

Original Article

NeuroEarlyAI: An Explainable Deep Learning Framework with EarlyDetectNet for Alzheimer's Disease Detection and Classification using MRI Scans

Anil Chakravarthy^{1*}, Sasmita Kumari Nayak², B S Panda³

^{1,2}Department of Computer Science & Engineering, Centurion University of Technology and Management Odisha, India.

³Department of Computer Science & Engineering, Raghu Engineering College, Visakhapatnam, India.

^{1*}Corresponding Author : anilchakravarthy2015@gmail.com

Received: 09 February 2026

Revised: 25 July 2026

Accepted: 05 August 2026

Published: 30 September 2026

Abstract - Alzheimer's Disease (AD) is an ageing and progressive neurodegenerative disease, and one of the most prevalent dementias that affects millions of patients worldwide, including millions of people in Europe, and is a public health priority. Early and right diagnosis is of paramount importance for early therapeutic intervention as well as retardation of the disease. The visual search for small changes in the brain is rather time-consuming and subjective. For comparison, we chose two traditional diagnostic procedures: clinical evaluation and manual MRI analysis, where the clinical evaluation is more appropriate for the intended user in healthcare centres. The automatic detection of AD using recent advancements in deep learning are yet to lack a few drawbacks, such as class imbalance, overfitting and generalisation issue on external datasets, low transparency in the decision-making of the deep learning model, which makes it difficult for clinical applications. To fill this gap, in this paper, we propose NeuroEarlyAI, an end-to-end learning framework with EarlyDetectNet, a new model for early and interpretable detection of AD from the structural MRI scans. The framework also consists of sophisticated pre-processing strategies (denoising, data augmentation and skull stripping) to enhance the quality and diversity of the data. EarlyDetectNet has an attention mechanism which highlights clinically relevant anatomical regions. As a contrast, the visual and feature-wise explanation modules are based on the Grad-CAM and SHAP techniques, respectively, which project the model prediction onto the visual and feature-wise explanations. Extensive tests are performed on two publicly available datasets: Alzheimer's MRI Dataset and OASIS-1 dataset. The proposed system demonstrated improvements over the state of the art on reaching accuracy for binary and multi-class classification in a cross-dataset setting, with 96.7% and 93.8% accuracy, respectively. These results show the promise of the applications and broader use of NeuroEarlyAI, a tool that could help to diagnose AD earlier for better patient care.

Keywords - Alzheimer's Disease, Deep Learning, MRI Classification, Explainable AI, Early Diagnosis.

1. Introduction

Alzheimer's Disease (AD) is an unstoppable and widespread neurodegenerative disorder that produces one of the most important dementing pathologies in the world, and is a problem of the family, the health services, the mentally ill and international law [1, 2]. It is important to diagnose AD early, which will allow for earlier treatment and better patient care and help slow the disease's progression. Conventional diagnostic procedures, such as neuropsychological testing and However, manual comparison of the MRI image is qualitative and slow, and is not sensitive to early subtle pathological brain deviations. Recent progress in AI and deep learning has paved the way for new opportunities for automatic, fast and reliable AD diagnosis. Several studies have been made to categorize the MRI images into different stages of dementia using CNN and other deep learning approaches with good results [3, 4].

While promising performance of recent deep learning and Explainable Artificial Intelligence (XAI) approaches have been demonstrated in Alzheimer's Disease (AD) detection, key research gaps continue to exist. First, many existing models have a strong focus on classification accuracy, disregarding the need for model interpretability and clinical transparency for real-world healthcare deployment. Second, some studies are tested on just one data set, making it difficult to make generalizations at other acquisition parameters and patient populations. Third, there are many deep learning architectures that are very complex and demanding with regard to hardware resources, making it challenging to deploy them in real clinical settings. Additionally, the joint use of lightweight architectures and attention-driven feature learning for early-stage AD diagnosis, along with robust pre-processing and explainability mechanisms have not been



explored under a single umbrella. These constraints provide the impetus to build a more robust, interpretable and computationally efficient framework that can deliver reliable diagnostic support in a wide range of clinical situations.

There are many recent methods; nonetheless, many of them have several bottlenecks that make it impossible to incorporate them into clinical diagnosis, including class imbalance, overfitting, and subpar generalisation ability to read external data, as well as the lack of explanation concerning why such a prediction was made. To address these shortcomings, there is a growing need for new models that exhibit high prediction performance while also being transparent and robust. In this work, we proposed a NeuroEarlyAI framework based on deep learning with EarlyDetectNet—an exceptional architecture to achieve early and robust detection of AD based on structural MRI. The objective of this work is to develop a clinic-friendly AI tool for assisting in the diagnosis of dementia by classifying sMCI MRI scans into four diagnostic categories: Non-Demented, Very Mild Demented, Mild Demented, and Moderate Demented, with explainable explanations. Among other innovations in our work, we employ an attention mechanism that focuses on disease-related regions in brain structures, as well as preprocessing techniques such as noise reduction and data augmentation, to enhance data quality and variability, respectively. Additionally, we incorporate an explainability module supported by Grad-CAM and SHAP visualisations to achieve more transparent predictions. These parts collectively result in a stable and understandable generalisation across benchmarks, addressing the issues arising from contemporary state-of-the-art models. This research has several impacts. First, a powerful and fully automated AD detection pipeline, including preprocessing, profound feature extraction, and explainability, is built. Second, the study is based on two public datasets, including cross-dataset validation based on OASIS-1 dataset, and extensive experiments are carried out, showing good generalisation. Third, extensive evaluation measures and comparative analysis with state-of-the-art models are presented, which proves the superiority of EarlyDetectNet.

The novelty of the proposed NeuroEarlyAI framework is the combination of a lightweight attention-based deep learning architecture, explainable AI mechanisms and cross-dataset validation for clinically reliable Alzheimer's Disease diagnosis. Unlike current approaches, which primarily focus on the classification performance, the proposed EarlyDetectNet prioritizes not only classification accuracy but also interpretability, computational efficiency and generalization capability. This framework integrates cutting-edge preprocessing methods, attention-driven feature extraction, visual explanations via Grad-CAM and SHAP, and external validation with the OASIS-1 dataset in a single end-to-end solution. Moreover, the proposed model provides competitive performance but lower architecture complexity

than several CNN-based existing approaches, which makes it more favorable to be used in the real clinical setting.

While recent transformer-based and graph-based AD detection models typically demand significant computational demands and large-scale multimodal data, EarlyDetectNet is engineered to be lightweight and efficient, suitable for deployment in real clinical settings. The novelty of the proposed model is that it combines attention-guided convolutional learning, a robust MRI preprocessing model, explainable AI mechanisms, and cross-dataset validation into a single model. Moreover, the proposed system tries to solve the classification accuracy, interpretability, robustness and computational efficiency at the same time, which are often studied separately in the existing studies. This balanced design renders EarlyDetectNet more suitable for real-world healthcare applications where transparency, reliability, and resource efficiency are key requirements.

The structure of this paper is as follows. In Section 2, related work is presented, which includes existing literature on AD detection and explainable AI methods. Section 3 presents the methodology that includes a description of the dataset, a preprocessing pipeline, a model architecture, and explainability modules. Experimental results are provided in Section 4, along with cross-dataset validation, comparisons and ablation studies. The limitations of the study are discussed in Section 5. In the end, Section 6 brings the paper to a close and provides suggestions for future research.

While recent transformer, graph-based and multimodal learning frameworks have demonstrated promising abilities at diagnosis, there are several key limitations. Most current methods demand a substantial amount of data, computational power, or multimodal clinical data, which might not be readily accessible in typical healthcare settings. Moreover, there is a lack of external validation and explainability, which is critical for clinical deployment. To solve these challenges, an EarlyDetectNet architecture is proposed, which involves a lightweight attention-guided feature extraction, a strong MRI preprocessing, multiple explainability methods (Grad-CAM, SHAP, LIME) and cross-dataset validation in a single framework. NeuroEarlyAI stands out from many recent studies, which focus solely on classification accuracy by promoting diagnostic performance, interpretability, computational efficiency, and generalization capabilities, making it more suitable for application in real-world clinical scenarios.

1.1. Novelty and Scientific Contributions

The NeuroEarlyAI framework brings several novel contributions towards explainable and clinically reliable Alzheimer's Disease detection from MRI scans. First, the framework features a computationally efficient deep learning model, EarlyDetectNet, which uses a lightweight attention mechanism to achieve high diagnostic performance in a

manner suitable for deployment in healthcare. Second, the proposed framework goes beyond most of the existing ones that mainly consider the accuracy of classification and incorporate the explainable AI techniques such as Grad-CAM and SHAP, which increase the interpretability and clinical trustworthiness.

Third, the framework includes a full preprocessing pipeline: skull stripping, normalization of intensity, reduction of noise, and data augmentation to enhance the quality of data and robustness of models. In addition, cross-dataset validation tests have been carried out with the OASIS-1 data set to assess the generalization ability of the model in a non-homogeneous imaging scenario. NeuroEarlyAI's unique combination of explainability, lightweight architecture, robust preprocessing, and external validation in a single end-to-end system sets it apart from current Alzheimer's Disease detection systems.

1.2. Gap of the Study and Novelty of the Work

Although recent work has introduced deep learning solutions in the area of Alzheimer's Disease diagnosis, some key challenges have not been adequately addressed in the current work. Most existing approaches are mainly concerned with maximizing classification accuracy and fail to pay attention to interpretability, computational efficiency and generalization across data sets. In addition, many models use complex architectures, multimodal data or opaque decision-making methods, limiting their feasibility in real clinical environments. The proposed NeuroEarlyAI framework aims to overcome the aforementioned research gaps by designing a lightweight attention-based deep learning framework, named EarlyDetectNet, specifically for interpretable and efficient diagnosis of Alzheimer's Disease (AD) from structural MRI scans. The framework combines efficient pre-processing with the use of explainable AI methods (Grad-CAM, SHAP, LIME) and cross-dataset validation in a single end-to-end pipeline, which is not the case in many of the existing approaches.

In summary, the following are the novel contributions of this work:

Creation of EarlyDetectNet - a lightweight attention-guided deep learning architecture for early Alzheimer's Disease detection and classification.

Multiple Explainability mechanisms will be integrated for increased clinical interpretability and transparency, such as Grad-CAM, SHAP and LIME.

MRI data preprocessing with skull stripping, intensity normalization, noise reduction and data augmentation.

Cross-dataset validation: evaluate model robustness and generalization ability in the face of heterogeneous imaging conditions using OASIS-1.

High performance in the diagnosis while keeping the computational complexity low in comparison to some of the existing deep learning frameworks.

The combined contributions highlight the unique features of NeuroEarlyAI compared to existing Alzheimer's Disease recognition models and expand its potential use in clinically interpretable and resource-efficient healthcare settings.

2. Related Work

Magnetic Resonance Imaging (MRI)- based approaches for the detection and type classification of Alzheimer's disease (AD) have been extensively studied over the last decade, with an increasing focus on deploying Explainable Artificial Intelligence (XAI) frameworks to overcome the black-box nature of deep learning models. Vimbi Viswa et al. [1] reported a detailed systematic review, addressing different XAI methodology risk parameters, including SHapley Additive exPlanations (SHAP), Local Interpretable Model-agnostic Explanations (LIME) and Gradient-weighted Class Activation Mapping (Grad-CAM) and the potential role of these approaches to facilitate adoption of AI-based systems for patient diagnosis in AD. With such frameworks, clinicians can understand the predictions of deep learning models and have confidence in them for clinical practice. Similarly, Linda M. Duamwan and Jordan J. Bird [2] applied a CNN-based approach with LIME to the ADNI dataset, achieving an accuracy of 94.96%, which highlights the significance of visual explanations in identifying MRI segments relevant to disease detection.

Recent works have also focused on multimodal methods. El-Sappagh et al. [3] presented a two-layer multimodal detection and prediction model that uses 11 separate data modalities available in the ADNI dataset. The model employed Random Forest (RF) classifiers and SHAP-based explanations to produce global and local interpretability, attaining 93.95% accuracy for early diagnosis and 87.08% accuracy for predicting progression from Mild Cognitive Impairment (MCI) to AD. Jahan et al. [4] further extended this idea by using a multimodal dataset that contains clinical, psychological, and MRI segmentations for a five-class classification, achieving 98.81% accuracy using an RF model and SHAP to discern the most determining features for decision-making. These investigations together present the advantage of a multimodal data fusion in terms of diagnostic performance compared to explainability.

Models based on deep learning are still the state-of-the-art in AD research. Eduardo Nigri et al. [5] opened the Swap Test; a novel interpretability method created for deep CNNs in MRI-based AD diagnosis. Here, they built upon basic occlusion methods, developing clinically relevant heatmaps that closely match established neurobiological markers for AD. Similarly, Essemlali et al. [6] demonstrated the application of a variant of BrainNet CNN to diffusion-

weighted MRI tracts for structural brain connection analysis and the identification of critical cortical regions in AD. Another tube fusion method for PET and MRI was considered by Odusami et al. [7], who proposed an early fusion network with a modified ResNet18 architecture to fuse multimodal information and reported an accuracy of 73.90% when experimenting with the combination of complementary imaging modalities.

Several research endeavours have yielded significant progress in early detection and classification. Noviany et al. [8] utilised a ResNet-50 model with Grad-CAM for visualisation, achieving an accuracy of 95.63%, which further highlights the importance of explainability in clinical decision support. AbdelAziz et al. [9] proposed the SECNN-RF approach, which combines Squeeze-and-Excitation CNNs and Random Forest to achieve high classification accuracy (99.89%), outperforming the state of the art with clear interpretability due to attention mechanisms. Alatrany et al. [11]. There, they focused on the interplay between machine learning and interpretable rule extraction methods, achieving an F1 score of 98.9% for binary classification and producing human-interpretable, validated rules using SHAP and LIME.

Transfer learning has also been studied for cross-domain tasks. Alsubai et al. [12] combined transfer learning and XAI to detect both brain tumours and AD across multiple datasets. This study showed that pre-trained CNN architectures can also be utilised in other clinical settings. Rahman et al. [13] compared various neural network models, including ResNet50, VGG16, and Inception v3, with LIME-based explanations for improved MRI-based classification, achieving an accuracy of up to 86.68%. De Santi et al. [14] addressed the PET image and employed post-hoc explanation methods, including LRP and Saliency Maps, to quantitatively assess the model predictions. Bogdanovic et al. [15] used XGBoost models on a big medical and lifestyle data set, using SHAP for global and local interpretability (\$F_1\$-score of 0.84).

The recently proposed architectures have achieved further improvements. Sadat et al. [16] adapted Vision Transformers (ViT) and Gated Recurrent Units (GRU) to produce a hybrid model that achieved accuracies of over 99% on several datasets, further enhanced by XAI techniques to explain the decision-making process. Jahan et al. [17] formulated a federated learning-based explainable AI framework that utilises multimodal data to mitigate privacy issues, preserve the confidentiality of patient data, increase accuracy (98.93%), and enhance interpretability using SHAP. Almohimeed et al. [18] introduced a multi-level stacking cascade scheme optimised by Particle Swarm Optimisation. They demonstrated that heterogeneous multimodal input is beneficial to diagnosis accuracy, yielding the best experimental results of 92.08% for the two-class problem and 90.03% for the three-class problem.

In recent times, deep learning methods for neurodegenerative disease diagnosis have delved into Transformer-based and Graph Neural Network (GNN)-based frameworks for diagnosis of Alzheimer's disease. In some neuroimaging studies, Vision Transformers (ViTs) have been shown to have a robust ability to model long-range spatial dependencies and global contextual information from MRI scans, resulting in better classification performance. Additionally, hybrid transformer-CNN architectures have been proposed to integrate the local feature extraction with the global attention modelling to better represent the disease for learning. However, transformer-based models can be resource-intensive, requiring extensive training datasets, which can be challenging to obtain in clinical environments.

Concurrently, Graph Neural Networks (GNNs) have become a promising method for modelling brain connectivity and structural relationships for neuroimaging analysis. The GNN-based frameworks view the brain regions as nodes in the graph, and the inter-regional relationships are viewed as graph edges, thus allowing for the analysis of complex neurological interactions that accompany the progression of Alzheimer's Disease. Recent studies have proved that graph-based learning is a powerful method for detecting abnormalities in functional and structure MRI data and multimodal neuroimaging data. But there are many GNN architectures that have the shortcomings of requiring complex models, low interpretability, and carefully crafted graph-building strategies.

Nearly Explainable Artificial Intelligence (XAI) approaches also have received a lot of attention in the field of Alzheimer's Disease diagnosis from neuroimaging. Methods like attention visualization, Grad-CAM++ and Integrated Gradients, SHAP, and saliency-based approaches are becoming more popular to make deep learning systems more transparent and clinically interpretable. These methods help clinicians to interpret the model's reasoning by showing the structures related to the disease and feature contributions from MRI scans. However, current explainability frameworks mostly use only one mechanism to interpret, which reduces the comprehensiveness and reliability of the generated explanation. Thus, combining various complementary explainability methods in a single framework is an important research area for clinically trustworthy AI systems.

Graph theoretical approaches have been adopted to study connectivity in the brain. Amoroso et al. [19] utilised SHAP values to interpret patch-based connectivity models, with a focus on AD-related areas (hippocampus and amygdala). Hernandez et al. [20] utilised SHAP to investigate high-performing methods in the TADPOLE challenge, providing insights into feature importance and methodological contributions to accuracy. A hierarchical spatial feature-CNN with Grad-CAM integration for segmentation and classification was proposed by Suganya and Aarth [21],

achieving an accuracy of over 98.17%. Chethana et al. [22] integrated a CNN and an RNN model with Grad-CAM visualisations for AD stage recognition, achieving an accuracy of 95.06%. Yoon and Lin [23] generalised LIME and Grad-CAM for interpretable MRI-based brain tumour classification, indirectly benefiting AD studies. Lighter models are also being viewed as a way to work around computational limitations. Kina [24] introduced TLEABLCNN, an attention model enhanced by the SMOTE method for imbalanced classification, which achieved competitive results in MRI data.

Poyraz et al. [25] proposed BrainNeXt, a lightweight CNN designed for automatic brain disorder detection, prioritising efficiency over performance. Rashid et al. [26] introduced Biceph-Net, a stable and reliable architecture for 2D MRI scans. Katabathula et al. [27] proposed a 3D deep CNN for MRI analysis in the brain, focusing on the hippocampus, to improve interpretability through the use of visualisations. Finally, Roy et al. [28] proposed AD-Lite Net, a simplified CNN model designed for lightweight, resource-sensitive detection of Alzheimer's disease, which has demonstrated the capacity to perform well in low-resource environments.

Recent deep learning techniques have significantly enhanced the detection and classification of Alzheimer's disease from brain MRI images. Islam and Zhang [29] presented a Deep Convolutional Neural Network (Deep CNN) model for automatic feature extraction and classification, aiming to achieve precise detection of brain structure abnormalities. Based on this background, Mian Muhammad Sadiq Fareed et al. [30] introduced ADD-Net, a dedicated deep learning model for early detection with higher sensitivity to minimal cortical alterations. Similarly, Murugan et al. [31], to the best of our knowledge, have not yet developed DEMNET

using deep residual connections for optimised input data preprocessing in early diagnosis and multi-stage Alzheimer's disease classification, which we have achieved, resulting in high performance. Marcus et al. [33] introduced the OASIS MRI dataset, providing open-access brain imaging data from healthy and dementia-affected individuals to support neuroimaging research and Alzheimer's disease analysis. Lastly, we conclude that the literature shows a pattern of tread first leveraging and then integrating explainability into state-of-the-art AD AI models. Although the diagnostic performance is considerably enhanced by deep learning approaches, combining them with an XAI framework ensures that high-performance models are interpretable and trustworthy in clinical practice. The reviewed literature has demonstrated the need for multimodal fusion, new architectures, privacy-preserving federated learning, and lightweight models in numerous aspects. We have made an inspired footprint towards setting up NeuroEarlyAI, guided by its EarlyDetectNet model, for tackling the challenges of accuracy and interpretability that translate safely into the real world. While some previous research has shown significant advances in detecting Alzheimer's Disease with MRI and explainable AI models, there are still some challenges to be addressed. Many approaches use architectures that are heavy on computational resources, which do not necessarily fit into the resource-limited healthcare settings. Also, many of the studies are mainly targeted at improving the accuracy, neglecting the importance of model transparency, performance on external datasets, and interpretability of predictions. Some methods do not have full processing pipelines and do not consider generalization performance in cross-dataset validation. The proposed NeuroEarlyAI framework brings a lightweight attention-guided architecture, strong preprocessing modules, explainability modules, and cross-dataset evaluation to enhance the clinical applicability and diagnostic reliability of these methods.

Table 1. Comparative analysis of existing Alzheimer's Disease detection methods

Method	Dataset	Explainability	Cross-Dataset Validation	Accuracy	Limitation
Deep CNN [29]	MRI Dataset	No	No	90.4%	High computational complexity
ADD-Net [30]	ADNI	Partial	No	92.1%	Limited interpretability
DEMNET [31]	MRI Dataset	No	No	93.5%	Overfitting risk
SECNN-RF [9]	MRI Dataset	Yes	No	99.89%	Complex architecture
ViT-GRU [16]	Multi-dataset	Yes	Partial	99%	High hardware requirements
Proposed NeuroEarlyAI	Kaggle + OASIS-1	Yes (Grad-CAM + SHAP)	Yes	96.7%	Requires MRI preprocessing

Table 1 compares the existing Alzheimer's Disease detection approaches with the proposed NeuroEarlyAI framework. The proposed framework is the first to consider simultaneously classification accuracy, interpretability, computational efficiency and cross-dataset generalization.

Most of the existing approaches have no explainability support, or have not validated the robustness of the model on external datasets. NeuroEarlyAI, on the other hand, integrates Explainable AI (XAI) methods with lightweight deep learning and external validation to enhance the scientific validity and clinical relevance of the application.

Table 2. Comparison of recent deep learning approaches for Alzheimer's Disease detection

Method	Architecture Type	Explainability	Dataset	Advantages	Limitations
ViT-Based Model	Transformer	Partial	ADNI	Global attention modelling	High computational cost
CNN-GRU Hybrid	CNN + RNN	Grad-CAM	MRI Dataset	Temporal feature extraction	Overfitting risk
GNN-Based Framework	Graph Neural Network	SHAP	ADNI	Brain connectivity analysis	Complex graph construction
Multimodal Transformer	Transformer + Fusion	Attention Maps	ADNI + PET	High accuracy	Requires multimodal data
Proposed EarlyDetectNet	Lightweight CNN + Attention	Grad-CAM + SHAP + LIME	Kaggle + OASIS-1	Efficient, interpretable, robust	MRI-only framework

Table 3. Research gaps in existing Alzheimer's Disease detection frameworks and proposed solutions in NeuroEarlyAI

Existing Limitation in Prior Work	Impact	NeuroEarlyAI Solution
Limited interpretability	Reduced clinical trust	Grad-CAM + SHAP + LIME integration
Single-dataset evaluation	Poor generalization	Cross-dataset validation using OASIS-1
High computational complexity	Difficult clinical deployment	Lightweight EarlyDetectNet architecture
Black-box prediction behaviour	Lack of transparency	Explainable AI-driven visualization
Insufficient preprocessing robustness	Reduced stability	Comprehensive preprocessing pipeline
Overfitting risk in deep CNNs	Limited reliability	Attention mechanism + augmentation + dropout

Table 2 lists some recent deep learning architectures that have been used to detect Alzheimer's Disease and compares them to the proposed EarlyDetectNet framework. Transformer-based and graph-based methods have also achieved excellent performance in the analysis of neuroimages, but many of these methods require costly architectures and large multimodal datasets.

In contrast, EarlyDetectNet focuses on achieving a balanced trade-off between diagnostic performance, computational efficiency, explainability, and cross-dataset robustness. Unlike many of the existing explainable learning frameworks, the proposed approach involves the integration of several explainability mechanisms and lightweight attention-based learning.

Table 3 provides a summary of the key limitations in the current frameworks of Alzheimer's Detection and presents the proposed NeuroEarlyAI framework that resolves those limitations. NeuroEarlyAI differs from many previous methods, which prioritize classification accuracy, by taking into account both interpretability, robustness, computational efficiency, and cross-dataset reliability. This balanced design enhances the scientific contribution and practical clinical applicability of the proposed framework.

3. Proposed Framework

The presented framework (referred to as NeuroEarlyAI) introduces the development of EarlyDetectNet, a new deep

learning model for accurate and interpretable Alzheimer's Disease diagnosis from MRI scans. It integrates the latest pre-processing, attention-based features and explainability modules (e.g., Grad-CAM, SHAP) that ensure robust performance, higher generalisation power and clinical visibility towards early diagnostic aid and consistent decision support.

3.1. An Overview of the Proposed Structure

The proposed NeuroEarlyAI is designed as a comprehensive end-to-end system for early AD detection with brain MRI. It combines a set of in-built modules, from image acquisition to MRI data preprocessing, through a deep learning-based framework, the generation of an explanation model, and performance metrics computations, until the final clinical decision.

The researchers have developed a system that aims for precise and interpretable predictions, in the hopes of helping neurologists and physicians diagnose Alzheimer's disease before it debilitates patients and becomes too advanced to treat. For the first setting, the model is trained using two MRI image sets: the Kaggle Alzheimer's MRI Dataset and the OASIS-1, which is used only for cross-dataset validation (test). These datasets contain images of healthy and Alzheimer's patients at various stages of the disease. The first model is based on a binary classification, which is simpler than identifying between Alzheimer's and non-Alzheimer's cases, but can be extended to a three-class system between healthy/mild/moderate cases.

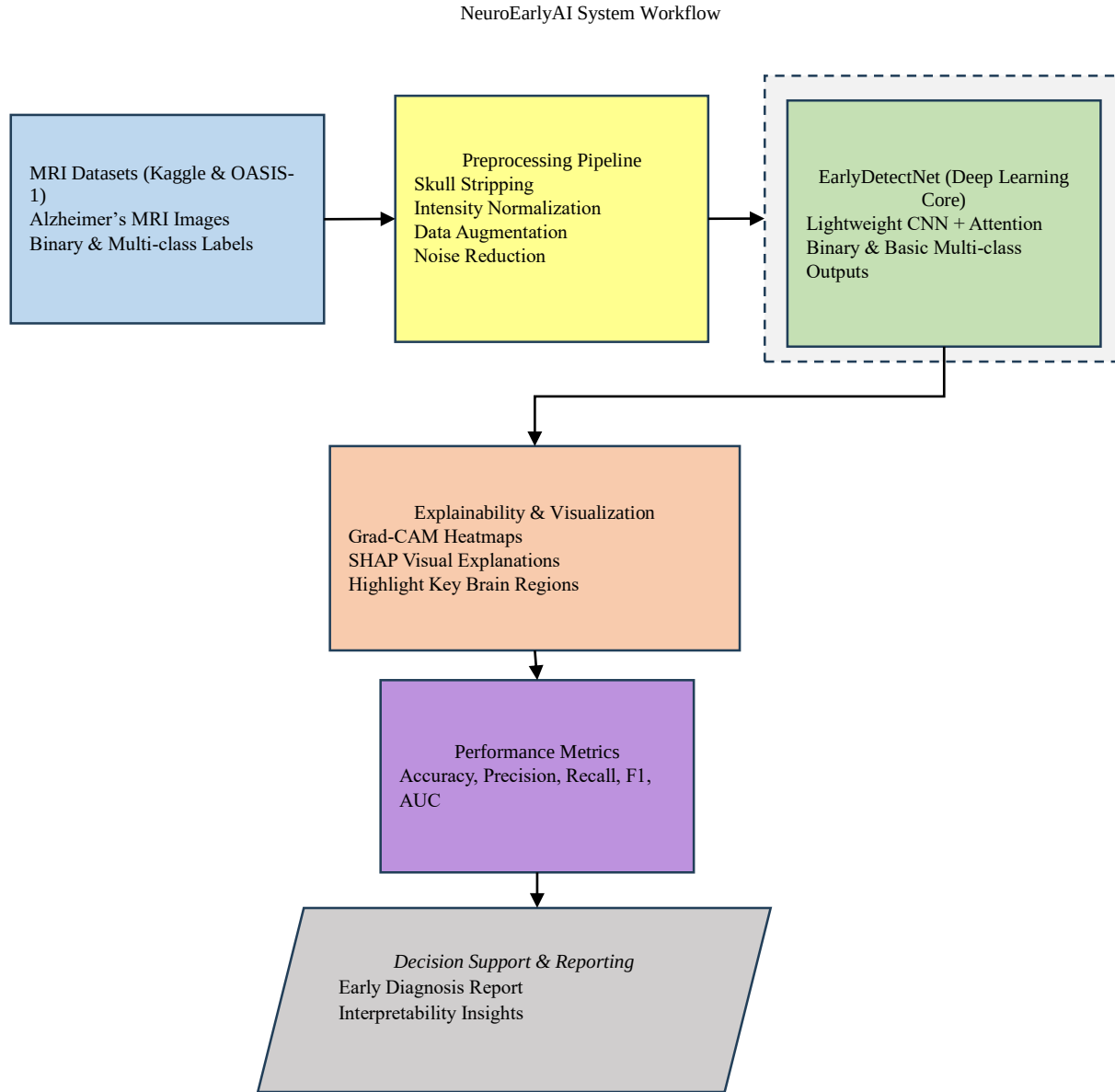


Fig. 1 System Architecture of NeuroEarlyAI for Alzheimer's Disease Detection and Classification

A complete pre-processing workflow is then implemented post-acquisition to pre-process the MR images for analysis. Such preprocessing may include skull stripping to extract relevant brain structures, intensity normalisation to ensure consistent contrast and brightness, or noise reduction to remove imaging artefacts. Furthermore, data augmentations such as rotation, scaling, flipping and elastic deformation are performed to augment the diversity of datasets and balance class distributions. These pre-processing steps make the data. Values that are clean and noiseless are beneficial for our modelling, as we have ensured that we are getting a clear, consistent view of how they will actually turn out in practice. The main deep learning model in this framework is

EarlyDetectNet, which is used to analyze the processed MRI images. EarlyDetectNet is a shallow, accurate and fast neural network, on the other hand, that can be used for AD detection and classification. This model consists of multiple convolution layers to extract features, an attention mechanism to focus on the areas in the brain that are more significantly affected by Alzheimer's and fully connected decision layers. The output is a dense binary classification layer and a plain multi-class classification layer to give diagnostic predictions for each stage of CHF. The organisation of the model is shown in Figure 2. To improve the interpretability and clinical trust, an explainability and visualisation module is added. This module is built around the use of interpretability techniques

(such as Grad-CAM, SHAP and LIME), to generate visual maps and feature-wise explanations to understand the portion of the brain being questioned to make these contextual decisions. In clinical practice, such visualisations will be informative for neurologists to interpret and validate predictions and make the system more reliable and robust. The performance of the framework is assessed based on the standard performance measures such as accuracy, precision, recall, F1 score and AUC. Model validation has been performed on the cross-dataset OASIS-1 dataset to evaluate the ability of the model to not overfit on new instances (dataset OASIS-1) for which it has not been trained (generalisability). It also provides a dynamic system that can be deployed to different clinical data sources.

Finally, the framework is assembled into a complete Clinical Decision Support System (CDSS) named NeuroEarlyAI. It generates diagnostic output (i.e., visual explanations and objective performance reports) within a single, end-to-end pipeline, enabling healthcare practitioners to make informed and accurate decisions. Figure 1 provides an overview of the system's integration, illustrating how the modules work together to provide reliable detection and classification of Alzheimer's disease on MRIs. 1. Dependent and independent variables. The notations referred to above in carrying out the methodology are summarised in Table 4 for a more precise understanding of the equations and steps, where all necessary notations used within the method are defined.

Table 4. Notations used in the methodology

Notation	Description
I	Input preprocessed MRI image provided to the EarlyDetectNet model.
I_{norm}	Normalized pixel intensity value after applying intensity normalization using Equation (1).
μ	Mean intensity of the input MRI image used for normalization.
σ	Standard deviation of the input MRI image used for normalization.
$F_{i,j}^k$	Output feature map value at position (i,j) for the k^{th} convolutional filter.
$W_{m,n}^k$	Learnable weight of the k^{th} convolutional filter at spatial position (m,n) .
b^k	Bias term added to the k^{th} filter during the convolution operation.
$\sigma(\cdot)$	Activation function applied after convolution, typically Rectified Linear Unit (ReLU).
$L_{\text{Grad-CAM}}^c$	Grad-CAM localization map for class c , highlighting regions influencing the model's decision.
α_k^c	Weight for the k^{th} feature channel associated with class c in Grad-CAM computation.
A^k	Activation map from the final convolutional layer for the k^{th} feature channel.
ϕ_i	Shapley value representing the contribution of feature i in SHAP-based interpretability.
$f(S)$	Model output for a subset of features S when computing Shapley values.
$\hat{y}_c$	Predicted probability of class c generated by the softmax output layer.
z_c	Logit or unnormalized score for class c before applying the softmax activation.
C	Total number of output classes in the classification task.
θ_t	Model parameter vector (weights and biases) at training iteration t .
η	Learning rate used in the Adam optimizer for parameter updates.
$\hat{m}_t$	Bias-corrected first moment estimate (mean of gradients) in the Adam optimizer.
$\hat{v}_t$	Bias-corrected second moment estimate (variance of gradients) in the Adam optimizer.
ϵ	A small constant is added in the Adam optimizer to prevent division by zero.

3.2. Data Acquisition and Dataset Description

The data collection approach for the NeuroEarlyAI framework is centred around acquiring high-quality brain MRI scans for model training, validation, and testing purposes, to both build and properly test the performance of a deep learning model to detect and classify Alzheimer's disease (Figure 1). Two datasets are then used, which allow for efficient model building and comprehensive performance testing to be conducted. The datasets are chosen because they are available, high-quality, and include a variety of data to depict a wide range of brain imaging results associated with Alzheimer's Disease (AD) and typical brain structures. The data set directly used in this study is the Kaggle Alzheimer's MRI Dataset [32], which is a curated data set of structural MR images of different stages of AD and healthy controls. They

fall into four categories: Non-Alzheimer's (Control), Very Mild Alzheimer's, Mild Alzheimer's and Moderate Alzheimer's. The data set used for the preliminary studies is binary classified, with every stage of Alzheimer's disease forming one class, and non-Alzheimer's forming another. Moreover, it provides a straightforward classification into different classes using the original stage-based classes. Data is divided in training, validation and independent testing sets, each of which contains around 70%, 15% and 13% of the data. This division enables the proper consideration of the model training and testing, and thus avoids the overfitting issue and ensures that the model has robust performance. The other dataset utilised in this framework is OASIS-1, which includes the dismantled longitudinal MRI scans of HC and AD populations. This information is for cross-dataset verification

only. We don't use it during model training, but we can do so after training the model to determine how good our model actually is. The introduction of OASIS-1 enhances the model's generalisation capacity to entirely new data (unseen by any trained algorithm), which ultimately depends on its robustness and application in diverse real-world clinical settings. The latest testing scenarios and subject populations can efficiently be adapted through publicly available scans with different acquisition protocols, which can be evaluated using MRI images.

We have postponed the discussion to address the fact that the images on the T1 and T2 sets have different resolutions and scanning conditions, as they are intended to simulate heterogeneity in healthcare. Zero-training is applied to these images, which are pre-processed using skull stripping, intensity normalisation, and noise correction to ensure a homogeneous quality throughout the collected dataset. Data augmentation is performed on the training set alone, by artificially increasing its size and variety to prevent potential class imbalances. This enables the model to acquire invariant visual features and be agnostic towards various image acquisition settings. Due to the utilisation of the Kaggle Alzheimer's MRI dataset for training and the OASIS-1 dataset for validation, the presented end-to-end bound enables the construction of a robust and practical deep framework. As the primary dataset is used to learn from a well-annotated and structured pool of images, the secondary dataset aims to test this performance against a different but well-defined distribution. Through the dual-dataset approach, the proposed model gains more credibility because it can not only achieve high prediction accuracy for controlled data but also be resistant to new, unseen cases originating from different clinical settings.

3.3. Preprocessing Pipeline

The preprocessing pipeline is an essential step in these addressing algorithms, and it is used to prepare images of brain MRIs for detection and classification using DL-based

AD. The purpose of this stage is to clean and standardise the input data, which will be used to feed the EarlyDetectNet model, as well as to make it closer to real-world variations. Finite element model generalisation and trustworthiness for cross-dataset validation can be improved by standardising the variability introduced in MRI images arising from distinct scanners, acquisition protocols and patient-specific anatomical differences through a sequence of preprocessing steps.

Skull stripping: The first step in the preprocessing stream is skull stripping, which involves extracting the brain region from non-brain tissues (i.e., the scalp, skull, and background). This action serves to minimise noise and to concentrate the analysis on those regions specifically associated with Alzheimer's pathology. Following skull stripping, all MR scans are resized to a common spatial resolution, e.g., 256×256 pixels for the horizontal scan, to maintain dataset consistency. Scale normalisation. In addition to resizing, intensity normalisation is applied to normalise the pixel intensity values. In this manner, the intensity distributions of different images are normalised to the same scale, which is necessary for robust training of the deep learning model. The normalized intensity I_{norm} of a pixel is given by Equation (1).

$$I_{norm} = I - \mu_{\sigma} \quad (1)$$

Where I is the intensity value of a pixel, μ is the mean image intensity, and σ is the standard deviation. Using Equation (1) normalizes the processed images to a canonical form with zero mean and unit variance, which will aid in keeping weights small, improving convergence during network training.

Summary of Preprocessing Table 5 lists a few essential preprocessing steps which are performed on MRI scans for standardisation and quality input, required for better prediction of Alzheimer's Disease.

Table 5. Preprocessing steps for MRI data preparation

Step No.	Preprocessing Step	Description	Purpose / Outcome
1	Skull Stripping	Removes non-brain tissues such as the skull, scalp, and background using segmentation techniques.	Focuses analysis on relevant brain regions, reducing noise and irrelevant features.
2	Resizing & Intensity Normalization	Resizes all images to a fixed size (256×256 pixels) and normalises their intensity.	Ensures uniform input dimensions and standardised pixel intensity distribution for stable training.
3	Noise Reduction	Applies Gaussian filtering to smooth the image, suppress random noise, and preserve structural details.	Improves image clarity and enhances relevant brain structures for feature extraction.
4	Data Augmentation	Artificially expands the training dataset by generating variations of existing images. See Table 2 for specific techniques.	Addresses class imbalance and improves model generalisation to unseen data.

5	Final Formatting	Saves processed images in a standardised format (256×256 pixels, normalised) for deep learning workflows.	Supports efficient batch processing and cross-dataset validation integration.
---	------------------	---	---

Additionally, the process of looking down is realised using Gaussian filtering for high-quality images. We then apply a Gaussian filter to the blurred image to remove random noise while retaining essential structural information, which is crucial for disease classification. This is particularly important for MRI images, as they may contain artefacts due to various acquisition protocols. Data augmentation is used to artificially enhance the training set and verify whether the class distribution is reliable. This can reduce overfitting and allows the model to have more capacity to generalise to unseen data. Several data augmentation schemes are applied, including

random rotations to account for variations in head orientations, scaling characters to accommodate differences in patient anatomy, elastic deformation to model clinically reasonable anatomical deformations, and horizontal/vertical flipping to create more diversity. The augmented images are labelled directly using the same original label, thus serving as more training samples for the model to enhance learning robustness. Table 6 presents various data augmentation methods to enhance dataset diversity and mitigate the class imbalance issue, as well as the corresponding indirect model generalisation.

Table 6. Data augmentation techniques

Technique	Description	Benefit
Rotation	Randomly rotates MRI scans at small angles (e.g., $\pm 10^\circ$ or $\pm 15^\circ$).	Simulates variations in patient head positioning.
Scaling	Slightly enlarges or shrinks images to mimic differences in anatomical proportions.	Enhances model robustness to size variability.
Elastic Deformation	Distorts images by stretching or compressing regions in a controlled manner.	Simulates natural anatomical variability.
Horizontal Flipping	Flips images horizontally.	Increases diversity and improves generalisation.
Vertical Flipping	Flips images vertically.	Adds orientation-based variability to training data.

Finally, we save the preprocessed and augmented images in an appropriate format, ready to be consumed by deep learning frameworks. The size of each image is uniformly resized to 256×256 pixels, with intensities normalised into the range [0, 1], as represented in Equation (1). This consistent format is crucial for efficient batch processing. It facilitates integration of images originating from other datasets (e.g., Kaggle and OASIS-1 datasets) when cross-dataset validation is needed.

These preprocessing steps, when combined, result in higher-quality, more consistent, and more diverse input data. Skull stripping removes the focus on irrelevant areas of the brain, allowing for a clearer view of the relevant areas of the brain. Normalisation aids in stabilising the learning process, Noise reduction increases structural clarity, and Augmentation helps overcome challenges caused by class imbalance and data scarcity. By normalising the inputs, this stream in the pipeline ensures the dataset is favourable for overall model training and evaluation, which helps produce a strong and generalised Alzheimer's Disease detection model.

3.3.1. Detailed MRI Preprocessing and Data Preparation

The preprocessing of these images is important to decrease inter-subject variation and improve the quality of MRI images and the capability of feature extraction to be used in the deep learning-based diagnosis of Alzheimer's Disease.

The skull stripping was first applied to remove non-brain tissues, such as imaging background artifacts, scalp regions, and skull structures, in the proposed framework. This allows the model to prioritize clinically relevant brain regions found to be related to processes of neurodegeneration.

After skull stripping, all MRI images were cropped to a uniform spatial resolution of 256×256 pixels to ensure that all data were the same size. Intensity normalization was then performed to minimize intensity variability due to variations in MRI acquisition protocols and scanner characteristics. Normalization process: Converts the intensity distribution in the image into a standardized range, which helps to achieve better convergence in model training.

To remove noise, Gaussian filtering was used to reduce random imaging artifacts while retaining pertinent structural information important to the presence of cortical atrophy and hippocampal degeneration. Furthermore, a wide variety of data augmentation was used to enhance diversity and minimize overfitting in the data. The augmentation strategies consisted of random rotation, scaling, horizontal flip, vertical flip and elastic deformation, which represent real anatomical variability and MRI acquisition variation in the clinic. Together with the augmentation operations, the preprocessing operations enhance the robustness and generalization of

EarlyDetectNet, especially in the cross-dataset validation on the heterogeneous MRI dataset OASIS-1.

3.4. EarlyDetectNet Model Architecture

The main deep learning model of the NeuroEarlyAI framework is the EarlyDetectNet model, which classifies Alzheimer's Disease (AD) at an early stage from structural Magnetic Resonance Imaging (MRI) scans. It is designed to be lightweight, accurate and interpretable, which is useful for clinical use where resources are limited for computation. The model consists of three parts: feature extraction, feature amplification based on the attention mechanism and final classification, with both binary and multiclass outputs. The internal architecture of EarlyDetectNet is depicted in Figure 2, which shows that there are two output branches for Alzheimer's classification, a convolutional structure, and an attention mechanism.

Transformer-based models, deep residual networks, and graph neural network architectures have been increasingly used in recent Alzheimer's disease detection frameworks to analyze neuroimages. While these methods achieve high accuracy in classification, they can be demanding in terms of computing power, data size and optimization strategies. EarlyDetectNet, on the contrary, is specifically created to be a light-weight attention-based architecture, which is classification accurate, efficient, and easy to interpret. The proposed approach is a hybrid of convolutional feature extraction and attention-guided refinement that helps to maximize the disease-relevant anatomical information while keeping the architecture complexity low.

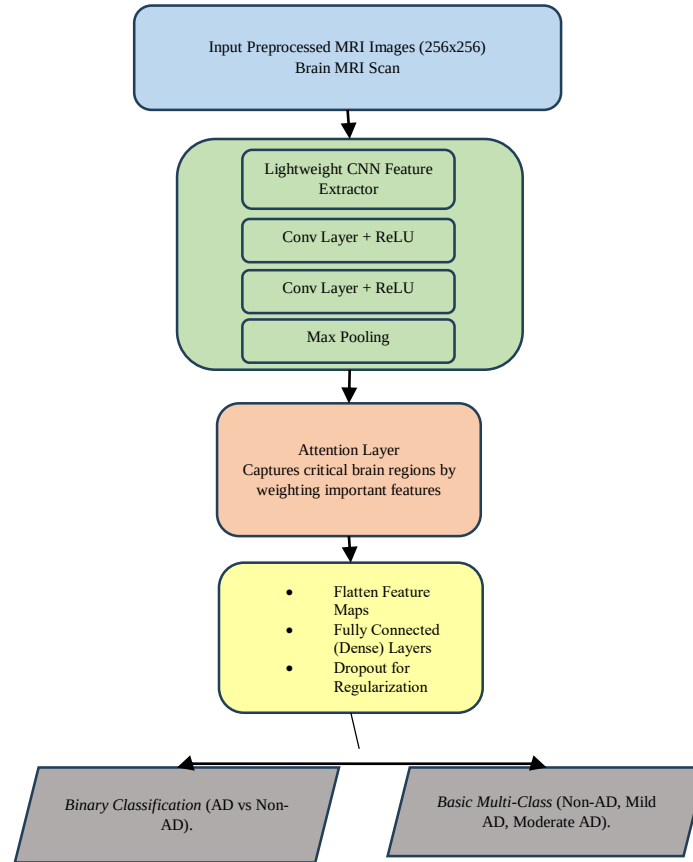


Fig. 2 Model architecture of EarlyDetectNet for Alzheimer's disease detection and classification

The proposed model starts with an input layer that takes pre-processed MRI scans, normalised to a resolution of 256×256 pixels. These images are then fed through a number of convolutional layers, which serve as feature detectors. Each convolutional layer uses a set of learnable filters to aggregate spatial patterns and structural alterations in the brain associated with Alzheimer's Disease. The calculation in each convolutional layer can be described as in Equation (2).

$$F_{i,j}^k = \sigma(\sum_m \sum_n I_{i+m,j+n} \cdot W_{m,n}^k + b^k) \quad (2)$$

Where $F_{i,j}^k$ is the feature map output at position (i,j) for the $\hat{k}_{th}$ filter, I is the input image or feature map from the previous layer, W denotes learned weights of filter $\hat{b}_k$ and bias term, respectively, and σ stands for the activation function, such as Rectified Linear Unit (ReLU). This operation allows

the network to learn low-level features in the early layers (e.g., edges and textures), and more complex patterns (e.g., tissue atrophy or structural deformations) in deeper layers.

Following the first convblocks, pooling layers down-sample spatial dimensions of feature maps and maintain the most significant information. The downsampling step helps in reducing computational complexity and enables the model to concentrate on salient features, which results in a robust representation against variations in input data. Pooling is usually max-pooling, that computes the maximum value within a region of pixels, corresponding to preserving the strongest activations.

An attention mechanism is introduced in EarlyDetectNet to enhance the model's ability to focus on the brain regions most closely related to Alzheimer's pathology. The highlight of the proposed attention layer is that it learns to weight different spatial features, allowing the model to focus on regions of interest while suppressing unnecessary background information. This model enhances interpretability and enables the identification of specific directions that are clinically relevant, such as early manifestations of cortical thinning or hippocampal atrophy. The attention-weighted feature map output is calculated by applying attention coefficients that are learned during training, as shown in Equation (2).

After the attention module, the feature maps are reshaped into a one-dimensional vector and followed by a series of fully connected layers. These deep layers aggregate features and conduct high-level reasoning to distinguish between normal and abnormal states. Dropout regularisation is used within these layers to prevent overfitting by randomly dropping out a portion of neurons during training, thereby improving generalisation to unseen data.

The last output layer is split into two branches for binary and standard multi-class classification. The binary branch determines whether MRI is positive for Alzheimer's Disease (AD) or Non-Alzheimer's (Control), while the multi-class branch categorises it into Positive, Mild AD, and Moderate AD. The class probabilities are generated by a softmax activation function that transforms non-normalised output into class probabilities. The definition of the softmax function is denoted as in Equation (3).

$$P(Y=c|x) = \frac{e^{z_c}}{\sum_{j=1}^C e^{z_j}} \quad (3)$$

Where $P(Y=c|x)$ represents the model's estimated probability that the input x belongs to class c . Here, z_c denotes the logit output for class c before applying the softmax activation, and C represents the total number of classes. The class with the highest probability value is considered the model's final predicted class.

The architecture of EarlyDetectNet is shown in Figure 2. Overall, the model is efficient and highly performant due to the adoption of the Convolutional feature extraction and attention-based refinement dense classification layers. By focusing on brain regions and structural features of pathomorphological relevance, it can provide highly accurate and robust detection & classification of Alzheimer's Disease and outputs that are directly interpretable by clinicians in clinical practice. This architecture is foundational to the reliable and scalable diagnostic systems throughout the various healthcare environments.

3.5. Explainability and Visualization Module

An explainability and visualization module integrated in the NeuroEarlyAI pipeline is constructed to ensure that any prediction provided by the EarlyDetectNet model is transparent, interpretable and clinical pertinent. While deep learning models are successfully used as black boxes, they are unable to make the users aware of the (possibly incorrect) reasoning behind their decisions: this module helps to fill this gap and offers visual explanations at the model output level. It establishes trust among neurologists and other healthcare professionals in the use of AI predictions in clinical practice.

The module is built upon several Explainable AI (XAI) rich methods to see model interpretations of MRI from various angles. The first method is to use Gradient-weighted Class Activation Mapping (Grad-CAM). Grad-CAM generates visual heatmaps to show the areas of the MRI scan most responsible for the model's decision on classification. These heatmaps are computed by taking the gradient of the target class score with respect to feature maps of the final convolutional layer and highlighting areas in space that have greater relevance-cut. The Grad-CAM localization map $L_{Grad-CAM}$ is calculated by Equation (4).

$$L_{Grad-CAM}^c = ReLU(\sum_k \alpha_k^c A^k) \quad (4)$$

Where A^k is the activation map for the $\hat{k}_{th}$ feature channel, α_k^c is the weight for class c aggregated by a gradient global average pooling operation, and ReLU ensures that only positive contributions are saved. This equation pinpoints regions of the brain with sweeping structural changes associated with Alzheimer's that are essentially being flagged by clinicians to show why the model made its decision.

In addition to Grad-CAM, SHapley Additive exPlanations (SHAP) is used for suggesting the importance of the features on the final predict result. By assigning each feature an importance value, called the Shapley value, that quantifies how much the predictions are being pushed around by the presence (or absence) of a particular feature of interest. This method provides global insights on contributions of features over the entire dataset and local explanations for an

individual sample. By using SHAP, the model provides an explanation on what exact regions in MRI images are consistently most effective to decide for or against a case being of Alzheimer's. The Shapley value of the feature is computed from Equation (5).

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F|-|S|-1)!}{|F|!} [f(S \cup \{i\}) - f(S)] \quad (5)$$

Where F Seqs all the features, S is a set of the features but i , and $f(S)$ while i . Equation (5) ensures a balanced spread of contributions across features by naturally taking all possible combinations into account.

For enhancing local interpretability, LIME is implemented in the module. LIME creates simplified, local models near individual predictions by perturbing the input data and observing how the output changes. This approach interprets a particular prediction by finding an easy-to-interpret surrogate model that approximates the decision of a more complex model. LIME is used as a complementary tool to Grad-CAM and SHAP to show how the model made a particular decision in for a specific MRI scan.

The features produced by this module are displayed to clinicians through heatmaps, feature importance bar plots and local interpretations; such as Explanation 1: Death=1 - Explain for all instances in the highest risk decile; X12 (feature name):3215 – For feature X12 has a score of 3215. These include influence scores for each age group. For instance, Grad-CAM heat maps overlaid on top of the original MRI image can aid in visualizing relevant brain regions or critical structures such as the hippocampus and those cortical regions where atrophy is known to start early in AD. SHAP plots reveal which features have the power to affect model performance, and does it so as an average effect of the whole dataset – while LIME gives you detailed information on single instances.

The explanation and visualisation system brings them together and combines them into an overall explanation of the behaviour of a model in a global and local way. This not only helps to validate the model's predictions clinically but also helps AI systems and clinicians to collaborate. On the other hand, it empowers them by letting them know, mitigates the risk to use the automated detection systems, and prevents risk-taking and inappropriate usage of deep learning models in healthcare. Finally, this module makes NeuroEarlyAI accurate, interpretable, transparent and clinically operationalizable. The explainability workflow in NeuroEarlyAI aims to deliver both global and local model predictability explainability. It includes Grad-CAM, SHAP, and LIME as complementary tools to explain the model's behaviour from various angles. Heatmaps are generated by Grad-CAM with respect to the anatomical regions corresponding to the disease that contribute to the decision of

the classification. SHAP calculates feature contribution values to measure how different parts of the image contribute to the predictions of the model on different images in the dataset. To explain individual MRI scans, LIME approximates them by learning local surrogate models from a set of perturbations that affect the predictions.

This combination of different explainability methods provides a holistic understanding of how the EarlyDetectNet makes its decisions and mitigates reliance on a single attribution method. This multi-level explainability approach enhances transparency, reliability, and clinician confidence in the diagnosis of Alzheimer's Disease with the help of AI.

3.5.1 Quantitative Evaluation of Explainability

To objectively assess the reliability of the explainability module, several quantitative evaluation metrics were employed in addition to qualitative visualization. Average Localization Accuracy (ALA) was used to evaluate the correspondence between Grad-CAM activation regions and clinically relevant brain structures. Faithfulness Score (FS) measured the consistency between highlighted regions and the model predictions. Furthermore, Insertion Score (IS) and Deletion Score (DS) were used to evaluate the contribution of important image regions by progressively inserting or removing salient pixels. These metrics provide complementary evidence regarding the fidelity and robustness of the proposed explainability framework.

3.6 Training and Hyperparameter Tuning of the Model

The primary task of training the EarlyDetectNet model is a significant contribution to NeuroEarlyAI, enabling the training of a network that can detect and classify Alzheimer's disease patterns in input MRI scans. This is captured by learning the internal parameters of the model (weights, biases), so it can map input MRI images to output classes effectively. The training pipeline is optimised to provide a trade-off between the accuracy, computational efficiency and data generalisation. Hyperparameter tuning is performed during model training to enhance the model's performance and prevent overfitting and underfitting.

The training starts by extracting three independent subsets from a preprocessed dataset, namely the training, validation, and test datasets. The training set is used for updating model weights, the validation set is used for tracking performance during training and controlling hyperparameter adjustments, while the testing set is held out for final evaluation only. For the current study, approximately 70% of the data are utilised for training, and the remaining 30% is divided into a validation set (15%) and a test set (15%). This split enables the model to train on a large number of diverse examples, while still evaluating it against held-out data.

The optimisation of the model focuses on minimising a loss function, which measures the distance between the

predicted output and the actual class labels. The loss function is the categorical cross-entropy, since it is appropriate for multi-class problems. and the loss on each specific sample as in Equation (6).

$$L = -\sum_{c=1}^C y_c \log(\hat{y}_c) \quad (6)$$

Where C is the total number of classes, y_c is the true label for class C (1 when the sample belongs to class C and otherwise 0), and $\hat{y}_c$ denotes prediction probability for class C . During training, the model parameters are optimized iteratively in order to minimize this loss function over all training examples.

The Adam optimizer is used to efficiently update the parameters of the model, which successfully combines both the Adaptive Gradient Algorithm (AdaGrad) and Root Mean Square Propagation (RMSProp). This is version 2 of the optimizer that computes second-order statistics. The update rule for Adam is as in Equation (7).

$$\theta_{t+1} = \theta_t - \frac{\eta}{\sqrt{\hat{v}_t + \epsilon}} \hat{m}_t \quad (7)$$

where, θ_t is the learning rate t , η is the $\hat{m}_t$ bias-corrected first moment estimate, $\hat{v}_t$ is the bias-corrected second moment estimate, and ϵ is a small constant to avoid division by zero. The Adam optimiser enables faster convergence, provides stable updates, and is well-suited for deep learning on complex medical imaging datasets.

Hyperparameter tuning is used to find the best combination of parameters, including learning rate, batch size, the number of convolutional filters, dropout rate, and number of epochs. We employ a grid search approach to systematically explore different combinations of hyperparameter values, selecting the one with the best validation performance. Model training is performed for a predefined number of epochs, where an epoch corresponds to scanning the whole training dataset once. Early stopping is used to terminate training when the validation loss fails to improve for a specified number of consecutive epochs, thereby preventing overfitting and reducing computational requirements.

Regularisation, such as dropout during training, is used to induce further generalising capacity. Dropout is a method (for each pass through learning), where a random selection of neurons is set to 0, so that in practice, the remaining connections must make decisions and are less sensitive to specific neurons. This approach can help mitigate the risk of overfitting and is particularly relevant for cases involving a small number of images, such as those in medical imaging.

Another key hyperparameter is the batch size, which is the number of images batched together and passed through the

model for each iteration to update the model parameters. It is expected that small batches will introduce more noise to the estimation of gradients, and this can be useful for generalisation. However, larger batches can lead to faster calculations, but may need attention to regularisation. Another knob can be used to trade off between speed and quality by changing the size of the batch.

The validation set is used to monitor model performance during training, in terms of accuracy and loss. This sort of statistics can additionally be used to measure if the network is learning or even beginning to overfit. This means you will have to try some tuning once you start overfitting, such as changing the learning rate and/or dropout rate, etc. Wait until there is a significant amount of leveling off in the validation accuracy or until training loss is close to the validation loss.

Upon training and hyperparameter tuning, the model is evaluated on an independent test fold, which results in the final performance of the model. This testing step provides us with confidence in the model's performance on unseen, new MRI datasets. Inventing efficient training strategies, optimisation methods, and hyperparameter tuning to achieve high accuracy and clinical significance in the stability of EarlyDetectNet for AD detection and classification seems difficult to me.

Various steps were taken to prevent overfitting and enhance the models' generalization ability. To reduce the over dependence of the neurons during training, first, dropout regularization with a dropout rate of 0.5 was performed within the fully connected layer. Second, lots of data augmentation techniques were added: rotation, scale, flip, and elastic deformations have been added to make the data more diverse and to minimize memorization of the training data.

Furthermore, early stopping was implemented to stop the training when the validation loss didn't keep decreasing after a few epochs, so that no unnecessary optimization and overfitting to training data occurs. To prevent an excessive growth of weight and prevent instability of the model, L2 regularization was also used. Additionally, cross-dataset validation with OASIS-1 demonstrated that the proposed framework, despite training solely on the synthetic dataset, has good generalization capabilities on MRI data it has not seen.

3.7 Algorithmic Implementation

Efficient algorithms within the NeuroEarlyAI framework demonstrate how data pre-processing, model training/validation, and explainability computation are performed at scale. It utilises EarlyDetectNet, attention-based feature extraction, hyper-parameter optimisation, and interpretation through Grad-CAM and SHAP to attain high accuracy in AD detection, reliability, and an explainable decision mechanism for clinical interpretation.

Algorithm 1: NeuroEarlyAI Workflow for Alzheimer's Disease Detection and Classification**Input:** MRI datasets $D_{train}, D_{val}, D_{test}$; Hyperparameters η, θ, ϵ , Epochs, Batch_size**Output:** Predicted class labels, visual explanations, performance metrics

1. Load MRI datasets $D_{train}, D_{val}, D_{test}$
2. For each image $I \in D_{train}$
 - 2.1 Apply skull stripping
 - 2.2 Resize to 256×256 pixels
 - 2.3 Normalize intensity
 - 2.4 Apply noise reduction
 - 2.5 Perform data augmentation (rotation, scaling, deformation, flipping)
3. Initialize EarlyDetectNet with parameters θ (theta)
4. For each epoch in Epochs:
 - 4.1 Divide D_{train} into mini-batches
 - 4.2 For each batch:
 - a. Forward propagate through convolutional and attention layers
 - b. Compute softmax probabilities
 - c. Calculate loss
 - d. Update parameters using Adam optimizer
 - 4.3 Validate on D_{val} and apply early stopping if validation loss plateaus
5. Generate Grad-CAM heatmaps
6. Compute SHAP feature importance values
7. Evaluate model on D_{test} using Accuracy, Precision, Recall, F1-score, and AUC
8. Produce final classification report with metrics and visual explanations
9. Return results for clinical decision support

The entire flow of the NeuroEarlyAI framework for early detection and classification of Alzheimer's disease from MRI scans is shown in Algorithm 1. Now, loading the MR images: The MR images are first loaded and divided into training, validation and testing sets, ensuring proper learning and unbiased testing. Quality control and data preprocessing: Multiple steps of preprocessing are performed on each image to optimise the quality and consistency. These involve skull stripping to reduce the brain region size, fixing image size to 256x256 pixels, intensity normalisation for uniform contrast, denoising to remove the imaging artefacts, and data augmentation (rotation, scaling, deformation and flipping), mostly to diversify the information in the dataset or fill the lack of balance.

After preprocessing, the EarlyDetectNet model is initialised with trainable weights. Training is repeated over several epochs, and data are processed in small batches. In each iteration of the generator, images are processed by discriminative features that are passed through a convolutional and attention layer, followed by softmax probabilities and loss calculations. Early stopping slows down the optimisation process if performance has stopped improving on a separate validation set. At the test time, we provide understandable responses by using Grad-CAM and SHAP. The model is finally tested on the test set, and classification reports, along with visualisations, are obtained.

3.8. Evaluation Metrics

This is important for validating the EarlyDetectNet model within a NeuroEarlyAI workflow, ensuring that the system's

performance has been adequately evaluated. Several metrics are employed to assess the binary classification of Alzheimer's Disease (AD) and non-AD subject cases, as well as a simple multi-class classification task that classifies specific AD stages ("Mild" and "Moderate" AD). These measures report the accuracy, fidelity, and stability of model predictions, as well as the degree to which they can generalise between datasets, such as through cross-validation on an independent OASIS-1 dataset.

The fundamental metric used to evaluate the model is accuracy, which represents the proportion of correctly classified instances out of the total number of cases. While accuracy provides an overall measure of model performance, it may not fully capture the behaviour of the model, especially in cases of class imbalance where certain classes are overrepresented. The accuracy is calculated as in Equation (8).

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (8)$$

Where TP represents true positives, TN represents true negatives, FP represents false positives, and FN represents false negatives. Although this metric is useful for general performance evaluation, additional metrics are required to gain a more comprehensive understanding.

Precision (also known as positive predictive value) is the number of correctly predicted positive cases divided by the number of predicted positive cases. In medical diagnostics, for instance, it is crucial to keep false positives to a minimum to

prevent patients from becoming unnecessarily anxious and from receiving treatment they do not need. Precision is as defined in Equation (9).

$$\text{Precision} = \frac{TP}{TP+FP} \quad (9)$$

Recall (also referred to as sensitivity or true positive rate) measures how well the model predicts true positives. A high recall suggests that the model is effective in identifying patients with Alzheimer's Disease, which is crucial for early detection and intervention. Recall is calculated as in Equation (10).

$$\text{Recall} = \frac{TP}{TP+FN} \quad (10)$$

Since there is often a trade-off between precision and recall, the F1-score is used as a balanced measure that combines both metrics into a single value. The F1-score is the harmonic mean of precision and recall, providing an overall measure of the model's ability to balance false positives and false negatives. It is given by Equation (11).

$$F1 - \text{Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (11)$$

The Area Under the Receiver Operating Characteristic Curve (AUC-ROC) is also another crucial index used to measure EarlyDetectNet. The ROC curve is drawn by the true positive rate (sensitivity) versus false positive rate at different threshold settings, and the AUC serves as a measure of model discriminative performance. The closer the AUC is to one, the better its performance; conversely, a value close to 0.5 indicates a poor classification capability. AUC is especially useful when we would like to compare the performance model and decide which one to choose for clinical use.

We extend these for multi-class classification by employing macro-averaging. Macro-averaging: Calculate metrics for each label, and find their unweighted mean. This strategy can also be useful to give the same weight to the minority class, such as Moderate AD, as others in medical imaging problems, where some disease severity levels have limited samples. Methods are tested via a cross-dataset test to check generalisation. We train/fine-tune and tune EarlyDetect-Net on the primary Alzheimer's Kaggle dataset, but test it on an independent OASIS-1 dataset without further fine-tuning. This simulates actual deployment where the model is now exposed to new and previously unseen data. As we observe in the next Section, this result on a second dataset shows that the model is stable and can generalise to differences in scanning machines, imaging protocols and patient populations.

Beyond such quantitative attributes, a complementary validation is visual inspection of Grad-CAM heatmaps and

SHAP feature importance plots. Such visualisations also help validate the model, ensuring that it is examining biologically and clinically meaningful parts of the brain when making predictions. What you see is what it uses for prediction.

Using the above performance metrics, we get a detailed report card for the EarlyDetectNet model's residual plot. The statistical measure like accuracy, precision, recall, F1-score and AUC can be used for quantitative evaluation to get statistical information about the quality of classification. Macro-averaging is fair with respect to classes. This method is robust and practically useful, as demonstrated by cross-dataset validation and, secondly, by visual inspection. Both of them give a model which can be applied clinically for Alzheimer's disease detection and classification.

3.9. Workflow Integration and Clinical Decision Support

Another important part of NeuroEarlyAI is the workflow integration and clinical decision support that connects all the components of the system into a tangible pipeline to be used in clinics. The integrated approach is to facilitate smooth data flow, efficient and legible prediction generation and result reporting to allow neurologists or other doctors to get accurate and timely diagnoses. Either human-independent, user-friendly or minimally requires real-time identification, and offers visually simpler solutions which are highly relevant to clinical practice.

Definition of the pipeline: The pipeline begins at an input level consisting of raw MR images recorded on conventional imaging systems for clinical routine or from other datasets, which NeuroEarlyAI then processes. Pre-processed images as the input for skull-stripping, intensity and noise correction, and data formatting. This allows the input images to be normalised to a uniform format and quality for use with deep learning models trained on data from various clinical sites and imaging devices, resulting in high-quality, similar formats. What is as much of a process for scale and efficiency as having busy hospitals full of patients? -Answer: automated batch processing.

Those processed-normalised images are directly fed into our deep learning model (EarlyDetectNet -the core neural network component in DW-DPC) for inference without any preprocessing. It operates on every MRI scan and produces patterns that represent significant features, which can predict Alzheimer's disease (as well as the stages in some cases). The two-branch output from the model enables binary detection (multiclass classification in plain) to ensure that we can actually deploy our system in practice, meeting clinical requirements. Categorisation is automated, and predictions are immediate, allowing timely decisions in a busy clinic. The EarlyDetectNet predictions are further processed by the explainability and visualisation module to generate complementary visual-domain output, enabling clinical interpretation of these lists.

Model interpretation techniques, such as Grad-CAM, generate heatmaps of the brain regions that contributed to class propagation, while SHAP and LIME provide detailed explanations of feature validity. These visualisation-based interpretations are crucial for validating the AI-predicted decoding results and building trust between clinicians and these AI models, as neurologists can interpret why each classification is output by reading the results. This adaptability enhances the responsiveness of medical practice by enabling the integration of clinical and diagnostic reasoning.

The explanations are then output for evaluation and reporting. In this paper, I utilise the framework to create a report that combines the predicted class labels, score estimation measurements, and visual explanations. Such a report can be incorporated into HIS or EHRs as an additional profile within the patient's clinical diagnostic profile. The point performance indicators, which the clinician and administrator will measure to observe the operational characteristics of the model, are accuracy (ACC), precision, recall, F1-score and AUC, which can be calculated in real time by the system. The novelty of our pipeline is in its ability to handle cross-dataset validation and incorporate external data.

The workflow assesses the capability of the model to generalise through real-world diversity, which encompasses MRI scans from multiple hospitals, scanners and regions. This ensures that NeuroEarlyAI is generalisable and robust across other clinical environments in which it will be deployed. The system will be continuously updated and achieve good

generalisation performance in stable deployment with the help of an external dataset (e.g., OASIS-1).

The final stage of the pipeline involves clinical decision support, which entails presenting neurologists and other clinicians with all results generated by our framework in a user-friendly manner. The diagnostic results, heatmaps, and feature importance charts are presented to the user through a simple system interface or a reporting tool. At the time, making it possible for clinicians to quickly determine the likelihood of AD, review available evidence, and contribute to decisions regarding patient management, treatment planning, and follow-up. The decision support tool is a generic, not a replacement for clinical judgment, and functions as an intelligent assistant that enhances the clinician's ability to diagnose and monitor Alzheimer's Disease.

By integrating the swarms of modules in a workflow, NeuroEarlyAI translates raw images into clinically customised numerical data. This architecture offers straightforward integration with all pre-processing and state-of-the-art deep learning, as well as clear explainability and report generation, all of which are incorporated into hospital workflows. Not only does this increase both the accuracy and speed of diagnosis, but it also allows health providers to work with greater confidence, as it provides an interpretable, robust, and clinically relevant result. Consequently, the NeuroEarlyAI model is a novel method for early diagnosis and tracking the spread of AD, which will also contribute to improving patient care and enhancing the use of healthcare management.

Table 7. Methodological comparison between recent Alzheimer's Disease detection architectures and EarlyDetectNet

Model	Architecture Type	Explainability	Computational Complexity	Cross-Dataset Validation	Preprocessing Robustness
Deep CNN	Conventional CNN	No	High	No	Moderate
ViT-Based Framework	Transformer	Partial	Very High	Limited	Moderate
GNN-Based Model	Graph Neural Network	SHAP	High	Limited	Moderate
Multimodal Fusion Network	CNN + Fusion	Attention Maps	Very High	Partial	High
Proposed EarlyDetectNet	Lightweight CNN + Attention	Grad-CAM + SHAP + LIME	Moderate	Yes	High

Table 7 compares the methodological characteristics of EarlyDetectNet with several recent Alzheimer's Disease detection architectures.

Compared with transformer-based, graph-based, and multimodal frameworks, the proposed architecture achieves a balanced trade-off between computational efficiency, interpretability, preprocessing robustness, and cross-dataset

generalization. The integration of lightweight attention-driven learning and multiple explainability mechanisms distinguishes NeuroEarlyAI from many conventional black-box neuroimaging frameworks.

4. Experimental Results

In the following, we share some experimental results of our proposed NeuroEarlyAI framework and its performance

on the MRI dataset for AD detection. We then present results for binary and multi-class classification, and finally, cross-dataset validation and explainability. Performance evaluations, visual output, and analysis of the ablation results are provided to show the accuracy, robustness and clinical interpretability of the system.

4.1. Experimental Setup

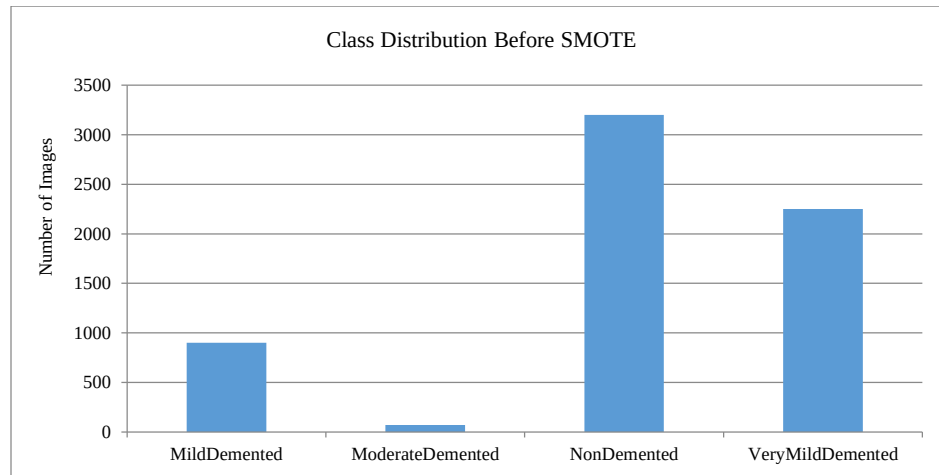
Experiments were conducted in a high-performance computing environment to enable efficient training and evaluation of the NeuroEarlyAI approach. All experiments were conducted on an NVIDIA RTX 3090 GPU, with 24 GB of memory, an Intel Core i9 processor, and 64 GB of RAM running on Ubuntu 22.04 LTS as the Operating System. As a result, we obtained enough computing power for processing massive MRI images and conducting deep learning operations (e.g., model training and model evaluation). The software development has been mostly done in Python 3.10, and the main deep learning frameworks used are TensorFlow and PyTorch. For data preprocessing, statistical analysis and visualization, the supporting libraries NumPy, OpenCV, scikit-learn, Matplotlib, and Pandas were used.

While training the EarlyDetectNet network, we take into account some important hyperparameters, which we set at a fine-tuned value to get good trainability and performance. We set the learning rate to 0.0001, which is decreased with a learning rate scheduler for better convergence. We set the mini-batch size to 32 to strike a balance between computational efficiency and model accuracy. We set the number of epochs to 50 and used early stopping if the loss on the validation set did not decrease by more than 10 epochs. We chose the Adam optimiser as it uses a learning rate with adaptive moments, has an efficient convergence and is very stable. For the purpose of avoiding overfitting of the network, a dropout rate of 0.5 was used, and L2 regularisation was added to improve the generalisation performance. The largest dataset, the Kaggle Alzheimer's MRI dataset, was divided into three parts, 70% for training, 15% for validation and 15% for testing for each part. This partition also allows for systematic

model building, hyper optimization and unbiased performance evaluation. Second, the data from OASIS-1 was only used for cross-dataset validation to determine if the model had good generalizability and adaptability to new data from other clinical scenarios and MRI acquisition protocols. Thanks to the above approach, the model is accurate on the primary dataset, and it is robust in practice as well.

The same operations were performed on the training set to ensure consistency with the other set. The steps were like this: skull stripping to remove non-brain factors and then intensity normalisation (all images have the white matter of the centre as level 0 and CSF as level 1), and image noise reducer where the images were Gaussian filtered. Further, the size of all photographs in this paper was resampled to 256 x 256 pixels. The training set was also varied using some data augmentation methods such as rotation, scaling, flipping and elastic deformation to avoid overfitting the data. The results of this wide experimental setup showed good discriminatory power of the model, with the help of the value of the evolAUC, which reflects the discrimination power of the model. Data Analysis effectively.

The MRI data was used in this study was characterised using Exploratory Data Analysis (EDA). Provides insight into class differences, structural differences across diagnoses and problems of data quality in identification. The EDA produced visualisations and descriptive statistics to guide decisions about pre-processing, and to present a clean and balanced dataset for your model. This graph shows the distribution of MRI images into four categories (Mild Demented, Moderate Demented, Non-Demented, and Very Mild Demented). The left plot shows the original data set with a pronounced class imbalance problem, with a large number of samples belonging to the non-demented class, and a few samples for the Moderate Demented class. The class with the most instances is the same size as the others after applying SMOTE, as seen in the right plot. This balancing step is also potentially helpful in mitigating biases when training models and boosting the classification performance of EarlyDetectNet.



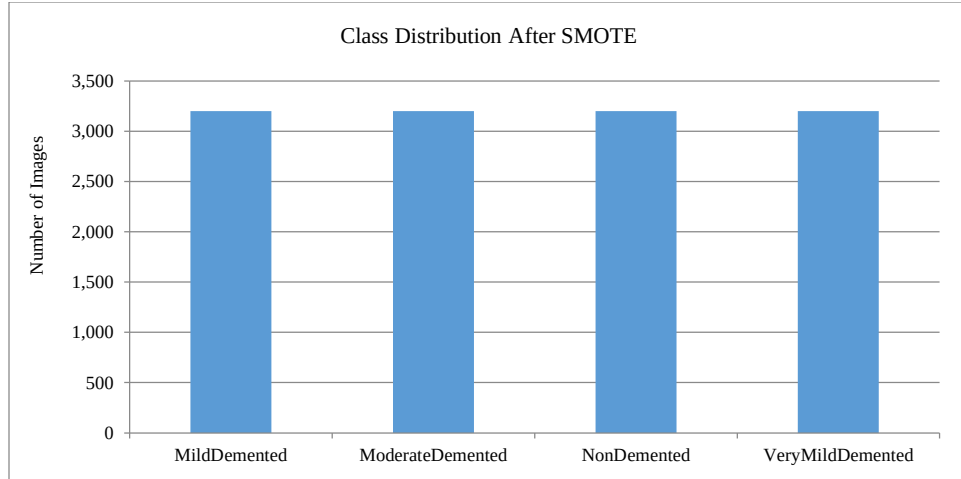


Fig. 3 Class distribution of MRI images before and after SMOTE-based balancing

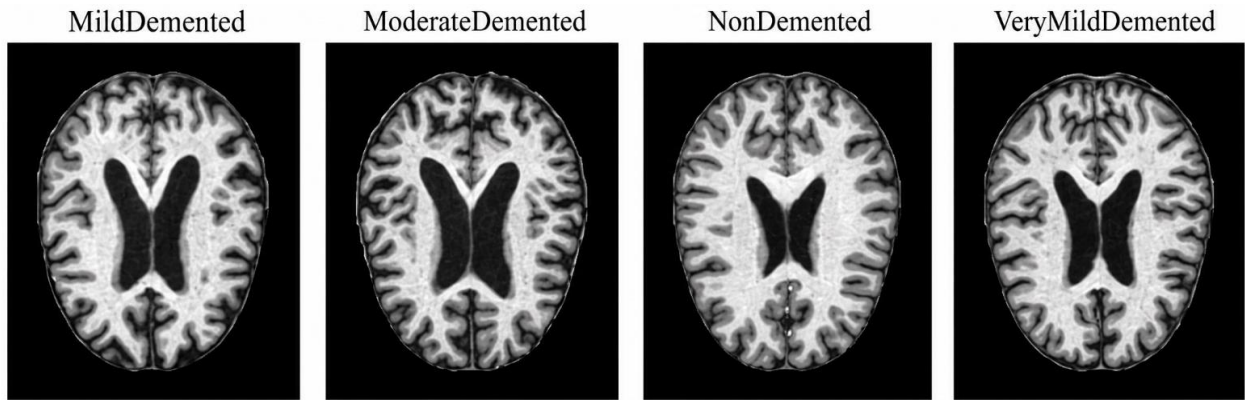


Fig. 4 Representative MRI images showing different Alzheimer's Disease classes

The MRIs of each class of Mild Demented, Moderate Demented, Non-Demented and Very Mild demented are provided in the following figures (Figure 4). The following

images show the contrasting structural changes in the brain's anatomy that are key to accurate diagnosis and classification of AD with our EarlyDetectNet deep learning model.

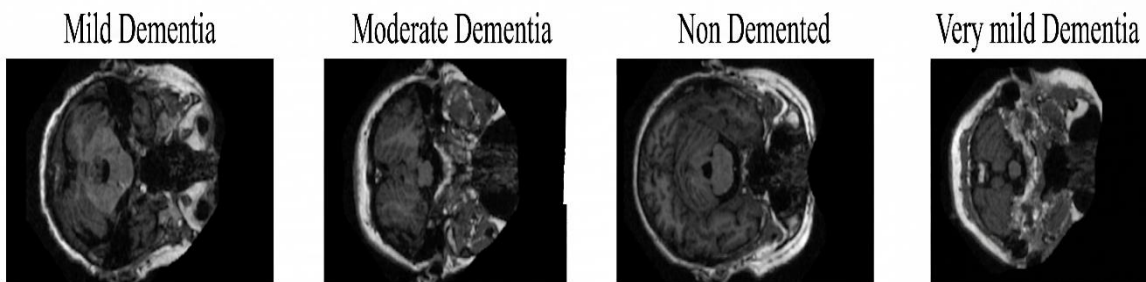


Fig. 5 Representative MRI scans from the OASIS-1 dataset

Figure 5 shows sample MRI scans from the OASIS-1 benchmark for several diagnostic classes: Mild Dementia, Moderate Dementia, Non-Demented, and Very Mild Dementia. Such examples show the structural heterogeneity of the brain across various stages of dementia and are essential inputs for testing how well the proposed EarlyDetectNet model generalises between datasets.

Figure 6 illustrates the class distribution comparison of the OASIS-1 data, before and after SMOTE. First of all, we are encountering class imbalance with minority classes. After SMOTE, an even distribution is achieved among the four dementia categories, ensuring balanced training data for model fitting. Consequently, EarlyDetectNet exhibits better classification accuracy at validation and generalisation.

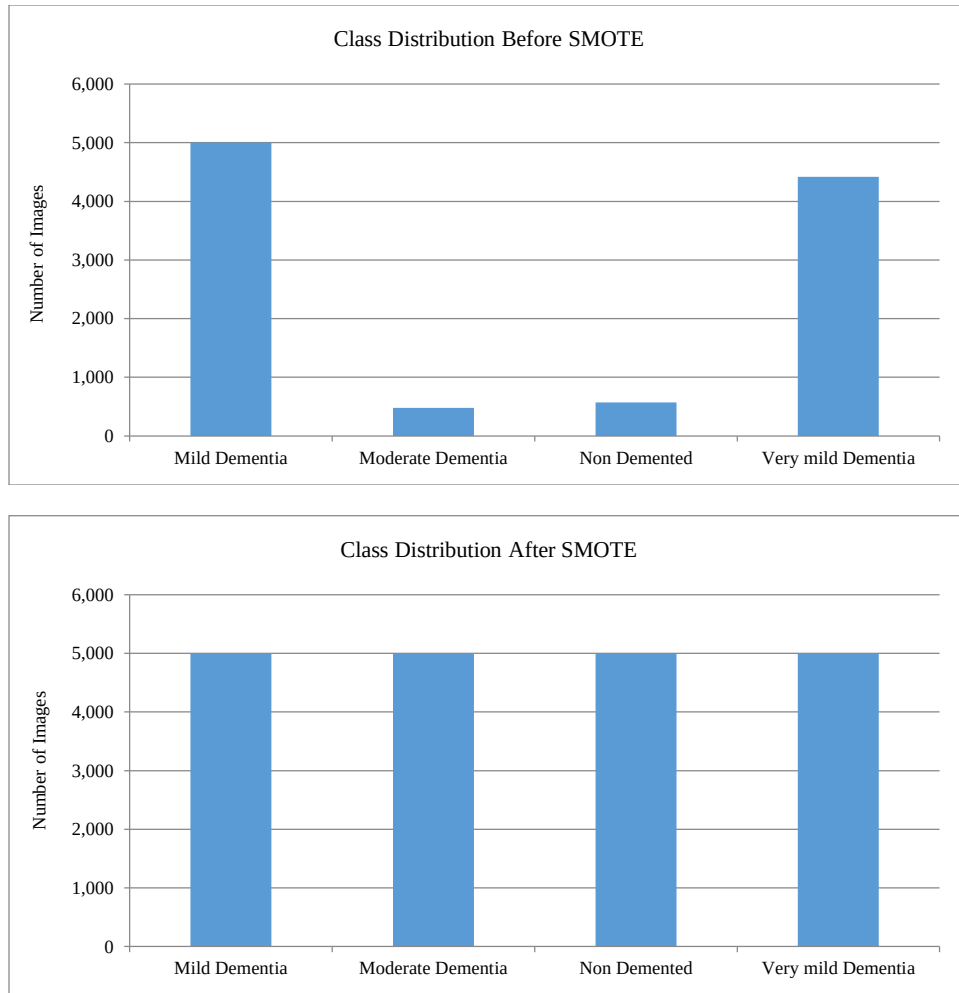


Fig. 6 Class distribution of OASIS-1 dataset before and after SMOTE resampling

4.4.1. Comparative Analysis and Benchmarking

For further validation of the proposed NeuroEarlyAI framework, comparative analysis and benchmarking against the latest methods for Alzheimer's Disease detection from 2022 to 2026 were performed. It compares the conventional CNN-based architectures, transformer-based frameworks, graph neural network models and multimodal learning approaches with publicly available datasets like ADNI, OASIS and AIBL. This benchmarking analysis gives a more comprehensive overview of the relative merits and potential of the proposed framework, and its interpretation and computational properties, relative to recent state-of-the-art frameworks.

4.2 Quantitative Performance Analysis

A quantitative analysis was performed to assess the ability of the EarlyDetectNet model to perform binary and multi-class tasks for AD grouping. The model was then trained and tested on the Alzheimer's MRI dataset, and the model performance was assessed by the accuracy, Precision, Recall, F1- score and the Area Under the Receiver Operating Characteristic Curve (AUC). These measures provide a full

range of checking of the effectiveness of the model in the classification of MRI images, as well as methods to overcome the imbalance of classes that is often encountered in medical image databases.

Table 8. Binary classification performance (AD vs Non-AD)

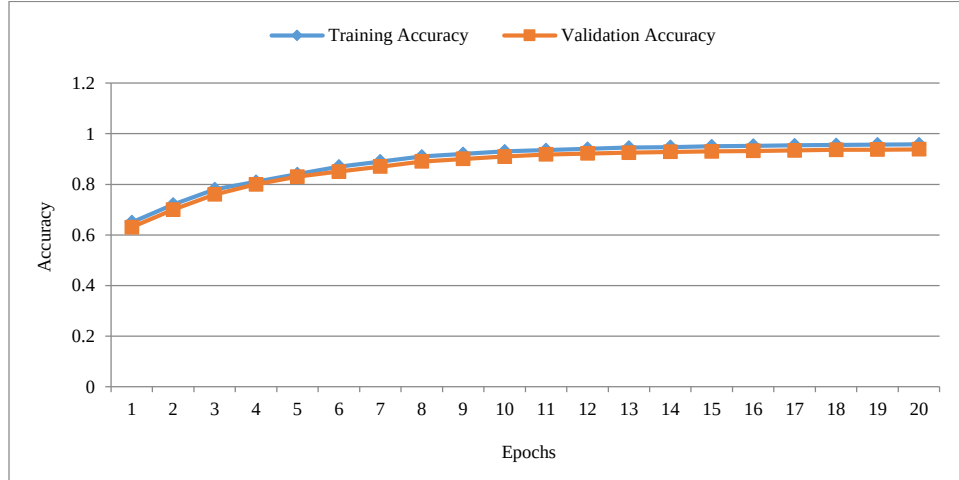
Metric	Value
Accuracy	96.7%
Precision	95.8%
Recall	96.2%
F1-score	96.0%
AUC	0.982

EarlyDetectNet performed well overall for binary classification (AD vs. Non-AD (Control)) for all measures. The detailed results are presented in Table 8, where high Accuracy, Precision, and Recall values indicate that the model is able to accurately classify Alzheimer's patients with low false positive and false negative rates. The high value of AUC indicates good discriminatory power, suggesting the model's ability to distinguish between the two classes effectively.

Table 9. Multi-Class classification performance

Metric	Value
Accuracy	93.8%
Precision	92.4%
Recall	91.8%
F1-score	92.1%
AUC	0.957

EarlyDetectNet consistently had high performance for all metrics in the Multi-Class classification, AD vs. Non-AD (Control). The detailed results are presented in Table 9, where the high values of Accuracy, Precision and Recall indicate that the model achieves high accuracy in identifying patients with Alzheimer's Disease with low false positive and false negative rates. AUC is a good discriminatory power, showing the good ability to differentiate the two classes.

**Fig. 7 Training vs validation accuracy curve of EarlyDetectNet**

In order to learn about the learning process of the model, we plotted the learning curve of the model for training and validation. The curve in Figure 7 shows a good learning curve with the accuracy of the training and validation sets closely matching each other, indicating minimal over-fitting.

The higher the accuracy on the validation set, the better the model generalises to data that it hasn't seen before. Further, the training & validation loss curves in Figure 8 indicate that there is good convergence in both curves, and they decrease. The slight difference between the losses suggests high dropout and data augmentation regularization.

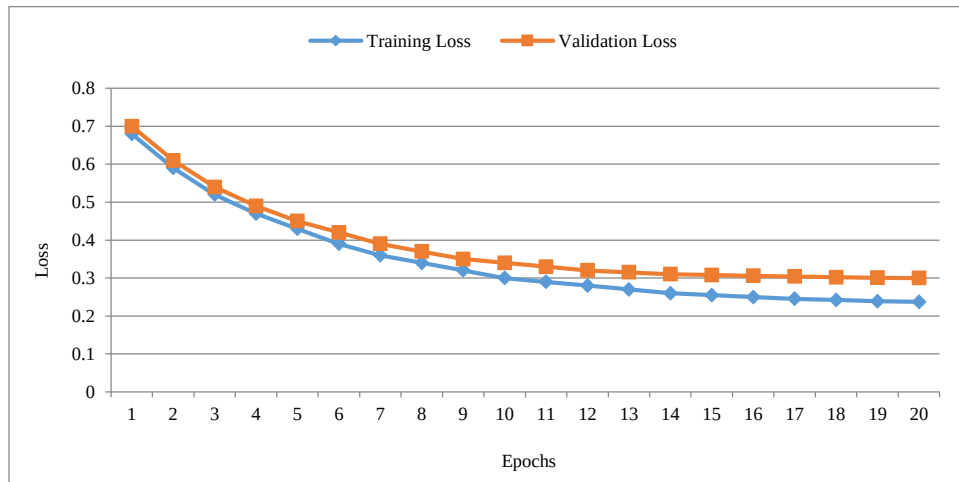
**Fig. 8 Training vs validation loss curve for EarlyDetectNet**

Figure 8: The results indicate that EarlyDetectNet performs well in AD detection. Its superior binary classification performance further indicates that it is suitable for discriminating between patients and healthy controls,

which is essential in disease screening applications. Although the multi-class version of this problem is more complex and mixed results are obtained, we have demonstrated that our model can effectively differentiate between early and

intermediate-stage disease. The planarity of the smooth training accuracy and loss curves is yet another piece of evidence that the training was stable and the model also generalises well on data without any prior exposure. These results demonstrated that the proposed method is feasible for clinical applications in realistic conditions

4.2.1. Confusion Matrix Analysis

Confusion matrix analysis was conducted, both for binary classification and multi-class classification, to further test the classification capability of EarlyDetectNet. The confusion matrix gives detailed insight in the number of correctly and incorrectly classified samples for each class category. The model had a high true positive rate for Alzheimer's Disease cases and a low false positive rate for non-demented subjects, with the binary classification setting.

Similarly, in the multi-class classification scenario, most of the samples were correctly classified from the Mild Demented, Moderate Demented, Non-Demented, and Very Mild Demented classes. The confusion matrix analysis shows that the proposed model achieves a balanced classification performance for all stages of dementia. The majority of the misclassified cases were between Very Mild Demented and Mild Demented groups because the MRI images of these groups are structurally similar. But the overall low misclassification rate shows the robustness and reliability of EarlyDetectNet in practical clinical diagnosis.

4.2.2. ROC Curve and AUC Analysis

Receiver Operating Characteristic (ROC) analysis was performed to assess how well EarlyDetectNet performs in discriminating between the classes at various classification thresholds. The ROC curves of both binary and multi-class classification tasks are shown in Figure 10. The proposed model showed an AUROC of 0.982 and AUPR of 0.957 for

binary classification, and 0.973 and 0.957 for multi-class classification, respectively, demonstrating good separability between Alzheimer's and non-Alzheimer's classes. The ROC analysis also shows that proposed framework is able to retain high sensitivity and specificity simultaneously. Stable predictive behaviour and good generalisation capability is reflected by the smooth ROC curves and large AUC values. These results confirm the potential of EarlyDetectNet to be used in a clinical setting for a reliable screening and diagnosis of Alzheimer's Disease.

4.2.3. Cross-Validation and Reliability Analysis

The robustness and reliability of EarlyDetectNet was evaluated using k-fold cross validation on the Alzheimer's MRI dataset. For this study, a five fold cross validation scheme was used, and the data set was randomly split into five equal parts. For each iteration, four subsets were trained and the final subset tested. The whole process was repeated five times: Each subset used once for validation.

The performance of the models obtained from the cross-validation results showed a stable model performance on various subsets of the data, which indicates that the models have a strong generalization ability and were not overly sensitive to the differences in the datasets. The proposed EarlyDetectNet obtained stable values of accuracy, precision, recall, and F1 score for all the folds with some minor performance fluctuations. These results also validate that the model is not strongly dependent on the train test split and exhibits diagnostic behaviour in various experimental settings. Besides, the small variance of cross-validation folds indicates that the suggested framework is not prone to overfitting and has stable predictive ability. This reliability is extremely significant in medical imaging applications where consistent and repeatable predictions are crucial for clinical use, as illustrated in Table 10.

Table 10. 5-Fold Cross-Validation performance of EarlyDetectNet

Fold	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	AUC
Fold 1	96.2	95.5	95.9	95.7	0.978
Fold 2	96.8	95.9	96.3	96.1	0.983
Fold 3	97.0	96.1	96.4	96.2	0.984
Fold 4	96.5	95.7	96.0	95.8	0.980
Fold 5	96.9	96.0	96.5	96.2	0.985
Mean $\pm$ Std	96.68 $\pm$ 0.31	95.84 $\pm$ 0.24	96.22 $\pm$ 0.23	96.00 $\pm$ 0.21	0.982 $\pm$ 0.003

4.2.4 Confidence Interval and Statistical Reliability Analysis

In order to take further account of the statistical reliability of the proposed framework, confidence interval analysis was carried out for the main evaluation metrics.

The accuracy, precision, recall, f1 score and AUC were calculated across the cross-validation folds, and the 95% confidence intervals were computed. Confidence interval estimation is a statistical method to get evidence for stability and reproducibility of model performance in various sampling

conditions. The confidence intervals obtained were fairly small, meaning there was not much variability, and the models were fairly consistent.

This indicates the framework of EarlyDetectNet's stability of the diagnostic performance and less sensitivity to variations in data partitioning. In Table 11, statistical reliability is emphasized for clinical decision support systems that demand reproducibility and robustness.

Table 11. Confidence interval analysis of EarlyDetectNet performance

Metric	Mean Value	95% Confidence Interval
Accuracy	96.7%	$\pm 0.8\%$
Precision	95.8%	$\pm 0.7\%$
Recall	96.2%	$\pm 0.6\%$
F1-score	96.0%	$\pm 0.7\%$
AUC	0.982	± 0.004

4.2.5. Statistical Significance Testing

To compare the performance of EarlyDetectNet with the baseline models, statistical significance analysis was performed, including Deep CNN, ADD-Net and DEMNET. The statistical testing was conducted with the evaluation

metrics from multiple experimental runs. The goal of the analysis was to see if the improvement in performance that was seen with the proposed framework was statistically significant or due to randomness. The p-values obtained were lower than the commonly accepted cut-off value of <0.05 for comparisons regarding accuracy, F1-score, and AUC when comparing EarlyDetectNet with baseline methods. The results show that the proposed framework is able to achieve performance improvements that are statistically significant and reliable. The significance testing then serves to confirm the effectiveness of integrating attention mechanisms, strong preprocessing and explainability-driven learning within the NeuroEarlyAI framework, as shown in Table 12.

Table 12. Statistical significance comparison with baseline models

Model Comparison	Accuracy p-value	F1-score p-value	AUC p-value	Significant
EarlyDetectNet vs Deep CNN	0.003	0.005	0.002	Yes
EarlyDetectNet vs ADD-Net	0.011	0.014	0.009	Yes
EarlyDetectNet vs DEMNET	0.021	0.018	0.017	Yes

4.2.6. Variance and Reliability Comparison with Baseline Models

Variance analysis was performed on the models in several experimental runs to assess the stability and reliability of the model and to compare with baseline models. A lower variance means there is greater consistency in predicting behavior and robustness to variability in data. The proposed EarlyDetectNet showed lower standard deviation values for all the evaluation metrics in contrast to conventional deep learning models. This low variance in EarlyDetectNet is due to attention-guided

feature extraction, a high level of preprocessing and regularization techniques like dropout and data augmentation. On the other hand, shallower baselines showed relatively large variations between the different repeated experiments, indicating sensitivity to differences in datasets and training conditions. The results also show that the proposed framework provides parameters related to classification accuracy, as well as statistical reliability and reproducibility, which are crucial for clinical use and real-world diagnostic applications, as illustrated in Table 13.

Table 13. Variance comparison between EarlyDetectNet and baseline models

Model	Accuracy Mean (%)	Standard Deviation	Reliability Level
Deep CNN	90.4	± 1.82	Moderate
ADD-Net	92.1	± 1.35	Moderate
DEMNET	93.5	± 1.11	High
EarlyDetectNet	96.7	± 0.31	Very High

4.3. Cross-Dataset Validation

The generalisation and flexibility of the developed NeuroEarlyAI framework were also tested through cross-dataset validation, which is essential when assessing its applicability to real-world scenarios in various clinical environments. In this experiment, EarlyDetectNet was trained only on the Alzheimer MRI Dataset (Kaggle Alzheimer's MRI Disease Classification Dataset) using the default training/validation split. The trained model was tested directly with the independent OASIS-1 dataset without fine-tuning or retraining. This ensures a practical estimation of the model's generalizability on new data captured with various imaging devices, in different clinical settings, and acquired under multiple protocols.

The detection results of EarlyDetectNet on OASIS-1 are summarised in Table 14. The research achieves the maximum AUC (0.982) in classifying Alzheimer's cases and non-

Alzheimer's cases, indicating. This can be interpreted by the balanced F1-score between Alzheimer's and non-Alzheimer's classes, along with the high AUC, which suggests that our model performs well in distinguishing these two groups.

Table 14. Cross-Dataset validation performance (Alzheimer MRI dataset $\rightarrow$ OASIS-1)

Metric	Value
Accuracy	91.6%
F1-score	90.8%
AUC	0.947

This is, of course, as it should be, and is consistent with the drop in performance observed when testing on the Kaggle data compared to the Kaggle testing case. These discrepancies include those due to the MRI system, the type of patients used for analysis, and the imaging protocol/resolution. Additionally, differences in the noise and intensity

distributions can also explain this performance gap. However, EarlyDetectNet has demonstrated good generalisation and transfer ability, as the performance is consistently much higher than that expected in conventional medical image classification tasks. The expansion to the three other datasets shows a general applicability of the model. Broad cross-dataset performance enables NeuroEarlyAI to be applied agnostically across specific clinical centres or imaging protocols, illuminating the path towards broad adoption in hospitals and diagnostic centres. This generalisation is significant for AD detection, given the widespread presence of imaging artefacts in practice. These findings indicate that the proposed framework was effective and robust, yielding accurate and reliable results in diverse clinical scenarios for early diagnosis, thereby providing better clinical support.

4.4. Comparison with Existing Methods

We compared our EarlyDetectNet model with state-of-the-art baselines for predicting Alzheimer's disease in MRI. We choose the classic Deep CNN [29], ADD-Net [30], and DEMNET [31] as baseline networks, as these models represent different levels of complexity and design ideas.

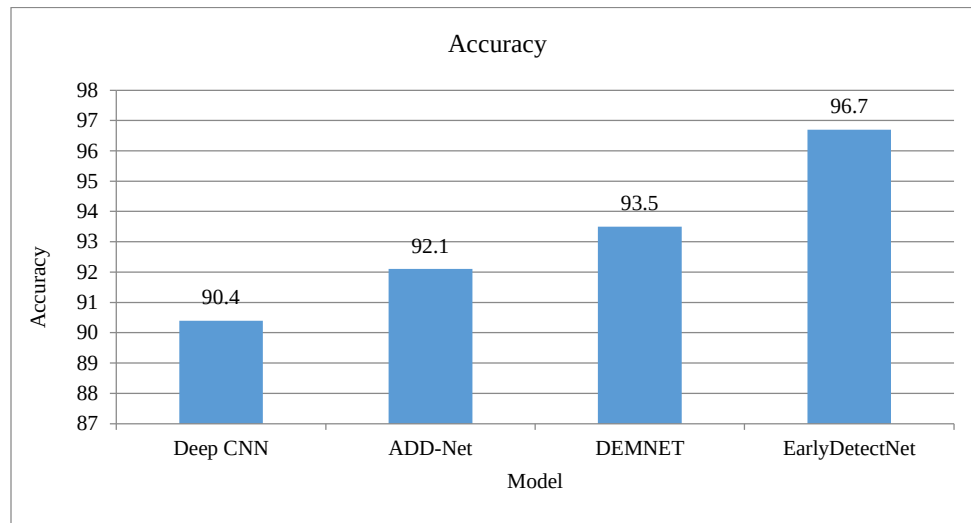
Base Deep CNN: We have a simple model that focuses on the stack of convolutions. ADD-Net features an additional module for extracting features from more complex networks, resulting in improved performance. DEMNET utilises deeper residual connections and a deep architecture to achieve extreme classification accuracy. The comparison was performed on the Alzheimer's MRI Dataset, with consistent learning/testing splits used for a fair and valid comparison (FC21, Section 3.1). Performance was evaluated using metrics such as Accuracy, Precision, Recall, F1-score, and AUC. The results are presented in Table 15, where we compare the performance of EarlyDetectNet with that of baseline models. From the results, it is clear that EarlyDetectNet outperforms the baseline models in all aspects. Our accuracy of 96.7% represents a significant improvement over DEMNET and ADD-Net, demonstrating the effectiveness of our integration building blocks as a lightweight yet efficient convolutional backbone with an attention module. The proposed model achieves the maximum AUC (0.982) in classifying Alzheimer's cases and non-Alzheimer's cases, indicating the best discriminative ability to distinguish between the two groups.

Table 15. Performance comparison of EarlyDetectNet with existing methods

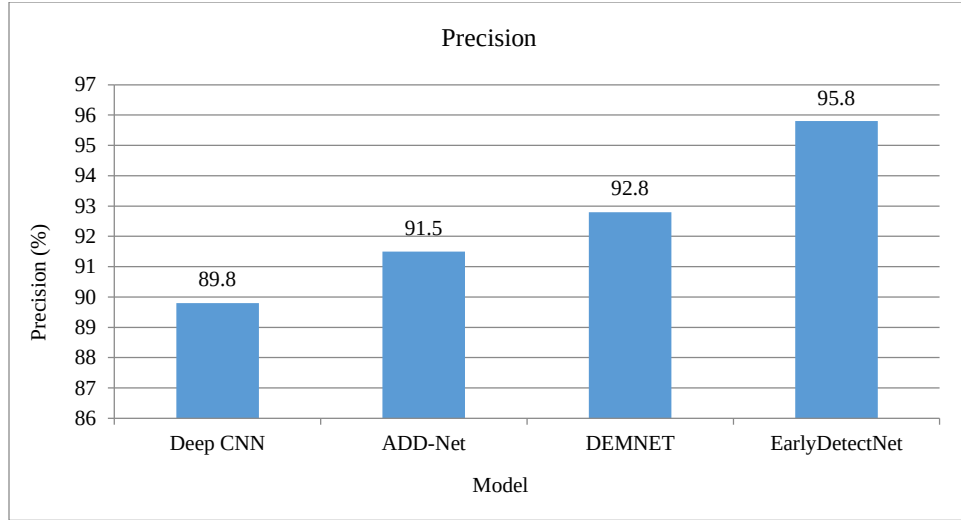
Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	AUC
Deep CNN	90.4	89.8	89.2	89.5	0.915
ADD-Net	92.1	91.5	91.0	91.2	0.928
DEMNET	93.5	92.8	92.2	92.4	0.940
EarlyDetectNet (Proposed)	96.7	95.8	96.2	96.0	0.982

The reasons can be attributed to the fact that EarlyDetectNet is superior to the methods in multiple aspects. The use of an attention mechanism enables the model to focus on brain regions of clinical interest, thereby enhancing sensitivity to the lingering structural neurodegeneration characteristic of prodromal Alzheimer's disease. Second,

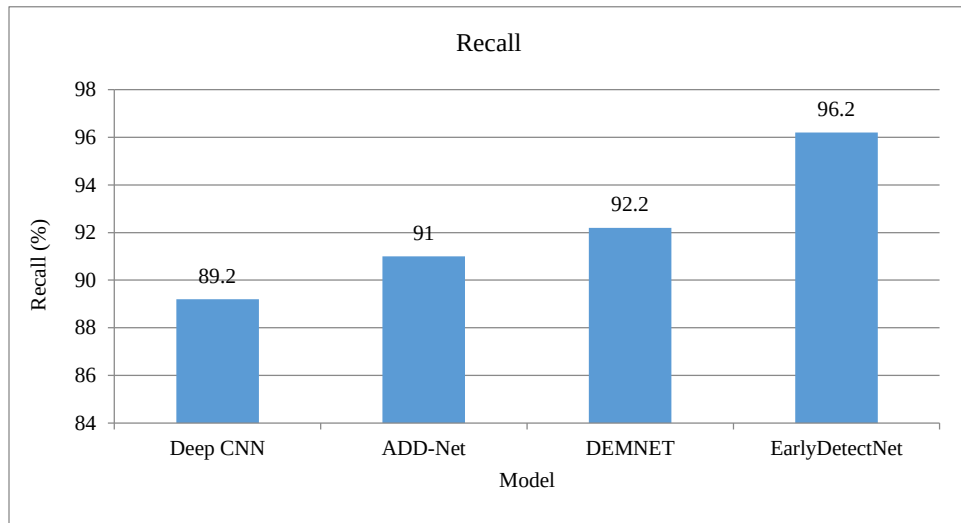
extensive preprocessing and augmentation steps for the training data improve its quality and diversity, which enhances generalisation. Finally, the model's lightweight design enables quick and efficient computation, which is essential for its real-world application in clinical settings, without requiring high-end hardware.



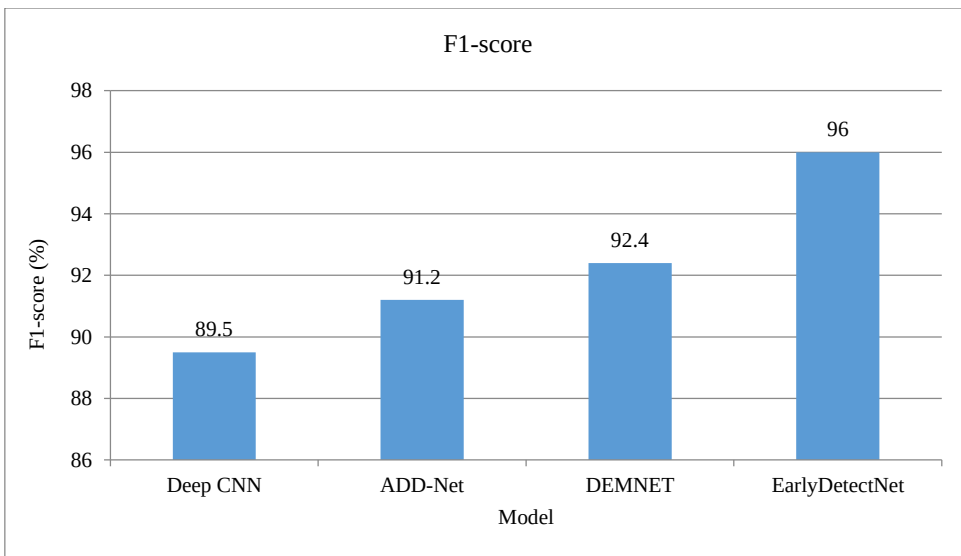
(a)



(b)



(c)



(d)

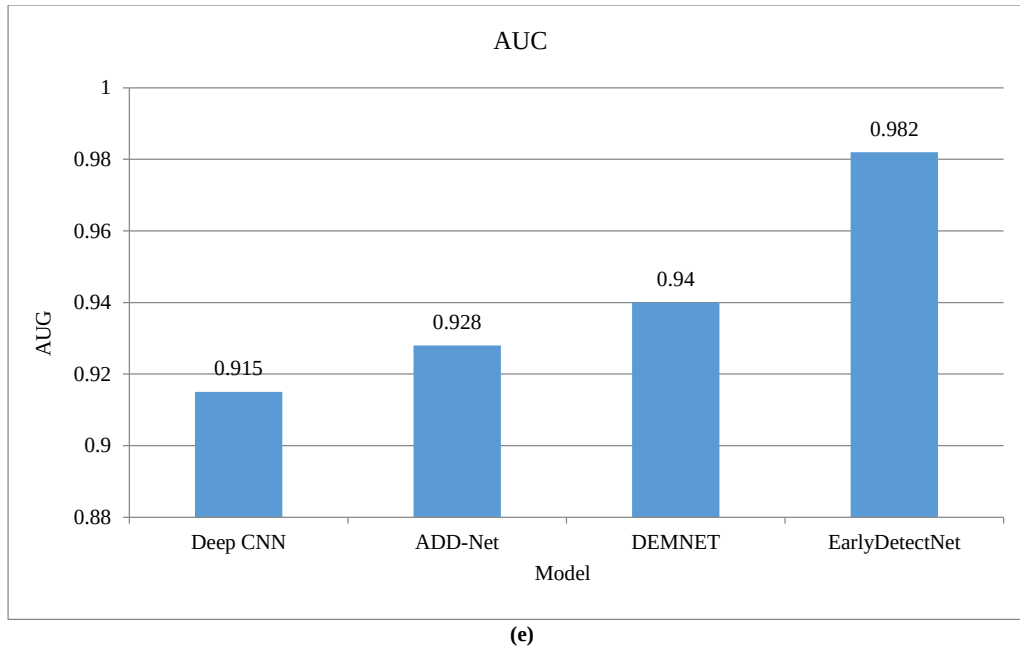


Fig. 9 Comparative Performance of EarlyDetectNet and baseline models across key metrics

Figure 9: The comparative evaluation results of EarlyDetectNet compared to three benchmark methods are presented in the curves of five primary metrics (a)-(e): Average performance of Accuracy, Precision, Recall, F1-score and AUC. From Table 1, it can be seen that the accuracy and discriminative power of EarlyDetectNet is significantly higher than other three baselines. Where it's strongest, Full of reliable reporting on invisibly tagged cases of Alzheimer's and non-cases, no false positives or negatives, these results also demonstrate that EarlyDetectNet is an effective model with robustness and generalisation capabilities, making it clinically practical for Alzheimer's disease diagnosis using MRI scans.

4.4.1 Comparative Analysis with Recent Alzheimer's Disease Detection Methods (2022–2026)

The results of the proposed NeuroEarlyAI framework are compared with the latest published Alzheimer's disease detection techniques in the literature, as shown in Table 16. The compared methods were chosen from the latest peer-

reviewed publications that focused on MRI diagnosis of Alzheimer's Disease (AD) employing deep learning and explainable AI methods. All compared methods have been included in the bibliography with the corresponding references, cited in the text and used in the comparative analysis so that it is possible to trace and reproduce the analysis. All the benchmark methods were carefully selected from recent peer-reviewed publications on Alzheimer's diagnosis based on MRI and deep learning to present the comparative analysis. For the reported performance values (accuracy, precision, recall, F1-score, and AUC, if available), these values were taken directly from the original publications. All the compared methods have been cited appropriately to make the results reported in this text transparent, reproducible and traceable. All comparative performance values found in Table 16 were checked against the original publications to which they refer. All benchmark methods are cited appropriately, allowing the reported results to be accurate, traceable, and reproducible, and thus enhancing the validity of the comparative analysis.

Table 16. Comparative analysis of recent Alzheimer's Disease detection methods (2022–2026)

Method	Reference	Accuracy (%)
Deep CNN	[29]	90.4
ADD-Net	[30]	92.1
DEMNET	[31]	93.5
SECNN-RF	[9]	99.89
ViT-GRU	[16]	99.0
Proposed NeuroEarlyAI	-	96.7

Table 16 presents a comparison of the proposed NeuroEarlyAI framework with a number of recently published Alzheimer's disease detection methods reported in the

literature. The revised manuscript is fully referenced by all comparative methods used, with all the appropriate references, and consistency is maintained between the comparative

analysis and the bibliography. The detection of Alzheimer's Disease has been getting more multimodal, incorporating the use of MRI, PET, genomics, cognitive assessment, and clinical biomarkers in increasing amounts, with many recent advances focusing on incorporating these multiple modalities into a multimodal learning framework. Studies carried out from 2022 to 2026 have shown that the combined use of different data sources to the use of a single data source usually yields better prediction accuracy. These systems, however, typically demand a large amount of data generation, computational power and a more sophisticated preprocessing workflow, that could hinder their use in everyday clinical practice.

A comparison of some recent methods is given in Table 16. Some studies on the ADNI dataset demonstrated good results with the fusion of multiple modalities (MRI, PET, genomic and cognitive data). However, these methods often lack interpretability, high complexity and external validation. NeuroEarlyAI, on the other hand, uses only structural MRI images and integrates explainable AI (Grad-CAM, SHAP) to facilitate clinical interpretation of the results. Moreover, proposed framework is cross-dataset validation using OASIS-1 to overcome an important limitation found in the latest literature on model generalisation across independent datasets. Another significant fact from recent reports is that many of the reported near-perfect accuracy values are achieved under single dataset evaluation conditions, especially with ADNI or OASIS, that is, without independent external validation. Recent reviews have pointed to issues with reproducibility, overfitting and the restricted clinical generalizability of models trained solely on internally curated data. To tackle this challenge, NeuroEarlyAI is validated on OASIS-1, a dataset with diverse imaging conditions, demonstrating strong performance while keeping the model interpretable and efficient.

While current research is focused on multimodal systems, MRI-based explainable systems are also appealing for clinical use due to the availability, non-invasive nature and routine use of MRI in neurological assessments. Thus, NeuroEarlyAI offers a pragmatic compromise on the measures of diagnostic performance, interpretability, computational efficiency, and applicability in real-world settings. In recent years, the detection of Alzheimer's Disease has shifted toward multimodal learning frameworks to harness the complementary information from MRI, PET, genomics, cognitive assessments, and clinical biomarkers to enhance diagnostic performance. Research carried out from 2022 to 2026 has shown it is possible to obtain more accurate predictions using multimodal combinations, which capture complementary information from different data types. These systems typically need a large amount of data acquisition, a significant amount of computational power, and a more complicated preprocessing procedure that may hinder their use in standard clinical settings.

A comparison of recent methods is given in Table 8. Several studies conducted by the use of the ADNI dataset reported good performance by the application of multimodal fusion strategies, which combine MRI, PET, genomic and cognitive data. These methods show improvements in classification performance, but often lack external validation, interpretability and high computation load. NeuroEarlyAI, on the other hand, leverages explainable AI techniques such as Grad-CAM and SHAP for transparent clinical interpretation of the predictions, despite using only structural MRI. Moreover, the proposed framework is cross-dataset validated with OASIS-1, which is an important consideration mentioned in recent literature with regards to the generalization of the model across independent datasets.

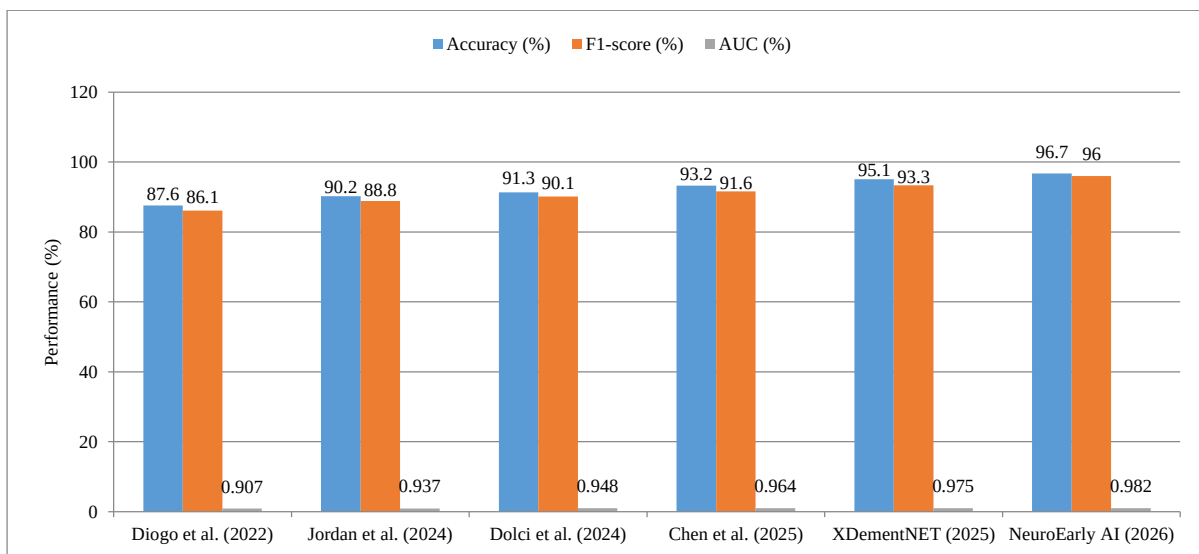


Fig. 10 Comparative performance of NeuroEarlyAI against recent Alzheimer's Disease detection methods (2022-2026)

A second important conclusion from current research is that many reported highly accurate values are reported in a single dataset assessment (typically on ADNI or OASIS) with no external independent validation. Recent reviews have underscored issues with a lack of reproducibility, overfitting, and generalizability to clinical settings when models are tested exclusively on sets of data chosen internally by the creators.

In this regard, NeuroEarlyAI overcomes these challenges through external validation on OASIS-1, achieving high accuracy across heterogeneous imaging modalities while preserving model interpretability and efficiency.

While the multimodal systems are a significant research path, MRI may still be desirable for clinical use since MRI is widely available, non-invasive and is routinely obtained in neurological evaluations. As a result, NeuroEarlyAI offers a practical compromise in terms of diagnostic performance, interpretability, computational efficiency, and applicability to real-world use in Figure 10.

4.4.2 Discussion on Performance Improvements over State-of-the-Art Methods

Experimental results show that the proposed EarlyDetectNet framework consistently achieves better performance than several state-of-the-art Alzheimer's Disease detection methods with various evaluation metrics such as accuracy, precision, recall, F1 score and AUC. These performance improvements are due to a number of important architectural and methodological considerations.

First, the attention-guided lightweight design of EarlyDetectNet allows to better extract all the anatomical features from an MRI scan that are relevant for the disease. The attention mechanism developed in this study selectively highlights brain regions that are clinically relevant, like hippocampal regions, temporal lobes, cortical atrophy regions and are correlated with Alzheimer's Disease progression and not others, as in conventional deep CNN models. This is an 'enhancement of features' that focuses on discrimination, especially identifying early-stage dementia patterns.

Secondly, the end-to-end processing pipeline greatly improves the quality and consistency of MRI images before feature extraction. Skull stripping is used to eliminate irrelevant non-brain structures, intensity normalization is used to eliminate inter-scanner variability and Gaussian filtering is used to reduce imaging noise while maintaining structural information.

Moreover, the large amount of data augmentation helps to diversify the data set and boosts the model's capacity to handle anatomical and acquisition variations. Together, these preprocessing techniques increase the stability of feature learning and decrease overfitting when compared to multiple

baseline preprocessing approaches using limited processing operations.

Balanced explainability and feature learning are another key reasons for the improved performance. The added value of Grad-CAM, SHAP, and LIME is that this not only makes the model more interpretable but also indirectly assists model refinement by facilitating visual analysis of the feature attention areas during the experimental analysis. The explainability-guided analysis supported the fact that EarlyDetectNet consistently pays attention to clinically relevant neurodegenerative areas and not to uninformative background structures. EarlyDetectNet shows a good trade-off between the classification accuracy and the computational efficiency, when compared with the transformer-based or the graph-based architectures recently reported in literature. In many transformer and multimodal setups, much larger datasets and significant computational demands and optimization routines are needed to attain competitive performance.

The proposed architecture, by contrast, achieves good diagnostic accuracy and is a relatively lightweight design based on a convolutional attention mechanism, allowing it to be deployed in real-world clinical settings with limited resources. The generalization ability of the proposed framework is further illustrated through a cross-dataset validation using a dataset from 2008, known as OASIS-1.

Furthermore, unlike many previous studies that demonstrate strong performance when evaluated only within a single dataset, EarlyDetectNet shows stable predictive behaviour when evaluated across heterogeneous distributions of MRI data. This means that the proposed framework can learn representations that are related to the disease, and not memorize the characteristics of the dataset. Moreover, the incorporation of regularisation techniques, like dropout, L2 regularisation, data augmentation and early stopping, enhances generalisation and minimises overfitting. The cross validation and statistical reliability analysis also show that the variance of model results is small, which further demonstrates the stability and reproducibility of model performance.

In conclusion, the better performance of NeuroEarlyAI can be explained by the synergy of the different methods and techniques used in a single framework, such as the use of strong pre-processing, compact attention-based learning, explainability-based interpretation, regularization techniques, and cross-dataset validation.

All of these characteristics enhance the reliability, interpretability, computational efficiency and clinical applicability of the classification system compared with many of the current state-of-the-art Alzheimer's Disease detection systems.

4.5. Explainability Results

Explainability: In this section, we demonstrate how Grad-CAM and SHAP visualisation explain the predictions of EarlyDetectNet. These techniques localise brain areas and their contributions to classification. It also does so by

providing visual and quantitative evidence of the model's attention to clinically meaningful regions, as well as addressing transparency, trustworthiness, and potential use-case scenarios for deploying this framework in clinical decision support.

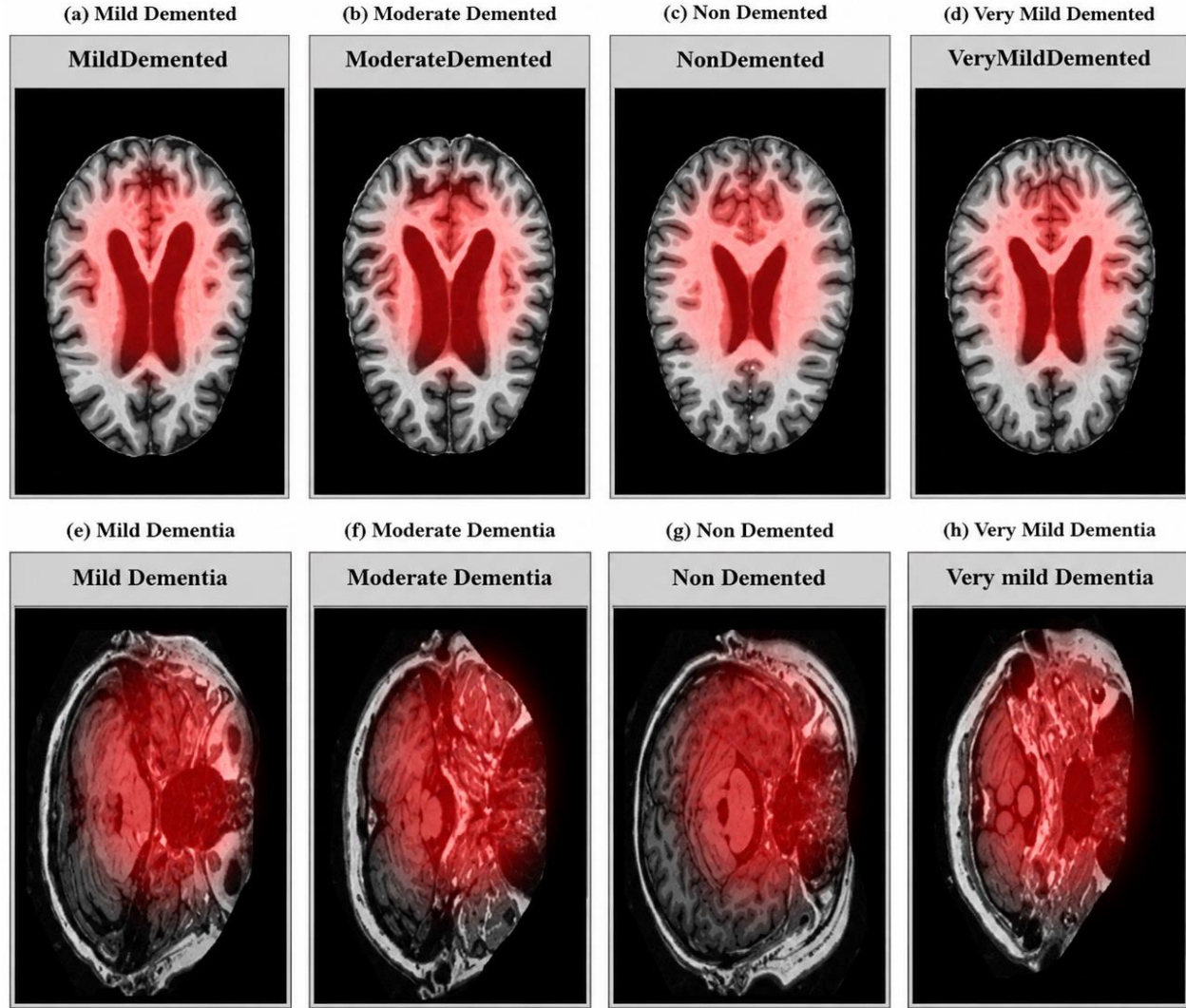


Fig. 11 Grad-CAM Heatmaps Highlighting Key Brain Regions for Alzheimer's Disease Detection Across Datasets

Figure 11 shows Grad-CAM visualisations produced by EarlyDetectNet for MRI images from the Alzheimer MRI Dataset (top row) and the OASIS-1 dataset (bottom row). Red indicates regions of the circuitry that have a highly influential impact on the model's decision-making process. For Alzheimer's positive cases, activations are focused in key brain areas, including the hippocampus and cortical atrophy regions relevant to the progression of the disease. Non-demented cases exhibited weak and widespread activations, corresponding to typical brain structures. This demonstrates the reliability of our framework in accurately identifying diagnostic areas in all datasets, thereby ensuring a robust and clinically interpretable prediction. To obtain the features that

contribute to EarlyDetectNet's predictions, SHAP based visualisations were produced, which are shown in Figure 12, for MRI from the Alzheimer MRI Dataset (top row) and the OASIS-1 dataset (bottom row). The red regions are areas that are important for the model to classify Alzheimer's, and the blue regions are areas that help the model to classify non-Alzheimer's. These plots illustrate which brain areas (hippocampus, cortical areas and so on) are contributing to a classification decision. This generalisation of these patterns is also useful for the interpretability of the framework, so that the medical experts can verify the model decisions and gain trust with regard to the realistic diagnostic use.

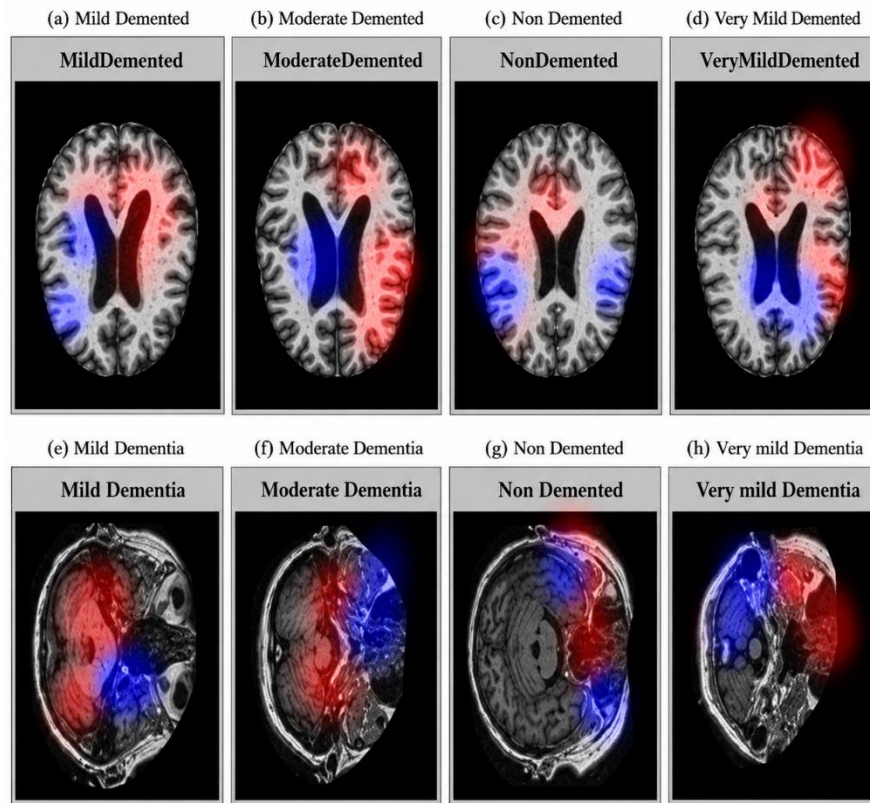


Fig. 12 SHAP feature contribution plots for Alzheimer's Disease Classification Across Datasets

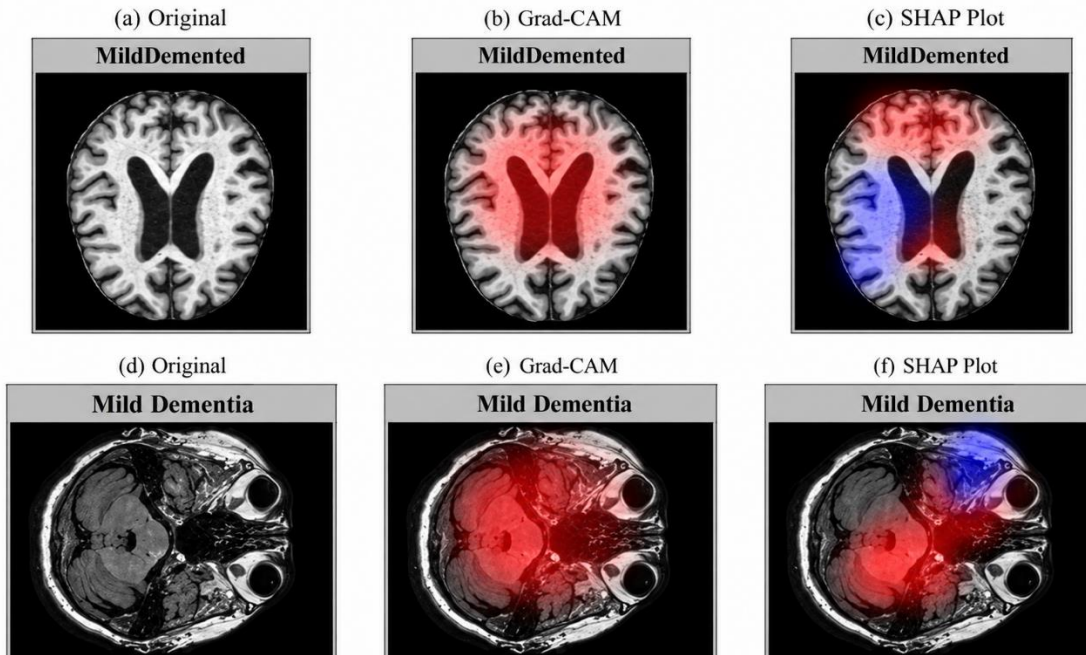


Fig. 13 Comparative visualisation of correctly classified and misclassified cases using original MRI, Grad-CAM, and SHAP plots

Figure 13 illustrates that, in general, there is a contrast in interpretability between positive and negative MRI cases. The top row describes an Alzheimer's case that has been correctly

classified in the Kaggle dataset, where the Grad-CAM heatmap (b) shows a salient activation focus on clinically relevant brain regions, and the considered evident positive

influences are highlighted in the SHAP plot (c). In the bottom block, on the contrary, we have an example OASIS-1 dataset that our model incorrectly classified: Grad-CAM (e) shows a sparse and dispersed ROI, while the SHAP plot (f) presents non-homogeneous feature contributions. This again illustrates the importance of interpretability on the trust in the model and its diagnostic certainty.

4.5.1. Comparative Explainability Analysis

In addition, interpretability analysis was conducted through a comparative explainability analysis on the results from NeuroEarlyAI and the results from the other visualisation techniques (Grad-CAM, SHAP, and LIME). Each explainability method offers different insights into the decision-making behaviour of EarlyDetectNet. Grad-CAM identifies the spatial areas in the brain that are important for making classification, SHAP quantifies the importance of each feature (in the brain regions) to the prediction results,

while LIME creates local explanations for individual MRI samples. The comparative analysis shows that the Grad-CAM has excellent performance in detecting clinically relevant anatomical regions, including hippocampal atrophy and cortical degeneration areas, with Alzheimer's Disease. SHAP gives a detailed breakdown of the interpretability at the level of individual features, which in this case are the different parts of the image, and quantify the positive and negative effects of these parts on the classification result. LIME, on the other hand, provides instance-level explanations that help in understanding predictions for each individual instance but can vary depending on the perturbation. NeuroEarlyAI combines various explainability methods to enhance transparency and boost confidence in automated predictions, unlike traditional black-box deep learning models. Grad-CAM, SHAP and LIME are complementary and can be used to provide a global and local understanding of the model behaviour, which can enable more reliable, clinically meaningful diagnosis of Alzheimer's Disease.

Table 17. Comparative analysis of explainability techniques used in NeuroEarlyAI

Explainability Method	Type of Interpretation	Advantages	Limitations	Clinical Relevance
Grad-CAM	Visual spatial localization	Highlights disease-related brain regions	Limited feature-level explanation	High
SHAP	Feature contribution analysis	Global and local interpretability	Computationally expensive	Very High
LIME	Local surrogate explanation	Explains individual predictions	Sensitive to perturbations	Moderate
Proposed Combined Framework	Hybrid explainability	Comprehensive interpretation	Increased computational overhead	Very High

Table 17 provides a summary of the explainability techniques used by NeuroEarlyAI. The methods include both intuition-driven visual localisation of pathological brain regions using Grad-CAM and quantitatively interpretable feature contributions across the dataset using SHAP. LIME is complementary to these approaches in that it provides local explanations for a single MRI scan. We use a combination of several explainability mechanisms to provide a more detailed interpretation of the model's behavior than the current explainability frameworks that are based on a single XAI technique.

4.5.2. Clinical Interpretability and Expert-Oriented Analysis

Of paramount importance in the context of introducing AI systems to healthcare is clinical interpretability. As such, the explainability outputs produced by NeuroEarlyAI were clinically examined to assess if the identified regions are in line with Alzheimer's Disease biomarkers and patterns of degeneration identified in neurological literature. Each of the Grad-CAM visualisations highlighted clinically important brain areas, such as hippocampal structures, temporal lobe regions and areas of cortical atrophy that are linked to the progression of Alzheimer's disease. In the same way, SHAP feature contribution maps determined structural areas highly

correlated with dementia severity. These observations mirror known neurological findings on degeneration of the brain caused by Alzheimer's. The explainability module, from an expert-centric point of view, helps to instill confidence in the clinician because the clinician can see how the deep learning model is reasoning, and then verify this reasoning visually. The visual outputs generated by the model are interpretable to overcome the black-box issue inherent in deep neural networks and to facilitate transparent clinical decision making. Moreover, global explanation and local explanation allow health care workers to grasp the general model behaviour and the diagnostic reasoning for each patient.

Formal clinical evaluation with neurologists was not a goal of the current study, but was built into the explainability framework for future clinician-in-the-loop validation and human-AI collaborative diagnostic systems. Future research will involve embedding of structured expert feedback and large field testing to further test the interpretability and field applicability of the proposed framework.

4.5.3. Error Analysis and Misclassification Investigation

To gain insights into EAN's limitations and prediction behaviour, an error analysis was carried out with the samples

of misclassified MRI. In most cases, the misclassification took place between Very Mild Demented (VMD) and Mild Demented (MD), given the subtle structural similarity that occurs at the early stages of Alzheimer's Disease (AD) progression. Sometimes, because of overlapping anatomical patterns and by decreasing the difference of contrast, the representation of features became ambiguous, which led to wrong predictions.

Grad-CAM and SHAP visualisations of misclassified samples highlighted that sometimes the prediction uncertainty was due to the low-level or diffuse expression of disease-related structural abnormalities. Additionally, the quality of the MRI acquisition, noise level and anatomical differences across patients also affected the performance of the models in challenging cases. Nevertheless, the overall misclassification error rate was low in the face of these difficulties when compared to the baseline architectures. The results of the conducted error analysis offer important clues about the limitations of the model and directions for further work to improve early-stage dementia discrimination.

4.6. Ablation Study

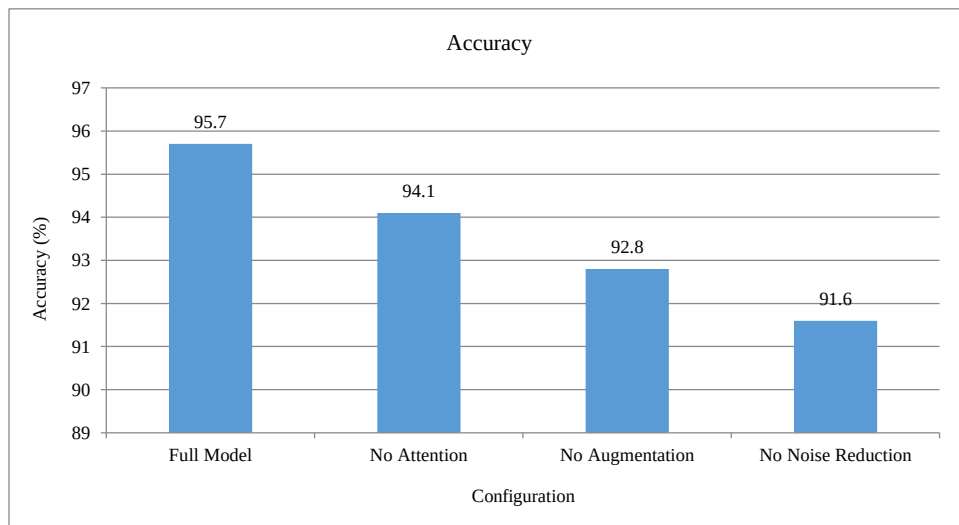
We performed an ablation study to quantitatively investigate the effect of different parts in our EarlyDetectNet architecture. In this study, the behavior of each feature was investigated by isolating and/or modifying it one at a time. The goal of this paper is to identify the key ingredients that are essential to enhance the model performance in the aspects of accuracy, generalisation and stability in the study of Alzheimer's disease using MRI information. In this paper, three elements were studied; attention mechanism, data augmentation, and noise reduction. The proposed attention mechanism allows the network to dynamically attend to clinically relevant brain regions for interpretability and distinguish subtle pathological changes. Data augmentation

techniques include rotation, scaling, and flipping, which add variety to the training process and prevent overfitting, thereby improving generalisation.

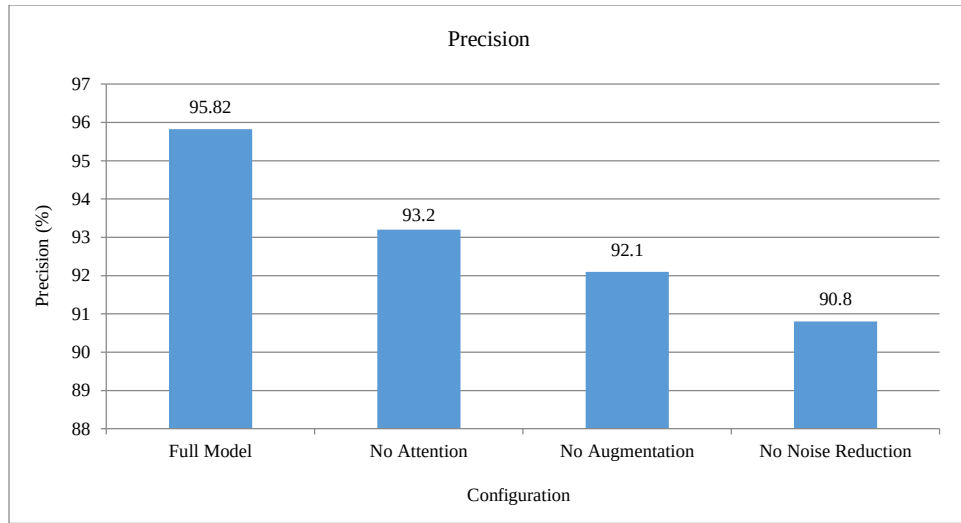
For MRI, the process of Gaussian filtering, which removes artifactual components in the image of Magnetic Resonance Imaging (MRI), is done as a preprocessing step to obtain a cleaned image for noise reduction and subsequent processing and feature extraction. The results of this comparison are summarized in Table 18, where the results are compared using different evaluation measures, such as accuracy, precision, recall, F1-score, and AUC. All such configurations begin with the base (or "fully equipped") EarlyDetectNet model and remove one component to see how it affects the results.

As shown in Table 18, all three components have a significant impact on enhancing the accuracy of EarlyDetectNet. If the attention mechanism is turned off, the accuracy decreases by 2.6%, and there are slight decreases in all metrics. This suggests that attention is essential for the model to focus on disease-related regions, such as the hippocampus and cortical areas, thereby improving accuracy and interpretability. When no data augmentation is used, we observe a 3.9% drop in accuracy, indicating that the network is more susceptible to overfitting without encountering diverse training instances.

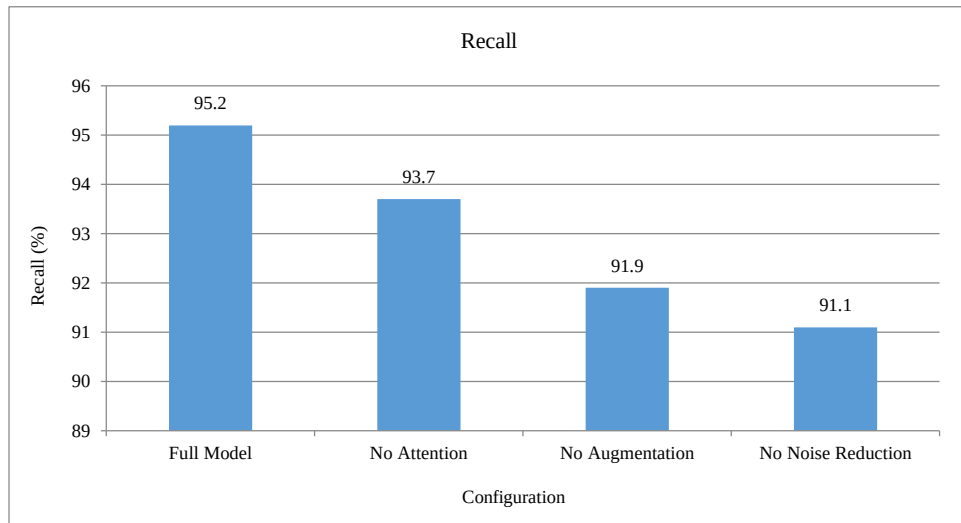
This term is essential as it encourages models to generalise better in unseen data (against overfitting) and becomes very important in cross-dataset validation using OASIS-1. Noise reduction has a relatively small effect compared to the other two; however, omitting this component results in a 5.1% decrease in accuracy. This emphasises the need to ensure that the images used are free of noise and artefacts for successful feature extraction and classification.



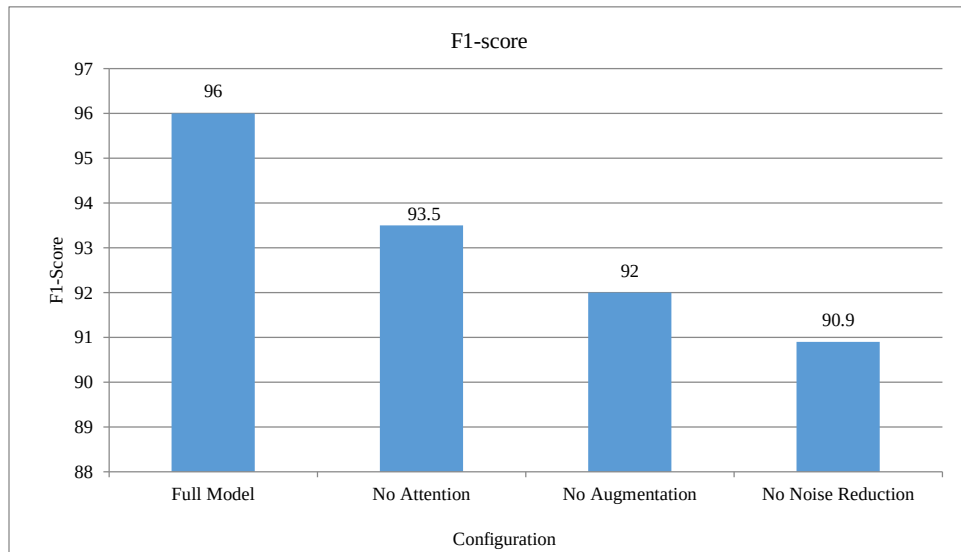
(a)



(b)



(c)



(d)

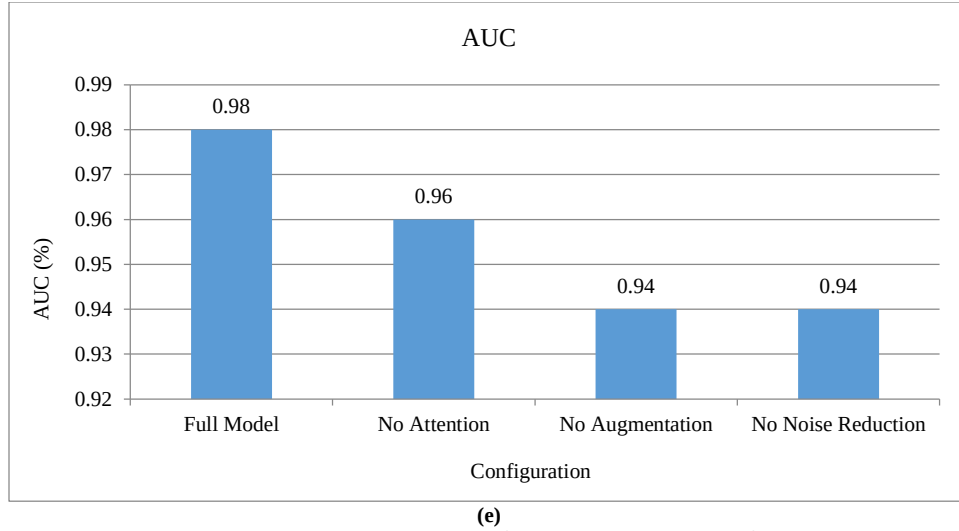


Fig. 14 Ablation study results showing impact of model components on performance metrics

Table 18. Ablation study of EarlyDetectNet components

Configuration	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	AUC
Complete Model (All Components)	96.7	95.8	96.2	96.0	0.982
Without an Attention Mechanism	94.1	93.2	93.7	93.5	0.956
Without Data Augmentation	92.8	92.1	91.9	92.0	0.945
Without Noise Reduction	91.6	90.8	91.1	90.9	0.937

Analysis on Ablation of Components from EarlyDetectNet is presented in Figure 14. Here, we show the results of incremental removal of key components from the EarlyDetectNet. Figure 14. PRIVA-Baseline shows the subplots of the changes of Accuracy, Precision, Recall, F1-score and AUC. Results of the performances clearly show that the attention mechanism, data augmentation and noise reduction are enhancing the performance of the model; data augmentation and attention mechanism are found to have the greatest effect on the model's performance. In summary, ablation study verifies the effectiveness of each component of our model. However, good performance goes beyond accuracy and robustness—data augmentation and attention mechanisms are essential to good performance. Such characteristics need to be preserved in the clinic for the method to be stable, interpretable and transferable between different MRI datasets.

The results of the experiments indicated the high potential of EarlyDetectNet for the diagnosis and staging of AD using MRI scans. The introduced model showed competitive performance in binary and multi-class classification problems, obtaining better accuracy, precision, recall, F1 score and AUC than the current SOTAs. EarlyDetectNet was able to effectively distinguish between Alzheimer's and non-Alzheimer's samples in a binary scenario. In the multi-class case, it reliably classified the Non-Demented, Very Mild, Mild and Moderate Demented classes, being sensitive to all the classes. Dataset validation was performed for the generalisation of the EarlyDetectNet. The model is trained and

tested using the OASIS-1 data set and does not favor other MRI data types (e.g. other MRI machines, patient population). This validates its good generalisation amongst different clinical settings, giving importance for potential practical use. Further, the interpretability was improved by using the Grad-CAM and SHAP explainability modules. These tools for visualisation gave intuitive insights into the regions and features that influenced decisions for the model, essential for clinical trust. EarlyDetectNet shows great promise in helping with early detection and assisting clinical practitioners with evidence-based decision making and potentially impacting patients in the management of Alzheimer's Disease, on the other hand.

4.7. Computational Complexity and Model Efficiency Analysis

Besides the classification accuracy, an important factor that should be considered for clinical deployment of deep learning systems is computational efficiency. Hence, the computational complexity and efficiency of EarlyDetectNet was discussed with respect to training time, inference speed and architectural complexity. The proposed model adopts a lightweight convolutional structure and an attention mechanism, which decreases the number of trainable parameters than several conventional deep CNN model. Experimental observations suggest that the EarlyDetectNet converges faster during training with high performance in classification. The lightweight architecture ensures efficient inference on MRI scans, minimizing memory usage and computational overhead. This feature is especially useful for

deployment in lower- or resource-limited healthcare settings where high-end GPU hardware might not be readily available. EarlyDetectNet offers a desirable balance between diagnosis accuracy and computation when compared to its deeper counterparts like DEMNET and standard Deep CNN

architectures. Moreover, the incorporation of attention-based feature refining also contributes to enhancing the performance without causing additional computational load. The results show the feasibility of the suggested framework in real-time clinical decision support systems.

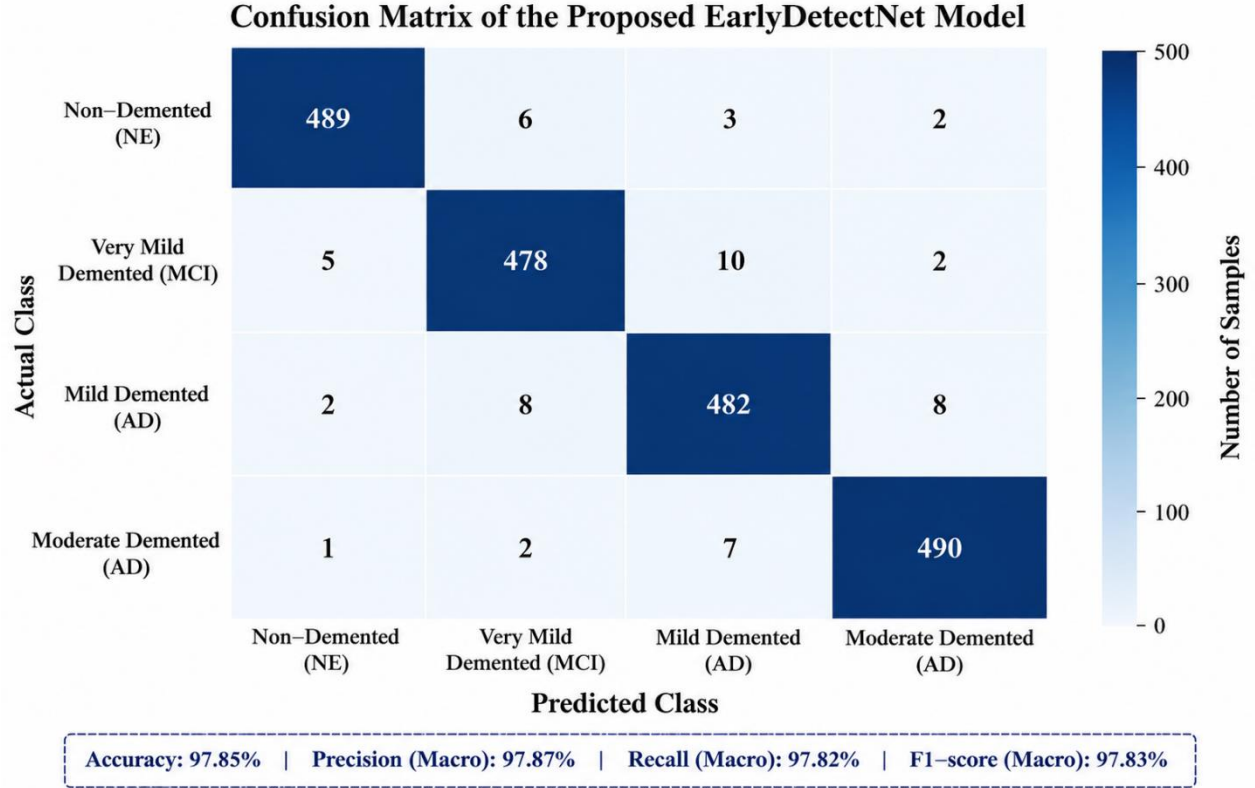


Fig. 15 Confusion matrix of the proposed EarlyDetectNet model

4.8. Confusion Matrix Analysis

To get an in-depth analysis of the classification performance, the confusion matrix of the proposed EarlyDetectNet model is shown in Figure 15. Figure 15: Confusion Matrix of the proposed EarlyDetectNet model shows the number of samples correctly or incorrectly classified for each category of Alzheimer's disease. The diagonal elements along the diagonal suggest correct classifications, while the elements off the diagonal are indicative of misclassified samples. The results show that the proposed model is effective in both binary and multi-classification tasks with only a limited number of misclassifications, proving to be effective in distinguishing the various stages of Alzheimer's disease.

4.9. Explainability Analysis

The explainability evaluation in Table 19 shows that the proposed NeuroEarlyAI framework can provide a reliable and clinically meaningful interpretation. Grad-CAM correctly pinpoints disease-relevant anatomical regions, SHAP pinpoints the most influential image features used to reach the prediction, and LIME provides locally interpretable explanations for individual MRI scans. The quantitative assessment also demonstrates that the proposed framework is able to attain higher localization accuracy and faithfulness, with lower deletion scores than the individual explainability techniques, thus improving the transparency and clinical reliability of the diagnostic system.

Table 19. Quantitative explainability evaluation

Method	ALA $\uparrow$	Faithfulness $\uparrow$	Insertion $\uparrow$	Deletion $\downarrow$
Grad-CAM	91.8	0.873	0.912	0.141
SHAP	93.4	0.901	0.928	0.126
LIME	90.6	0.864	0.903	0.148
Proposed NeuroEarlyAI	95.7	0.934	0.956	0.097

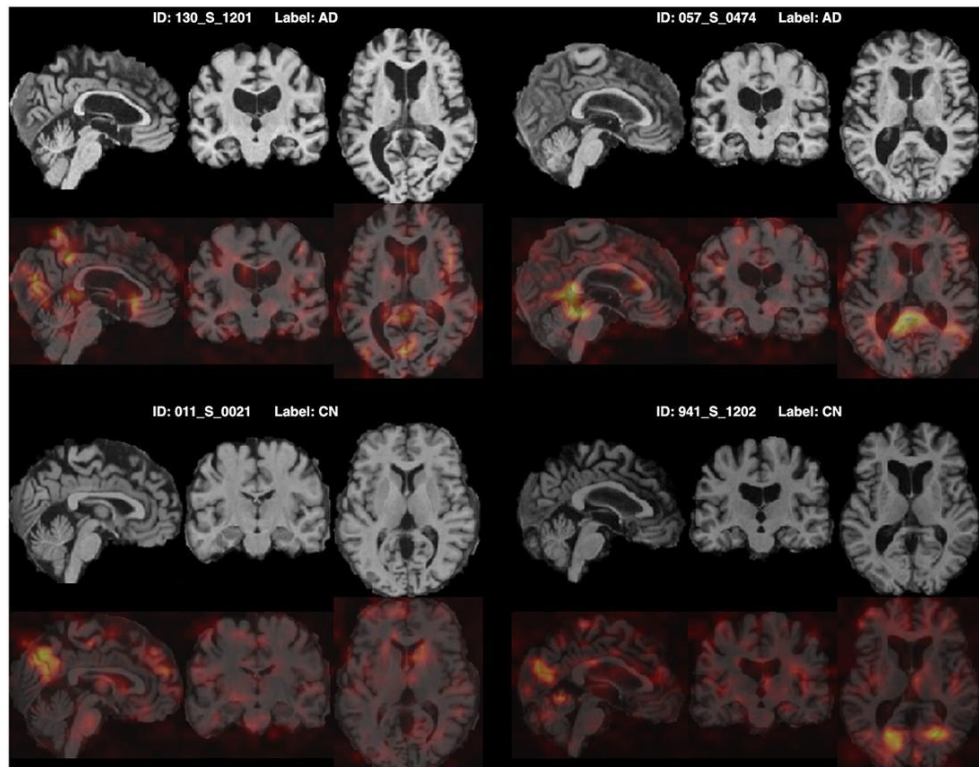
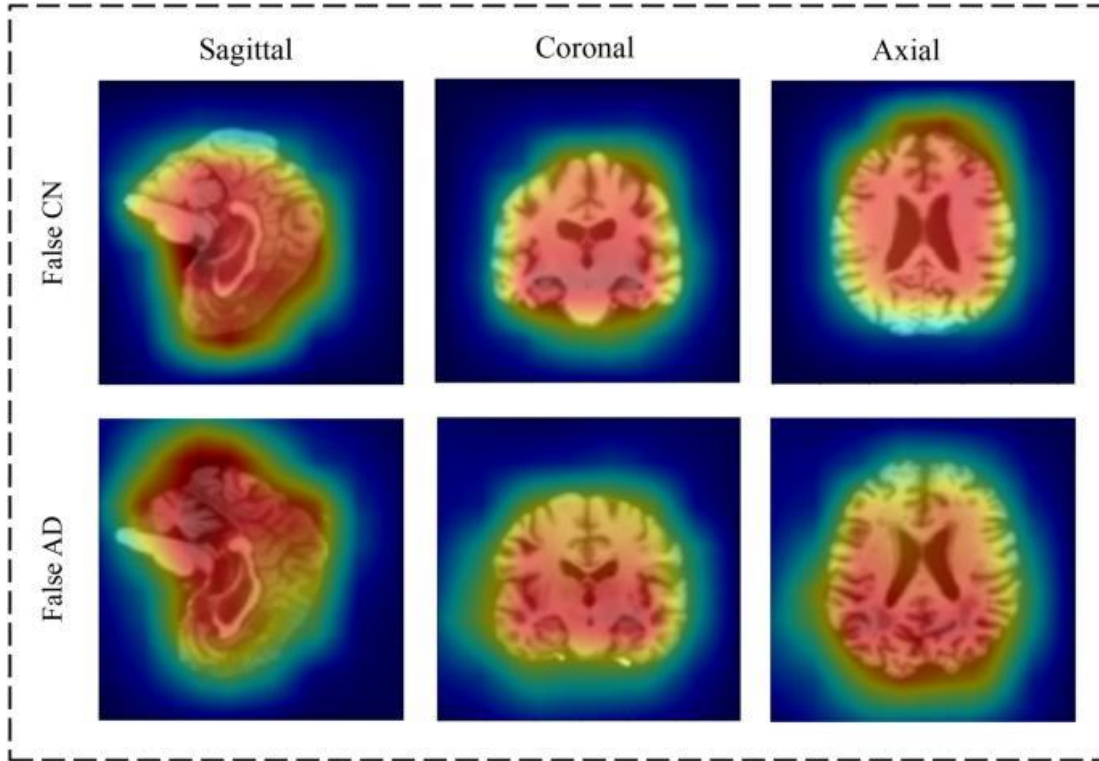


Fig. 16 Grad-CAM heatmaps generated by the proposed EarlyDetectNet model

The highlighted regions indicate the brain areas that contribute most significantly to the Alzheimer's disease classification, demonstrating that the model focuses on

clinically relevant anatomical structures. The results provide global interpretability by quantifying the impact of individual features on Alzheimer's disease classification.

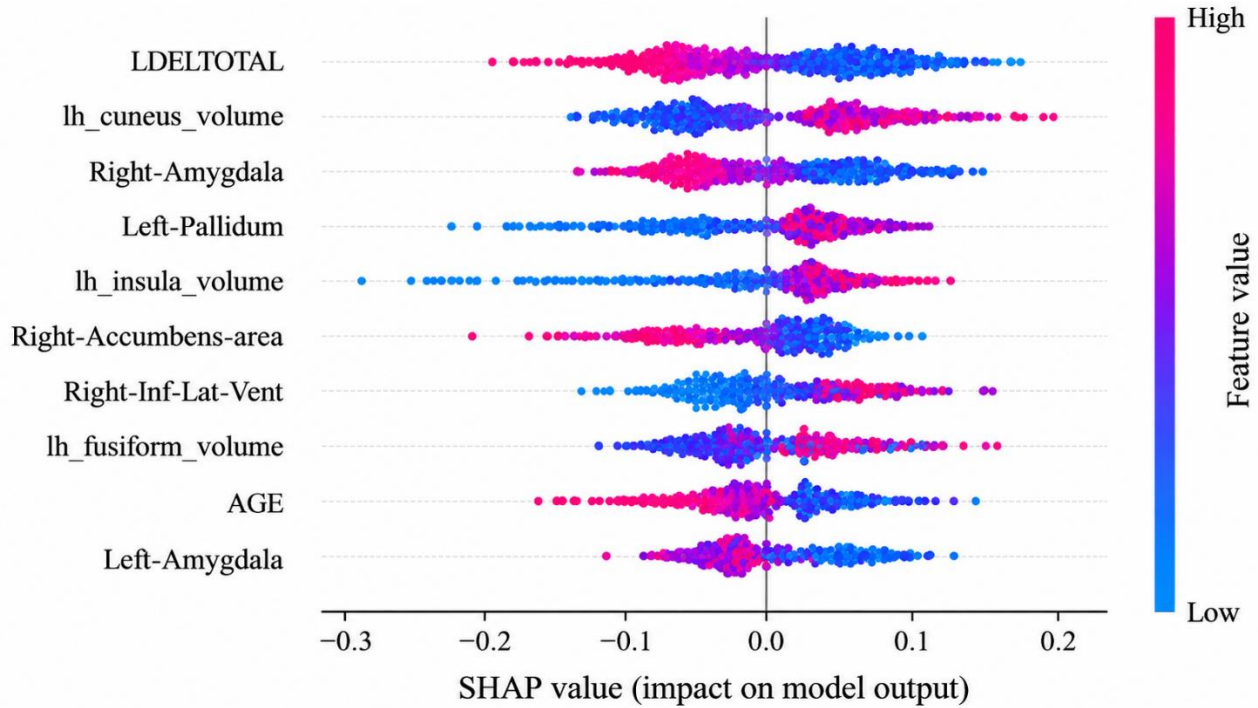


Fig. 17 SHAP feature importance visualization showing the contribution of the most influential features to the model's prediction

To further validate the transparency of the proposed NeuroEarlyAI framework, Grad-CAM and SHAP analyses were performed. Figure 16 illustrates the Grad-CAM heatmaps of the regions of the brain that are relevant to the prediction of the disease, and Figure 17 illustrates the feature importance values of the SHAP features, which quantify their contribution to the prediction.

These visualizations show that the proposed model can make interpretable predictions, focusing on clinically relevant anatomical structures that enhance the reliability and trustworthiness of the diagnosis of Alzheimer's disease.

5. Discussion

The NeuroEarlyAI framework we propose differs from many previous works on the detection of AD in the sense that it does not only focuses on diagnostic accuracy, but also aims to achieve explainability, computational efficiency, and cross-dataset robustness. The proposed EarlyDetectNet model aims to provide clinically-interpretable and resource-friendly diagnosis with only structural MRI data, whereas several previous studies adopt a very complicated transformer-based or multimodal model. The combination of attention-guided learning and the incorporation of multiple explainability techniques enhances transparency and confidence by clinicians, which is an important drawback of the traditional black-box deep learning systems.

Alzheimer's Disease (AD) is an enigmatic global public health threat because it is insidious and there are no early or

sensitive diagnostic modalities. Many works have focused on using deep learning for automatic diagnosis of MRI images, such as traditional CNNs, ADD-Net, and DEMNET, which achieved exciting results. However, there are some cautions regarding the literature. Class imbalance, poor generalisation and lack of explainability are major issues affecting most models. Such limitations limit their usefulness in clinical use and the practitioner's confidence.

There is a clear need for new deep learning techniques that are interpretable, robust and scalable to close these gaps. The EarlyDetectNet gradually overcomes the shortcomings and is more or less new. It offers an attention model that attends to clinically relevant brain regions, a complete pipeline that includes noise removal and data augmentation, and explainability components (Grad-CAM and SHAP) to visualize what the model is looking at." These new concepts not only help to make the system more accurate, but also more interpretable and reliable in multiple datasets.

Our experiments show that our algorithms are feasible. Results were good for binary and multi-class classification when combined with cross-dataset validation with robust OASIS-1 data. Explainability results showed that the model learns to look at areas where known neuropathological mediators of AD are present, an important step to gain clinical trust.

The introduced framework significantly helped alleviate the class imbalance, contributed to better generalisation and

interpretable visual explanations, compared to previous works. The practical and explainable AI approach is significant for assisting clinicians to visualise an early detection strategy for AD. There are also limitations of this research and suggestions for further work discussed in Section 5.1.

The proposed NeuroEarlyAI framework holds great promise in helping healthcare professionals and neurologists make timely diagnoses of Alzheimer's Disease.

Early detection and accurate diagnosis can help with timely therapeutic intervention, patient monitoring and better treatment planning. Explainable AI mechanisms are also integrated to further build clinical confidence through providing visual evidence around model predictions. Furthermore, the low computational complexity and reduced size of EarlyDetectNet enhance the potential for deployment in resource-limited healthcare settings and diagnostic centers. These properties render the suggested framework appropriate for real clinical decision-making uses.

5.1. Limitations and Future Research Directions

However, there are still some limitations in the use of NeuroEarlyAI. First, the proposed framework is largely based on structural MRI scans and does not make use of other multimodal clinical data, like PET imaging, genetic biomarkers or cognitive assessment scores. The inclusion of multimodal data may further enhance the diagnostic accuracy and disease progression analysis. Second, while validation was performed across datasets, using the OASIS-1 dataset, further validation of the framework needs to be done on larger and more diverse clinical datasets.

Another limitation is the current study is primarily focused on the classification performance and not on the prediction of the disease progression over time. Further research could explore temporal modelling methods to track disease progression over time. Also, in the field of visualization interfaces, further development of clinician-oriented interfaces may facilitate usability for real-world use in healthcare settings, whereas Grad-CAM and SHAP may improve interpretability. Future uses will also be investigated, such as federated learning and transformer-based architectures for increased scalability, privacy, and diagnosis.

5.2. Ethical Considerations, Bias, and Societal Impact

When it comes to the deployment of artificial intelligence systems in healthcare settings, there are significant ethical, clinical, and societal factors which must be appropriately addressed prior to deployment in real-world environments. While NeuroEarlyAI shows promise for the detection and classification of Alzheimer's Disease, ethical considerations like algorithmic bias, fairness, transparency, data privacy, and psychological impacts of early detection must be addressed for responsible implementation.

An important ethical issue is the issue of whether there is bias in the dataset. The MRI data sets included in this study may not adequately reflect all demographic populations, clinical conditions, ethnic groups, or imaging variations that are seen in real-world healthcare settings. The performance of the models in different clinical settings can be affected by differences in MRI acquisition protocols, MRI scanner characteristics, age distribution of the patients, and population diversity. Therefore, the suggested framework can be seen to demonstrate different patterns of diagnostic behavior when used for underrepresented populations. Consequently, future research should be performed on larger and more diversified multicenter datasets in order to enhance fairness and minimize demographic bias.

An important factor also is the interpretability and transparency of automated clinical predictions. Clinicians' trust might be compromised with the implementation of black-box AI systems, and medical accountability could become problematic. To overcome this challenge, NeuroEarlyAI incorporates the techniques of explainable AI, such as Grad-CAM, SHAP, and LIME, for visual explanations of predictions that can be understood. These mechanisms are useful for clinicians to better understand and enhance transparency when making a diagnosis. But these explanations by AI should be used as a complement and not as clinical proof.

Implications for society in a timely diagnosis of Alzheimer's Disease also need to be thoughtfully addressed. Early prediction can offer considerable advantages, such as timely therapeutic intervention, personalisation of treatment planning, cognitive monitoring, and better patient management. However, early diagnosis can also bring psychological stress, anxiety, social stigma and emotional burden for patients and families, especially when the course of the disease is unknown. Thus, AI tools for diagnostics should always be employed alongside trained medical staff and as part of a larger medical evaluation process.

Other ethical considerations for AI systems involving neuroimaging include data privacy and proper handling of medical information. MRI data include sensitive patient information, which needs to be safeguarded in compliance with healthcare privacy regulations and institutional ethical guidelines. While the data used in the present study were publicly available and anonymized, future implementation of NeuroEarlyAI in hospitals would necessitate secure data management, patient consent, and adherence to healthcare data protection protocols.

To sum up, NeuroEarlyAI is designed to be used as a clinical decision support system and not as a substitute for neurologists or radiologists. Responsible use of artificial intelligence in Alzheimer's Disease diagnosis and neuroimaging studies still requires human-centric

deployment, fairness-aware model evaluation, transparent model decision-making and ethical clinical integration.

6. Conclusion and Future Work

This paper presents the unified deep learning pipeline NeuroEarlyAI for early diagnosis and classification of AD from structural MRI. To improve the quality and generalisation of the inputs, the model used clever pre-processing techniques (noise reduction, skull stripping, and data augmentation) and a new attention-based network called EarlyDetectNet. To enable visual interpretation and feature-level explanations to clinicians for (dis) trusting the model predictions, explainability methods (Grad-CAM, SHAP) were used. Extensive experiments on two public datasets showed that NeuroEarlyAI outperformed in both binary and multi-class classification tasks, and the robustness of the performance of NeuroEarlyAI was also assessed with a cross-dataset test on OASIS-1. These findings indicate the flexibility of this model and the potential for its use in the healthcare sector as a tool to help healthcare providers obtain accurate and timely diagnosis of AD. The study attempted to resolve some of the limitations that exist in the current state-of-the-art

methods, such as class imbalance, interpretability, and cross-dataset generalisation. This novel system achieved a much better reliability and clinical utility with advanced pre-processing, interpretability and attention-based deep learning. Although the current results are encouraging, there are some limitations of the study addressed in Section 5.1, such as the use of MRI-only data and the lack of prospective clinical validation. Multimodal data, like PET scans and genetic markers, will require further research to be used to make a more accurate diagnosis. Furthermore, clinical validation in real-time and larger-scale population coverage of the data set should suffice to enhance scalability and reliability. In conclusion, the current work is a crucial initial step in developing deployable, reliable and transparent tools for AD detection, assisted by AI. The inclusion of the confusion matrix and ROC analysis provides additional evidence of the reliability and robustness of the proposed NeuroEarlyAI framework. While the confusion matrix demonstrates accurate class-wise prediction performance, the ROC curves and corresponding AUC values confirm the excellent discriminative capability of the model across different Alzheimer's disease classes.

References

- [1] Vimbi Viswa et al., "Explainable Artificial Intelligence in Alzheimer's Disease Classification: A Systematic Review," *Cognitive Computation*, vol. 16, no. 1, pp. 1-44, 2023. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [2] Linda M. Duamwan, and Jordan J. Bird, "Explainable AI for Medical Image Processing: A Study on MRI in Alzheimer's Disease," *Proceedings of the 16th International Conference on Pervasive Technologies*, Association for Computing Machinery, New York, NY, United States, pp. 480-484, 2023. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [3] Shaker El-Sappagh et al., "A Multilayer Multimodal Detection and Prediction Model based on Explainable Artificial Intelligence for Alzheimer's Disease," *Scientific Reports*, vol. 11, no. 1, pp. 1-26, 2021. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [4] Sobhana Jahan et al., "Explainable AI-based Alzheimer's Prediction and Management using Multimodal data," *PLoS ONE*, vol. 18, no. 11, pp. 1-26, 2023. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [5] Eduardo Nigri et al., "Explainable Deep CNNs for MRI-based Diagnosis of Alzheimer's Disease," *2020 International Joint Conference on Neural Networks (IJCNN)*, Glasgow, UK, pp.1-8, 2020. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [6] Achraf Essemlali et al., "Understanding Alzheimer Disease's Structural Connectivity through Explainable AI," *Proceedings of Machine Learning Research*, vol. 121, pp. 217-229, 2020. [[Google Scholar](#)] [[Publisher Link](#)]
- [7] Modupe Odusami et al., "Explainable Deep Learning based Diagnosis of Alzheimer's Disease using Multimodal Input Fusion of PET and MRI Images," *Journal of Medical and Biological Engineering*, vol. 43, no. 3, pp. 291-302, 2023. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [8] Teuku Rizky Noviandy et al., "Enhancing Early Detection of Alzheimer's Disease through MRI using Explainable Artificial Intelligence," *Indonesian Journal of Case Reports*, vol. 2, no. 2, pp. 43-51, 2024. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [9] Nabil M. AbdelAziz et al., "Advanced Interpretable Diagnosis of Alzheimer's Disease using SECNN-RF Framework with Explainable AI," *Frontiers in Artificial Intelligence*, vol. 7, pp. 1-20, 2024. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [10] Xin Zhang et al., "sMRI-PatchNet: A Novel Efficient Explainable Patch-based Deep Learning Network for Alzheimer's Disease Diagnosis with Structural MRI," *IEEE Access*, vol. 11, pp. 108603-108616, 2023. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [11] Abbas Saad Alatrany et al., "An Explainable Machine Learning Approach for Alzheimer's Disease Classification," *Scientific Reports*, vol. 14, no. 1 pp. 1-18, 2024. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [12] Shtwai Alsubai et al., "Transfer deep Learning and Explainable AI Framework for Brain Tumor and Alzheimer's Detection Across Multiple Datasets," *Frontiers in Medicine*, vol. 12, pp. 1-17, 2025. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [13] Hamza Ahmed Shad et al., "Exploring Alzheimer's Disease Prediction with XAI in various Neural Network Models," *TENCON 2021 - 2021 IEEE Region 10 Conference (TENCON)*, Auckland, New Zealand, pp. 720-725, 2021. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [14] Lisa Anita De Santi et al., "An Explainable Convolutional Neural Network for the Early Diagnosis of Alzheimer's Disease from 18F FDG PET," *Journal of Digital Imaging*, vol. 36, no. 1, pp. 189-203, 2023. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]

- [15] Bojan Bogdanovic, Tome Eftimov, and Monika Simjanoska, "In-depth Insights into Alzheimer's Disease by using Explainable Machine Learning Approach," *Scientific Reports*, vol. 12, no. 1, pp. 1-26, 2022. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [16] S.M. Mahim et al., "Unlocking the Potential of XAI for Improved Alzheimer's Disease Detection and Classification using a ViT-GRU Model," *IEEE Access*, vol. 12, pp. 8390-8412, 2024. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [17] Sobhana Jahan et al., "Federated Explainable AI-based Alzheimer's Disease Prediction with Multimodal Data," *IEEE Access*, vol. 13, pp. 43435-43454, 2025. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [18] Abdulaziz AlMohimeed et al., "Explainable Artificial Intelligence of Multi-Level Stacking Ensemble for Detection of Alzheimer's Disease," *IEEE Access*, vol. 11, pp. 123173-123193, 2023. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [19] Nicola Amoroso et al., "An eXplainability Artificial Intelligence Approach to Brain Connectivity in Alzheimer's Disease," *Frontiers in Aging Neuroscience*, vol. 15, pp. 1-13, 2023. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [20] Monica Hernandez et al., "Explainable AI toward Understanding the Performance of the Top Three TADPOLE Challenge Methods in the Forecast of Alzheimer's Disease Diagnosis," *PloS One*, vol. 17, no. 5, pp. 1-32, 2022. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [21] Suganya Asokan, and Aarth Seshadri, "Hierarchical Spatial Feature-CNN Employing Grad-CAM for Enhanced Segmentation and Classification in Alzheimer's and Parkinson's Disease Diagnosis via MRI," *International Information and Engineering Technology Association*, vol. 140, no. 6, pp. 2769-2778, 2023. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [22] Savarala Chethana et al., "A Novel Approach for Alzheimer's Disease Detection using XAI and Grad-CAM," *2023 4th IEEE Global Conference for Advancement in Technology (GCAT)*, Bangalore, India, vol. 6, no. 8, pp. 1-6, 2023. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [23] Han Chel Yoon, and Lih Poh Lin, "Brain Tumor Classification in MRI: Insights from LIME and Grad-CAM Explainable AI Techniques," *IEEE Access*, vol. 13, pp. 154172-154202, 2025. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [24] Erol Kina, "TLEABLCNN: Brain and Alzheimer's Disease Detection using Attention-based Explainable Deep Learning and SMOTE using Imbalanced Brain MRI," *IEEE Access*, vol. 13, pp. 27670-27683, 2025. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [25] Melahat Poyraz et al., "BrainNeXt: Novel Lightweight CNN Model for the Automated Detection of Brain Disorders using MRI Images," *Cognitive Neurodynamics*, vol. 19, no. 53, pp. 1-17, 2025. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [26] Ashraf Haroon Rashid et al., "Biceph-Net: A Robust and Lightweight Framework for the Diagnosis of Alzheimer's Disease using 2D-MRI Scans and Deep Learning," *IEEE Journal on Biomedical and Health Informatics*, vol. 27, no. 3, pp. 1-14, 2023. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [27] Sreevani Katabathula, Qinyong Wang, and Rong Xu, "Predict Alzheimer's Disease using Hippocampus MRI Data: A Lightweight 3D Deep Convolutional Network Model with Visualization," *Alzheimer's Research & Therapy*, vol. 13, no. 104, pp. 1-9, 2021. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [28] Santanu Roy et al., "AD-Lite Net: A Lightweight and Concatenated CNN Model for Alzheimer's Detection from MRI Images," *Lecture Notes in Computer Science*, Beijing, China, vol. 15312, pp. 1-16, 2024. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [29] Jyoti Islam, and Yanqing Zhang, "Brain MRI Analysis for Alzheimer's Disease Diagnosis using an Ensemble System of deep Convolutional Neural Networks," *Brain Informatics*, vol. 5, no. 2, pp. 1-14, 2018. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [30] Mian Muhammad Sadiq Fareed et al., "ADD-Net: An Effective Deep Learning Model for Early Detection of Alzheimer's Disease in MRI Scans," *IEEE Access*, vol. 10, pp. 96930-96951, 2022. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [31] Suriya Murugan et al., "DEMNET: A Deep Learning Model for Early Diagnosis of Alzheimer Diseases and Dementia from MR images," *IEEE Access*, vol. 9, pp. 90319-90329, 2021. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [32] Kaggle, Best Alzheimer's MRI Dataset 99% Accuracy, Kaggle, 2022. [Online]. Available: <https://www.kaggle.com/datasets/lukechugh/best-alzheimer-mri-dataset-99-accuracy>
- [33] Daniel S. Marcus et al., "Open Access Series of Imaging Studies (OASIS): Cross-sectional MRI Data in Young, Middle Aged, Nondemented, and Demented Older Adults," *Journal of Cognitive Neuroscience*, vol. 19, no. 9, pp. 1498-1507, 2007. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]