of segmentation deepens, it was proven that the method greatly improves the training speed and efficiency. DL models have demonstrated superior performance in many medical image analysis tasks, including AD diagnosis, due to their ability to capture nonlinear relationships and hierarchical features. In the diagnosis of AD, CT-based volumetric measurements correlate closely with MRI-based measurements, showing a comparable prognosis suggesting the possibility of using CT as a primary screening tool for the diagnosis of neurodegenerative diseases after further validation [52]. However, CT has rarely been used for tissue classification because the contrast between gray

matter and white matter has been considered insufficient [53]. The results in [54] indicate that improved DL methods significantly outperform other ML techniques in terms of effectiveness, efficiency, and scalability. In [55] each layer of the neural network is decoded, revealing the decision-making processes using DL and then extracting results for each layer of the model. Different transfer learning models were created, namely VGG-19 with 58% training accuracy, MobileNet V2, Inception V3 with 55% training accuracy, ResNet-50 with 45% training accuracy, and a custom model developed by DL algorithms with 62% training accuracy. Each model produced accurately distinct outputs, with MobileNet V2 outperforming the others, with a training accuracy of 67% and a test accuracy of 60%. To provide a deeper understanding of the internal workings of the model, principal component analysis and Grade CAM techniques were used. A novel framework is proposed in [56] to detect features of AD taking advantage of ViT's ability to capture complex features and GRU's effectiveness in modeling temporal dependencies, which is crucial considering AD develops over time.

2.2.1. Convolutional Neural Network (CNN)

Studies by [14,16,21,22,25,30] have shown that CNNs can outperform traditional ML methods in classifying AD stages by extracting high-level abstract features from MRI data. CNNs consist of multiple layers, such as convolutional layers, pooling layers, and fully connected layers. The convolution operation in CNNs can be represented as:

$$f_{i,j} = \sum_{m} \sum_{n} x_{i+m,j+n} \cdot w_{m,n} + b \tag{3}$$

where X is the input feature map, W is the filter (or kernel), b is the bias term, and is the output feature map at position (i,j). This operation allows CNNs to learn spatial hierarchies of features in the input data. CNN architectures such as VGGNet and ResNet have shown impressive results in distinguishing between healthy brains, MCI, and brains with AD as in [35]. The use of transfer learning, where pre-trained CNN models are fine-tuned on AD datasets, has improved diagnostic accuracy and reduced the need for large labeled datasets. Using brain MRI scans, [57] used two CNN models, MobileNetV3 and DenseNet121, to detect AD. MobileNetV3 achieved 93% accuracy, while DenseNet121 achieved 88% accuracy. The GAN-augmented dataset in article [58] achieved an accuracy of 81% using a conventional CNN model.

2.2.2. Attention Networks

Attention mechanisms have emerged as a powerful augmentation of traditional neural network architectures, enabling models to selectively focus on relevant parts of the input data [59]. Attention mechanisms in the context of AD detection can significantly improve the performance of CNNs by allowing the model to prioritize regions of MRI images that are most indicative of neurodegenerative changes. A dual-attention multi-instance DL (DA-MIDL) model is proposed for discriminative pathological localization and diagnosis of AD in [60]. Research by [19] demonstrates the effectiveness of attention mechanisms in various applications, including image classification and natural language processing. In the detection of AD, attention-based models have been shown to improve feature representation, leading to improved classification accuracy. For example, the integration of spatial and channel attention mechanisms can help focus on important areas of the brain, thereby improving the diagnostic process. Adaptive Hybrid Attention Network (AHANet) is proposed in the article [61]. The hybrid network operates on two attention modules, namely enhanced non-local attention (ENLA) and focal attention. The Adaptive Features Fusion (AFA) module is also proposed to fuse features from both global and local levels. The proposed network demonstrated better performance compared to existing works. The network was trained and tested on the ADNI dataset and produced 98.53% classification accuracy. Although attention mechanisms improve explainability, they do not completely solve the problem of the "black box" nature of DL models [26].

2.2.3. Recurrent Neural Networks (RNNs)

Recurrent neural networks (RNNs) are a type of neural network designed to deal with sequential data by maintaining a hidden state that captures information about previous elements in the sequence. RNNs repeat connections, allowing information to persist, which makes them suitable for tasks where temporal dynamics are important. The equation for updating the hidden state in an RNN can be represented as [62]:

$$h_t = f(W_x x_t + W_h h_{t-1} + b) \quad (4)$$

where the hidden state at the time step is the input at the time step and are weight matrices, and b is the bias term. The function f is typically a non-linear activation function like tanh or ReLU. RNNs are particularly useful in scenarios where sequences of inputs can provide additional context, such as analyzing changes in brain scans over time. They can be used to analyze MRI scan sequences, helping to identify patterns or changes that indicate progression from MCI to AD. By tracking patient data over time, RNNs provide insights into the effectiveness of treatment or the progression of the disease [5].

Table 3 provides an overview of previous articles and summarizes the models used and their limitations and advantages. The primary limitation in all models is "inadequate explanation", suggesting a need to improve the interpretability of AI models for AD detection. Despite this, each model offers distinct advantages, and each method addresses specific challenges but also highlights the ongoing need for more interpretable AI models.

Table 3. Overview of Previous Articles.

References	Model Used	Limitations	Advantages
[10,14–16,26,27,30,50]	CNN	Insufficient Explanation	Achieved high accuracy in extracting key features from MRI.
[14,25,45,50]	SVM	Insufficient Explanation	Demonstrated strong performance in handling high-dimensional data with smaller sample sizes.
[5]	RNNs	Insufficient Explanation	Effectively captured temporal dependencies, improving prediction reliability.
[16,28]	ResNet50	Insufficient Explanation	Allowed deeper architecture while avoiding vanishing gradient issues.
[16]	Xception	Insufficient Explanation	Efficiently captured complex spatial features from MRI data.
[16]	SENet50	Insufficient Explanation	Enhanced the detection accuracy through channel-wise attention mechanisms in the articles.
[24,25,37,45]	RF	Insufficient Explanation	Robust against overfitting.
[10,24,25,37]	DT	Insufficient Explanation	Capable of capturing non-linear relationships between features relevant to Alzheimer's progression.
[10,14,25]	KNN	Insufficient Explanation	Adaptability in both classification and regression tasks.
[45,50,63]	XGB	Insufficient Explanation	Showed superior predictive performance through boosting.
[13,16,28]	VGG16	Insufficient Explanation	Provided consistent feature extraction for hierarchical image structures in MRI analysis.
[25,37]	AdaBoost	Insufficient Explanation	Improved prediction accuracy by combining multiple classifiers.

2.3. Ensembles Learning in AD

Combining multiple models—known as ensemble learning—often improves speed and performance. Multiple model predictions can be combined either through averaging or through more complex methods such as stacking or boosting. Ensemble strategies have evolved to improve the generalization of learning models, including techniques like bagging, boosting, stacking, and negative correlation learning (NCL) [64], while RF employs bagging to prevent overfitting by picking random feature subsets, bagging improves per-

formance by producing multiple predictors from independent samples. By concentrating on incorrectly identified data, boosting techniques like AdaBoost and Gradient Boosting transform weak learners into strong ones. With the help of a meta-learning strategy, stacking combines several models to get better predictions. By reducing empirical risk, NCL encourages variation across ensemble models, while explicit/implicit ensembles seek to emulate ensemble behavior without appreciably raising computational costs. As described in the article [45] DPE can integrate different neural network architectures, such as CNNs and attention-based models, to take advantage of their complementary strengths. The study showed that the ensemble approach can achieve superior diagnostic accuracy compared to individual models, as it averages out the biases and variances of individual predictors. The study [65] presents a multi-task learning algorithm to predict AD progression using a similarity measure-based multi-task learning (MTL) approach to the spatiotemporal variation of brain biomarkers to model AD progression.

2.4. XAI Methods Applied to Alzheimer's Detection

XAI techniques, which meet the need for transparency and interpretability in AI models used in clinical settings, have become essential tools in the identification of AD. The increasing complexity and capability of AI models, especially DL models, has caused academics and general physicians to become concerned about these "black box" algorithms since they need to know exactly how these models make predictions. Methods such as Pixel Density Analysis (PDA) and Probabilistic Graphical Models (PGM) have been used to enhance interpretability [66]. Additionally, model-agnostic techniques like SHAP and LIME, which explain model outputs in terms of input features, are also widely used to ensure that these models provide clear and clinically relevant explanations.

2.4.1. SHAP (Shapley Additive exPlanations)

A popular XAI technique called SHAP, which is not dependent on any specific model, explains individual predictions by allocating their cause to the contribution of each feature. SHAP has been utilized in AD detection to pinpoint particular characteristics that are most important in predicting the beginning or course of the disease on MRI images as in [14,25,27,67]. SHAP is a popular XAI technique that provides model-agnostic explanations by attributing individual predictions to the contributions of different features. While SHAP helps in interpreting DL models for AD detection, offering some insights into which features influence predictions, it still functions as a black box in many cases as in [5,43]. For example, applying SHAP to a model trained on brain MRI data reveals feature importance scores, but it does not always provide a clear, intuitive understanding of how these features interact or contribute to the overall decision-making process [25]. According to [25] the Shapley value for a feature value is its effect on the batch, weighted and summed over all conceivable feature value combinations:

$$\phi_j(\text{val}) = \sum_{\{S \subseteq \{1,\ldots,p\} \setminus \{j\}\}} \frac{|S|!(p-|S|-1)!}{p!} (\text{val}(S \cup \{j\}) - \text{val}(S)) \quad (5)$$

where S is a subset of features utilized in the model, p is the number of features, and x is the vector of feature values for the example that requires explanation. Val(S) is the predicted value of the features in the set S that are prioritized over features that are not in the set S:

$$\text{val}_x(S) = \int \hat{f}(x_1, \ldots, x_p)\, \mathrm{dP}_{x_S = S} - \mathbb{E}_X(\hat{f}(X)) \quad (6)$$

The sole attribution method that meets the requirements of efficiency, symmetry, illusory, and summation—all of which are useful in determining equitable compensation—is Shapley's value. Consequently, the contribution of a variable (or several variables) to the difference between the value predicted by the model and the average of all individual projections is its Shapley value for a particular individual.

2.4.2. Local Interpretable Model-Agnostic Explanations (LIME)

Another model-agnostic AI technique is called LIME, which generates individual predictions by locally approximating the AI model around the particular instance that is being forecasted. LIME can be used to explain why a brain MRI scan was determined to be suggestive of AD in the context of the disease's identification. LIME determines which features had the greatest influence on the choice as in [43,57] by varying the input data and examining changes in the model's output. Nevertheless, despite these difficulties, LIME is still a useful technique for ensuring interpretability in intricate AI models. According to [31], this equation is central to the LIME methodology and is used to find the best interpretable model:

$$\epsilon(x) = \arg\min_{g \in G}[L(f, g, \pi x) + \Omega(g)] \quad (7)$$

where and should be minimized to provide local integrity and interpretability.

2.4.3. Granular Neuron-Level Explainer (LAVA)

LAVA is an advanced AI technique that focuses on neuron-level explanations as demonstrated in the study by [30], LAVA was applied in AD detection using retinal imaging. LAVA delves into the individual activations of neurons, using attention mechanisms to identify neurons that contribute most to a prediction. This layer-level analysis of variance allows for more granular explanation capabilities, providing insights into how deep neural networks work internally.

This method is particularly useful for complex tasks where understanding the features that lead to a particular prediction is essential for clinical confidence. LAVA has the advantage of providing explanations from fundus images that can be traced back to specific layers and neurons in the network, making it easier to interpret and validate model decisions from a clinical perspective. This level of detail in explainability ensures that clinicians can trust AI systems to help diagnose or monitor disease progression. A notable application of LAVA was demonstrated in the study by [30], which used the technique to analyze fundus images to assess AD. The neuron-level explainability provided by LAVA ensures that model predictions can be linked to specific neural activity in a deep learning network, potentially enhancing the reliability of the AI system.

2.5. Summary & Future Directions in AD Detection Techniques

The various techniques used to detect AD have distinct advantages and limitations. Traditional ML methods such as SVM and RF are easier to implement and interpret, as they provide reliable results with relatively low computational costs. However, it may face difficulties in handling high-dimensional data and lack the feature extraction capabilities of DL models. DL models, especially CNNs, have shown superior performance in classifying stages of AD due to their ability to learn complex patterns from data. Attention mechanisms enhance the ability of these models to focus on critical areas, thereby improving diagnostic accuracy. However, DL models are often criticized due to their "black box" nature, which limits their applicability in clinical decision-making without additional interpretive measures. Ensemble learning methods can improve diagnostic performance further by combining the strengths of multiple models. While this improves accuracy and robustness, ensemble approaches tend to be computationally expensive and challenging to interpret. XAI techniques such as SHAP, LIME, LAVA Score-CAM, Grad-CAM, and others, attempt to fill the interpretability gap, providing transparency into model predictions. However, they still do not completely solve the interpretability problem in DL models, and their interpretations may sometimes lack clinical nuance. While DL models have the highest diagnostic potential, the need to improve explainability remains critical.

3. Overview of Explainable Artificial Intelligence (XAI)

XAI refers to a set of technologies and methods that make the behaviors and decision-making processes of AI models more transparent and understandable to humans [32]. The

primary goal of AI is to enable users to understand how an AI model arrived at a particular decision or prediction, whether the users are developers, data scientists, or end users. This is critical in applications where trust, accountability, and insight into decision-making in healthcare, finance, legal systems, etc. are important. While traditional AI functions as a "black box", the internal processes of the model cannot be easily understood by humans, especially in complex models. Users can see inputs and outputs, but the reasoning behind the model's decisions remains hidden. In medical diagnosis, especially for conditions such as AD, XAI is essential to gain general physicians' and non-specialists' trust and facilitate the integration of AI tools into clinical practice. XAI techniques provide insights into the decision-making processes of complex models such as CNNs. However, while CNNs excel at feature extraction and classification, they often act as "black boxes", providing little visibility into the decision-making process [68]. This lack of transparency poses a significant challenge in articles [45]. where understanding the rationale behind the diagnosis is crucial, XAI aims to open the black box by providing explanations for the model's decisions, making the processes and factors involved in decision-making more transparent. XAI models are designed to provide insights into how model inputs are converted into outputs, making it easier for users to understand and trust model decisions.

3.1. Types of XAI Methods

By providing insights into the decision-making process of AI models, XAI strategies seek to improve the transparency and understandability of these models. These techniques are essential for building trust and empowering people to appropriately understand AI's predictions, especially in high-stakes industries like banking and healthcare. Model-specific and model-agnostic techniques are the two primary groups into which XAI methods are typically separated. Model-specific techniques provide in-depth explanations based on an AI model's internal operations and are linked to its design. model-agnostic approaches, on the other hand, are more adaptable and can be used with any kind of model, regardless of its architecture. Both strategies have benefits and drawbacks; model-agnostic techniques give more adaptability, while model-specific techniques offer more precision. Furthermore, XAI techniques can be post hoc, in which justifications are given after the model has been trained, or intrinsic, in which case the model is interpretable by design. This section explores the mechanics, applications, and trade-offs between interpretability, flexibility, and accuracy of the many types of XAI approaches.

3.1.1. Model-Specific Techniques in XAI

Model-specific approaches are closely related to an AI model's architecture and are made to capitalize on certain features or functions. While model-specific offers better integration with the model structure, they still focus primarily on machine-learned features, which may not align with domain-specific clinical knowledge. For example, Grad-CAM techniques are employed in CNNs to view picture regions that impact the model's judgment, offering insights into the inner workings of the network. A model described in [55] uses techniques like principal component analysis and Grad-CAM to decode a neural network's layers and expose decision-making processes. However, the main drawback of procedures that are exclusive to a model is that they are not transferable. These methods might not work with other models because they are specialized to certain architectures. Model-specific techniques are nevertheless quite effective even in XAI when explanations closely match the model's internal operations. One such example is the VGG16 model, which uses its architecture to make precise and understandable judgments when combined with transfer learning for early AD in [13]. While model-specific approaches offer accurate and tailored insights, they are not universally applicable across many models. The article [32] emphasizes how flexible in explaining different models, but it also exposes the possibility of explanations that are too general or imprecise.

3.1.2. Drawbacks of Model-Specific XAI

While model-specific XAI techniques provide comprehensive and highly relevant explanations of specific models, they can be challenging for humans to fully understand their decision-making process, as these models often operate as black boxes. This ambiguity means that while the internal processes of a model may be complex and difficult to interpret, the resulting explanations are tailored to the specific architecture and do not necessarily provide insight into the overall logic behind decisions. Additionally, complex dependencies between model parameters can obscure the underlying logic, making it difficult for users to understand why certain predictions are made. Furthermore, these strategies are limited in their applicability to models with diverse architectures because they largely depend on the structure of the model to which they are applied. For example, Grad-CAM may work best on CNNs but less effectively on recurrent neural networks (RNNs) or other models with very different operational structures. Due to their limited adaptability, these approaches can provide deep insights into a single model but are unable to generalize across different AI systems. Furthermore, because they rely on specific layers and actions within a model, developers often have to design unique solutions for different models, which can be resource-intensive. As demonstrated with the VGG16 model in [13], alternative approaches are required when working with different architectures, even if the technique yielded exact insights because of the model's distinct structure.

3.1.3. Model-Agnostic Techniques in XAI

Model-agnostic techniques, as opposed to model-specific approaches, are made to work with any AI model, regardless of its architecture. These adaptable techniques offer a means of comprehending how various inputs impact a model's output, enabling explanations to be applied across several models without change. LIME, which generates explanations by approximating the model's behavior locally, is an example of a model-agnostic technique. Another technique that is independent of models is SHAP, which rates each feature according to how much it contributes to the prediction. Because these techniques are flexible, they have been frequently used. The work in [43] demonstrated the flexibility of model-agnostic methodologies by showing how SHAP and LIME may be applied to various models to assess their dependability. But this flexibility has a price: since they don't take advantage of the model's internal structure, the explanations they offer might not be as accurate as those offered by model-specific techniques. This generality occasionally results in less precise explanations, particularly for more intricate models.

3.1.4. Challenges in Model-Agnostic XAI

Model-agnostic methods such as SHAP and LIME are useful because of their flexibility, but they also have significant drawbacks. These techniques' explanations can occasionally be imprecise because they are not customized to a model's internal design. This problem is especially noticeable when using complicated models, as broad model-agnostic approaches could overlook the subtle interactions between layers or features. For instance, the study in [57] employed LIME to show the predictions of a MobileNetV3 model; nevertheless, it highlighted a potential discrepancy between the explanation given and the actual decision-making of the model by identifying non-brain regions as relevant. Users may be misled by this lack of accuracy if irrelevant qualities are mistakenly marked as important. Moreover, when applied to more complex AI models, these methods may produce general explanations that fall short of capturing the subtleties of the model's decision-making process since they do not take advantage of particular model architecture features.

3.1.5. Model-Specific vs. Model-Agnostic Techniques

Some XAI models are designed to be inherently interpretable due to their simple structures, but even then, it can be difficult for humans to fully understand the decision-making process in results. For example, while DT and linear models are inherently transparent, users may still find it difficult to understand the logic behind the prediction. A study in [55]

attempts to uncover the decision-making processes within each layer of the neural network using principal component analysis and Grad-CAM, but the intrinsic interpretability remains limited. On the other hand, interpretability is often added via post hoc procedures after the model is trained. While [57] used LIME to explain MobileNetV3 predictions, ref. [43] employed SHAP and LIME post-training to interpret model predictions. These experiments show that although post hoc methods can be applied to any model, they sometimes highlight features that are not relevant. However, post hoc methods remain useful in improving the interpretability of complex models, and compensating for the ambiguous nature of black-box models. The article [50] highlights intrinsic interpretability, especially in models where decisions are linked to biologically significant features. For example, white matter features could serve as biomarkers of AD, linking model decisions to biologically meaningful factors.

3.2. Challenges in Implementing XAI

Implementing XAI in the context of AD detection poses several challenges, as demonstrated in the articles reviewed. The availability of large, high-quality datasets is a major challenge, as the study in [50] pointed out the limited data available to validate ML models, especially in using DTI data and white matter characteristics as biomarkers. The article in [13] also pointed out the need to further verify the high accuracy achieved by the VGG16 model using external datasets, emphasizing that robust XAI requires large-scale data to ensure the reliability and generalizability of interpretations. Balancing the precision of complex models with the need for interpretability is a recurring theme. The article in [31] concluded that an XAI framework should provide clear and understandable explanations of the model's predictions to guide general physicians and non-specialists in making informed decisions. However, achieving this balance is difficult, as evidenced by the results in [57], where LIME highlighted non-cerebral regions as influential, raising concerns about the reliability of subsequent interpretations in very precise but complex models such as MobileNetV3. Translating XAI from research into clinical practice is another key challenge. While studies such as [55] propose cutting-edge methods for decoding neural networks, the practical application of these methods in real-world clinical settings remains untested. Likewise, the contribution of providing truly XAI tools, as discussed in [32], has not yet been implemented due to the lack of large datasets and real XAI data. This gap highlights the need for further research and development to make XAI tools practical and useful for clinical decision-making.

4. Role of XAI in Alzheimer's Disease Detection

Because clinical adoption of XAI is dependent on the transparency and understandability of AI-driven judgments, XAI is essential to the diagnosis of AD [6,32]. Because AD is characterized by a steady deterioration in cognitive function, early identification is essential to successful treatments. Through medical imaging, AI models have shown potential in correctly diagnosing AD [12,31]. However, general physicians are hesitant to rely only on these instruments due to the "black box" aspect of many AI models. To overcome this difficulty, XAI offers precise explanations of how AI models make their predictions, such as pinpointing the precise parts of the brain that the illness affects [5,55]. This openness boosts clinician confidence in addition to facilitating early diagnosis by making small changes in brain structure more intelligible [13,57]. Furthermore, XAI encourages more in-depth conversations about diagnosis and available treatments between patients and healthcare professionals [45,68]. XAI facilitates clinical decision-making by allowing the validation of AI-driven insights against additional clinical data, resulting in more dependable and knowledgeable outcomes [7,43]. Moreover, XAI plays a crucial role in creating and enhancing AI models by offering perceptions that boost the model's resilience and functionality [20,50]. Ultimately, XAI improves the use of AI in AD detection, helping patients and promoting medical research [38,41].

5. Critical Review of Existing Methods

AI systems are divided into two primary categories: traditional AI and XAI as shown in Figure 3. The decision-making capabilities of systems fall into two categories: based on the use of unique features acquired during learning and those based on features that are input into the model for it to learn. Both approaches yield explainable results through XAI frameworks (such as SHAP, LAVA, and Grad-CAM). However, they lack an understandable explanation when displaying results, making them less suitable for clinical decision-making despite achieving high levels of accuracy. SHAP or LIME provides local explanations that measure the impact of each feature on a given prediction, but these features are often abstract or machine-generated, making it difficult for non-specialists to interpret clinical relevance. Table 4 presents the different XAI techniques used in the detection of AD across different types of data, distinguishing between model-agnostic and model-specific techniques.

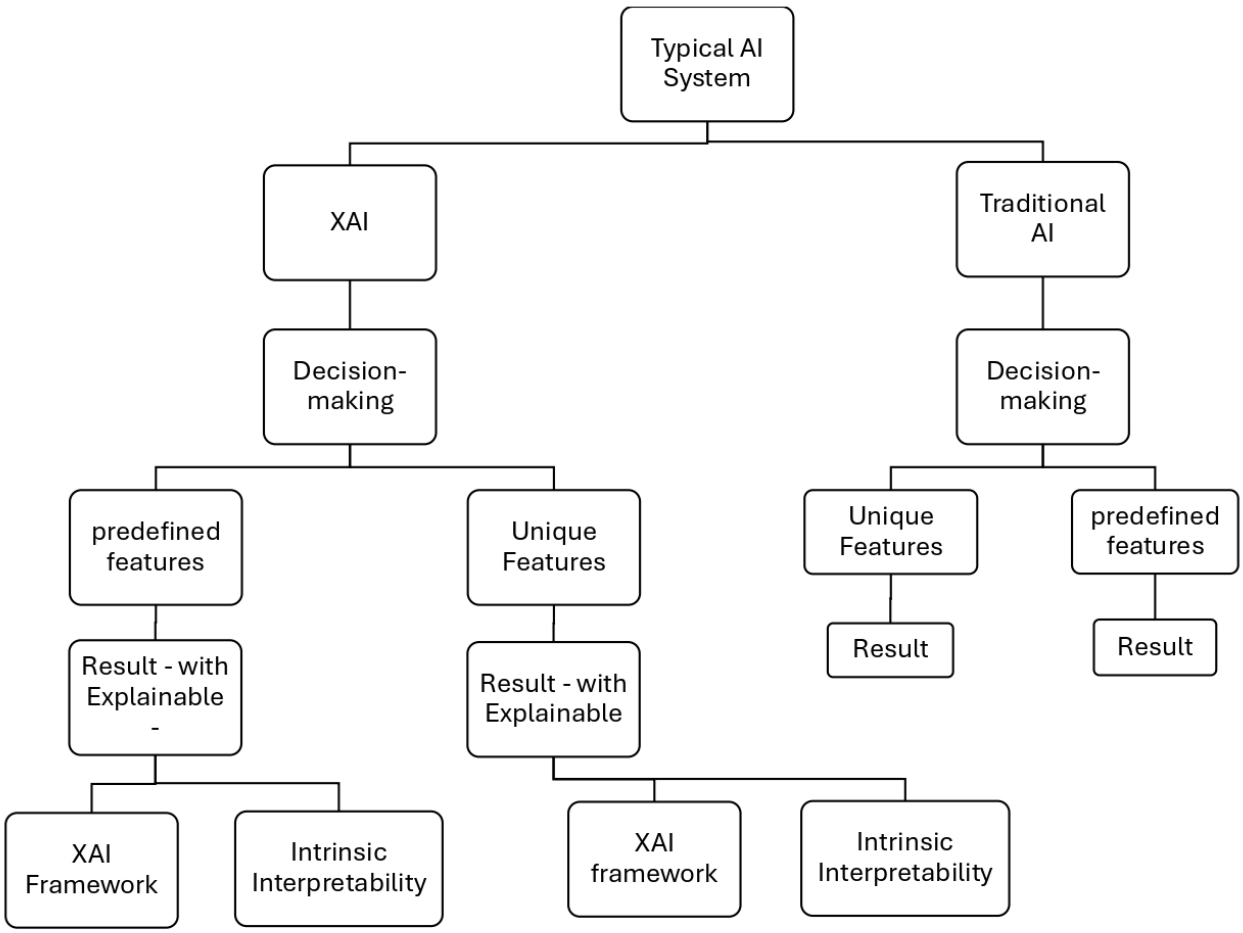


Figure 3. Typical AI Systems. Traditional AI provides results without explanations, while XAI focuses on providing, understandable insights.

It has been understood that directive explanations in XAI are provided in two types: the first is model-agnostic, and the second is model-specific, further classifications can be made regarding how the properties of output interpretations are presented [69]. As a future work, we propose a third type of XAI which can be described as feature-augmented, it can be specifically designed to support clinical decision-making and may assist non-specialist doctors in detecting Alzheimer's disease. In contrast to model-agnostic approaches such as in article [14] or model-specific approaches such as in article [37], which rely primarily on inputs, outputs, and estimates or explore internal processes without providing a direct connection to clinical significance, this third approach incorporates clinically meaningful

features—such as brain volume and cortical thickness measurements—and displays them in the interpretation process. This type of contribution has not been implemented so far due to the unavailability of large datasets and real XAI data [31,32].

Table 4. XAI techniques are used in AD detection across different types of data.

Author	Data Type	Model-Agnostic	Model-Specific
[16]	Facial images	x	✓
[24]	MRI	x	✓
[30]	Fundus image	✓	x
[25]	Clinical data, MRI	✓	✓
[26]	sMRI	✓	x
[14]	MRI	✓	x
[28]	MRI	✓	x
[29]	MRI	✓	x
[27]	MRI	✓	x
[37]	MRI	x	✓
[15]	MRI	✓	x

6. Future Directions and Research Opportunities

There are many promising avenues for advancing the field of explainable AI in AD detection. One major direction is to provide interpretability methods that provide clearer insights into the decision-making processes of complex AI models. While methods such as SHAP, Grad-CAM, and LIME have made great strides in improving model interpretability, their ability to capture the complexity of medical image analysis remains limited [43,55]. As the medical community increasingly relies on AI tools to support diagnosis, it is essential that these tools not only provide accurate predictions but also provide clinically meaningful interpretations. The development of more accurate AI methods that can interpret multi-modal data—such as MRI, PET, and clinical biomarkers—would greatly enhance diagnostic accuracy and confidence in AI-driven decisions [7,32]. Still, deciphering such intricate models is difficult, and the next study ought to concentrate on creating AI methods that can efficiently handle multimodal data [13,57]. Extensive research is needed to develop AI techniques that can function in these settings without sacrificing interpretability [40]. The challenge, however, is to make these advanced models interpretable in a way that is useful to non-specialists in busy clinical settings. While XAI's potential to improve diagnostic accuracy has been discussed in previous studies, this review contributes a unique angle by proposing the "Feature-Augmented" explanation as a future-oriented solution. This type of explanation addresses the black box issue in current XAI models, which is something not tackled effectively by existing methods.

6.1. Development of Feature-Augmented

The novel contribution proposed in this review can directly address the black-box issue of current models. The proposed method will provide a more transparent explanation of the model's predictions, allowing general physicians and non-specialists to make informed decisions based on clinically meaningful data. The diagram in Figure 4 illustrates the types of explanations in XAI, branching into two main categories: Model-Agnostic, and Model-Specific. A third type, Feature-Augmented, is the proposed type; a future approach that displays clinically relevant features, such as brain volume and cortical thickness, into the explanations. This approach should have the potential to greatly enhance both interpretability and accuracy by linking AI-generated insights to well-understood clinical signs. Feature-augmented can increase their ability to interpret model output, leading to more accurate diagnoses and better patient outcomes. Since this approach integrates directly with existing AI frameworks, it provides a practical way to enhance the utility of AI models without sacrificing their fundamental complexity.

For example, when predicting AD, rather than explaining the model's prediction based solely on abstract features learned during training, feature-augmented explicitly

highlights known clinical biomarkers that are directly associated with disease progression. This new method can play an important role in bridging the gap between AI and clinical practice by providing explanations that are consistent with the medical expertise required in diagnosing AD. By displaying clinically relevant features within the model's explanation, this approach provides clinicians with a more meaningful context for decision-making. Rather than relying on machine-learned patterns alone, they can see how traditional biomarkers influence the model's output.

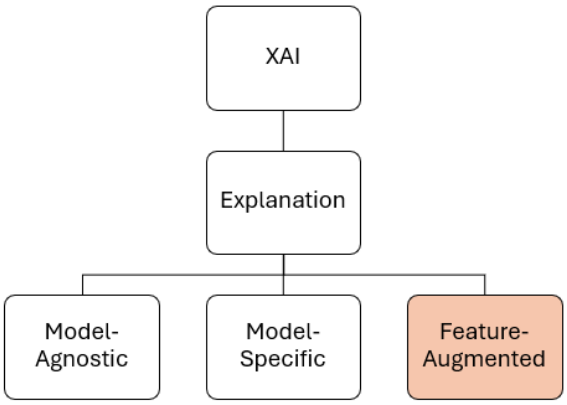


Figure 4. Explanation of AI. The diagram illustrates the types of explanations in XAI, branching into two main categories: Model-Agnostic, and Model-Specific, with the proposed future approach, Feature-Augmented.

6.2. Advancements in Multimodal AI Interpretability

Another promising future direction is working on multimodal XAI systems capable of processing and interpreting data from different sources – such as imaging, clinical vital signs, and patient history–simultaneously and explaining the decision process. While single-modal models, such as those that focus solely on MRI data, have shown promise in detecting AD, integrating data from multiple modalities improves diagnostic accuracy by providing a more comprehensive view of disease progression [7,32]. This level of complexity poses challenges for XAI, as it becomes increasingly difficult to explain the decisions of models that process large amounts of diverse data. The goal of future research in this area should be to develop multimodal AI methods that can balance interpretability and performance. This requires creating technologies that can effectively combine and explain insights from different types of data without overwhelming the general physicians and non-specialists with irrelevant or confusing information. By focusing on practical applications, researchers can ensure that these methods are both technically robust and clinically useful, ultimately improving diagnostic workflows and patient outcomes [13,30,50]. Future work could explore how feature-augmented can be extended to multimodal settings, providing improved explanations that incorporate the rich data available in real-world clinical settings.

6.3. Addressing Implementation Challenges in Clinical Settings

Although significant progress has been made in the field of XAI for AD detection, significant barriers to implementation in real-world clinical settings remain. A major challenge is the lack of standardized guidelines for evaluating the interpretability of AI models in clinical practice. The lack of standardized evaluation methods hinders the broader adoption of XAI technologies in healthcare [41,50]. Future research should focus on creating frameworks that can effectively evaluate the interpretability and clinical utility of AI methods while ensuring that they meet the stringent requirements of medical professionals. Developing standards to compare different XAI techniques in terms of their interpretability and accuracy will be critical for their widespread clinical adoption. Furthermore, researchers must address practical concerns associated with integrating these XAI systems into existing

clinical workflows. XAI tools must not only provide interpretable outputs but also fit seamlessly into the diagnostic processes already in place in hospitals and clinics [30,32,68]. With continued innovation, these challenges can be overcome, ultimately leading to XAI tools of clinical value and ease of use [13,30,50].

7. Summary and Challenges

The diversity of data used as in [25] to detect AD greatly enhances the learning and decision-making capabilities of XAI tools. A major limitation of currently available models is the lack of clear explanations, which may be attributed to the absence of large datasets [32]. Existing interpretable models, in addition to being a black box, are typically rated as less accurate or efficient than more complex data-based approaches, meaning there is a trade-off between interpretability and classification or prediction accuracy [69]. Study [13] has shown that an increased amount of data can help AI tools improve the accuracy of their predictions. However, the process speed tends to decrease with larger data sets. To address this problem, parallel deep learning methods have been proposed as an effective solution, which is supported by the results presented in paper [70], where high speedup was achieved using such techniques on large datasets. In large networks, greater efficiency can be achieved as the segmentation level deepens in large trainable parameters [51]. Despite the progress made in XAI models to detect AD, a major challenge remains the lack of interpretability. Ensuring future XAI models provide understandable explanations is critical for their effective use in clinical settings. The challenge is to provide clear explanations through XAI frameworks. Current AD detection tools, when explaining the decision process, identify areas outside the brain as important factors in the decision-making process, which raises questions about the reliability of these explanations.

8. Conclusions

XAI has the potential to significantly improve diagnostic accuracy and the confidence of general practitioners and non-specialists in detecting AD when applied in crowded clinical settings. By providing understanding explanations for AI models, XAI enables general practitioners and non-specialists to make more informed decisions in the absence of specialists. This is particularly valuable in busy clinical settings, where the ability to accurately diagnose is critical. A "black box" in XAI refers to a model that makes decisions without providing any clear explanation or insight into how those decisions were reached. While current XAI models, such as model-agnostic and model-specific, offer varying degrees of explainability, they still face challenges in fully explaining the decision-making process of complex AI models, so they are considered black boxes. This review contributes to the ongoing debate by providing a comprehensive overview of current AI techniques and identifying areas for future research. A key contribution of this work is the proposal of a novel type of explanation for future research, which we have termed "Feature-Augmented". This approach incorporates clinically relevant biomarkers, such as brain volume and cortical thickness, into the interpretation process and displays them when presenting the results, to address the limitations of "black box" models. By visually incorporating these clinical features, the model's predictions become more interpretable and consistent with clinical understanding, greatly improving the ability of non-specialist clinicians to make accurate diagnoses. Incorporating feature-augmented into the clinical workflow could enhance the reliability of AI-driven diagnoses, making them an invaluable tool in improving patient outcomes in the detection of AD.

Author Contributions: Conceptualization, F.H.S.; methodology, F.H.S.; validation, M.N.A.-A. and W.M.Y.W.B.; formal analysis, F.H.S.; investigation, W.M.Y.W.B.; writing—original draft preparation, F.H.S. writing—review and editing, F.H.S. and M.N.A.-A.; visualization, W.M.Y.W.B.; supervision, M.N.A.-A. All authors have read and agreed to the published version of the manuscript.

Funding: Centre for Research and Innovation Management (CRIM), Universiti Teknikal Malaysia Melaka (UTeM), and Research Management Center (RMC), Multimedia University.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: The authors express their gratitude to the Centre for Research and Innovation Management (CRIM), Universiti Teknikal Malaysia Melaka (UTeM), and Research Management Center (RMC), Multimedia University, for their valuable support in this research. The authors also would like to acknowledge the support of the Kesidang Scholarship in this research.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Fabrizio, C.; Termine, A.; Caltagirone, C.; Sancesario, G. Artificial Intelligence for Alzheimer's Disease: Promise or Challenge? *Diagnostics* **2021**, *11*, 1473. [CrossRef]
2. Majd, S.; Power, J.; Majd, Z. Alzheimer's Disease and Cancer: When Two Monsters Cannot Be Together. *Front. Neurosci.* **2019**, *13*, 155. [CrossRef] [PubMed]
3. Vrahatis, A.G.; Skolariki, K.; Krokidis, M.G.; Lazaros, K.; Exarchos, T.P.; Vlamos, P. Revolutionizing the Early Detection of Alzheimer's Disease Through Non-Invasive Biomarkers: The Role of Artificial Intelligence and Deep Learning. *Sensors* **2023**, *23*, 4184. [CrossRef] [PubMed]
4. Villain, N.; Fouquet, M.; Baron, J.C.; Mézenge, F.; Landeau, B.; de La Sayette, V.; Viader, F.; Eustache, F.; Desgranges, B.; Chételat, G. Sequential Relationships Between Grey Matter and White Matter Atrophy and Brain Metabolic Abnormalities in Early Alzheimer's Disease. *Brain* **2010**, *133*, 3301–3314. [CrossRef]
5. Al Olaimat, M.; Martinez, J.; Saeed, F.; Bozdag, S.; Initiative, A.D.N. PPAD: A Deep Learning Architecture to Predict Progression of Alzheimer's Disease. *Bioinformatics* **2023**, *39*, i149–i157. [CrossRef] [PubMed]
6. Arafa, D.A.; Moustafa, H.E.D.; Ali-Eldin, A.M.T.; Ali, H.A. Early Detection of Alzheimer's Disease Based on the State-of-the-Art Deep Learning Approach: A Comprehensive Survey. *Multimed. Tools Appl.* **2022**, *81*, 23735–23776. [CrossRef]
7. Martínez-Murcia, F.; Górriz, J.; Ramírez, J.; Puntonet, C.; Salas-González, D. Computer Aided Diagnosis Tool for Alzheimer's Disease Based on Mann–Whitney–Wilcoxon U-Test. *Expert Syst. Appl.* **2012**, *39*, 9676–9685. [CrossRef]
8. Woodbright, M.D.; Morshed, A.; Browne, M.; Ray, B.; Moore, S. Toward Transparent AI for Neurological Disorders: A Feature Extraction and Relevance Analysis Framework. *IEEE Access* **2024**, *12*, 37731–37743. [CrossRef]
9. Chandra, A.; Dervenoulas, G.; Politis, M. Magnetic resonance imaging in Alzheimer's disease and mild cognitive impairment. *J. Neurol.* **2018**, *266*, 1293–1302. [CrossRef]
10. Alarjani, M. Alzheimer's Disease Detection Based on Brain Signals Using Computational Modeling. In Proceedings of the 2024 Seventh International Women in Data Science Conference at Prince Sultan University (WiDS PSU), Riyadh, Saudi Arabia, 3–4 March 2024. [CrossRef]
11. Ting, C.P.; Ma, M.C.; Chang, H.I.; Huang, C.W.; Chou, M.C.; Chang, C.C. Diet Pattern Analysis in Alzheimer's Disease Implicates Gender Differences in Folate–B12–Homocysteine Axis on Cognitive Outcomes. *Nutrients* **2024**, *16*, 733. [CrossRef]
12. Bazarbekov, I.; Razaque, A.; Ipalakova, M.; Yoo, J.; Assipova, Z.; Almisreb, A. A Review of Artificial Intelligence Methods for Alzheimer's Disease Diagnosis: Insights from Neuroimaging to Sensor Data Analysis. *Biomed. Signal Process. Control* **2024**, *92*, 106023. [CrossRef]
13. Rehman, S.U.; Tarek, N.; Magdy, C.; Kamel, M.; Abdelhalim, M.; Melek, A.; Mahmoud, L.N.; Sadek, I. AI-Based Tool for Early Detection of Alzheimer's Disease. *Heliyon* **2024**, *10*, e29375. [CrossRef] [PubMed]
14. Yilmaz, D. Development and Evaluation of an Explainable Diagnostic AI for Alzheimer's Disease. In Proceedings of the 2023 International Conference on Artificial Intelligence Science and Applications in Industry and Society (CAISAIS), Galala, Egypt, 3–5 September 2023. [CrossRef]
15. Mansouri, D.; Echtioui, A.; Khemakhem, R.; Hamida, A.B. Explainable AI Framework for Alzheimer's Diagnosis Using Convolutional Neural Networks. In Proceedings of the 2024 IEEE 7th International Conference on Advanced Technologies, Signal and Image Processing (ATSIP), Sousse, Tunisia, 11–13 July 2024. [CrossRef]
16. Umeda-Kameyama, Y.; Kameyama, M.; Tanaka, T.; Son, B.K.; Kojima, T.; Fukasawa, M.; Iizuka, T.; Ogawa, S.; Iijima, K.; Akishita, M. Screening of Alzheimer's Disease by Facial Complexion Using Artificial Intelligence. *Aging* **2021**, *13*, 1765. [CrossRef] [PubMed]
17. Arnold, T.C.; Freeman, C.W.; Litt, B.; Stein, J.M. Low-Field MRI: Clinical Promise and Challenges. *J. Magn. Reson. Imaging* **2023**, *57*, 25–44. [CrossRef]
18. Srikrishna, M.; Heckemann, R.A.; Pereira, J.B.; Volpe, G.; Zettergren, A.; Kern, S.; Westman, E.; Skoog, I.; Schöll, M. Comparison of Two-Dimensional and Three-Dimensional-Based U-Net Architectures for Brain Tissue Classification in One-Dimensional Brain CT. *Front. Comput. Neurosci.* **2021**, *15*, 785244. [CrossRef]
19. Yao, Z.; Wang, H.; Yan, W.; Wang, Z.; Zhang, W.; Wang, Z.; Zhang, G. Artificial Intelligence-Based Diagnosis of Alzheimer's Disease with Brain MRI Images. *Eur. J. Radiol.* **2023**, *165*, 110934. [CrossRef] [PubMed]

20. Arya, A.D.; Verma, S.S.; Chakarabarti, P.; Chakrabarti, T.; Elngar, A.A.; Kamali, A.M.; Nami, M. A Systematic Review on Machine Learning and Deep Learning Techniques in the Effective Diagnosis of Alzheimer's Disease. *Brain Inform.* **2023**, *10*, 17. [CrossRef]
21. Krishnamoorthy, P.; Swetha, D.; Geetha, P.S.; Karunambiga, K.; Ayyasamy, R.K.; Kiran, A. Revolutionizing Medical Diagnostics: Exploring Creativity in AI for Biomedical Image Analysis. In Proceedings of the 2024 International Conference on Signal Processing, Computation, Electronics, Power and Telecommunication (IConSCEPT), Karaikal, India, 4–5 July 2024. [CrossRef]
22. Khan, P.; Kader, M.F.; Islam, S.M.R.; Rahman, A.B.; Kamal, M.S.; Toha, M.U.; Kwak, K.S. Machine Learning and Deep Learning Approaches for Brain Disease Diagnosis: Principles and Recent Advances. *IEEE Access* **2021**, *9*, 37622–37655. [CrossRef]
23. Ben Hassen, S.; Neji, M.; Hussain, Z.; Hussain, A.; Alimi, A.M.; Frikha, M. Deep Learning Methods for Early Detection of Alzheimer's Disease Using Structural MR Images: A Survey. *Neurocomputing* **2024**, *576*, 127325. [CrossRef]
24. Achilleos, K.; Leandrou, S.; Prentzas, N.; Kyriacou, P.; Kakas, A.; Pattichis, C. Extracting Explainable Assessments of Alzheimer's disease via Machine Learning on brain MRI imaging data. In Proceedings of the 2020 IEEE 20th International Conference on Bioinformatics and Bioengineering (BIBE), Cincinnati, OH, USA, 26–28 October 2020. [CrossRef]
25. Jahan, S.; Taher, K.A.; Kaiser, M.S.; Mahmud, M.; Rahman, M.S.; Hosen, A.S.M.S.; Ra, I.H. Explainable AI-Based Alzheimer's Prediction and Management Using Multimodal Data. *PLoS ONE* **2023**, *18*, e0294253. [CrossRef]
26. Guan, H.; Wang, C.; Cheng, J.; Jing, J.; Liu, T. A parallel attention-augmented bilinear network for early magnetic resonance imaging-based diagnosis of Alzheimer's disease. *Hum. Brain Mapp.* **2022**, *43*, 760–772. [CrossRef] [PubMed]
27. Haddada, K.; Khedher, M.I.; Jemai, O.; Khedher, S.I.; El-Yacoubi, M.A. Assessing the Interpretability of Machine Learning Models in Early Detection of Alzheimer's Disease. In Proceedings of the 2024 16th International Conference on Human System Interaction (HSI), Paris, France, 8–11 July 2024. [CrossRef]
28. Shad, H.A.; Rahman, Q.A.; Asad, N.B.; Bakshi, A.Z.; Mursalin, S.; Reza, M.T.; Parvez, M.Z. Exploring Alzheimer's Disease Prediction with XAI in Various Neural Network Models. In Proceedings of the TENCON 2021-2021 IEEE Region 10 Conference (TENCON), Auckland, New Zealand, 7–10 December 2021. [CrossRef]
29. Tima, J.; Wiratkasem, C.; Chairuean, W.; Padongkit, P.; Pangkhiao, K.; Pikulkaew, K. Early Detection of Alzheimer's Disease: A Deep Learning Approach for Accurate Diagnosis. In Proceedings of the 2024 21st International Joint Conference on Computer Science and Software Engineering (JCSSE), Phuket, Thailand, 19–22 June 2024. [CrossRef]
30. Yousefzadeh, N.; Tran, C.; Ramirez-Zamora, A.; Chen, J.; Fang, R.; Thai, M.T. Neuron-Level Explainable AI for Alzheimer's Disease Assessment from Fundus Images. *Sci. Rep.* **2024**, *14*, 7710. [CrossRef] [PubMed]
31. Vimbi, V.; Shaffi, N.; Mahmud, M. Interpreting Artificial Intelligence Models: A Systematic Review on the Application of LIME and SHAP in Alzheimer's Disease Detection. *Brain Inform.* **2024**, *11*, 10. [CrossRef] [PubMed]
32. Viswan, V.; Shaffi, N.; Mahmud, M.; Subramanian, K.; Hajamohideen, F. Explainable Artificial Intelligence in Alzheimer's Disease Classification: A Systematic Review. *Cogn. Comput.* **2023**, *16*, 1–44. [CrossRef]
33. Yokoi, T. Alzheimer's Disease is a Disorder of Consciousness. *Gerontol. Geriatr. Med.* **2023**, *9*, 2. [CrossRef]
34. Sadegh-Zadeh, S.A.; Fakhri, E.; Bahrami, M.; Bagheri, E.; Khamsehashari, R.; Noroozian, M.; Hajiyavand, A.M. An Approach Toward Artificial Intelligence Alzheimer's Disease Diagnosis Using Brain Signals. *Diagnostics* **2023**, *13*, 477. [CrossRef]
35. Savarala, C.; Charan, S.S.; Vemula, S.; Palaniswamy, S.; Pati, P.B. A Novel Approach for Alzheimer's Disease Detection Using XAI and Grad-CAM. In Proceedings of the 2023 4th IEEE Global Conference for Advancement in Technology (GCAT), Bangalore, India, 6–8 October 2023. [CrossRef]
36. AlMohimeed, A.; Saad, R.M.A.; Mostafa, S.; El-Rashidy, N.M.; Farrag, S.; Gaballah, A.; Elaziz, M.A.; El-Sappagh, S.; Saleh, H. Explainable Artificial Intelligence of Multi-Level Stacking Ensemble for Detection of Alzheimer's Disease Based on Particle Swarm Optimization and the Sub-Scores of Cognitive Biomarkers. *IEEE Access* **2023**, *11*, 123173–123193. [CrossRef]
37. Alvarado, M.; Gómez, D.; Nuñez, A.; Robles, A.; Marecos, H.; Ticona, W. Implementation of an Early Detection System for Neurodegenerative Diseases Through the Use of Artificial Intelligence. In Proceedings of the 2023 IEEE XXX International Conference on Electronics, Electrical Engineering and Computing (INTERCON), Lima, Peru, 2–4 November 2023. [CrossRef]
38. Abadir, P.; Oh, E.; Chellappa, R.; Choudhry, N.; Demiris, G.; Ganesan, D.; Karlawish, J.; Marlin, B.; Li, R.M.; Dehak, N.; et al. Artificial Intelligence and Technology Collaboratories: Innovating aging research and Alzheimer's care. *Alzheimer's Dement.* **2024**, *20*, 3074–3079. [CrossRef]
39. Fabietti, M.; Mahmud, M.; Lotfi, A.; Leparulo, A.; Fontana, R.; Vassanelli, S.; Fasolato, C. Early Detection of Alzheimer's Disease From Cortical and Hippocampal Local Field Potentials Using an Ensembled Machine Learning Model. *IEEE Trans. Neural Syst. Rehabil. Eng.* **2023**, *31*, 2839–2848. [CrossRef]
40. Battista, P.; Salvatore, C.; Berlingeri, M.; Cerasa, A.; Castiglioni, I. Artificial Intelligence and Neuropsychological Measures: The Case of Alzheimer's Disease. *Neurosci. Biobehav. Rev.* **2020**, *114*, 211–228. [CrossRef]
41. de la Fuente Garcia, S.; Ritchie, C.W.; Luz, S. Artificial Intelligence, Speech, and Language Processing Approaches to Monitoring Alzheimer's Disease: A Systematic Review. *J. Alzheimer's Dis.* **2020**, *78*, 1547–1574. [CrossRef] [PubMed]
42. Eke, C.S.; Jammeh, E.; Li, X.; Carroll, C.; Pearson, S.; Ifeachor, E. Early Detection of Alzheimer's Disease with Blood Plasma Proteins Using Support Vector Machines. *IEEE J. Biomed. Health Inform.* **2021**, *25*, 218–226. [CrossRef] [PubMed]
43. Brusini, L.; Cruciani, F.; Dall'Aglio, G.; Zajac, T.; Galazzo, I.B.; Zucchelli, M.; Menegaz, G. XAI-Based Assessment of the AMURA Model for Detecting Amyloid-β and Tau Microstructural Signatures in Alzheimer's Disease. *IEEE J. Transl. Eng. Health Med.* **2024**, *12*, 569–579. [CrossRef] [PubMed]

44. Khater, T.; Ansari, S.; Alatrany, A.S.; Alaskar, H.; Mahmoud, S.; Turky, A.; Tawfik, H.; Almajali, E.; Hussain, A. Explainable Machine Learning Model for Alzheimer Detection Using Genetic Data: A Genome-Wide Association Study Approach. *IEEE Access* **2024**, *12*, 95091–95105. [CrossRef]
45. Syed, M.R.; Kothari, N.; Joshi, Y.; Gawade, A. EADDA: Towards Novel and Explainable Deep Learning for Early Alzheimer's Disease Diagnosis Using Autoencoders. *J. Intell. Syst. Appl. Eng.* **2023**, *11*, 234–246.
46. Kim, J.; Lee, M.; Lee, M.K.; Wang, S.M.; Kim, N.Y.; Kang, D.W.; Um, Y.H.; Na, H.R.; Woo, Y.S.; Lee, C.U.; et al. Development of Random Forest Algorithm Based Prediction Model of Alzheimer's Disease Using Neurodegeneration Pattern. *Psychiatry Investig.* **2021**, *18*, 69. [CrossRef]
47. Rashmi, U.; Singh, T.; Ambesange, S. MRI Image-Based Ensemble Voting Classifier for Alzheimer's Disease Classification with Explainable AI Technique. In Proceedings of the 2023 IEEE 8th International Conference for Convergence in Technology (I2CT), Lonavla, India, 7–9 April 2023. [CrossRef]
48. Weber, C.J.; Carrillo, M.C.; Jagust, W.; Jack, C.R.; Shaw, L.M.; Trojanowski, J.Q.; Saykin, A.J.; Beckett, L.A.; Sur, C.; Rao, N.P.; et al. The Worldwide Alzheimer's Disease Neuroimaging Initiative: ADNI-3 Updates and Global Perspectives. *Alzheimer's Dement. Transl. Res. Clin. Interv.* **2021**, *7*, e12226. [CrossRef]
49. LaMontagne, P.J.; Benzinger, T.L.S.; Morris, J.C.; Keefe, S.; Hornbeck, R.; Xiong, C.; Grant, E.; Hassenstab, J.; Moulder, K.; Vlassenko, A.G.; et al. OASIS-3: Longitudinal Neuroimaging, Clinical, and Cognitive Dataset for Normal Aging and Alzheimer Disease. *medRxiv* **2019**, 1. [CrossRef]
50. Qu, Y.; Wang, P.; Liu, B.; Song, C.; Wang, D.; Yang, H.; Zhang, Z.; Chen, P.; Kang, X.; Du, K.; et al. AI4AD: Artificial Intelligence Analysis for Alzheimer's Disease Classification Based on a Multisite DTI Database. *Brain Disord.* **2021**, *1*, 100005. [CrossRef]
51. Al-Andoli, M.; Cheah, W.P.; Tan, S.C. Deep Learning-Based Community Detection in Complex Networks with Network Partitioning and Reduction of Trainable Parameters. *J. Ambient. Intell. Humaniz. Comput.* **2020**, *12*, 2527–2545. [CrossRef]
52. Srikrishna, M.; Ashton, N.J.; Moscoso, A.; Pereira, J.B.; Heckemann, R.A.; van Westen, D.; Volpe, G.; Simrén, J.; Zettergren, A.; Kern, S.; et al. CT-Based Volumetric Measures Obtained Through Deep Learning: Association with Biomarkers of Neurodegeneration. *Alzheimer's Dement.* **2024**, *20*, 629–640. [CrossRef] [PubMed]
53. Srikrishna, M.; Pereira, J.B.; Heckemann, R.A.; Volpe, G.; van Westen, D.; Zettergren, A.; Kern, S.; Wahlund, L.O.; Westman, E.; Skoog, I.; et al. Deep Learning from MRI-Derived Labels Enables Automatic Brain Tissue Classification on Human Brain CT. *NeuroImage* **2021**, *244*, 118606. [CrossRef] [PubMed]
54. Al-Andoli, M.N.; Tan, S.C.; Sim, K.S.; Lim, C.P.; Goh, P.Y. Parallel Deep Learning with a Hybrid BP-PSO Framework for Feature Extraction and Malware Classification. *Appl. Soft Comput.* **2022**, *131*, 109756. [CrossRef]
55. Shukla, A.; Upadhyay, S.; Bachan, P.R.; Bera, U.N.; Kshirsagar, R.; Nathani, N. Dynamic Explainability in AI for Neurological Disorders: An Adaptive Model for Transparent Decision-Making in Alzheimer's Disease Diagnosis. In Proceedings of the 2024 IEEE 13th International Conference on Communication Systems and Network Technologies (CSNT), Jabalpur, India, 6–7 April 2024. [CrossRef]
56. Mahim, S.M.; Ali, M.S.; Hasan, M.O.; Nafi, A.A.N.; Sadat, A.; Hasan, S.A.; Shareef, B.; Ahsan, M.M.; Islam, M.K.; Miah, M.S.; et al. Unlocking the Potential of XAI for Improved Alzheimer's Disease Detection and Classification Using a ViT-GRU Model. *IEEE Access* **2024**, *12*, 8390–8412. [CrossRef]
57. Deshmukh, A.; Kallivalappil, N.; D'souza, K.; Kadam, C. AL-XAI-MERS: Unveiling Alzheimer's Mysteries with Explainable AI. In Proceedings of the 2024 Second International Conference on Emerging Trends in Information Technology and Engineering (ICETITE), Vellore, India, 22–23 February 2024. [CrossRef]
58. Jain, V.; Nankar, O.; Jerrish, D.J.; Gite, S.; Patil, S.; Kotecha, K. A Novel AI-Based System for Detection and Severity Prediction of Dementia Using MRI. *IEEE Access* **2021**, *9*, 154324–154346. [CrossRef]
59. Guo, M.H.; Lu, C.Z.; Liu, Z.N.; Cheng, M.M.; Hu, S.M. Visual Attention Network. *Comput. Vis. Media* **2023**, *9*, 733–752. [CrossRef]
60. Zhu, W.; Sun, L.; Huang, J.; Han, L.; Zhang, D. Dual Attention Multi-Instance Deep Learning for Alzheimer's Disease Diagnosis With Structural MRI. *IEEE Trans. Med. Imaging* **2021**, *40*, 2354–2366. [CrossRef]
61. Illakiya, T.; Ramamurthy, K.; Siddharth, M.V.; Mishra, R.; Udainiya, A. AHANet: Adaptive Hybrid Attention Network for Alzheimer's Disease Classification Using Brain Magnetic Resonance Imaging. *Bioengineering* **2023**, *10*, 714. [CrossRef]
62. Nguyen, M.; He, T.; An, L.; Alexander, D.C.; Feng, J.; Yeo, B.T. Predicting Alzheimer's Disease Progression Using Deep Recurrent Neural Networks. *NeuroImage* **2020**, *222*, 117203. [CrossRef]
63. Fujita, K.; Katsuki, M.; Takasu, A.; Kitajima, A.; Shimazu, T.; Maruki, Y. Development of an Artificial Intelligence-Based Diagnostic Model for Alzheimer's Disease. *Aging Med.* **2022**, *5*, 167–173. [CrossRef]
64. Ganaie, M.; Hu, M.; Malik, A.; Tanveer, M.; Suganthan, P. Ensemble Deep Learning: A Review. *Eng. Appl. Artif. Intell.* **2022**, *115*, 105151. [CrossRef]
65. Zhang, Y.; Liu, T.; Lanfranchi, V.; Yang, P. Explainable Tensor Multi-Task Ensemble Learning Based on Brain Structure Variation for Alzheimer's Disease Dynamic Prediction. *IEEE J. Transl. Eng. Health Med.* **2023**, *11*, 1–12. [CrossRef] [PubMed]
66. Kamal, M.S.; Chowdhury, L.; Nimmy, S.F.; Rafi, T.H.H.; Chae, D.K. An Interpretable Framework for Identifying Cerebral Microbleeds and Alzheimer's Disease Severity Using Multimodal Data. In Proceedings of the 2023 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), Sydney, Australia, 24–27 July 2023. [CrossRef]

67. Xu, X.; Yan, X. A Convenient and Reliable Multi-Class Classification Model Based on Explainable Artificial Intelligence for Alzheimer's Disease. In Proceedings of the 2022 IEEE International Conference on Advances in Electrical Engineering and Computer Applications (AEECA), Dalian, China, 20–21 August 2022. [CrossRef]
68. Zhao, X.; Ang, C.K.E.; Acharya, U.R.; Cheong, K.H. Application of Artificial Intelligence Techniques for the Detection of Alzheimer's Disease Using Structural MRI Images. *Biocybern. Biomed. Eng.* **2021**, *41*, 456–473. [CrossRef]
69. González-Alday, R.; García-Cuesta, E.; Kulikowski, C.A.; Maojo, V. A Scoping Review on the Progress, Applicability, and Future of Explainable Artificial Intelligence in Medicine. *Appl. Sci.* **2023**, *13*, 778. [CrossRef]
70. Al-Andoli, M.N.; Tan, S.C.; Cheah, W.P. Distributed Parallel Deep Learning with a Hybrid Backpropagation-Particle Swarm Optimization for Community Detection in Large Complex Networks. *Inf. Sci.* **2022**, *600*, 94–117. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.