BMBAD BJESR Mühendislik Bilimleri ve Araştırmaları Dergisi
Journal of Engineering Sciences and Researches

Stage Classification of Alzheimer's Disease MRI Data via Deep Neural Network Architectures

Derin Sinir Ağı Mimarileri ile Alzheimer Hastalığı MR Görüntülerinin Evre Sınıflandırılması

[1]Aslıhan GÜVEN, [2]Abdurrahim Hüseyin EZİRMİK, [3]Mustafa Furkan CEYLAN, [4]Fatih AYDIN

[1,2,3,4]Balıkesir Üniversitesi, Mühendislik Fakültesi, Bilgisayar Mühendisliği Bölümü, Balıkesir, Türkiye

[1]aslihangvn@outlook.com, [2]huseyin.ezirmik@balikesir.edu.tr,
[3]furkan.ceylan@balikesir.edu.tr, [4]fatih.aydin@balikesir.edu.tr

Araştırma Makalesi/Research Article

ARTICLE INFO

Article history

Received : 22 January 2026
Accepted : 27 March 2026

Keywords:

Deep Learning, Transfer Learning, Alzheimer's Disease, MRI, Model Benchmarking

ABSTRACT

This study compares five deep learning architectures for Alzheimer's disease stage classification using structural magnetic resonance imaging (MRI) data. The experiments are conducted using the OASIS dataset, which comprises brain MRI scans representing multiple clinical stages of Alzheimer's disease obtained from real-world clinical scenarios. Deep learning models are trained and assessed using the same parameters to maintain consistency in comparison. Accuracy, F1 score, and macro-averaged ROC-AUC are used to measure performance on the test set. InceptionV3 achieves the highest performance (test accuracy: 0.86, macro-AUC: 0.97). MobileNetV2 obtains 0.74 test accuracy and 0.90 macro-AUC while requiring significantly lower computational cost. These findings indicate that InceptionV3 delivers the most reliable overall performance, whereas MobileNetV2 represents a practical choice for deployment in resource-constrained environments.

© 2026 Bandirma Onyedi Eylul University, Faculty of Engineering and Natural Science. Published by Dergi Park. All rights reserved.

MAKALE BİLGİSİ

Makale Tarihleri

Gönderim : 22 Ocak 2026
Kabul : 27 Mart 2026

Anahtar Kelimeler:

Derin Öğrenme, Transfer Öğrenme, Alzheimer Hastalığı, MR, Model Karşılaştırması

ÖZET

Bu çalışma, yapısal manyetik rezonans görüntüleme (MRI) verileri kullanılarak Alzheimer hastalığının evre sınıflandırması için beş derin öğrenme mimarisini karşılaştırmaktadır. Deneylerde, dört klinik Alzheimer evresini temsil eden 416 bireye ait MRI görüntülerini içeren OASIS veri seti kullanılmıştır. Tüm modeller, adil karşılaştırma sağlamak amacıyla aynı ön işleme adımları ve hiperparametre ayarları altında eğitilmiştir. Performans, test kümesi üzerinde doğruluk, F1 skoru ve makro ortalamalı ROC-AUC metrikleri ile değerlendirilmiştir. InceptionV3 modeli %0.86 test doğruluğu ve 0.97 makro AUC değeri ile en yüksek performansı göstermiştir. MobileNetV2 ise %0.74 test doğruluğu ve 0.90 makro AUC değeri ile daha düşük hesaplama maliyeti sunmuştur. Bulgular, InceptionV3'ün en güçlü sınıflandırma performansını sağladığını; MobileNetV2'nin ise kaynak kısıtlı ortamlarda uygulanabilir bir alternatif olduğunu göstermektedir.

© 2026 Bandırma Onyedi Eylül Üniversitesi, Mühendislik ve Doğa Bilimleri Fakültesi. Dergi Park tarafından yayınlanmaktadır. Tüm Hakları Saklıdır.

ORCID ID: [1]0009-0001-9263-0215
[2]0000-0002-1154-1537
[3]0009-0005-5609-395X
[4]0000-0001-9679-0403

1. INTRODUCTION

Population aging has made Alzheimer's disease (AD) increasingly common worldwide, and it is now a major public health concern [1]. The World Health Organization reports that about 57 million people worldwide were living with dementia in 2021. New cases are estimated to be close to 10 million each year and most occur in low- and middle-income countries which points to a rapidly growing public health burden linked to Alzheimer's disease and related dementias [2]. AD progressively impairs memory, cognition, language and decision-making as neural structures are damaged over time [3]. As the disease advances, individuals experience reduced ability to maintain daily activities, decreasing independence and increasing care needs [4]. Because AD often begins slowly and progresses subtly, early diagnosis is critical to slow progression, improve the effectiveness of pharmacological and non-pharmacological interventions and enhance quality of life for patients and caregivers [5]. This creates demand for reliable, repeatable, and sensitive approaches that can capture early neurodegenerative changes.

Clinical diagnosis of AD typically follows a multi-stage evaluation that includes medical history, observation of cognitive symptoms, neuropsychological tests and neuroimaging analysis [4]. Common screening tools such as MMSE and MoCA can guide staging decisions, but they may show limited sensitivity for early-stage cases [5]. Neuroimaging modalities such as MRI, PET, and CT can reveal structural and functional changes however, interpretation requires expert knowledge and may include observer-dependent subjectivity [4]. In addition, advanced imaging methods can be expensive and are not equally accessible in all healthcare systems, especially in low- and middle-income settings [2]. These limitations increase interest in automated systems that are fast, objective, repeatable, and clinically feasible.

In recent years interest has increased in imaging-based methods that can reveal morphological changes linked to early Alzheimer's disease [6]. Among these approaches magnetic resonance imaging (MRI) is commonly used because it is noninvasive and provides high tissue contrast with detailed anatomical information. Structural MRI can help identify AD related patterns such as brain volume loss hippocampal atrophy and cortical thinning [7]. Still manual review of MRI scans is slow and depends strongly on expert experience which can differ between raters. To ease these limitations machine learning has become more common in medical image analysis to support faster and more consistent diagnostic workflows [8]. Deep learning methods are particularly useful since they can learn meaningful representations from complex imaging data [9]. Convolutional neural networks (CNNs) have shown strong results in tasks like classification and anomaly detection, and they have shaped much of recent medical imaging research [8]. In AD studies CNN based models can examine MRI data at fine spatial detail and may capture subtle structural patterns that matter in early-stage differentiation [10]. Beyond structural MRI AI methods can also combine multi modal inputs such as fMRI PET and DTI to support broader clinical insight including progression estimation and risk assessment [11]. At the same time real world use still faces challenges tied to explainability, transparency data quality and reliable generalization across different clinical sites.

In this study we compare five widely used deep learning architectures EfficientNet-B0, ResNet50, MobileNetV2, a U-Net encoder-based classifier, and InceptionV3 under a consistent preprocessing and training pipeline. Performance is evaluated using standard classification metrics and final ranking is intentionally test-driven rather than validation driven. Validation metrics are used to control training and select checkpoints but model selection for reporting is based on held-out test performance, emphasizing macro-AUC and F1-score alongside accuracy. In addition, we report practical efficiency indicators (model size and training time) to support deployment-oriented decisions. This is important because a model with slightly higher accuracy may be less attractive if it requires substantially higher computational resources, while a smaller model may be preferable for real-world use. The primary goal is to identify practical deep learning options for MRI-based AD staging by quantifying both generalization performance on a held-out test set and operational cost factors that affect applicability. The study also contributes a systematic comparison under a consistent evaluation logic, addressing a gap in the literature where architecture-level comparisons are often reported under non-uniform settings.

The contributions introduced in this study can be summarized as follows:

- A consistent benchmarking pipeline for four-class AD staging using five commonly used deep learning architectures.
- A test-driven evaluation that emphasizes generalization metrics (macro-AUC and F1-score) rather than validation accuracy.
- An applied engineering interpretation linking predictive performance to model size, training cost, and deployability.

The remainder of this article is organized as follows: Section 2 reviews related work on MRI-based Alzheimer's disease analysis and deep learning approaches. Section 3 describes the materials and methods, including the dataset, preprocessing pipeline, model configurations, and evaluation metrics. Section 4 presents experimental results, reporting the test-driven benchmarking outcomes and best-model diagnostic analyses. Finally, Section 5 concludes the paper and outlines potential directions for future work.

2. RELATED WORKS

This section reviews recent work on deep learning methods for Alzheimer's disease diagnosis with a focus on neuroimaging approaches. Early studies examined the biological basis of the disease and the value of medical imaging. Over time the focus moved toward deep learning models for image analysis. In this line of research common convolutional neural network architectures such as EfficientNet, ResNet, U-Net, MobileNetV2 and Inception have been widely studied [12]. These models are often compared based on diagnostic accuracy computational cost and their suitability for use in clinical settings.

Recent progress in artificial intelligence has changed how medical diagnosis and clinical decision support are developed. Deep learning methods are now widely used in medical image analysis because they can learn spatial patterns directly from imaging data. Many studies report strong results across different diagnostic tasks. In some cases, model performance is similar to that of human experts especially for tasks that involve complex visual pattern recognition in medical images [8], [9].

Earlier machine learning approaches relied on hand crafted features such as texture descriptors gray level histograms or GLCM based measures. These techniques can work well in controlled environments, but they depend heavily on expert knowledge during feature design. They also tend to show weaker scalability and generalization. Deep learning models address these limits by combining feature extraction and classification within a single learning framework. This allows useful representations to be learned directly from data. Several studies report that CNN based approaches outperform traditional machine learning methods in Alzheimer's disease classification with higher accuracy and AUC values [13], [14].

Alzheimer's disease develops gradually and shows large variation between individuals. Because of this relying on a single imaging modality can reduce diagnostic sensitivity especially in early stages. To address this issue researchers have proposed multimodal deep learning models that combine MRI with other data sources such as PET functional MRI cerebrospinal fluid biomarkers or cognitive test scores. These models usually process each modality separately and later combine the learned features. This design supports joint modeling of structural and functional changes and has been shown to improve early-stage discrimination compared to MRI only approaches [14], [15]. Deep learning-based image analysis has also shown promise in detecting subtle structural changes linked to neurodegeneration. CNN models can identify small anatomical differences in MRI scans that are difficult to observe through visual inspection alone. This supports early diagnosis disease staging and long-term monitoring with reasonable reliability [11]. Many studies show that these models can separate Alzheimer's patients from healthy controls and from individuals with mild cognitive impairment which highlights their clinical relevance [10]. In recent years, several studies have specifically addressed multi-class Alzheimer's disease stage classification using structural MRI datasets such as OASIS and ADNI. Narnaware and Mulla proposed a Hybrid Capsule Autoencoder (HCapsNet) framework combining capsule networks and handcrafted feature fusion, reporting accuracy above 99% on multi-class classification tasks [16]. Lakshmanan et al. introduced an attention-enhanced deep CNN evaluated on subject-level OASIS data, achieving 97% accuracy while improving interpretability through spatial attention maps [17]. Similarly, Kumar et al. developed Neuro_CBAM-ADNet, integrating convolutional block attention modules and GAN-style noise injection, and reported a mean accuracy of 98.28% for four-stage classification [18]. While these approaches demonstrate strong numerical performance, they often rely on high-capacity architectures and computationally intensive pipelines.

A major challenge in medical imaging research is the limited availability of labeled clinical data [9]. This increases the risk of overfitting, especially for deep models with many parameters. Transfer learning has become a common strategy to reduce this problem. In this approach models pretrained on large natural image datasets such as ImageNet are adapted to medical imaging tasks. This allows effective learning even when datasets are small and have been shown to improve performance in Alzheimer's disease studies [19]. Beyond research settings deep learning systems are increasingly being adopted in clinical practice. The number of artificial intelligence based medical devices approved by the U.S. Food and Drug Administration has increased steadily [20]. This trend is especially visible in areas such as radiology dermatology mammography and retinal imaging where CNN based tools support clinicians by improving diagnostic consistency [21]. Similar developments are now emerging in neuroimaging applications related to Alzheimer's disease and other neurological disorders [13].

In Alzheimer focused imaging studies CNN architectures have been shown to detect important structural biomarkers. These include hippocampal atrophy cortical thinning and ventricular enlargement which are closely linked to disease progression [22]. These capabilities support more objective and repeatable diagnostic workflows and may allow earlier clinical intervention. At the same time the growing use of deep learning in medical imaging raises important concerns. Issues related to data privacy, ethical approval and model reliability remain critical. MRI data is highly sensitive and requires strict anonymization, secure storage and compliance with ethical standards. Questions about responsibility in cases of misdiagnosis transparency of model decisions and regulatory consistency are still actively discussed [23]. For these reasons deep learning based diagnostic systems should be evaluated not only for technical performance but also for ethical suitability, safety and clinical trust.

3. MATERIAL AND METHOD

This study introduces a structured deep learning framework for automatic classification of Alzheimer's disease stages using structural MRI data. The overall workflow is shown in Figure 1 and outlines the full process from data acquisition to evaluation and deployment related considerations. The pipeline starts with T1 weighted structural MRI scans obtained from the openly available OASIS-1 dataset [24]. The original three-dimensional brain volumes are converted into two dimensional coronal slices and then normalized to ensure consistency across subjects. After preprocessing, the images are divided into training validation and test sets using a stratified split so that class distributions are preserved.

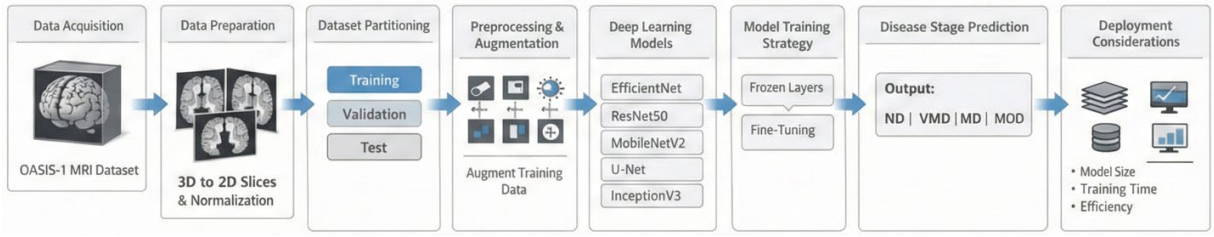


Figure 1. Workflow of the MRI based deep learning for Alzheimer's disease stage classification.

Data augmentation and preprocessing steps are applied only to the training data to improve robustness and reduce overfitting while the validation and test sets remain unchanged. Several transfer learning based convolutional neural network architectures are evaluated including EfficientNet B0, ResNet50, MobileNetV2, an encoder based U-Net classifier and InceptionV3. Training follows a two stage approach with an initial phase where the backbone is frozen followed by selective fine tuning of deeper layers. Each model outputs four classes representing the clinical stages of Alzheimer's disease which are non demented, very mild demented, mild demented and moderate demented. Model performance is evaluated on the held-out test set using standard classification metrics and additional analyses examine model size parameter count and training time to assess suitability for practical deployment as summarized in Figure 1.

3.1. Dataset and Data Preparation

This study uses the openly accessible OASIS-1 structural MRI dataset to classify Alzheimer's disease stages from T1-weighted brain scans [24]. The dataset was created to support research on aging- and dementia-related neurodegeneration and contains one anonymized MRI scan per participant. In total, dataset includes 416 individuals aged 18–96. Diagnostic labels follow the clinical dementia rating (CDR) scale and are grouped into four classes: Non demented (CDR = 0), very mild demented (CDR = 0.5), mild demented (CDR = 1) and moderate dementia (CDR = 2). Figure 2 presents representative coronal MRI slices from each class in the OASIS-1 dataset to illustrate typical visual appearances across dementia stages.

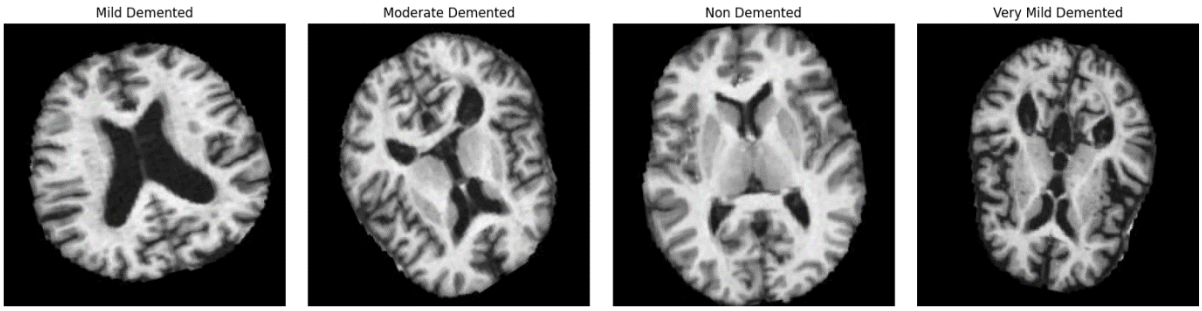


Figure 2. Representative coronal MRI slices from the OASIS-1 dataset illustrating the four diagnostic classes.

The original MRI data consist of three-dimensional volumes with a spatial resolution of 176 × 208 × 176. In this study, two-dimensional coronal MRI slices derived from these volumes were used to facilitate benchmarking with widely adopted CNN architectures while avoiding the substantially higher computational cost associated with 3D deep learning frameworks. The resulting folder-based dataset comprised 8,268 images organized by class and divided into three predefined subsets: training, validation, and testing. Specifically, 5,141 images were allocated to the training set, 1,848 to the validation set, and 1,279 to the test set. The detailed class-wise distribution of MRI slices across the training, validation, and test subsets is presented in Table 1.

Class representation was preserved across all subsets using a stratified distribution strategy. To strictly prevent data leakage, dataset partitioning was performed at the subject level rather than at the slice level. All slices originating from a given subject were assigned exclusively to a single subset (training, validation, or testing). Consequently, no subject contributed slices to more than one subset. This subject-wise splitting protocol

Table 1. Class-wise distribution of MRI slices across the training, validation, and test sets.

Class	Train	Validation	Test	Total
Non-Demented	2560	914	640	4114
Very Mild Demented	1812	657	448	2917
Mild Demented	717	261	179	1157
Moderate Demented	52	16	12	80
Total	5141	1848	1279	8268

ensures a realistic, unbiased, and clinically meaningful evaluation of model generalization performance.
Although Alzheimer's disease is inherently a three-dimensional neurodegenerative process, the use of 2D slice-based modeling in this study was a deliberate methodological choice. Several previous studies have demonstrated that 2D CNN architectures can effectively capture discriminative structural patterns from MRI data while maintaining lower computational complexity and improved training stability, particularly when dataset size is limited. In our case, the OASIS-1 dataset includes 416 subjects, which may not be sufficient to reliably train high-capacity 3D CNN models without a substantial risk of overfitting. By converting 3D volumes into coronal slices, we enabled benchmarking across widely adopted ImageNet-pretrained 2D backbones under a unified transfer learning framework. This approach allows for more controlled architectural comparison while reducing GPU memory requirements and training time. Nevertheless, we acknowledge that 3D models preserve volumetric context and may provide complementary spatial information. Future work will investigate 3D CNN and hybrid 2D–3D architectures to further explore volumetric feature integration.

3.2. Preprocessing and Data Augmentation

All MRI slices were processed using a standardized pipeline intended to provide consistent model inputs and to improve training stability and generalization. Images were resized using bilinear interpolation and aligned with the expected input shapes of each architecture. EfficientNetB0, ResNet50, and MobileNetV2 used 224 × 224 × 3 inputs, InceptionV3 used 299 × 299 × 3, and the U-Net encoder classifier used 256 × 256 × 3. This model-aligned resizing strategy supports more comparable benchmarking by matching each backbone's standard resolution.
Normalization was applied according to each backbone's expected preprocessing convention. For EfficientNetB0, ResNet50, and InceptionV3, images were first rescaled by 1/255 and then passed through the corresponding preprocess input function to match the input distribution of ImageNet-pretrained weights. The U-Net encoder classifier used only 1/255 rescaling without ImageNet-specific normalization. MobileNetV2 was loaded using image dataset from directory and normalized using MobileNetV2 preprocess input implemented through a Lambda layer within the model. These steps ensured that inputs remained compatible with each architecture's pretrained feature extractor and supported fair comparison across models.
Overfitting was handled by applying data augmentation only to the training set in the presence of limited labelled medical imaging data. The validation and test sets were kept fixed to maintain an unbiased evaluation process. For EfficientNetB0, ResNet50, InceptionV3 and the U-Net encoder classifier, augmentation was carried out using ImageDataGenerator. The applied transformations included small random rotations, translations, zoom, brightness changes, shear operations and horizontal flipping. Nearest neighbour filling was used to handle pixels created by geometric transformations. For MobileNetV2, augmentation was performed through Keras augmentation layers that were integrated directly into the model pipeline. This allowed random flips, rotations, zoom operations, translations and contrast adjustments to be applied during training. These transformations were chosen to improve robustness to small acquisition related variations while preserving the diagnostic information contained in structural MRI slices.

3.3. Deep Learning Models and Transfer Learning Setup

Five deep learning architectures were benchmarked to represent diverse trade-offs between predictive capability and computational efficiency: EfficientNetB0, ResNet50, MobileNetV2, an encoder-based U-Net classifier, and InceptionV3[25]. EfficientNet is based on compound scaling [26]. ResNet leverages residual learning for stable optimization in deep networks [19], MobileNetV2 is designed for resource-constrained settings using inverted residual and linear bottleneck blocks [27], U-Net follows an encoder–decoder design commonly used in biomedical imaging [28], and InceptionV3 applies parallel multi-scale convolutional paths to capture heterogeneous feature patterns [25].
For EfficientNetB0, ResNet50, MobileNetV2, and InceptionV3, a transfer learning strategy was used by initializing each backbone with ImageNet-pretrained weights and removing the original classification layers (include_top=False). A task-specific classification head was then attached to produce four-class predictions (ND, VMD, MD, MOD). The head design was kept consistent across backbones to support fair comparison and consisted of global pooling followed by regularized fully connected layers and a final softmax layer. In contrast, the U-Net model was implemented as an encoder-based classifier, where the encoder output was pooled into a compact feature representation and passed to a classification head, avoiding the original segmentation-oriented decoder pathway.

Although U-Net was originally designed for biomedical image segmentation, it was deliberately included in this benchmark to evaluate whether its encoder structure could effectively capture fine-grained spatial patterns relevant to Alzheimer's disease staging. Structural MRI data contain subtle and spatially localized atrophy patterns, particularly in hippocampal and cortical regions, which are clinically meaningful. The hierarchical convolutional blocks of the U-Net encoder are capable of extracting multi-scale spatial representations that may support stage-level discrimination. By incorporating U-Net as an encoder-based classifier, this study investigates whether a segmentation-oriented spatial architecture can provide complementary representational benefits compared to classification-native backbones such as VGG or DenseNet within the same experimental framework.

InceptionV3 represents the high-capacity end of the benchmark and is intended to capture heterogeneous morphological patterns through multi-branch convolutional modules operating at different receptive-field scales. As shown in Figure 3, the architecture is organized into repeated Inception modules separated by grid-size reduction blocks, enabling progressively abstract feature extraction while maintaining multi-scale sensitivity [25]. This design is particularly suitable for MRI-based classification where discriminative cues may appear at different spatial resolutions.

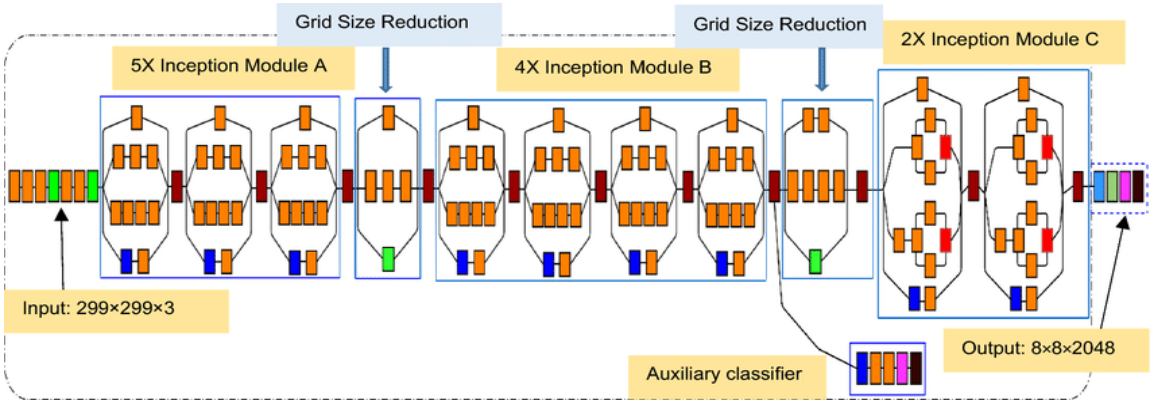


Figure 3. Overview of the InceptionV3 architecture [25].

MobileNetV2 serves as the efficiency focused model in the benchmark and is aimed at deployment settings with limited computing power and memory. As shown in Figure 4, the architecture uses lightweight building blocks based on inverted residual connections and linear bottlenecks, which lower computational cost while still preserving key feature representations [27]. This design leads to faster training and a smaller model size, making MobileNetV2 a practical choice when deployment feasibility and runtime limits are important considerations.

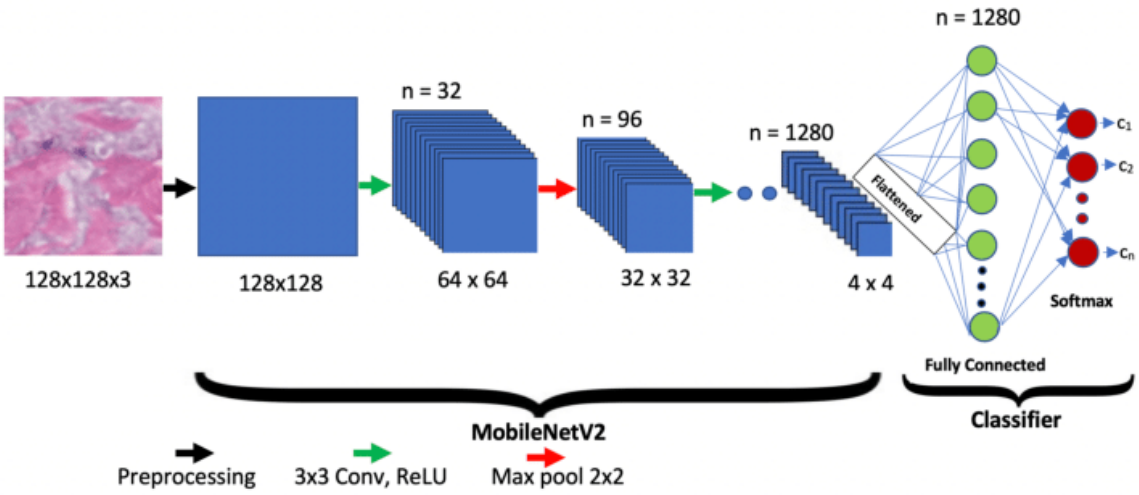


Figure 4. Schematic of the MobileNetV2 architecture [27].

3.4. Training Protocol and Optimization Settings

All experimental evaluations were performed on the Google Colab Pro platform using an NVIDIA Tesla T4 graphics processor with 16 GB of video memory and 25 GB of system RAM within an Ubuntu-based execution environment. The implementation used Python 3.10, TensorFlow 2.15, and Keras, with NumPy, Pandas, Matplotlib, and scikit-learn libraries used for data handling and evaluation. Training was performed with a batch size of 16 using a unified two-stage strategy to support controlled comparison across models. In Stage 1, the pretrained backbone (when applicable) was frozen and only the classification head was trained to obtain stable

convergence and reduce overfitting risk. In Stage 2, selective fine-tuning was performed by unfreezing the last layers of the backbone while keeping Batch Normalization layers non-trainable to preserve stable running statistics.

Training was controlled using EarlyStopping (monitor = validation loss, patience = 12) to prevent unnecessary epochs and reduce overfitting, ModelCheckpoint to preserve the best-performing model on the validation set, and ReduceLROnPlateau to adapt the learning rate when validation loss stagnated. Gradient clipping (clipnorm = 1.0) was applied to improve optimization stability. Class imbalance was addressed using class weights and, for several models, focal loss with label smoothing [29]. The architecture-specific optimization configuration used in this study is summarized in Table 2.

Table 2. Optimization and fine-tuning parameters used for each deep learning architecture.

Model Architecture	Optimizer and Learning Rate (Stage 1 → Stage 2)	Loss Function Components	Fine-Tuning Depth (Active Layers)
EfficientNet-B0	Adam ($1 \times 10^{-3} \rightarrow 1 \times 10^{-5}$)	Focal Loss ($\gamma = 2.0$) with Label Smoothing (0.05) and Class Weights	Last 120 layers excluding Batch Normalization
ResNet50	Adam ($8 \times 10^{-4} \rightarrow 1 \times 10^{-5}$)	Focal Loss ($\gamma = 2.0$) with Label Smoothing (0.05) and Class-Frequency Weighting	Last 100 layers excluding Batch Normalization
MobileNetV2	Adam ($1 \times 10^{-3} \rightarrow 3 \times 10^{-5}$)	Categorical Cross-Entropy with Label Smoothing ($\varepsilon = 0.1$) and Class Weights	Last 120 layers excluding Batch Normalization
InceptionV3	Adam ($8 \times 10^{-4} \rightarrow 1 \times 10^{-5}$)	Focal Loss ($\gamma = 2.0$) with Label Smoothing (0.05) and Class Weights	Last 140 layers excluding Batch Normalization
U-Net Encoder	Adam ($5 \times 10^{-4} \rightarrow 2 \times 10^{-5}$)	Focal Loss ($\gamma = 2.0$) with Label Smoothing (0.05) and Class Weights	All layers are unfrozen excluding Batch Normalization

3.5. Evaluation Metrics

The performance of the proposed models was evaluated on the held-out test set using several complementary metrics. These metrics were selected to reflect both overall classification accuracy and the ability of the models to distinguish between different Alzheimer's disease stages. The evaluation includes Accuracy Precision Recall F1-score and macro-averaged ROC-AUC. Macro averaging was preferred because it reduces the influence of dominant classes and provides a more balanced view of performance across all disease stages. In addition to predictive results the study also reports practical indicators such as model size parameter count approximate FLOPs and total training time. Final model selection is based on test-set performance especially F1-score and macro-AUC rather than validation accuracy alone in order to better reflect generalization to unseen data.

Accuracy represents the proportion of correctly classified samples among all test samples and provides a general measure of model correctness. It is defined as:

$$\text{Accuracy} = \frac{\sum_{i=1}^{k} T\,P_i}{\sum_{i=1}^{k} (T\,P_i + FP_i + FN_i + TN_i)} \tag{1}$$

Here TP_i refers to the number of correctly predicted samples of class i. FP_i represents incorrectly predicted positive samples. FN_i denotes missed positive samples and TN_i indicates correctly predicted negative samples. Although accuracy is easy to interpret it can be misleading when class distributions are imbalanced. This situation is common in multi-stage Alzheimer's disease datasets.

Precision measures how reliable the positive predictions of the model are. It shows the proportion of correctly predicted positive samples among all samples predicted as positive and is given by:

$$\text{Precision}_i = \frac{TP_i}{TP_i + FP_i} \tag{2}$$

A high precision value indicates that the model produces fewer false positive results. This is important in clinical applications where incorrect disease labeling may cause unnecessary medical procedures or stress.

Recall also known as sensitivity measures how well the model identifies actual positive cases. It is defined as:

$$\text{Recall}_i = \frac{TP_i}{TP_i + FN_i} \tag{3}$$

Recall is especially critical in medical diagnosis because low recall means that true patients may be classified as healthy. In the context of Alzheimer's disease this may delay diagnosis and treatment.

The F1-score combines precision and recall into a single value using their harmonic mean. This metric provides a balanced evaluation when both false positives and false negatives are important:

$$F1_i = 2 \times \frac{\text{Precision}_i \times \text{Recall}_i}{\text{Precision}_i + \text{Recall}_i} \quad (4)$$

To ensure fair evaluation across all classes the Macro-F1 score was calculated by averaging the F1-scores of all classes:

$$\text{Macro-F1} = \frac{1}{k}\sum_{i=1}^{k} F\,1_i \quad (5)$$

This approach treats each disease stage equally and is well suited for datasets with class imbalance. Finally, ROC-AUC analysis was used to assess the discriminative ability of the models across different decision thresholds. Class-wise ROC curves were computed, and macro-averaged ROC-AUC values were reported. This metric is independent of a specific threshold and provides a robust measure of classification quality for clinical decision support systems.

4. RESULTS AND DISCUSSION

This section examines the performance of deep learning models for four class Alzheimer's disease stage classification using structural MRI data. The main results are summarized in Table 3, which includes validation accuracy, test accuracy, test macro-AUC, test F1 score, model size, storage use, and total training time. Instead of focusing mainly on validation results, this study gives more weight to test set metrics. These values better show how the models behave on unseen data and they reflect conditions closer to real clinical practice where generalization matters.

The results show a clear difference between validation accuracy and test performance for several models. Some architectures reach very high validation accuracy, but their test accuracy and F1 score drop clearly. This suggests overfitting, especially when the model is too complex for the size and variability of the dataset. Because of this, model performance is judged mainly using test macro-AUC and test F1 score. These metrics describe class level behavior more clearly and they remain more reliable in multi class problems where class imbalance is common.

4.1. Comparative Performance Across Models

As shown in Table 3, InceptionV3 achieves the highest overall classification performance on the test set, with a test accuracy of 75.29%, a macro-AUC of 0.9224, and the highest test F1-score (76.41%). These results indicate strong discriminative power across all Alzheimer stages and confirm the effectiveness of multi-scale feature extraction for capturing heterogeneous neuroanatomical patterns in MRI data. However, this performance comes at the cost of a large model capacity (22M parameters), a substantial storage footprint (92 MB), and the longest training time among the CNN-based classifiers.

Table 3. Model performance and efficiency comparison on the OASIS test set.

Deep Learning Model	Validation Accuracy (%)	Test Accuracy (%)	Test Macro-AUC	Test F1-Score (%)	Parameter Size	Model Size (MB)	Training Time (min.)
InceptionV3	99.95	75.29	0.9224	76.41	22M	92	~100
MobileNetV2	97.35	73.10	0.9165	68.81	3.4M	14	~55
ResNet50	98.92	73.10	0.9140	69.97	23M	98	~90
EfficientNet-B0	76.03	65.68	0.8338	62.08	5.3M	20	~65
U-Net Encoder	55.52	52.07	0.7889	37.47	31M	120	~110

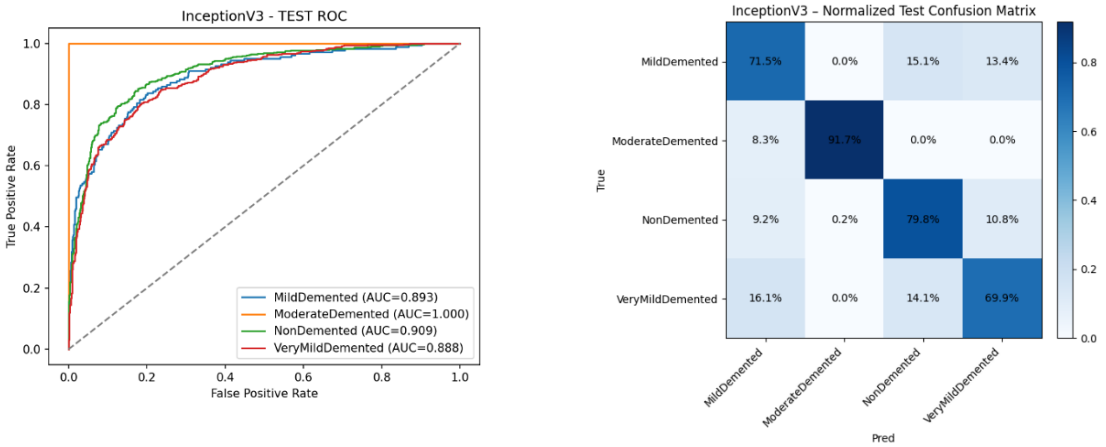
MobileNetV2 shows a different balance between performance and efficiency. Its test accuracy is 73.10% and its macro-AUC is 0.9165 which are slightly lower than InceptionV3. At the same time, it uses only 3.4 million parameters and requires just 14 MB of storage. This reduces training time to about 55 minutes and lowers memory demand. These characteristics make MobileNetV2 a practical choice when deployment limits are important. The small drop in test accuracy compared to InceptionV3 also suggests that large savings in computation can be achieved with only a modest loss in predictive strength.

ResNet50 reaches results that are close in accuracy with a macro AUC of 0.9140 and the same test accuracy as MobileNetV2. This performance comes at a higher cost. The model includes 23 million parameters and needs 98 MB of storage which increases both training time and memory usage. For this dataset the deeper residual structure does not provide a clear advantage that would justify the added computational load. EfficientNet B0 follows a more compact design with 5.3 million parameters and a model size of 20 MB. Even with this efficiency, its test accuracy of 65.68% and macro AUC of 0.8338 are clearly lower than those of InceptionV3 MobileNetV2 and

ResNet50. This suggests that its representational capacity is not sufficient to capture the subtle structural differences between nearby Alzheimer stages in this dataset.
The U-Net encoder-based classifier performs weakest across all metrics. It has the largest parameter count at 31 million and the highest storage need at 120 MB yet it shows the lowest test accuracy at 52.07% and the lowest F1 score at 37.47%. These results suggest that architectures designed for pixel level segmentation do not transfer well to image level classification unless they are carefully redesigned for that task.

4.2. Best-Performing Models Analysis

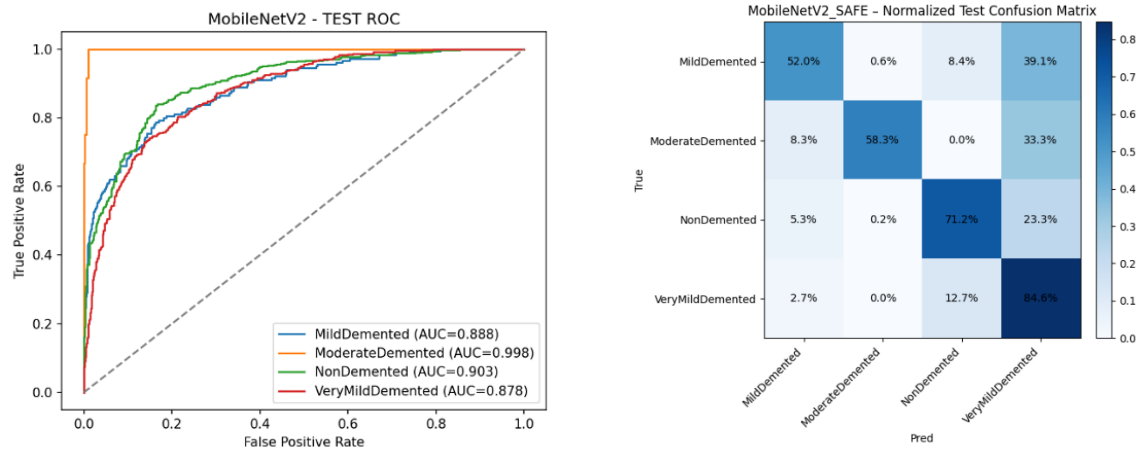
Based on the comparison in Table 3, InceptionV3 appears as the most effective model on the carried-out test set. Several architectures reach similar validation accuracy, but the differences become clearer when the focus shifts to test results and robustness related metrics like macro AUC and F1 score. InceptionV3 achieves the highest test accuracy at 75.29% along with the highest macro AUC at 0.9224 and the highest test F1 score at 76.41%. These results suggest a more balanced and reliable classification behavior across all dementia stages.
The ROC analysis for InceptionV3 shown in Figure 5-a helps explain this behavior in more detail. The class wise ROC curves show strong separation across all four categories. The Moderate Demented class reaches an AUC of 1.000 which points to near perfect separation from the other classes. The Non Demented class also shows strong separability with an AUC of 0.909 which indicates that the model captures MRI patterns linked to normal cognitive ageing. The very mild demented and mild demented classes show slightly lower AUC values at 0.888 and 0.893. This is expected because early-stage disease shows subtle anatomical changes and clear boundaries are harder to define.



(a) (b)

Figure 5. (a) ROC curves (b) confusion matrix of the InceptionV3 model on the test dataset.

The confusion matrix shown in Figure 5-b shows how the model behaves across classes and reveals error patterns that matter in a clinical sense. Correct predictions appear mainly on the diagonal entries. This is especially clear for the non demented and very mild demented groups where true positives are much higher than wrong predictions. Most errors happen between nearby clinical stages such as very mild demented and mild demented. Errors rarely appear between distant categories. This pattern matches the gradual and progressive nature of Alzheimer's disease. It suggests that many mistakes come from borderline clinical cases rather than clear weaknesses in the model. Confusion between Non Demented and Moderate Demented cases remains low, and this supports the model's ability to separate advanced pathology from normal aging.
Among the remaining models MobileNetV2 stands out as the most effective lightweight option. As shown in Table 3 it achieves a test accuracy of 73.10% and a macro AUC of 0.9165 and a test F1 score of 68.81%. These results are lower than those of InceptionV3 but they remain stable and competitive when the much lower model complexity is considered.
The ROC curves in Figure 6-a show very strong separability for the Moderate Demented class with an AUC of 0.998 and solid discrimination for the Non Demented class with an AUC of 0.903. Lower AUC values for the very mild demented at 0.878 and mild demented at 0.888 again point to the challenge of separating early disease stages.
The confusion matrix in Figure 6-b indicates that MobileNetV2 mainly confuses adjacent stages which mirrors the behavior observed for InceptionV3. Misclassification between Non Demented and Moderate Demented cases remains limited. This pattern supports the clinical plausibility of the predictions and suggests that performance loss is concentrated around subtle stage boundaries rather than clear diagnostic errors.
From a computational point of view InceptionV3 is a high capacity model with about 22 million parameters and a model size of 92 MB as shown in Table 3. This footprint is clearly larger than lightweight models like MobileNetV2. The added capacity allows the network to model multi scale anatomical variations present in



(a) (b)

Figure 6. (a) ROC curves (b) confusion matrix of the MobileNetV2 model on the test dataset.

structural MRI data more effectively. The training time of roughly 100 minutes reflects this added complexity and highlights the tradeoff between performance and computational cost. When deployment is considered, InceptionV3 fits best in settings where diagnostic accuracy and robustness are more important than strict resource constraints, such as hospital servers or cloud-based clinical systems. More compact models like MobileNetV2 are better suited for edge-based or resource-limited environments, although this comes with a noticeable drop in classification performance. Overall, the results indicate that InceptionV3 achieved the highest numerical performance under the current experimental setting among the evaluated models, while MobileNetV2 remains a practical choice when efficiency is the primary concern.

5. CONCLUSION

This study presented a fair comparison of several deep learning architectures for classifying four clinical stages of Alzheimer's disease using structural MRI data. By analyzing prediction performance alongside computational characteristics, it is intended to identify models that are both diagnostically reliable and easy to deploy in practice. The experimental results show that InceptionV3 delivers the best overall performance with the highest macro-AUC and F1-score among all the compared models. Analysis of the ROC curves and confusion matrix indicates satisfactory discrimination between clinically neighboring classes rather than distant disease categories. These results imply that InceptionV3 captures multi-scale anatomical features that are connected to Alzheimer's progression. MobileNetV2 provides a complementary option for the deployment. Its accuracy rates are slightly lower than those of InceptionV3. However, its much smaller parameter count, reduced model size and shorter training time make it attractive for use in environments with limited resources, like mobile or edge-based healthcare systems. ResNet50 reaches similar accuracy levels, but it demands considerably higher computational resources without delivering a clear performance benefit. EfficientNet-B0 and the encoder-based U-Net classifier show lower effectiveness for multi-class Alzheimer stage classification under the experimental setup. According to these results, it is stated that the model choice for Alzheimer's disease diagnosis should not be limited to classification accuracy. Generalization capability on unseen data and practical deployment constraints should be considered together to maintain both clinical importance and technical feasibility. The findings support the use of transfer learning–based convolutional neural networks for MRI-based Alzheimer stage classification.

Future research will aim to extend the proposed framework to larger and more diverse multi-center datasets to better examine generalization across different imaging protocols and demographic groups. Incorporating three dimensional MRI representations and additional modalities such as PET or functional MRI may allow a more complete characterization of neurodegenerative changes. Another important direction involves the integration of explainable artificial intelligence methods to improve model transparency and strengthen clinician trust which is essential for real-world clinical adoption.

Author's Contributions

Aslıhan Güven: conceptualization, methodology, investigation. **Abdurrahim Hüseyin Ezirmik:** technical support, writing assistance. **Mustafa Furkan Ceylan:** writing, editing, manuscript organization. **Fatih Aydın:** supervision, scientific guidance. All authors reviewed and approved the final manuscript.

Conflict of Interest

Author declares that he has no conflicts of interest.

REFERENCES

[1] M. Prince, A. Wimo, M. Guerchet, G.-C. Ali, Y.-T. Wu, and M. Prina, "World Alzheimer Report 2015: The Global Impact of Dementia," Alzheimer's Disease International, 2015.

[2] World Health Organization, "Global status report on the public health response to dementia," World Health Organization, Geneva, 2021. Accessed: Jan. 17, 2026. [Online]. Available: https://www.who.int/health-topics/dementia

[3] P. Scheltens et al., "Alzheimer's disease," The Lancet, vol. 397, no. 10284, pp. 1577–1590, 2021.

[4] B. Dubois et al., "Preclinical Alzheimer's disease: Definition, natural history, and diagnostic criteria," Alzheimer's & Dementia, vol. 12, no. 3, pp. 292–323, 2016.

[5] C. R. Jack et al., "NIA-AA Research Framework: Toward a biological definition of Alzheimer's disease," Alzheimer's & Dementia, vol. 14, no. 4, pp. 535–562, 2018.

[6] E. Westman, A. Simmons, J. S. Muehlboeck, and C. Tunnard, "AddNeuroMed and ADNI: Similarities and differences in MRI analysis results for Alzheimer's disease," Neuroimage, vol. 58, no. 3, pp. 818–828, 2011.

[7] L. Sorensen, M. Nielsen, and A. D. N. Initiative, "Ensemble support vector machine classification of dementia using structural MRI and mini-mental state examination," J. Neurosci. Methods, vol. 302, pp. 66–74, 2018.

[8] G. Litjens et al., "A survey on deep learning in medical image analysis," Med. Image Anal., vol. 42, pp. 60–88, 2017.

[9] D. Shen, G. Wu, and H. I. Suk, "Deep learning in medical image analysis," Annu. Rev. Biomed. Eng., vol. 19, pp. 221–248, 2017.

[10] H.-I. Suk, S.-W. Lee, D. Shen, and A. D. N. I. (ADNI), "Hierarchical feature representation and multimodal fusion with deep learning for AD/MCI diagnosis," Neuroimage, vol. 101, pp. 569–582, 2014.

[11] E. E. Bron et al., "Standardized evaluation of algorithms for computer-aided diagnosis of dementia based on structural MRI: The CADDementia challenge," Neuroimage, vol. 111, pp. 562–579, 2015.

[12] A. Abedalla, M. Abdullah, M. Al-Ayyoub, and E. Benkhelifa, "Chest X-ray pneumothorax segmentation using U-Net with EfficientNet and ResNet architectures," PeerJ Comput. Sci., vol. 7, pp. 1–36, Jun. 2021.

[13] A. Lundervold and A. S. Lundervold, "An overview of deep learning in medical imaging focusing on MRI," Z. Med. Phys., vol. 29, no. 2, pp. 102–127, 2019.

[14] H. Wen, Y. Li, W. Chen, S. Song, Y. Qiao, and X. Li, "Convolutional Neural Networks for Classification of Alzheimer's Disease: A Review," Neurocomputing, vol. 394, pp. 41–53, 2020.

[15] S. Vieira, W. H. L. Pinaya, and A. Mechelli, "Using deep learning to investigate the neuroimaging correlates of psychiatric and neurological disorders: Methods and applications," Neurosci. Biobehav. Rev., vol. 74, pp. 58–75, 2017.

[16] M. B. Narnaware and T. A. Mulla, "Early detection of Multi-Class Alzheimer's disease using Hybrid capsule auto encoder model," Biomed. Signal Process. Control, vol. 110, no. 5, p. 108330, Dec. 2025.

[17] S. K. Lakshmanan et al., "Effective deep convolutional neural network with attention mechanism for Alzheimer disease classification," Frontiers in Radiology, vol. 5, p. 1698760, Jan. 2026.

[18] S. Kumar, S. Shastri, V. Mansotra, and R. Salgotra, "MRI neuroimaging-based Alzheimer's disease stage classification using deep neural network with convolutional block attention module and GAN-style noise injection," Scientific Reports 2026 16:1, vol. 16, no. 1, pp. 6946-, Feb. 2026.

[19] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778, 2016.

[20] V. Muralidharan et al., "A scoping review of reporting gaps in FDA-approved AI medical devices," NPJ Digit. Med., vol. 7, no. 1, p. 273, Dec. 2024.

[21] US Food and Drug Administration, "Artificial Intelligence and Machine Learning (AI/ML)-Enabled Medical Devices: List of Cleared or Approved Devices," FDA, 2023. [Online]. Available: https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-aiml-enabled-medical-devices

[22] G. B. Frisoni, N. C. Fox, C. R. Jack, P. Scheltens, and P. M. Thompson, "The Clinical Use of Structural MRI in Alzheimer Disease," Nat. Rev. Neurol., vol. 6, no. 2, pp. 67–77, 2010.

[23] E. Topol, Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again. Basic Books, 2019.

[24] D. S. Marcus, T. H. Wang, J. Parker, J. G. Csernansky, J. C. Morris, and R. L. Buckner, "Open Access Series of Imaging Studies (OASIS): Cross-sectional MRI Data in Young, Middle Aged, Nondemented, and Demented Older Adults," J. Cogn. Neurosci., vol. 19, no. 9, pp. 1498–1507, Sep. 2007.

[25] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning," in Proceedings of the AAAI Conference on Artificial Intelligence, 2016. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/11231

[26] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in Proceedings of the 36th International Conference on Machine Learning (ICML), 2019.

[27] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted Residuals and Linear Bottlenecks," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 4510–4520, 2018.

[28] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in Medical Image Computing and Computer-Assisted Intervention (MICCAI), Springer, Cham, 2015, pp. 234–241, 2015.

[29] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal Loss for Dense Object Detection," in Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2980–2988, 2017.