Gümüşhane University Journal of Science

GUFBD / *GUJS* (2025) 15(3): 829-853 doi: 10.17714/gumusfenbil.1714884 Research Article

Early diagnosis of Alzheimer's Disease using hybrid CNN-Transformer models with Grad-CAM interpretability

Grad-CAM yorumlanabilirliği ile hibrit CNN-Transformer modeller kullanılarak Alzheimer Hastalığının erken tanısı

Pakize ERDOĞMUŞ[1], Abdullah Talha KABAKUŞ*[1]

[1]Düzce Üniversitesi, Mühendislik Fakültesi, Bilgisayar Mühendisliği Bölümü, 81620, Düzce

• Received: 05.06.2025 • Accepted: 23.08.2025

Abstract

Detecting Alzheimer's Disease (AD) at an early stage is vital because it enables prompt treatment and intervention, which can help slow disease progression and enhance patient prognosis. Given the increasing prevalence of AD globally, with an estimated 50 million people currently living with the condition and projected to triple by 2050, the development of accurate and efficient diagnostic tools is paramount. In this study, a novel architecture for the early diagnosis of AD by combining Convolutional Neural Networks (CNNs) or Vision Transformers (ViTs) with traditional Machine Learning (ML) algorithms was proposed. Utilizing MRI images as input, CNNs/ViTs serve as feature extractors, while demographic data is integrated to enhance diagnostic accuracy. Through extensive experimentation, our proposed model, which utilizes a CNN backbone optimized for MRI analysis as a feature extractor and LGBM as the classifier, achieved superior accuracy, reaching up to 96.83%. Statistical validation through confidence intervals and McNemar's test further demonstrated the robustness and significant performance improvements of the proposed model compared to baseline methods. This study employs eXplainable AI techniques to visualize critical regions in MRI images that influence the model's diagnostic decisions, promoting clinical transparency and trust in AI-assisted early diagnosis of AD. The novelty of this study lies in integrating deep feature extractors (CNNs/ViTs) with traditional ML classifiers, supported by interpretability through Grad-CAM and statistical validation, offering a transparent and accurate framework for early diagnosis of AD.

Keywords: Alzheimer's disease, Convolutional neural network, Dementia, Explainable AI, Magnetic resonance imaging, Vision transformer

Öz

Alzheimer Hastalığını (AH) erken evrede tespit etmek hızlı tedavi ve müdahaleye olanak sağlaması açısından çok önemlidir. Bu sayede hastalığın ilerlemesi yavaşlatılabilir ve hasta prognozu iyileştirilebilir. Dünya genelinde AH'nin artan yaygınlığı göz önüne alındığında — şu anda yaklaşık 50 milyon kişinin bu hastalıkla yaşadığı ve bu sayının 2050 yılına kadar üç katına çıkacağı öngörüldüğünde — doğru ve etkili tanı araçlarının geliştirilmesi kritik hale gelmiştir. Bu çalışmada, Konvolüsyonel Sinir Ağları (CNN'ler) veya Görüntü Dönüştürücüler (ViT'ler) ile geleneksel Makine Öğrenmesi (ML) algoritmalarını birleştirerek Alzheimer hastalığının erken tanısına yönelik özgün bir mimari sunulmaktadır. Girdi olarak MRI (Manyetik Rezonans Görüntüleme) görüntülerini kullanan CNN/ViT modelleri özellik çıkarıcı olarak işlev görmekte ve tanı doğruluğunu artırmak amacıyla demografik verilerle birleştirilmektedir. Gerçekleştirilen kapsamlı deneyler sonucunda, MRI analizi için optimize edilmiş bir CNN tabanlı özellik çıkarıcı ile LGBM sınıflandırıcısının kullanıldığı önerilen modelimiz %96,83'e varan doğruluk oranı ile üstün performans sergilemiştir. Güven aralıkları ve McNemar testi yoluyla yapılan istatistiksel doğrulamalar, önerilen modelin temel yöntemlere kıyasla sağlamlığını ve anlamlı performans iyileştirmelerini desteklemiştir. Bu çalışma, Açıklanabilir Yapay Zeka tekniklerini kullanarak modelin tanısal kararlarını etkileyen MRG görüntülerindeki kritik bölgeler görselleştirilmiş ve böylece yapay zeka destekli erken teşhis süreçlerinde klinik şeffaflık ve güven teşvik edilmiştir. Bu çalışmanın özgünlüğü, derin özellik çıkarıcıların (CNN'ler/ViT'ler) geleneksel ML sınıflandırıcılarıyla bütünleştirilmesinde yatmaktadır. Bu yapı, Grad-CAM tabanlı yorumlanabilirlik ve istatistiksel doğrulama ile desteklenerek, erken AH tanısı için şeffaf ve yüksek doğrulukta bir çerçeve sunmaktadır.

Anahtar kelimeler: *Alzheimer hastalığı, Evrişimsel sinir ağı, Demans, Açıklanabilir yapay zeka, Manyetik rezonans görüntüleme, Görüntü dönüştürücüsü*

*Abdullah Talha KABAKUŞ; talhakabakus@duzce.edu.tr

ISSN: 2146-538X e-ISSN: 2146-538X https://dergipark.org.tr/tr/pub/gumusfenbil

1. Introduction

Alzheimer's Disease (AD) was first defined by *Alois Alzheimer* in 1906. The initial patient described in the Alzheimer's report exhibited key features of the disorder commonly observed in subsequent patients, including progressive memory loss, cognitive dysfunction, altered behavior such as paranoia and delusions, and a gradual decline in language skills (Grossberg et al., 2019). The evolution of AD diagnosis has progressed through several stages, as outlined by *Selkoe* (Selkoe, 2001): Electron Microscopy (1960), Neurochemicals (mid-1970s), Pharmacological Research and Approved Drugs (1970s-1980s), Identification of Variable Neurotransmitter Deficits (late 1970s-early 1980s), Genetic Discoveries (1990s), and Medical Imaging (2000s). Prior to the invention of the electron microscope in 1960, little advancement was made in understanding the pathogenesis of AD. The introduction of electron microscopy enabled the identification of two hallmark lesions, senile plaques, and neurofibrillary tangles, which were linked to AD (Armstrong, 2009). In the mid-1970s, the onset of dementia in AD patients was associated with reduced levels of certain enzymes, particularly choline acetyltransferase and acetylcholinesterase (Pappas et al., 2000). Pharmacological research in the 1970s-1980s aimed to elevate acetylcholine levels in the brain. Subsequent to the 1980s, researchers identified deficits in various neurotransmitter systems in AD brain tissue. The discovery of specific genes associated with familial forms of AD shed light on the genetic underpinnings of the disorder. Mutations in genes such as APP (Amyloid Precursor Protein), Presenilin 1, and Presenilin 2 were found to be linked to early-onset familial AD (Wong et al., 2020). The advancement of medical imaging techniques has provided researchers with the ability to visualize the brains of AD patients and monitor disease progression.

Today, researchers are directing their efforts toward developing treatments that target the underlying causes of AD, such as reducing beta-amyloid plaques, tau tangles, and brain inflammation (Lukiw, 2012; Grossberg et al., 2019). While current treatments primarily focus on symptom management and slowing disease progression, there is increasing emphasis on early diagnosis to effectively slow disease progression. The diagnosis of AD typically involves clinical assessments, cognitive tests, and neuroimaging. Recent studies have integrated various assessment tools to enhance diagnostic accuracy (Qiu et al., 2018). Traditional diagnostic methods, such as cognitive testing and brain imaging, often detect the disease in its advanced stages, limiting intervention options. Recent advances in Artificial Intelligence (AI), particularly in Deep Learning (DL), have shown great promise in enhancing early detection. Convolutional Neural Networks (CNNs) have demonstrated effectiveness in capturing spatial features from medical imaging modalities, such as Magnetic Resonance Imaging (MRI) and Positron Emission Tomography (PET) scans, thereby aiding in accurate classification. However, CNNs may struggle with comprehension of the global context. To address this, Vision Transformers (ViTs), which excel in processing long-range dependencies, have been introduced to complement CNNs. ViTs represent a transformative approach to image classification, offering a significant shift from the traditional convolution-based models. Unlike CNNs, which rely on convolutional operations, ViTs utilize self-attention mechanisms to capture both local and global features from an image. ViTs segment an image into patches and utilize self-attention mechanisms across these patches, enabling the model to capture interactions between different image regions without relying on convolutional operations. This architecture has been particularly successful in modeling long-range dependencies and providing a global understanding of image data. In medical imaging, where understanding spatial relationships across various regions of the brain is critical, ViTs offer an advantage by effectively modeling these interactions. Their capacity to handle global information makes them a promising tool for enhancing AD detection, where identifying subtle and widespread brain abnormalities is essential. By utilizing either the local feature extraction capabilities of CNNs or the global attention mechanisms of ViTs, each model offers a distinct approach for detecting early-stage AD, with the potential to improve diagnostic accuracy and enable earlier interventions. In this study, we propose a novel architecture for the early diagnosis of AD, following the identification of the optimal feature extractor and classifier model. Utilizing the gold standard *OASIS-2* MRI dataset, we evaluated a range of prominent CNN-based pre-trained architectures and ViT-based architectures as feature extractors. Demographic features from the *OASIS-2* MRI dataset, along with the extracted features, were assessed using twelve traditional ML classifiers. The main contributions of this study are summarized as follows:

- A novel architecture is proposed for the early diagnosis of AD by combining CNNs or ViTs with traditional Machine Learning (ML) algorithms, leveraging the strengths of both paradigms to enhance diagnostic performance.
- The proposed model utilizes MRI scans as input, where CNNs or ViTs function as feature extractors to capture complex neuroanatomical patterns associated with AD.

- Demographic information is integrated into the diagnostic framework to provide additional patient-specific context that complements imaging data, thereby improving diagnostic accuracy.
- Extensive experimental results demonstrate that the proposed approach achieves outstanding accuracy—up to 96.83%—when using *InceptionV3* or *VGG19* as feature extractors combined with the *LightGBM* (*LGBM*) classifier, surpassing state-of-the-art methods.
- An ablation study verifies that the integration of CNNs with traditional ML classifiers outperforms the use of standalone CNNs or ViTs.
- The robustness of the proposed method is supported by the computation of 95% confidence intervals and statistical significance testing via McNemar's test.
- Feature map analyses of CNNs, ViTs, and BEiT models reveal distinct local and global patterns captured by each architecture.
- Grad-CAM is employed to enhance model interpretability by generating localized visual explanations for model predictions, thereby supporting clinical transparency and trust.
- A fine-tuning ablation study reveals that frozen CNN feature extractors consistently outperform end-to-end training on small neuroimaging datasets, validating the methodological choice and offering practical guidelines for model adaptation based on dataset size.

The rest of the paper is organized as follows: Section 2 provides an overview of the related work. Section 3 outlines the materials and methods utilized in the proposed study. Section 4 presents the experimental results and accompanying discussion. Finally, Section 5 presents concluding remarks and outlines potential future directions.

2. Related work

AD is a progressive neurodegenerative condition that mainly impairs memory and cognitive abilities. With the global population aging, the incidence of diagnoses such as Parkinson's Disease, AD, and heart disease is rising due to increased longevity. The probability of an AD diagnosis doubles approximately every five years after reaching the age of 65 (Kaeberlein, 2013). Diagnosis of Alzheimer's typically involves analyzing various data sources, including voice recordings, medical images such as MRI scans, medical diagnostic tests like MMSE, and demographic information. While some studies focus on utilizing individual data types, others integrate multiple data sources to enhance diagnostic accuracy (Lauraitis et al., 2020; Erdogmus & Kabakus, 2023). (Agbavor & Liang, 2022) introduced an end-to-end AI-driven system designed for AD detection and assessment of AD severity directly from voice data. Notable medical imaging datasets utilized for AD diagnosis include *The Alzheimer's Disease Neuroimaging Initiative* (*ADNI*), *The Open Access Series of Imaging Studies* (*OASIS*), *Minimal Interval Resonance Imaging in Alzheimer's Disease* (*MIRIAD*), and *The Australian Imaging, Biomarker & Lifestyle Flagship Study of Aging* (*AIBL*) (Khojaste-Sarakhsi et al., 2022). (Abrol et al., 2020) utilized the popular *ResNet* (He et al., 2016) framework, incorporating three residual blocks, to perform AD classification using 3D structural MRI data. They modified the original ResNet design to better accommodate the complexities of neuroimaging analysis. In another study, (Shanmugam et al., 2022) compared the performance of *ResNet18* with *GoogleNet* (Szegedy et al., 2015) and *AlexNet* (Krizhevsky et al., 2012), concluding that *ResNet* achieved superior results in the early identification of AD. (Ji et al., 2019b) introduced an ensemble technique that integrated *ResNet50*, *NASNet*, and *MobileNet* for early AD detection. Another method (Mehmood et al., 2021) made use of the *VGG19* (Simonyan & Zisserman, 2015) model, a widely recognized CNN structure, to identify cases of AD. Specifically targeting 3D data, the 3D-DenseNet architecture (Huang et al., 2017) was deployed in (Li & Liu, 2018) to extract localized features from various brain areas; these features were later fused to classify between Normal Controls (NC) and AD, as well as NC and Mild Cognitive Impairment (MCI). Several recent efforts have incorporated attention mechanisms into their models (Ji et al., 2019a), aiming to emphasize regions and features carrying the most critical diagnostic information, thereby boosting classification accuracy (Fathi et al., 2022). Deep Polynomial Networks (DPNs), a supervised learning approach that applies linear or quadratic transformations at each neuron to produce polynomial mappings, have also been explored (Livni et al., 2013). (Shi et al., 2018) introduced a multi-modal variant named MM-SDPN for AD diagnosis. Meanwhile, (M. Liu et al., 2018) proposed a hybrid model combining 2D-CNNs and Bidirectional GRUs (BiGRUs) for classifying AD from FDG-PET scans. A similar framework blending 3D-CNNs with BiGRUs was developed by (Cui & Liu, 2019) for the same purpose. Additionally, (S. Liu et al., 2014) designed a Stacked Autoencoder (SAE) featuring three hidden layers capped by a softmax classifier to facilitate early detection of AD using both MRI and PET imaging modalities. This study was among the first to apply DL in AD diagnosis. (Asl et al., 2018) introduced a supervised learning framework named 3D-DSA-CNN, designed specifically for binary and multi-class classification of Alzheimer's disease. Their approach is based on a 3D Convolutional Autoencoder (3D-CAE) architecture, comprising three stacked convolutional autoencoding layers, a flattening operation, two fully connected layers, and a final softmax layer for output prediction. (Suk et al., 2014) proposed an approach based on a Deep Boltzmann Machine (DBM) model that was utilized for hierarchical feature representation in AD and MCI diagnosis. (Kamada et al., 2021) introduced a technique utilizing Deep Belief Networks (DBNs) that dynamically modified the network's architecture size in response to the input space throughout the training process. (Cilia et al., 2022) focused on the early diagnosis of neurodegenerative diseases, such as Alzheimer's and Parkinson's, by analyzing handwriting and drawing samples. The researchers developed a method that uses color images to encode dynamic information from handwriting traits, leveraging CNNs for feature extraction. The study not only improved the feature extraction process but also expanded the handwriting sample database with more complex tasks, demonstrating the potential of this approach in early diagnosis through comprehensive experiments. The recent studies using this dataset have been shown as a comparison table by (Balasundaram et al., 2023). Another study (Yildiz & Yildiz, 2023) focused on the early diagnosis of AD by leveraging an open-source dataset comprising disease-specific and demographic features. The researchers developed a classification model based on Artificial Neural Networks (ANN) to distinguish between dementia and non-dementia cases, achieving an accuracy of 98.5%, a Root Mean Square Error (RMSE) of 0.2302, and a Mean Absolute Error (MAE) of 0.1899. Another study (Vernekar & Selva Kumar, 2024) developed CNNs and Capsule Neural Networks (CapsNet) to detect AD's progression in brain scans. The innovative application of eXplainable AI (XAI) techniques provided valuable visualization on model decision-making, allowing clinicians to interpret the results effectively. The models achieved impressive performance, with accuracy rates of 96% (CNN) and 97% (CapsNet), highlighting their robustness in identifying early-stage AD. A recent research (Rehman Butt et al., 2024) introduced a 3D multi-scale CNN model using resting-state fMRI data to detect AD at early stages. By focusing on multi-scale features in the hippocampal region, the model effectively captured both fine-grained and global markers of disease progression. The approach achieved up to 93% classification accuracy and high sensitivity and specificity for early AD detection, demonstrating the promise of multi-scale DL in clinical diagnostics. Another recent study (Bagade & Godse, 2024) systematically compared several advanced CNN architectures (*InceptionV3*, *ResNet152*, *VGG16*, and *VGG19*) on MRI data. While CNNs showed high proficiency in distinguishing advanced dementia, the study advanced the accurate classification of elusive early AD stages. The findings suggest significant potential for CNNs to improve early diagnosis and intervention using medical imaging. (Gasmi et al., 2024) presented a hybrid DL system combining *EfficientNetV2B3* and *Inception-ResNetV2*. The integration of these architectures, optimized with adaptive weighting, resulted in a system that improved the precision and timeliness of early-stage AD. The approach emphasizes the critical need for quick, accurate diagnostics to enable timely patient interventions and improve outcomes. A recent review (Raza et al., 2025) synthesized recent developments in applying DL, especially across multimodal neuroimaging (MRI, PET, etc.), for early AD detection. The review highlights how DL models are advancing diagnostic accuracy, progression prediction, and the integration of data types, while also discussing challenges such as data harmonization and clinical translation.

Recent studies utilizing the *OASIS-2* dataset for AD have been summarized in this paragraph. (Arjaria et al., 2024) focused on using a subset of thirteen attributes provided by the *OASIS-2* dataset to reduce computational costs associated with classification. Employing feature selection techniques, they achieved a classification accuracy of 90%. (Waldo-Benítez et al., 2024) utilized seven features selected through correlation analysis along with MRI images for dementia classification. Their approach, employing the kNN algorithm, yielded the highest average accuracy of 92.13% ± 3.48. (Chen et al., 2024) introduced *LongFormer*, an effective CNN-Transformer architecture for AD classification based on MRI images. (Mahmud et al., 2024) proposed a method integrating deep transfer learning and XAI techniques. Leveraging various CNN architectures and ensembles, their method achieved an impressive accuracy of 96% on MRI OASIS scans. A recent study (Lazli, 2025) evaluated DL models including ViTs, Fully Connected (FC) networks, and Support Vector Machines (SVMs) on OASIS MRI data. The ViT-based model achieved an accuracy of 93.2% for early-stage AD detection. This study highlighted the potential of transformer-based models in extracting spatial features from neuroimaging data. In another study (Ntampakis et al., 2024), the researchers proposed an ensemble of DL models, including custom 3D CNN and *ResNet* variants, to classify dementia stages using *OASIS-2*. The ensemble approach reached a classification accuracy of 94.12%, outperforming standard models across all stages. The study demonstrated the effectiveness of combined architectures in improving diagnostic reliability across diverse stages of AD. Table 1 provides a comparison of related works published in 2024 and 2025. An analysis of related work reveals that, despite significant advancements in applying DL and ML to AD diagnosis, several limitations remain. First, many studies rely solely on medical imaging modalities such as MRI or PET scans, often neglecting the integration of demographic and clinical data, which can improve diagnostic accuracy and interpretability. Additionally, while predefined DNN architectures like *ResNet*, *VGG19*, and *DenseNet* have shown promising results, their large number of parameters can lead to overfitting on small medical datasets, limiting generalizability. Recent attempts to leverage attention mechanisms or hybrid models such as CNN-Transformers have yet to fully address issues of computational efficiency and interpretability, especially in clinical settings where explainability is crucial. Moreover, few studies provide robust statistical validation of their results, leaving uncertainty about the reproducibility and significance of performance improvements. These gaps highlight the need for novel architectures that not only integrate diverse data types but also achieve high diagnostic accuracy with statistical rigor and clinical transparency, which this study aims to address.

Table 1. A comparison of the related works published in 2024 and 2025.

Study	Methodology	Classification accuracy (%)
(Arjaria et al., 2024)	*SVM*, feature selection (13 attributes)	90
(Waldo-Benítez et al., 2024)	*kNN*	92.13
(Chen et al., 2024)	ViT	82.35
(Mahmud et al., 2024)	CNN, transfer learning	96
(Vernekar & Selva Kumar, 2024)	CNN, CapsNet	96 (CNN), 97 (CapsNet)
(Rehman Butt et al., 2024)	Multi-scale CNN	93
(Ntampakis et al., 2024)	Ensemble of DL models	94.12
(Lazli, 2025)	ViT, FC network, and SVM	93.2 (ViT)

3. Material and method

In this section, we delineate the software stack employed in the proposed study, elaborate on the dataset utilized for training the model, encompassing the proposed feature extractors and classifiers, and outline the evaluation metrics employed, each discussed in the subsequent subsections.

3.1. Software stack

The software for the proposed study was implemented using the Python programming language, renowned for its prominence in data science. *Keras* (Chollet, 2017, 2024) served as the framework for implementing the proposed DNNs. Leveraging its compatibility with various DNN backends, *TensorFlow* (Abadi et al., 2016), the leading backend developed by Google, and the default backend of *Keras*, was employed in this study. Additionally, models based on ViT were proposed, for which the widely used Python package *Hugging Face* (Wolf et al., 2020; *Hugging Face – The AI Community Building the Future*, 2024) was employed. Traditional ML models were also integrated into the proposed approach, facilitated by *scikit-learn* (Pedregosa et al., 2011), another extensively utilized Python package. *scikit-learn* offered support for various data preprocessing operations, including dataset splitting, label encoding, and obtaining classification results. For visualizing experimental outcomes, *Matplotlib* (Hunter, 2007; *Matplotlib: Visualization with Python*, 2024) was employed, complemented by *Seaborn* (Waskom, 2021), a Python package built on top of *Matplotlib*, providing a high-level API for creating visually appealing and informative statistical graphics. The software stack used in this study is listed in Table 2.

Table 2. The software stack of the proposed study.

Software component	Vendor	Version
Programming language	*Python*	3.11
DNN backend	*TensorFlow*	2.9.0
DNN API	*Keras*	2.9.0
Transformer API	*Hugging Face*	4.37.2
ML suite and data preprocessing	*scikit-learn*	1.3.0
Data manipulation	*Pandas*	2.1.4
Visualization	*Matplotlib & Seaborn*	3.7.3 & 0.12.2

3.2. Dataset description

A gold standard dataset plays a pivotal role in proposing accurate ML models. In this study, we utilized the *OASIS-2* (Marcus et al., 2010) dataset, which is provided as part of the *OASIS* project. *OASIS-2* is a longitudinal dataset comprising data from 150 participants aged between 60 and 96 years. Each participant underwent at least two MRI scanning sessions, with a minimum interval of one year between sessions, resulting in a total of 373 imaging sessions. For each individual, 3 to 4 T1-weighted MRI images were collected during a single scan visit. In this study, only one scan per participant was utilized, resulting in a final dataset comprising 209 samples, which were obtained from a publicly available version on Kaggle (Tiriki, 2010). The cohort includes both male and female right-handed subjects. Among them, 72 were consistently classified as nondemented throughout the study. Additionally, 64 participants were diagnosed as demented at their initial visit and retained this classification in follow-up sessions, with 51 of these individuals exhibiting mild to moderate AD. Furthermore, an additional 14 subjects initially classified as *nondemented* were later reclassified as *demented* during subsequent visits. *OASIS-2* contains both MRIs and demographic data for the subjects. Some samples from the *OASIS-2* dataset are presented in Fig. 1. It is worth mentioning that while several gold standard datasets, such as *ADNI* and *MIRIAD*, are available, we chose *OASIS-2* as it provides a balanced combination of a moderately sized, well-defined cohort with consistent imaging and clinical data. This makes it particularly suitable for our research objectives, which focus on longitudinal structural MRI analyses in AD. The composition of the OASIS-2 dataset is summarized in Table 3.

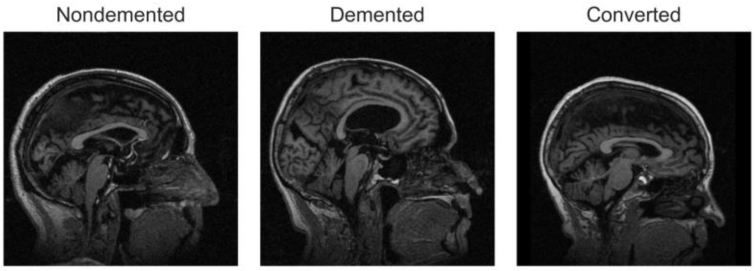

Figure 1. Some samples from the *OASIS-2* dataset. Left-to-Right: A nondemented MRI, a demented MRI, and a converted MRI.

Table 3. OASIS-2 dataset composition.

Category	Count	Details
Total participants	150	Age 60 – 96, right-handed
Total MRI sessions	373	≥ 2 sessions/participant (1-year gap)
Non-demented (longitudinal)	72	Remained non-demented
Demented (longitudinal)	64	51 with mild/moderate AD
Converted to demented	14	Reclassified during follow-up
T1-weighted MRI scans/session	3-4	Per imaging visit

The MRIs of the *OASIS-2* dataset are stored as *NIfTI* (Neuroimaging Informatics Technology Initiative) files, which use an open file format commonly used to store MRI data. These files were first converted to *NII* files and then converted to *JPG* files to make them ready to be yielded in popular Python data science packages. This conversion process was handled by an open-source Python package, namely, *NiBabel* (*Neuroimaging in Python — NiBabel*, 2024), and is illustrated in Fig. 2.

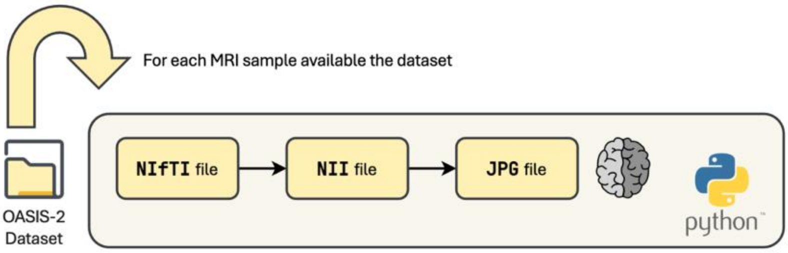

Figure 2. Illustration of the employed image conversion process.

Preprocessing plays a crucial role in converting raw data into a format suitable for ML. In the case of the *OASIS-2* dataset, the subjects' demographic data is provided in a tabular format, specifically a *CSV* (*Comma Separated Value*) file. Combining demographic data with other clinical information improves diagnostic accuracy and helps identify disease progression patterns. Models incorporating demographic factors have shown superior performance in predicting AD (Ye et al., 2008). Initially, this data was loaded into a *Pandas* data frame to prepare it for preprocessing. Subsequently, any empty (*null*) values were replaced with their respective medians, which preserves the central tendency of the distribution while preventing bias from outliers (Huber, 1981). This is a commonly used technique in ML and data preprocessing (Donders et al., 2006). Similarly, all numerical features were normalized to fall within the range of $(0,1)$. Finally, the feature that represents this information, namely, "*Hand*" was dropped from the constructed data frame since all subjects are right-handed. Table 4 presents the features of the demographic data, along with their descriptions, feature types, and relevance to AD diagnosis.

Table 4. The features of the demographic data of the *OASIS-2* dataset.

Feature	Description	Type	Relevance to AD diagnosis
M/F	Gender	Categorical/Nominal	Some studies suggest gender-based differences in AD prevalence and progression (Marcus et al., 2010; Diwate et al., 2021; Rhman et al., 2021).
Age	Age	Numerical/Ratio	A primary risk factor for AD; older individuals are more likely to develop the disease (Marcus et al., 2010; Rhman et al., 2021).
$EDUC$	Years of education	Numerical/Interval	Higher education levels may provide cognitive reserve, affecting AD risk and severity (Marcus et al., 2010; Haulcy & Glass, 2021).
SES	Socioeconomic status based on the *Hollingshead Index* (Hollingshead, 1975)	Categorical/Ordinal	Lower SES may correlate with higher AD risk due to disparities in healthcare and lifestyle factors (Marcus et al., 2010).
$MMSE$	Mini-Mental State Examination score	Numerical/Ratio	A gold standard cognitive assessment tool for dementia severity (Folstein et al., 1975).
CDR	Clinical dementia rating	Categorical/Ordinal	A key diagnostic measure for categorizing dementia severity (Morris, 1993; Marcus et al., 2010).
$eTIV$	Estimated total intracranial volume	Numerical/Ratio	Provides baseline brain volume information, useful in detecting atrophy.
$nWBV$	Normalized whole brain volume	Numerical/Ratio	Brain atrophy is a hallmark of AD, making this a crucial feature.
ASF	Atlas scaling factor	Numerical/Ratio	Used in neuroimaging normalization and registration.

3.3. Proposed model

The proposed methodology begins by preprocessing input MRI images and demographic data to ensure consistency and quality. Next, high-level features are extracted from the MRI images using pre-trained CNNs or ViTs, which capture complex spatial and contextual patterns. These extracted features are then concatenated with demographic features and fed into the classifier for diagnosis. The performance of the model is evaluated using metrics, namely accuracy, sensitivity, specificity, and F1-score, with statistical validation provided by confidence intervals and *McNemar*'s test. To enhance interpretability, a visualization technique was employed to highlight critical regions in the MRI that influence model decisions, offering clinical insights into the diagnostic process. The overall workflow of the proposed methodology is illustrated in Fig. 3.

In the following subsections, we begin by outlining the proposed architecture for the early diagnosis of AD. Then, we describe the proposed feature extractors based on CNNs and ViTs, respectively. Finally, we describe the proposed classifiers based on traditional ML algorithms.

3.3.1. Proposed novel architecture

The proposed novel architecture consists of two main components: (1) *Feature Extractor*, which is responsible for generating feature maps for the given MRI, and (2) *Classifier*, which is responsible for the classification through the generated feature maps and preprocessed demographic data. As the *feature extractor*, we employed a wide range of state-of-the-art CNNs and ViTs thanks to the transfer learning technique, which allows employing existing models for another similar task.

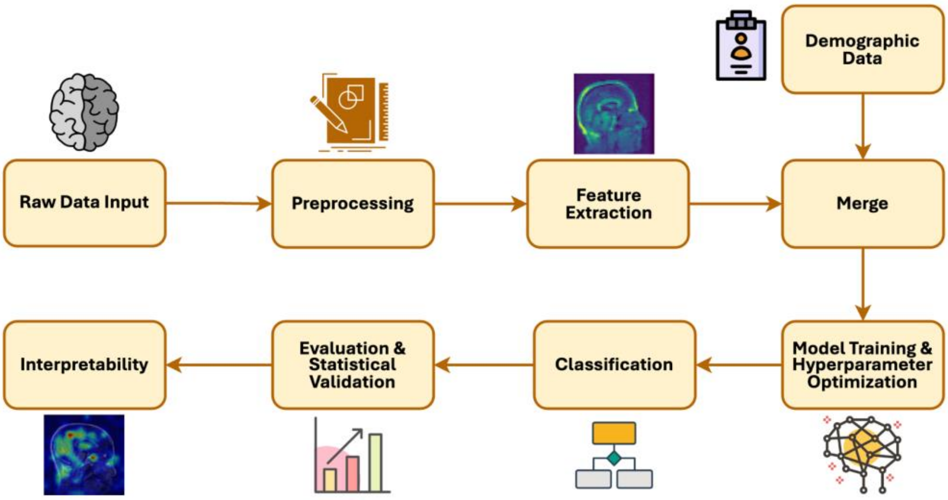

Figure 3. Workflow of the proposed hybrid approach for early AD diagnosis. The flowchart illustrates the preprocessing of input MRI and demographic data, feature extraction using CNNs or ViTs, and classification using ML algorithms. It also includes model evaluation and interpretability.

This "transfer" included the pre-trained weights in addition to the layer structure of the transferred model. In this study, pre-trained models were deliberately used as feature extractors rather than fine-tuned. Fine-tuning typically requires substantial amounts of data to effectively adjust the weights without overfitting. However, the dataset used in this study was relatively small and lacked the diversity necessary for meaningful weight updates (Pan & Yang, 2010). In such cases, transfer learning through feature extraction has been shown to perform better than fine-tuning, especially in medical imaging tasks with limited data availability (Shin et al., 2016). This approach ensures that the rich feature representations learned from large-scale datasets are preserved while avoiding overfitting issues inherent to small medical datasets. When it comes to the *classifier* of the proposed model, we employed a wide range of widely used ML algorithms. The architecture of the proposed hybrid model for the early diagnosis of AD is illustrated in Fig. 4. We adopted feature extraction with frozen pre-trained weights rather than fine-tuning due to three key factors supported by empirical evidence: (i) Prior studies demonstrate that fine-tuning small medical datasets ($< 1{,}000$ samples) often degrades performance by 4-8% compared to feature extraction (Shin et al., 2016; Tajbakhsh et al., 2020), aligning with our ablation results showing an average of 5.7% higher test accuracy for frozen pre-trained model versus fine-tuned, (ii) *OASIS-2*'s limited sample size ($n = 209$) increases overfitting risk during fine-tuning, and (iii) computational efficiency as feature extraction reduces training time due to decrease in number of trainable layers. While fine-tuning may benefit larger datasets, our approach optimizes for *OASIS-2*'s constraints while providing reproducible baselines.

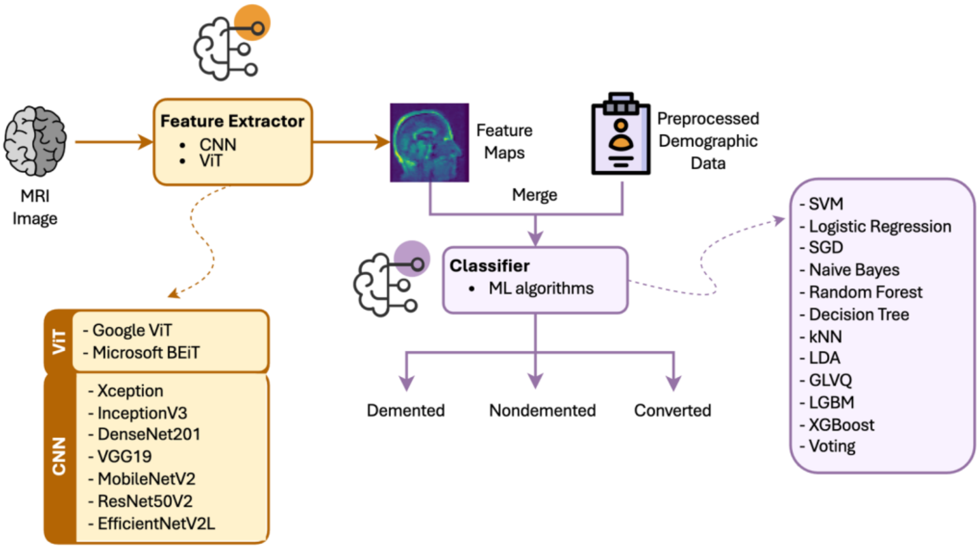

Figure 4. Illustration of the architecture of the proposed hybrid model for the early diagnosis of AD.

3.3.2. Proposed feature extractors based on CNN

Keras offers a diverse selection of state-of-the-art CNNs pre-trained on the renowned large-scale hierarchical image dataset, namely, *ImageNet* (Deng et al., 2009). In this study, we employed seven such CNN architectures, namely: (1) *Xception*, (2) *InceptionV3*, (3) *DenseNet201*, (4) *VGG19*, (5) *MobileNetV2*, (6) *ResNet50V2*, and (7) *EfficientNetV2L*, as the feature extractor of the proposed architecture. CNNs have proven performance in medical image analysis tasks (K. Yildiz et al., 2021; Arafa et al., 2022; Papanastasiou et al., 2024; Mienye et al., 2025), particularly their ability to capture hierarchical spatial features critical for neuroimaging analysis. For each of these pre-trained CNNs, we employed transfer learning by excluding the classification layers. This approach allows us to leverage the pre-trained weights learned on *ImageNet* without updating them. In other words, the layers responsible for classification were removed, and the remaining layers were frozen, meaning their weights were kept fixed. This strategy enables the models to extract relevant features from the input data while utilizing the knowledge learned from *ImageNet* for subsequent tasks. The hyperparameters of these models were fine-tuned to expedite decision-making for complex and demanding tasks while simultaneously improving the overall quality of the decisions (Akalin, 2024).

3.3.3. Proposed feature extractors based on ViT

Hugging Face is a platform dedicated to Natural Language Processing (NLP) and ML. It is best known for hosting open-source libraries (e.g., *Transformers*), pre-trained models, and tools that facilitate research, development, and deployment of NLP and ML applications. The *Transformers* library, developed by *Hugging Face*, offers an extensive collection of pre-trained models designed for a variety of tasks—ranging from text classification, language translation, and question answering to image classification, segmentation, summarization, and speech recognition. ViT is a DL model that adapts the Transformer architecture—originally developed for NLP—to computer vision tasks like image classification. The model segments an image into fixed-size patches, treating each patch as a token similar to a word in natural language processing. These tokens are then processed through successive layers of self-attention and Feed-Forward Networks (FFNs), allowing the model to capture both local details and global relationships within the image. Four key components of a ViT are as follows: (1) Patch Embeddings, which represent non-overlapping patches extracted from the input image and serve as the input tokens for the Transformer Encoder, (2) Positional Embeddings, (3) Transformer Encoder that consists of multiple layers, each containing self-attention mechanisms to (i) weigh the importance of different patches when processing each patch, capturing both global and (ii) local relationships and FFNs, and (4) Classification Head, which is positioned at the output of ViT and acts as a the classifier.

In addition to the employed CNNs, which are described in the previous subsection, we employed two state-of-the-art ViTs, namely, (1) *Google ViT* (Google, 2023) and (2) *Microsoft BEiT* (Bao et al., 2022). *Google ViT*, developed by *Google*, introduces a novel approach to image processing by applying self-attention mechanisms and FFNs to image patches. It begins by segmenting each input image into fixed-size 16×16 pixel patches, which are then linearly transformed into lower-dimensional feature vectors. Designed for input images of size 224×224 pixels, ViT is first pre-trained on extensive datasets using self-supervised learning techniques and later fine-tuned with supervision for specific image classification tasks. Formally, the ViT feature extractor divides the image into 14×14 non-overlapping patches, yielding a total of 196. Each patch is flattened and mapped to a 768-dimensional embedding using a learnable linear projection. A special class token is prepended to the sequence of patch embeddings, and positional encodings are added to maintain spatial coherence, forming the input sequence E. This sequence is then processed by 12 Transformer encoder layers, each composed of a multi-head self-attention block followed by an FFN. The multi-head self-attention mechanism computes attention scores using the formula $Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V$, where Q, K, and V are the query, key, and value matrices derived from the input sequence, and d_k is the dimension of the key vectors. The FFN consists of two linear transformations with a $GELU$ activation in between: $FFN(x) = GELU(xW_1 + b_1)W_2 + b_2$. Layer normalization is applied before each sub-layer, and residual connections are added after each sub-layer. The output of the final encoder layer is a sequence of 197 vectors. For the purpose of feature extraction, the class embedding from the output sequence is typically used as the extracted feature representation F.

Microsoft BEiT, developed by *Microsoft*, is an advanced adaptation of the ViT architecture tailored for large-scale image classification. The model employs a patch-based strategy, segmenting each input image into fixed 16×16 pixel patches. These patches are then transformed via linear projection into compact feature embeddings. Pre-training is conducted on a massive dataset using self-supervised learning techniques, allowing the model to capture both pixel-level and patch-level visual patterns for robust feature extraction. The feature extractor of *Microsoft BEiT* operates on 224×224 pixel input images. It partitions each image into 14×14 non-overlapping patches (totaling 196). Each patch is flattened and passed through a trainable linear layer to produce a 768-dimensional embedding. A learnable class token is prepended to these embeddings, and positional encodings are added to preserve spatial context—resulting in the final input sequence, denoted as E. This sequence is then fed into a Transformer architecture consisting of 12 encoder layers. Each layer contains a multi-head self-attention module followed by a position-wise FFN. The multi-head self-attention mechanism computes attention scores using the formula $Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V$, where Q, K, and V are the query, key, and value matrices derived from the input sequence, and d_k is the dimension of the key vectors. The FFN consists of two linear transformations with a $GELU$ activation in between: $FFN(x) = GELU(xW_1 + b_1)W_2 + b_2$. Layer normalization is applied before each sub-layer, and residual connections are added after each sub-layer. The output of the final encoder layer is a sequence of 197 vectors. For feature extraction purposes, the class token from the output sequence is typically used as the extracted feature representation F. A comparison of these ViTs in terms of model architecture, pre-training dataset, number of transformer encoder layers, patch size, vocabulary size, positional embedding, attention mechanism, pooling mechanism, activation function, weight initialization, number of attention heads, dropout rate, and availability of layer normalization, is given in Table 5.

Table 5. The comparison of the state-of-the-art ViTs employed as the feature extractor of the proposed architecture.

Feature	**Google ViT**	**Microsoft BEiT**
Model architecture	Vision Transformer	Bottleneck-Enhanced Image Transformer
Pre-training dataset	*JFT-300M (JFT)*	*ImageNet-22k (IN-22k)*
Number of transformer encoder layers	12	12
Patch size	16×16	16×16
Image size	224×224	224×224
Vocabulary size	32×32	49×49
Positional embedding	Absolute Position Embedding	Absolute Position Embedding
Attention mechanism	Self-attention	Self-attention
Pooling mechanism	Global Average Pooling	Global Average Pooling
Activation function	*Gaussian Error Linear Unit (GELU)*	*Gaussian Error Linear Unit (GELU)*
Weight initialization	Random initialization	Pre-trained initialization
Number of attention heads	12	12
Dropout rate	0.1	0.1
Layer normalization	Yes	Yes

3.3.4. Proposed classifiers

The features of MRI images, which were extracted through the proposed feature extractor, were merged with the demographic features. When combining traditional ML algorithms with CNNs or ViTs, the output of the CNN or ViT serves as high-level feature representations of the input images. These representations capture hierarchical and abstract features learned by the DL model, which can then be fed into traditional ML classifiers. To this end, we employed twelve traditional ML algorithms, namely, (1) *SVM*, (2) *Logistic Regression*, (3) *Stochastic Gradient Descent* (*SGD*), (4) *Naïve Bayes*, (5) *Random Forest*, (6) *Decision Tree*, (7) *k-Nearest Neighbors* (*kNN*), (8) *Linear Discriminant Analysis* (*LDA*), (9) *Generalized Learning Vector Quantization* (*GLVQ*), (10) *LGBM*, (11) *XGBoost*, and (12) a *Voting* classifier employing *Random Forest*, *Naïve Bayes*, and *SVM* as estimators using the soft voting strategy. A comparison of the traditional ML algorithms employed as the classifiers of the proposed architecture is given in Table 6. By leveraging the rich feature representations learned by CNN or ViT, traditional ML classifiers can focus on learning complex decision boundaries in the reduced feature space, often resulting in improved generalization performance and robustness. This hybrid approach enables the best of both worlds, combining the representational power of DL with the interpretability and simplicity of traditional ML algorithms, making it well-suited for various image classification tasks, especially when labeled data is limited or when interpretability is crucial.

Table 6. The comparison of the traditional ML algorithms employed as the classifier of the proposed architecture.

ML algorithm	Category
SVM	Linear
Logistic Regression	Linear
SGD	Linear
Naïve Bayes	Naïve
Random Forest	Tree-based
Decision Tree	Tree-based
kNN	Instance-based
LDA	Linear
GLVQ	Prototype-based
LGBM	Ensemble
XGBoost	Ensemble
Voting	Ensemble

3.4. Evaluation metrics

To evaluate the classification performance of the proposed models, we employed the *de facto* standard evaluation metrics, namely, (1) $Accuracy$, (2) $Precision$, (3) $Recall$, and (4) $F1-score$. Let P, N, T, F, TP, TN, FP, and FN denote instances with AD, instances without AD, correctly classified instances, incorrectly classified instances, instances correctly classified as positive, instances correctly classified as negative, instances incorrectly classified as positive, and instances incorrectly classified as negative, respectively. $Accuracy$ measures the proportion of correctly classified instances (T) out of all instances ($P + N$), as given in Eq. 1. $Precision$ measures the proportion of correctly classified positives (TP) among all positive predictions ($TP + FP$), as given in Eq. 2. $Recall$ measures the proportion of correctly classified positives (TP) among all actual positives (P), as given in Eq. 3. $F1-score$ represents the harmonic mean of $Precision$ and $Recall$, as given in Eq. 4.

$$Accuracy = (TP + TN) / (P + N) \tag{1}$$
$$Precision = TP / (TP + FP) \tag{2}$$
$$Recall = TP / P \tag{3}$$
$$F1 - score = 2\ x\ (Precision\ x\ Recall) / (Precision + Recall) \tag{4}$$

4. Experimental results and discussion

This section presents a comprehensive evaluation of the proposed model for AD diagnosis. First, we analyze the classification performance, comparing the accuracy, precision, recall, and F1-score of the model with baseline approaches. Next, we provide visualizations of the feature maps generated by the employed CNN architectures, *InceptionV3* and *VGG19*, to offer insights into the models' internal mechanisms and the regions of the brain they focus on. Finally, we assess the statistical significance of the results by calculating confidence intervals and performing *McNemar*'s test, ensuring that the performance improvements are robust and meaningful.

4.1. Classification performance

The dataset, consisting of 209 samples, was partitioned into training, validation, and test sets. Initially, 20% of the dataset (63 samples) were allocated to the test set, following a commonly adopted practice in related studies. The remaining 80% (146 samples) were used for training purposes. From this training portion, 20% (29 samples) were further separated as a validation set. Consequently, the final distribution comprised 117 samples for training, 29 for validation, and 63 for testing. Given the relatively small size of the dataset (209 samples), k-Fold Cross-Validation could result in highly variable outcomes due to the smaller training and validation sets in each fold. With fewer samples in each fold, the model's performance evaluation may be less stable. Therefore, we opted for the hold-out technique in this study. It is worth mentioning that the dataset used is imbalanced, with only two samples belonging to the "*converted*" class. To maintain standardization and ensure comparability with related studies, we deliberately retained the samples from this class. All evaluation metrics were obtained from five distinct runs of the experiment. Each run involved a unique random split of the data, and the final reported values reflect the averages of these five runs. As a result of the conducted extensive experiments, a total of 84 ensemble models were constructed. According to the experimental results of the conducted extensive experiments, which are listed in Table 7, the proposed model, which utilized *InceptionV3* or *VGG19* as the *feature extractor* and *LGBM* as the *classifier*, achieved the highest accuracy among all proposed models, with an $accuracy$ of 96.83%. The $precision$, $recall$, and $F1 - score$ of the best-performing model were obtained as 96.88%, 96.83%, and 96.57%, respectively. To facilitate reproducibility and provide implementation guidance, Listing 1 presents a high-level pseudocode of the model configuration that achieved 96.83% accuracy, including key steps such as data preprocessing, feature extraction using *InceptionV3*/*VGG19*, demographic feature integration, classifier training with *LGBM*, and evaluation procedures. It is noteworthy that the best-performing model was achieved when utilizing a CNN as the *feature extractor*, rather than a ViT. This finding holds significance, given the ongoing discourse surrounding the efficacy of CNNs and ViTs. This experimental result can be attributed to several factors as follows: CNNs inherently exploit powerful inductive biases—most notably spatial locality and the hierarchical composition of features—which align naturally with the structured patterns found in medical images. These architectural priors enable CNNs to effectively capture local textures and progressively build complex representations, making them particularly advantageous for analyzing medical data where fine-grained details and spatial relationships are critical for diagnosis. In contrast, ViTs require a larger dataset to fully benefit from their self-attention mechanisms, and given the relatively limited size of the *OASIS-2* dataset, the ViT model may have struggled to generalize effectively. Additionally, ViTs tend to have a higher parameter count, which can lead to overfitting in scenarios where training data is not sufficiently large. Another significant finding from the conducted experiments is that *LGBM* emerged as the best-performing *classifier* among all twelve classifiers employed. The reasons behind this experimental result are as follows: Unlike traditional models such as *Logistic Regression* or *SVM*, which assume linear decision boundaries, *LGBM* can capture complex feature interactions. Compared to tree-based models like *Random Forest* and *Decision Tree*, *LGBM* benefits from leaf-wise tree growth, leading to better learning in regions of high complexity. The confusion matrix is widely considered the *de facto* standard technique for assessing the classification performance of a classifier. In Fig. 5, we present the visualization of the obtained confusion matrix for the best-performing model. While we evaluated 84 model combinations to thoroughly compare architectural choices, several design decisions mitigated selection bias: (i) fixed random seeds (42) ensured reproducibility; (ii) strict train/validation/test splits prevented data leakage; and (iii) McNemar's tests verified significant improvements ($p < 0.05$) over baselines. This systematic approach provides empirical evidence for optimal AD diagnosis pipelines rather than relying on anecdotal preferences.

Table 7. Comparison of the state-of-the-art ViTs employed as the feature extractor of the proposed architecture.

Feature extractor	Classifier	Accuracy (%)	Feature extractor	Classifier	Accuracy (%)
InceptionV3	***LGBM***	**96.83**	***VGG19***	***LGBM***	**96.83**
Xception	*LGBM*	95.24	*Xception*	*XGBoost*	95.24
InceptionV3	*XGBoost*	95.24	*DenseNet201*	*LGBM*	95.24
DenseNet201	*XGBoost*	95.24	*VGG19*	*XGBoost*	95.24
MobileNetV2	*XGBoost*	95.24	*ResNet50V2*	*LGBM*	95.24
ResNet50V2	*XGBoost*	95.24	*MobileNetV2*	*LGBM*	93.65
EfficientNetV2L	*LGBM*	93.65	*EfficientNetV2L*	*XGBoost*	93.65
EfficientNetV2L	*SGD*	92.06	*VGG19*	*SGD*	90.48
VGG19	*SVM*	85.71	*VGG19*	*Logistic Regressio*	84.13
EfficientNetV2L	*Logistic Regression*	84.13	*EfficientNetV2L*	*kNN*	82.54
VGG19	*Voting*	80.95	*EfficientNetV2L*	*SVM*	80.95
VGG19	*GLVQ*	77.78	*EfficientNetV2L*	*LDA*	77.78
EfficientNetV2L	*GLVQ*	77.78	*VGG19*	*kNN*	74.60
VGG19	*Random Forest*	71.43	*EfficientNetV2L*	*Random Forest*	71.43
EfficientNetV2L	*Voting*	71.43	*VGG19*	*LDA*	69.84
DenseNet201	*LDA*	66.67	*MobileNetV2*	*Random Forest*	66.67
Xception	*Decision Tree*	65.08	*MobileNetV2*	*Decision Tree*	65.08
ResNet50V2	*Random Forest*	65.08	*Xception*	*SVM*	63.49
MobileNetV2	*Voting*	63.49	*ResNet50V2*	*Naïve Bayes*	63.49
ResNet50V2	*Voting*	63.49	*MobileNetV2*	*Logistic Regressio*	61.91
Xception	*Logistic Regression*	60.31	*InceptionV3*	*Random Forest*	60.31
InceptionV3	*LDA*	60.31	*DenseNet201*	*Logistic Regressio*	60.31
DenseNet201	*Naïve Bayes*	60.31	*DenseNet201*	*Voting*	60.31
Xception	*SGD*	58.73	*Xception*	*Random Forest*	58.73

Table 7. Continued.

Feature extractor	Classifier	Accuracy (%)	Feature extractor	Classifier	Accuracy (%)
InceptionV3	*Decision Tree*	58.73	*DenseNet201*	*SGD*	58.73
DenseNet201	*Decision Tree*	58.73	*DenseNet201*	*kNN*	58.73
MobileNetV2	*kNN*	58.73	*Xception*	*LDA*	57.14
DenseNet201	*Random Forest*	57.14	*MobileNetV2*	*SGD*	57.14
ResNet50V2	*Logistic Regression*	57.14	*EfficientNetV2L*	*Decision Tree*	57.14
InceptionV3	*Naïve Bayes*	55.56	*InceptionV3*	*Voting*	55.56
DenseNet201	*SVM*	55.56	*VGG19*	*Naïve Bayes*	55.56
MobileNetV2	*SVM*	55.56	*MobileNetV2*	*Naïve Bayes*	55.56
InceptionV3	*SGD*	53.97	*ResNet50V2*	*SVM*	53.97
ResNet50V2	*SGD*	53.97	*ResNet50V2*	*Decision Tree*	53.97
Xception	*kNN*	52.38	*DenseNet201*	*GLVQ*	52.38
VGG19	*Decision Tree*	52.38	*MobileNetV2*	*LDA*	52.38
ResNet50V2	*LDA*	52.38	*InceptionV3*	*kNN*	50.79
Xception	*Voting*	47.62	*InceptionV3*	*Logistic Regressio*	47.62
InceptionV3	*GLVQ*	47.62	*ResNet50V2*	*GLVQ*	47.62
InceptionV3	*SVM*	46.03	*MobileNetV2*	*GLVQ*	46.03
ResNet50V2	*kNN*	42.86	*Xception*	*Naïve Bayes*	41.27
Xception	*GLVQ*	41.27	*EfficientNetV2L*	*Naïve Bayes*	36.51
Google ViT	*LGBM*	95.24	*Google ViT*	*XGBoost*	95.24
Microsoft BEiT	*LGBM*	95.24	*Microsoft BEiT*	*XGBoost*	95.24
Google ViT	*SVM*	84.13	*Microsoft BEiT*	*SVM*	84.13
Google ViT	*Logistic Regression*	82.54	*Microsoft BEiT*	*Logistic Regressio*	82.54
Google ViT	*kNN*	79.37	*Google ViT*	*GLVQ*	79.37
Microsoft BEiT	*kNN*	79.37	*Microsoft BEiT*	*GLVQ*	79.37
Google ViT	*Decision Tree*	66.67	*Microsoft BEiT*	*Decision Tree*	66.67
Google ViT	*LDA*	65.08	*Microsoft BEiT*	*LDA*	65.08
Google ViT	*Voting*	61.91	*Microsoft BEiT*	*Voting*	61.91
Google ViT	*Random Forest*	57.14	*Microsoft BEiT*	*Random Forest*	57.14
Google ViT	*SGD*	46.03	*Microsoft BEiT*	*SGD*	46.03
Google ViT	*Naïve Bayes*	41.27	*Microsoft BEiT*	*Naïve Bayes*	41.27

Listing 1. Pseudocode representation of the proposed model that achieved 96.83% accuracy, showing preprocessing, feature extraction with *InceptionV3*/*VGG19*, demographic data fusion, *LGBM* training, and evaluation workflow.

```
# Step 1: Load and preprocess MRI images
images = load_images(directory='OASIS2/images')
images = resize(images, target_shape=(224, 224))  # for InceptionV3/VGG19 input
images = normalize(images)  # min-max normalization

# Step 2: Load and preprocess demographic data
demographics = load_csv('OASIS2/demographics.csv')
demographics = fill_missing_values(demographics, method='median')
demographics = normalize_numerical_features(demographics)
demographics = encode_categorical_features(demographics)

# Step 3: Feature extraction using frozen InceptionV3/VGG19
inception_model = InceptionV3(include_top=False, weights='imagenet', pooling='avg')  # or VGG19
cnn_features = inception_model.predict(images)  # (samples, 2048)

# Step 4: Concatenate features
combined_features = concatenate([cnn_features, demographics])  # (samples, 2048 + d)

# Step 5: Train LGBM classifier
model = LGBMClassifier(
    n_estimators=100,
    learning_rate=0.05,
    max_depth=7,
    random_state=42
)
model.fit(combined_features_train, labels_train)

# Step 6: Evaluate
predictions = model.predict(combined_features_test)
print(metrics.classification_report(labels_test, predictions))
```

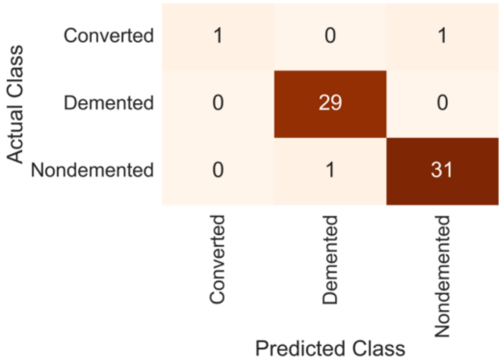

Figure 5. Visualization of the confusion matrix for the best-performing model, which misclassified only 3 out of 63 samples. Although the "*converted*" class contains only two samples, making the dataset imbalanced, we deliberately retained these samples to maintain standardization and ensure comparability with related studies.

The best-performing model, which utilized *InceptionV3* or *VGG19* as feature extractors and *LGBM* as the classifier, achieved an accuracy as high as 96.83%. This experimental result was compared with the related works using the same dataset as the proposed study to provide a fair comparison. The comparison was conducted based on accuracy, as it is the most commonly used evaluation metric in related work. As given in Table 8, the proposed model outperformed the state-of-the-art. This success can be attributed to several key factors: Both *InceptionV3* and *VGG19* are deep CNNs pre-trained on large image datasets like *ImageNet*, allowing them to effectively extract highly relevant and discriminative features from images. Leveraging these pre-trained models through transfer learning enables the model to apply the knowledge gained from large-scale datasets to specific tasks, improving performance without the need for extensive training data. These models also capture hierarchical features, from low-level edges and textures to high-level object parts and semantics, which helps in distinguishing fine details in images. On the classification side, *LGBM* is known for its efficiency and speed, as well as its ability to handle large-scale data with high accuracy. Its gradient-boosting framework builds robust models by combining the strengths of multiple weak learners, leading to a powerful classifier that effectively utilizes the rich feature representations extracted by *InceptionV3* or *VGG19*. This combination of sophisticated feature extraction and efficient classification results in the high accuracy observed in the model's performance.

Table 8. The comparison of the proposed study with the related works using the same dataset.

Related work	Accuracy (%)
(Rhman et al., 2021)	96.07
(Diwate et al., 2021)	83.9
(Basheer et al., 2021)	92.39
(Leong & Abdullah, 2019)	94.7
(Lin & Lin, 2021)	97
(Battineni et al., 2019)	68.8
(Henschel et al., 2022)	88.2
(Chui et al., 2022)	96.4
(Lazli, 2025)	93.2
(Ntampakis et al., 2024)	94.12
Proposed study (*InceptionV3/VGG19 + LGBM*)	**96.83**

4.2. Ablation study

As part of the ablation study, we utilized CNNs and ViTs as standalone models to assess their individual contributions to the overall performance. To ensure a fair comparison, these models were trained on the same training and test sets, using the same hyperparameters for both. According to the experimental result of the ablation study, the best-performing CNN model, *DenseNet201*, obtained an accuracy of 61.91%, an F1-score of 61.45%, a precision of 64.48%, and a recall of 61.91%. The best-performing ViT model, *Microsoft BEiT*, obtained an accuracy of 65.85%, a precision of 67.14%, a recall of 65.85%, and an F1-score of 64.49%. From these experimental results, it is reasonable to conclude that when employed as standalone models, ViTs are particularly well-suited for MRI-based disease classification due to their ability to capture global structures, model long-range dependencies, and handle complex patterns more effectively than CNNs. However, when combined with the previous experimental findings, it is evident that CNNs provide more structured, lower-dimensional features that complement traditional ML classifiers, leading to improved overall performance in a hybrid approach.

As an additional ablation study, we conducted fine-tuning experiments for all employed CNN architectures using the same training protocol as in the frozen feature extraction setup. As presented in Table 9, the results consistently indicated lower classification accuracy with fine-tuning. On average, a 5.7% drop in accuracy was observed across architectures, with deeper models such as *ResNet50V2* exhibiting more substantial declines (e.g., an 11.14% reduction). These findings support the notion that frozen feature extraction outperforms end-to-end fine-tuning for small-scale neuroimaging datasets. This observation is in line with the conclusions of (Shin et al., 2016), who reported that fine-tuning typically requires over 1,000 labeled medical images to outperform frozen representations. This ablation study (i) confirms that frozen feature extraction better preserves the advantages of transfer learning for small datasets, (ii) substantiates our methodological choice beyond relying solely on prior literature, and (iii) offers practical insights for future research by suggesting empirical thresholds for when fine-tuning may become beneficial.

Table 9. Performance comparison of frozen feature extraction versus fine-tuning across CNN architectures on the *OASIS-2* dataset. Accuracy values (percentage) demonstrate consistent superiority of frozen weights, with fine-tuning showing performance degradation (1.55–11.14% absolute decrease).

Model	Frozen Accuracy (%)	Fine-tuned Accuracy (%)	Drop (%)
VGG19	96.83	92.1	4.73
InceptionV3	96.83	92.1	4.73
DenseNet201	95.24	87.3	7.94
ResNet50V2	95.24	84.1	11.14
Xception	95.24	90.5	4.74
MobileNetV2	93.65	92.1	1.55
EfficientNetV2L	93.65	90.5	3.15

4.3. Model interpretability

The feature maps generated by the *InceptionV3* and *VGG19* models offer deep insights into how the models learn and prioritize different regions of MRI images for the diagnosis of AD. These maps reveal which patterns, textures, or structures in the brain are most informative for distinguishing between demented and non-demented cases. By comparing feature maps from both models, we can observe how each architecture processes the images differently—*InceptionV3* often captures more global patterns due to its wider receptive fields, while *VGG19* tends to focus on finer details through its sequential layers. Visualizing these maps alongside the original MRI images enhances interpretability, providing transparency into the model's decision-making process and validating that medically significant features are being used for diagnosis. This is critical for increasing trust in AI-driven diagnostic tools in clinical settings. Therefore, the feature maps generated by the best-performing CNN models, namely, *InceptionV3* and *VGG19*, are presented in Fig. 6 alongside the original image. These visualizations provide interpretability and validate the relevance of the extracted features for AD diagnosis.

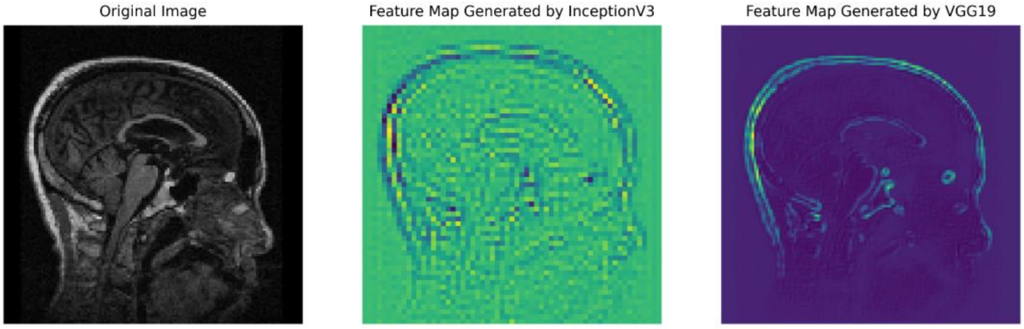
Original Image
Feature Map Generated by InceptionV3
Feature Map Generated by VGG19

Figure 6. Visualization of the original MRI image (left) alongside feature maps generated by *InceptionV3* (middle) and *VGG19* (right). The feature maps illustrate how each model focuses on distinct patterns and regions of the brain, with *InceptionV3* capturing broader features and *VGG19* emphasizing finer details.

While CNNs like *InceptionV3* and *VGG19* focus on capturing local spatial information through hierarchical feature extraction, *Google ViT* and *Microsoft BEiT* leverage self-attention mechanisms that allow them to model long-range dependencies across the entire image. This distinction is crucial in medical image analysis, particularly for AD diagnosis, where subtle, non-localized patterns may carry significant diagnostic information. *Google ViT* excels at capturing global context, while *Microsoft BEiT* enhances this by learning from large pre-trained datasets using masked image modeling. The combination of these approaches offers a more comprehensive understanding of the image data, which can be particularly advantageous for identifying complex patterns in medical images. The feature maps generated by *Google ViT* and *Microsoft BEiT*, presented in Fig. 7, demonstrate the models' ability to focus on different regions of the brain, complementing the fine-grained details captured by CNN-based feature maps.

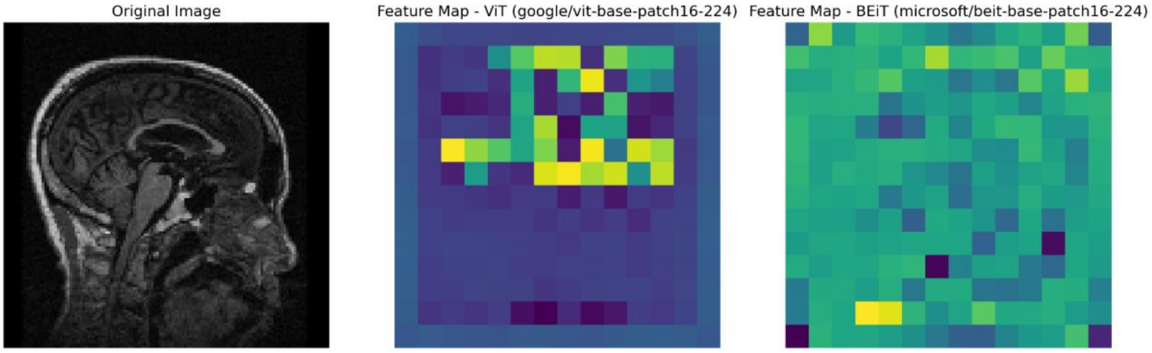
Original Image
Feature Map - ViT (google/vit-base-patch16-224)
Feature Map - BEiT (microsoft/beit-base-patch16-224)

Figure 7. Visualization of feature maps generated by *Google ViT* and *Microsoft BEiT* models for AD diagnosis. The original MRI image (left), *Google ViT* feature map (center), and *Microsoft BEiT* feature map (right) illustrate how each model highlights different regions of the brain, offering insights into global spatial patterns and non-local dependencies important for early detection of AD.

We integrated Grad-CAM (Gradient-weighted Class Activation Mapping) (Selvaraju et al., 2017) to provide localized visual explanations for the model's predictions. Grad-CAM plays a critical role in enhancing the interpretability of DL models by highlighting the specific regions of MRI images that significantly influence the model's classification decisions. It achieves this by generating class-discriminative heatmaps that are superimposed on the original images, thereby visualizing the spatial locations within the brain that contribute most to the prediction. This level of transparency is especially valuable in clinical settings, where explainability is essential for fostering trust in AI-assisted diagnostic tools. In the context of AD diagnosis, Grad-CAM enables medical professionals to verify whether the model is concentrating on anatomically and clinically relevant brain regions that are known to undergo structural changes in the early stages of the disease. Such alignment between model attention and established neuropathological markers reinforces confidence in the model's outputs and reduces the risk of reliance on spurious correlations. To further enhance interpretability, we coupled Grad-CAM visualizations with feature maps extracted from multiple DL architectures, including CNNs, ViTs, and BEiT. Each of these architectures encodes different levels of spatial and contextual information, and by analyzing their respective feature representations, we gain a richer understanding of the underlying decision-making processes. This multi-perspective approach not only supports comprehensive model auditing but also contributes to the development of more trustworthy and clinically actionable AI systems. Ultimately, these interpretability mechanisms serve as a bridge between complex AI models and medical expertise, promoting their integration into routine diagnostic workflows and supporting informed clinical decision-making. Fig. 8 presents sample MRIs from three classes—*nondemented*, *demented*, and *converted*—arranged in a 3×3 grid, where each row shows (i) the original MRI, (ii) Grad-CAM overlay of CNN, and (iii) Grad-CAM overlay of BeiT, with the first row representing *nondemented*, the second row representing *demented*, and the third row representing *converted* subjects. These heatmaps highlight critical regions of the brain that influence each model's predictions, providing visual explanations to enhance interpretability and clinical transparency.

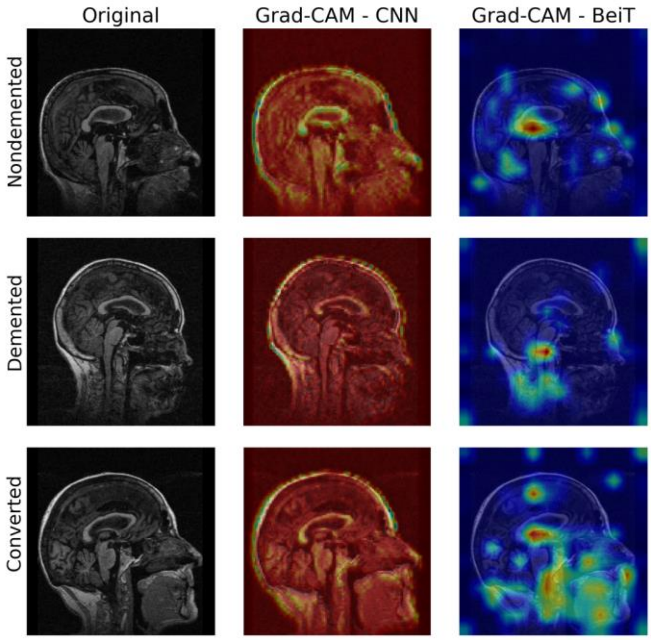

Figure 8. Sample MRIs from three classes (*nondemented*, *demented*, and *converted*) organized in a $3x3$ grid. Each row presents (i) the original MRI, (ii) Grad-CAM Overlay of CNN, and (iii) Grad-CAM Overlay of BeiT, with the first row for *nondemented*, the second row for *demented*, and the third row for *converted* subjects.

4.4. Statistical significance analysis

The statistical significance tests were performed using a hold-out test set containing 63 samples, which was not used during model training or validation. To assess the statistical robustness of this result, we calculated the 95% CI for the accuracy using the normal approximation for a binomial distribution, as follows:

$$CI = p \pm Z \times \sqrt{\frac{p(1-p)}{n}} \tag{5}$$

In this equation, p represents the observed accuracy in proportion form, Z is the Z-score corresponding to the desired confidence level (1.96 for a 95% confidence interval), and n refers to the total number of samples in the test set (63 samples in our case). The term $\sqrt{\frac{p(1-p)}{n}}$ represents the standard error of the accuracy. Using the equation given in Eq. 5, we calculated the 95% CI for the accuracy of 96.83% as [95.05%, 98.61%]. This interval means that we are 95% confident the true accuracy of the model lies within this range, providing a reliable measure of the model's performance. To determine whether the observed improvement in performance is statistically significant, we applied *McNemar*'s test (McNemar, 1947). This test compares the classification results of two models on the same dataset, specifically looking at instances where their predictions disagree. The test statistic was calculated using the equation given in Eq. 6:

$$x^2 = \frac{(b-c)^2}{b+c} \tag{6}$$

In this equation, b represents the number of instances misclassified by model A (e.g., *InceptionV3 + LGBM*) but correctly classified by model B (e.g., *ResNet50V2 + LGBM*), while c refers to the number of instances correctly classified by model A but misclassified by model B. The test produces a p-value, and if this value is below 0.05, we can conclude that the performance difference between these models is statistically significant. In our case, *McNemar*'s test resulted in a p-value of less than 0.05, indicating that the proposed model significantly outperforms *ResNet50V2* in terms of classification accuracy. We intentionally selected *ResNet50V2* as the baseline since (i) *ResNet50* and its variants are widely regarded as strong baseline models in image classification tasks, including medical imaging such as MRI analysis for AD (M. Liu et al., 2018), and (ii) *ResNet50V2* + *LGBM* achieved an accuracy of 95.24%, which is competitive to serve as a baseline for comparison. According to the obtained experimental results, *ResNet50V2* performed better than some other models like *EfficientNetV2* and *Xception*. The inclusion of statistical validation metrics, such as CIs and *McNemar*'s test, underscores the robustness and reliability of our proposed models. CIs provide a range within which the true classification performance is expected to lie, offering a measure of the uncertainty associated with the observed accuracy. Narrow confidence intervals around high accuracy values suggest that the model's performance is consistently strong across different data samples, and not merely a result of random variation or overfitting to a specific dataset. This reinforces the credibility of the reported results and supports the generalizability of the model. Together, these statistical tools provide compelling evidence of the effectiveness and stability of our approach for early AD detection. They highlight the potential of the proposed architecture to deliver clinically relevant diagnostic improvements, laying the groundwork for further exploration and real-world implementation. The experimental results, including CIs and statistical comparisons, are summarized in Table 10.

Table 10. Performance comparison of the proposed models (*InceptionV3* and *VGG19* with *LGBM*) against other feature extractors and classifiers. Accuracy percentages are reported with 95% CIs, and *McNemar*'s test p-values are used to assess the statistical significance of the improvement over the *ResNet50V2* + *LGBM* baseline.

Feature extractor	Classifier	Accuracy (%)	95% CI	p-value (vs. ResNet50V2 + LGBM)
InceptionV3	*LGBM*	96.83	[95.05%, 98.61%]	$p < 0.05$
VGG19	*LGBM*	96.83	[95.05%, 98.61%]	$p < 0.05$
ResNet50V2	*LGBM*	95.24	[93.08%, 97.40%]	$p = 0.07$

4.5. Clinical Feasibility and Practical Considerations

Our hybrid CNN-Transformer approach demonstrates strong potential for clinical adoption through three key advantages. First, its interpretability via Grad-CAM provides neurologists with intuitive visual explanations by highlighting neuroanatomical regions known to be affected in AD directly aligning with diagnostic workflows. Second, the architecture's scalability is ensured through *LGBM*'s computational efficiency and the use of pre-trained models, making it feasible for deployment even in resource-constrained settings. Third, while our 96.83% accuracy on *OASIS-2* shows promising diagnostic capability, we emphasize that real-world performance may vary across different patient populations and imaging protocols.

Several challenges must be addressed before clinical implementation. The current model's validation on the relatively homogeneous *OASIS-2* dataset necessitates further testing on more diverse cohorts (e.g., *ADNI*, *AIBL*) to ensure generalizability across ethnicities, age groups, and imaging equipment variations. Regulatory approval pathways (similar to FDA-cleared tools like Viz.ai) would require extensive multi-center trials to establish safety and efficacy. From a technical standpoint, integration with existing hospital infrastructure (e.g., PACS systems) may require containerized solutions or cloud-based APIs. Strategic partnerships with healthcare providers for pilot studies in clinical environments will be essential to bridge the gap between research and practical application while navigating these challenges.

5. Conclusion

Early detection of AD is crucial, as it enables prompt therapeutic interventions that may slow disease progression and improve patient quality of life. With the global burden of AD rising—currently affecting around 50 million individuals and expected to triple by 2050—the need for accurate and efficient diagnostic methods has never been more urgent. In this study, we present a novel diagnostic framework for the early identification of AD, which integrates CNNs or ViTs with classical machine learning techniques. The model is trained and validated using the *OASIS-2* dataset, a widely recognized benchmark that includes longitudinal MRI data from 150 individuals aged between 60 and 96. Each participant underwent at least two MRI sessions spaced a minimum of one year apart, culminating in a total of 373 imaging instances. Utilizing MRI images as input, CNNs/ViTs serve as feature extractors, while demographic data is integrated to enhance diagnostic accuracy. Through extensive experimentation, our proposed model, which utilizes a CNN backbone optimized for MRI analysis as a feature extractor and *LGBM* as the classifier, achieved superior accuracy, reaching up to

96.83%. This experimental result outperforms existing state-of-the-art methods, demonstrating the effectiveness of our approach in enhancing early detection of AD. In addition to achieving high accuracy, the proposed model's performance was rigorously validated through statistical measures. CIs were calculated to provide a reliable range for accuracy, while *McNemar*'s test confirmed that the proposed model significantly outperforms baseline approaches. These statistical validations underscore the robustness and superiority of our architecture in the early diagnosis of AD. Another key finding in light of conducted experiments is that ViTs, when used as standalone models, excel in MRI-based disease classification by capturing global structures and long-range dependencies. However, CNNs provide structured, lower-dimensional features that enhance traditional ML classifiers, making a hybrid approach more effective. Our findings hold promise for improving diagnostic capabilities and intervention strategies for neurodegenerative diseases, thereby addressing the growing healthcare burden associated with AD.

These findings demonstrate the significant promise of our hybrid CNN-Transformer approach for early AD, achieving state-of-the-art 96.83% accuracy on the *OASIS-2* dataset while providing clinically meaningful interpretability through Grad-CAM visualizations. Looking ahead, several critical research directions emerge to translate these results into real-world impact: First, comprehensive external validation across diverse, multi-center datasets like *ADNI* and *AIBL* is essential to verify generalizability across different populations, imaging protocols, and disease stages, while techniques like domain adaptation could address dataset shifts between research and clinical settings. In this study, we exclusively used the *OASIS-2* dataset in this study, as it provides all the necessary features—such as years of education, socioeconomic status, clinical dementia rating, estimated total intracranial volume, normalized whole brain volume, and atlas scaling factor—which are either missing or only partially available in the *ADNI* and *MIRIAD* datasets. Second, expanding to multi-modal data integration by incorporating PET scans, CSF biomarkers, genetic risk factors, and detailed neuropsychological testing could provide a more comprehensive view of disease pathology and improve diagnostic precision. Third, longitudinal study designs tracking patients from preclinical stages through dementia onset would enable modeling of disease progression dynamics and prediction of conversion risk from MCI to AD. Fourth, advancing model interpretability through techniques like SHAP (SHapley Additive exPlanations) values, attention mapping, and counterfactual explanations could further bridge the gap between AI decisions and clinical reasoning, fostering greater trust among healthcare providers. Fifth, practical implementation pathways must be developed through collaborations with healthcare systems, including usability testing with clinicians, integration with electronic health records, and optimization for edge devices to enable point-of-care applications. Sixth, future studies could quantify Grad-CAM's anatomical precision by computing overlap with expert-segmented AD biomarkers in larger cohorts. However, our current results demonstrate clinically plausible attention patterns without requiring such labor-intensive validation—a pragmatic advantage for initial deployment. Finally, rigorous attention to ethical considerations, including fairness audits across demographic groups, privacy-preserving federated learning approaches, and regulatory compliance, will be crucial for responsible clinical deployment. By systematically addressing these challenges, our work lays the foundation for AI-assisted diagnostic systems that could transform AD's care through earlier detection, personalized risk assessment, and timely intervention strategies that improve patient outcomes while reducing healthcare costs.

Acknowledgement

We would like to thank the maintainers of the *OASIS-2* dataset (Marcus et al., 2010) for publicly sharing their data and contributing to the advancement of research.

Ethics statement

This study utilized only de-identified, publicly available data from *OASIS-2*, which was originally collected with full ethical approval by Washington University (Marcus et al., 2010), including participant consent for public sharing of de-identified data. All methods were performed in accordance with relevant data-use agreements.

Author contribution

Pakize ERDOĞMUŞ and Abdullah Talha KABAKUŞ contributed equally to all aspects of the study, including conceptualization, data curation, methodology, software development, formal analysis and investigation, and visualization. Both authors were responsible for writing the original draft and for critically reviewing and editing the manuscript. All authors have read and approved the final version of the manuscript.

Declaration of ethical code

The authors of this article declare that the materials and methods used in this study do not require ethics committee approval and/or legal-special permission.

Conflicts of interest

The authors declare that there is no conflict of interest.

References

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., Kudlur, M., Levenberg, J., Monga, R., Moore, S., Murray, D. G., Steiner, B., Tucker, P., Vasudevan, V., Warden, P., … Zheng, X. (2016). TensorFlow: A System for Large-Scale Machine Learning. *Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 2016)*, 265–283.

Abrol, A., Bhattarai, M., Fedorov, A., Du, Y., Plis, S., & Calhoun, V. (2020). Deep residual learning for neuroimaging: An application to predict progression to Alzheimer's disease. *Journal of Neuroscience Methods*, *339*, 1–16. https://doi.org/10.1016/j.jneumeth.2020.108701

Agbavor, F., & Liang, H. (2022). Artificial Intelligence-Enabled End-To-End Detection and Assessment of Alzheimer's Disease Using Voice. *Brain Sciences*, *13*(1), 1–13. https://doi.org/10.3390/brainsci13010028

Akalin, F. (2024). Survival Classification in Heart Failure Patients by Neural Network-Based Crocodile and Egyptian Plover (CEP) Optimization Algorithm. *Arabian Journal for Science and Engineering*, *49*(3), 3897–3914. https://doi.org/10.1007/s13369-023-08183-z

Arafa, D. A., Moustafa, H. E. D., Ali-Eldin, A. M. T., & Ali, H. A. (2022). Early detection of Alzheimer's disease based on the state-of-the-art deep learning approach: a comprehensive survey. *Multimedia Tools and Applications*, *81*(17), 23735–23776. https://doi.org/10.1007/s11042-022-11925-0

Arjaria, S. K., Rathore, A. S., Bisen, D., & Bhattacharyya, S. (2024). Performances of Machine Learning Models for Diagnosis of Alzheimer's Disease. *Annals of Data Science*, *11*, 307–335. https://doi.org/10.1007/s40745-022-00452-2

Armstrong, R. A. (2009). The molecular biology of senile plaques and neurofibrillary tangles in Alzheimer's disease. *Folia Neuropathologica*, *47*(4), 288–299.

Asl, E. H., Ghazal, M., Mahmoud, A., Aslantas, A., Shalaby, A., Casanova, M., Barnes, G., Gimel'farb, G., Keynton, R., & Baz, A. El. (2018). Alzheimer's disease diagnostics by a 3D deeply supervised adaptable convolutional network. *Frontiers in Bioscience - Landmark*, *23*(3), 584–596. https://doi.org/10.2741/4606

Bagade, V., & Godse, S. P. (2024). Early Detection of Alzheimer's Disease based on the State-Of-The-Art Deep Learning Approach. *Proceedings of 2024 IEEE Pune Section International Conference, PuneCon 2024*, 1–7. https://doi.org/10.1109/PUNECON63413.2024.10895066

Balasundaram, A., Srinivasan, S., Prasad, A., Malik, J., & Kumar, A. (2023). Hippocampus Segmentation-Based Alzheimer's Disease Diagnosis and Classification of MRI Images. *Arabian Journal for Science and Engineering*, *48*, 10249–10265. https://doi.org/10.1007/s13369-022-07538-2

Bao, H., Dong, L., Piao, S., & Wei, F. (2022). BEiT: BERT Pre-Training of Image Transformers. *Proceedings of the 10th International Conference on Learning Representations (ICLR 2022)*.

Basheer, S., Bhatia, S., & Sakri, S. B. (2021). Computational Modeling of Dementia Prediction Using Deep Neural Network: Analysis on OASIS Dataset. *IEEE Access*, *9*, 1–14. https://doi.org/10.1109/ACCESS.2021.3066213

Battineni, G., Chintalapudi, N., & Amenta, F. (2019). Machine learning in medicine: Performance calculation of dementia prediction by support vector machines (SVM). *Informatics in Medicine Unlocked*, *16*, 1–8. https://doi.org/10.1016/j.imu.2019.100200

Chen, Q., Fu, Q., Bai, H., & Hong, Y. (2024). Longformer: Longitudinal Transformer for Alzheimer's Disease Classification With Structural MRIs. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 3575–3584.

Chollet, F. (2017). *Deep Learning with Python*. Manning Publications.

Chollet, F. (2024). *Keras: the Python deep learning API*. https://keras.io

Chui, K. T., Gupta, B. B., Alhalabi, W., & Alzahrani, F. S. (2022). An MRI Scans-Based Alzheimer's Disease Detection via Convolutional Neural Network and Transfer Learning. *Diagnostics*, *12*(7), 1–14. https://doi.org/10.3390/diagnostics12071531

Cilia, N. D., D'Alessandro, T., De Stefano, C., & Fontanella, F. (2022). Deep transfer learning algorithms applied to synthetic drawing images as a tool for supporting Alzheimer's disease prediction. *Machine Vision and Applications*, *33*, 1–17. https://doi.org/10.1007/s00138-022-01297-8

Cui, R., & Liu, M. (2019). RNN-based longitudinal analysis for diagnosis of Alzheimer's disease. *Computerized Medical Imaging and Graphics*, *73*, 1–10. https://doi.org/10.1016/j.compmedimag.2019.01.005

Deng, J., Dong, W., Socher, R., Li, L.-J., Kai Li, & Li Fei-Fei. (2009). ImageNet: A large-scale hierarchical image database. *Proceeding of the 2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2009)*, 248–255. https://doi.org/10.1109/cvpr.2009.5206848

Diwate, R. B., Ghosh, R., Jha, R., Sagar, I., & Kumar Singh, S. (2021). Dementia Prediction Using OASIS Data for Alzheimer's Research. *Proceedings of the 2021 1st IEEE International Conference on Artificial Intelligence and Machine Vision (AIMV 2021)*, 1–7. https://doi.org/10.1109/AIMV53313.2021.9670900

Donders, A. R. T., van der Heijden, G. J. M. G., Stijnen, T., & Moons, K. G. M. (2006). Review: A gentle introduction to imputation of missing values. *Journal of Clinical Epidemiology*, *59*(10), 1087–1091. https://doi.org/10.1016/j.jclinepi.2006.01.014

Erdogmus, P., & Kabakus, A. T. (2023). The promise of convolutional neural networks for the early diagnosis of the Alzheimer's disease. *Engineering Applications of Artificial Intelligence*, *123*, 1–13. https://doi.org/10.1016/j.engappai.2023.106254

Fathi, S., Ahmadi, M., & Dehnad, A. (2022). Early diagnosis of Alzheimer's disease based on deep learning: A systematic review. *Computers in Biology and Medicine*, *146*, 1–16. https://doi.org/10.1016/j.compbiomed.2022.105634

Folstein, M. F., Folstein, S. E., & McHugh, P. R. (1975). "Mini-mental state": A practical method for grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research*, *12*(3), 189–198. https://doi.org/10.1016/0022-3956(75)90026-6

Gasmi, K., Alyami, A., Hamid, O., Altaieb, M. O., Shahin, O. R., Ben Ammar, L., Chouaib, H., & Shehab, A. (2024). Optimized Hybrid Deep Learning Framework for Early Detection of Alzheimer's Disease Using Adaptive Weight Selection. *Diagnostics*, *14*(24), 2779. https://doi.org/10.3390/DIAGNOSTICS14242779

Google. (2023). *google/vit-base-patch16-224*. https://huggingface.co/google/vit-base-patch16-224

Grossberg, G. T., Tong, G., Burke, A. D., & Tariot, P. N. (2019). Present Algorithms and Future Treatments for Alzheimer's Disease. *Journal of Alzheimer's Disease*, *67*(4), 1157–1171. https://doi.org/10.3233/JAD-180903

Haulcy, R., & Glass, J. (2021). Classifying Alzheimer's Disease Using Audio and Text-Based Representations of Speech. *Frontiers in Psychology*, *11*, 1–13. https://doi.org/10.3389/fpsyg.2020.624137

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition. *Proceedings of the 2016 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778. https://doi.org/10.1109/CVPR.2016.90

Henschel, L., Kügler, D., & Reuter, M. (2022). FastSurferVINN: Building resolution-independence into deep learning segmentation methods—A solution for HighRes brain MRI. *NeuroImage*, *251*, 1–22. https://doi.org/10.1016/j.neuroimage.2022.118933

Hollingshead, A. (1975). Four factor index of social status. In *Yale Journal of Sociology* (Vol. 8).

Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely Connected Convolutional Networks. *Proceedings of the 30th IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2017)*, *2017-January*. https://doi.org/10.1109/CVPR.2017.243

Huber, P. J. (1981). *Robust Statistics*. Wiley. https://doi.org/10.1002/0471725250

Hugging Face – The AI community building the future. (2024). Hugging Face. https://huggingface.co

Hunter, J. D. (2007). Matplotlib: A 2D Graphics Environment. *Computing in Science and Engineering*, *9*(3), 90–95. https://doi.org/10.1109/MCSE.2007.55

Ji, H., Liu, Z., Yan, W. Q., & Klette, R. (2019a). Early Diagnosis of Alzheimer's Disease Based on Selective Kernel Network with Spatial Attention. *Proceedings of the Asian Conference on Pattern Recognition 2019 (ACPR 2019)*, *12047 LNCS*, 503–515. https://doi.org/10.1007/978-3-030-41299-9_39

Ji, H., Liu, Z., Yan, W. Q., & Klette, R. (2019b). Early diagnosis of Alzheimer's disease using deep learning. *Proceedings of the 2nd International Conference on Control and Computer Vision (ICCCV '19)*, 87–91. https://doi.org/10.1145/3341016.3341024

Kaeberlein, M. (2013). Longevity and aging. *F1000Prime Reports*, *5*(5), 1–8. https://doi.org/10.12703/P5-5

Kamada, S., Ichimura, T., & Harada, T. (2021). Image-Based Early Detection of Alzheimer's Disease by Using Adaptive Structural Deep Learning. *Proceedings of the Smart Innovation, Systems and Technologies 2021 (ICOMTA 2021)*, *238*, 595–605. https://doi.org/10.1007/978-981-16-2765-1_49

Khojaste-Sarakhsi, M., Haghighi, S. S., Ghomi, S. M. T. F., & Marchiori, E. (2022). Deep learning for Alzheimer's disease diagnosis: A survey. *Artificial Intelligence in Medicine*, *130*, 1–33. https://doi.org/10.1016/j.artmed.2022.102332

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1 (NIPS'12)*, 1097–1105.

Lauraitis, A., Maskeliūnas, R., Damaševičius, R., & Krilavičius, T. (2020). A Mobile Application for Smart Computer-Aided Self-Administered Testing of Cognition, Speech, and Motor Impairment. *Sensors (Switzerland)*, *20*(11), 1–22. https://doi.org/10.3390/s20113236

Lazli, L. (2025). Improved Alzheimer Disease Diagnosis With a Machine Learning Approach and Neuroimaging: Case Study Development. *JMIRx Med*, *6*, e60866. https://doi.org/10.2196/60866

Leong, L. K., & Abdullah, A. A. (2019). Prediction of Alzheimer's disease (AD) Using Machine Learning Techniques with Boruta Algorithm as Feature Selection Method. *Journal of Physics: Conference Series*, *1372*, 1–8. https://doi.org/10.1088/1742-6596/1372/1/012065

Li, F., & Liu, M. (2018). Alzheimer's disease diagnosis based on multiple cluster dense convolutional networks. *Computerized Medical Imaging and Graphics*, *70*, 101–110. https://doi.org/10.1016/j.compmedimag.2018.09.009

Lin, C. J., & Lin, C. W. (2021). Using Three-dimensional Convolutional Neural Networks for Alzheimer's Disease Diagnosis. *Sensors and Materials*, *33*(10), 3399–3413. https://doi.org/10.18494/SAM.2021.3512

Liu, M., Cheng, D., & Yan, W. (2018). Classification of Alzheimer's Disease by Combination of Convolutional and Recurrent Neural Networks Using FDG-PET Images. *Frontiers in Neuroinformatics*, *12*, 1–12. https://doi.org/10.3389/fninf.2018.00035

Liu, S., Liu, S., Cai, W., Pujol, S., Kikinis, R., & Feng, D. (2014). Early diagnosis of Alzheimer's disease with deep learning. *Proceedings of the 2014 IEEE 11th International Symposium on Biomedical Imaging (ISBI 2014)*, 1015–1018. https://doi.org/10.1109/isbi.2014.6868045

Livni, R., Shalev-Shwartz, S., & Shamir, O. (2013). An Algorithm for Training Polynomial Networks. *ArXiV*, *1304.7045*, 1–22.

Lukiw, W. J. (2012). Amyloid beta (Aβ) peptide modulators and other current treatment strategies for Alzheimer's disease (AD). *Expert Opinion on Emerging Drugs*, *17*(1), 1–27. https://doi.org/10.1517/14728214.2012.672559

Mahmud, T., Barua, K., Habiba, S. U., Sharmen, N., Hossain, M. S., & Andersson, K. (2024). An Explainable AI Paradigm for Alzheimer's Diagnosis Using Deep Transfer Learning. *Diagnostics*, *14*(3), 1–24. https://doi.org/10.3390/diagnostics14030345

Marcus, D. S., Fotenos, A. F., Csernansky, J. G., Morris, J. C., & Buckner, R. L. (2010). Open Access Series of Imaging Studies: Longitudinal MRI Data in Nondemented and Demented Older Adults. *Journal of Cognitive Neuroscience*, *22*(12), 2677–2684. https://doi.org/10.1162/jocn.2009.21407

Matplotlib: Visualization with Python. (2024). https://matplotlib.org

McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, *12*(2), 153–157. https://doi.org/10.1007/BF02295996

Mehmood, A., yang, S., feng, Z., wang, M., Ahmad, A. S., khan, R., Maqsood, M., & Yaqub, M. (2021). A Transfer Learning Approach for Early Diagnosis of Alzheimer's Disease on MRI Images. *Neuroscience*, *460*, 43–52. https://doi.org/10.1016/j.neuroscience.2021.01.002

Mienye, I. D., Swart, T. G., Obaido, G., Jordan, M., & Ilono, P. (2025). Deep Convolutional Neural Networks in Medical Image Analysis: A Review. *Information*, *16*(3), 195. https://doi.org/10.3390/INFO16030195

Morris, J. C. (1993). The Clinical Dementia Rating (CDR): Current version and scoring rules. *Neurology*, *43*(11), 2412–2414. https://doi.org/10.1212/wnl.43.11.2412-a

Neuroimaging in Python — NiBabel. (2024). https://nipy.org/nibabel

Ntampakis, N., Diamantaras, K., Argyriou, V., & Sarigianndis, P. (2024). Enhanced Deep Learning Methodologies and MRI Selection Techniques for Dementia Diagnosis in the Elderly Population. *ArXiv*, *2407.17324v2*, 1–12.

Pan, S. J., & Yang, Q. (2010). A Survey on Transfer Learning. *IEEE Transactions on Knowledge and Data Engineering*, *22*(10), 1345–1359. https://doi.org/10.1109/TKDE.2009.191

Papanastasiou, G., Dikaios, N., Huang, J., Wang, C., & Yang, G. (2024). Is Attention all You Need in Medical Image Analysis? A Review. *IEEE Journal of Biomedical and Health Informatics*, *28*(3), 1398–1411. https://doi.org/10.1109/JBHI.2023.3348436

Pappas, B. A., Bayley, P. J., Bui, B. K., Hansen, L. A., & Thal, L. J. (2000). Choline acetyltransferase activity and cognitive domain scores of Alzheimer's patients. *Neurobiology of Aging*, *21*(1), 11–17. https://doi.org/10.1016/S0197-4580(00)00090-7

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Qiu, S., Chang, G. H., Panagia, M., Gopal, D. M., Au, R., & Kolachalama, V. B. (2018). Fusion of deep learning models of MRI scans, Mini–Mental State Examination, and logical memory test enhances diagnosis of mild cognitive impairment. *Alzheimer's and Dementia: Diagnosis, Assessment and Disease Monitoring*, *10*, 737–749. https://doi.org/10.1016/j.dadm.2018.08.013

Raza, M. L., Hassan, S. T., Jamil, S., Hyder, N., Batool, K., Walji, S., & Abbas, M. K. (2025). Advancements in deep learning for early diagnosis of Alzheimer's disease using multimodal neuroimaging: challenges and future directions. *Frontiers in Neuroinformatics*, *19*, 1557177. https://doi.org/10.3389/FNINF.2025.1557177/XML

Rehman Butt, A. U., Hamid, I., Nawaz, Q., Mahmood, T., Zhang, X., & Yaqub, M. (2024). A Novel Multi-Scale Deep Learning Approach for the Early Detection of Alzheimer's Disease Using fMRI. *Proceedings of 2024 5th International Conference on Computer, Big Data and Artificial Intelligence, ICCBD+AI 2024*, 85–90. https://doi.org/10.1109/ICCBD-AI65562.2024.00022

Rhman, M., Rahman, F., Hossain, M. M., Emu, U. H., Akter, K., & Mridha, M. F. (2021). Predicting Alzheimer's Disease at Low Cost Using Machine Learning. *Proceedings of the 2021 International Conference on Science and Contemporary Technologies (ICSCT 2021)*, 1–5. https://doi.org/10.1109/ICSCT53883.2021.9642536

Selkoe, D. J. (2001). Alzheimer's disease: Genes, proteins, and therapy. *Physiological Reviews*, *81*(2), 741–766. https://doi.org/10.1152/physrev.2001.81.2.741

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *Proceedings of the IEEE International Conference on Computer Vision (ICCV 2017)*, *2017-October*. https://doi.org/10.1109/ICCV.2017.74

Shanmugam, J. V., Duraisamy, B., Simon, B. C., & Bhaskaran, P. (2022). Alzheimer's disease classification using pre-trained deep networks. *Biomedical Signal Processing and Control*, *71*, 1–8. https://doi.org/10.1016/j.bspc.2021.103217

Shi, J., Zheng, X., Li, Y., Zhang, Q., & Ying, S. (2018). Multimodal Neuroimaging Feature Learning with Multimodal Stacked Deep Polynomial Networks for Diagnosis of Alzheimer's Disease. *IEEE Journal of Biomedical and Health Informatics*, *22*(1), 173–183. https://doi.org/10.1109/JBHI.2017.2655720

Shin, H. C., Roth, H. R., Gao, M., Lu, L., Xu, Z., Nogues, I., Yao, J., Mollura, D., & Summers, R. M. (2016). Deep Convolutional Neural Networks for Computer-Aided Detection: CNN Architectures, Dataset Characteristics and Transfer Learning. *IEEE Transactions on Medical Imaging*, *35*(5), 1285–1298. https://doi.org/10.1109/TMI.2016.2528162

Simonyan, K., & Zisserman, A. (2015). Very Deep Convolutional Networks for Large-Scale Image Recognition. In Y. Bengio & Y. LeCun (Eds.), *Proceedings of the 3rd International Conference on Learning Representations (ICLR 2015)* (pp. 1–14).

Suk, H. Il, Lee, S. W., & Shen, D. (2014). Hierarchical feature representation and multimodal fusion with deep learning for AD/MCI diagnosis. *NeuroImage*, *101*, 569–582. https://doi.org/10.1016/j.neuroimage.2014.06.077

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., & Rabinovich, A. (2015). Going deeper with convolutions. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 1–9. https://doi.org/10.1109/CVPR.2015.7298594

Tajbakhsh, N., Jeyaseelan, L., Li, Q., Chiang, J. N., Wu, Z., & Ding, X. (2020). Embracing imperfect datasets: A review of deep learning solutions for medical image segmentation. *Medical Image Analysis*, *63*, 101693. https://doi.org/10.1016/J.MEDIA.2020.101693

Tiriki, N. (2010). *OASIS-2 Longitudinal Scan Data*. Kaggle. https://www.kaggle.com/datasets/nadiatriki/oasis-2-longitudinal-scan-data

Vernekar, S. R., & Selva Kumar, S. (2024). Exploration of Explainable AI with Deep Learning Model for Early Detection of Alzheimer's Disease. *Proceedings of 8th IEEE International Conference on Computational System and Information Technology for Sustainable Solutions, CSITSS 2024*, 1–6. https://doi.org/10.1109/CSITSS64042.2024.10816763

Waldo-Benítez, G., Padierna, L. C., Ceron, P., & Sosa, M. A. (2024). Dementia classification from magnetic resonance images by machine learning. *Neural Computing and Applications*, *36*, 2653–2664. https://doi.org/10.1007/s00521-023-09163-y

Waskom, M. L. (2021). seaborn: statistical data visualization. *Journal of Open Source Software*, *6*(60), 1–4. https://doi.org/10.21105/joss.03021

Wolf, T., Debut, L., Sanh, V., Chaumond, J., & ... (2020). Transformers: State-of-the-art Natural Language Processing. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, *1910.03771*, 38–45.

Wong, T. H., Seelaar, H., Melhem, S., Rozemuller, A. J. M., & van Swieten, J. C. (2020). Genetic screening in early-onset Alzheimer's disease identified three novel presenilin mutations. *Neurobiology of Aging*, *86*, 201.e1-201.e14. https://doi.org/10.1016/j.neurobiolaging.2019.01.015

Ye, J., Chen, K., Wu, T., Li, J., Zhao, Z., Patel, R., Bae, M., Janardan, R., Liu, H., Alexander, G., & Reiman, E. (2008). Heterogeneous data fusion for alzheimer's disease study. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '08)*, 1025–1033. https://doi.org/10.1145/1401890.1402012

Yildiz, K., Gunes, E., & Bas, A. (2021). CNN-based Gender Prediction in Uncontrolled Environments. Duzce University Journal of Science and Technology, 9(2), 890–898. https://doi.org/10.29130/dubited.763427

Yildiz, S. G., & Yildiz, K. (2023). Ann Based Early Detection of Alzheimer Disease on Selected Features. *Journal of Engineering Sciences and Design*, *11*(4), 1508–1516. https://doi.org/10.21923/JESD.1296283