

A DUAL-BACKBONE SWIN-TRANSFORMER AND EFFICIENTNET FRAMEWORK FOR ACCURATE ALZHEIMER'S DISEASE STAGE PREDICTION

¹M K V ANVESH, ²PRAJNA BODAPATI

¹Research Scholar, Department of Computer Science and Systems Engineering, Andhra University, A.P, India.

²Professor, Department of Computer Science and Systems Engineering, Andhra University, A.P, India.

E-mail: ¹mkvanvesh@gmail.com, ²prof.bprajna@andhrauniversity.edu.in

ABSTRACT

Alzheimer's disease (AD) is a serious health problem throughout the world that gradually damages memory and thinking abilities in human beings. Identifying the precise stage of Alzheimer's disease (AD) is a complex task for clinicians, as the disorder progresses through stages such as Non-Demented, Very Mild Demented, Mild Demented, and Moderate Demented. However, current diagnostic tools often struggle to correctly sort patients into these groups, these kind of problems frequently lead to wrong or late diagnoses. To address these challenges, this study proposes a dual-backbone deep learning framework that integrates Swin Transformer with EfficientNet-B7. The model was trained using the Alzheimer's Synthesized Dataset, which provides 8,000 pseudoRGB MRI images (roughly 2,000 images per disease stage). All images were prepared by being resized to pixels, put through various augmentations (changes), and then normalized. The model itself uses two specialized components: the Swin Transformer to pick up large-scale brain patterns, and EfficientNet-B7 to focus on fine, small local details. The special features from the two models are merged and passed to a single classifier, which makes the final, accurate prediction across the four disease classes. The model showed best results, achieved an accuracy of 99% and strong F1-scores after being trained with an 85% training and 15% validation data split. Further, Grad-CAM heatmaps are used to provide clear explanations and confirm that the results are reliable. This combination of high performance and clarity makes the overall system a promising and dependable method for quick Alzheimer's diagnosis.

Keywords: Alzheimer's Disease Classification, *Swin Transformer*, *EfficientNet-B7*, *pseudoRGB MRI*, Dual-Backbone Deep Learning, *Grad-CAM*.

1. INTRODUCTION

The brain is an essential organ that governs our memories, emotions, ideas, and general day-to-day activities [1]. Alzheimer's disease (AD) is a neurodegenerative disorder that slowly impairs memory and thinking, making daily activities harder over time and eventually leading from mild forgetfulness to severe dementia and complete dependence, mainly in people over 65. [2, 3]. Today, around 6.9 million Americans aged 65 and older live with AD, a number expected to double by 2060. Death rates from Alzheimer's disease have increased significantly over the past 20 years, making it one of the major causes of death for older persons. Dementia care costs hundreds of billions of dollars each year, and a shortage of trained healthcare workers emphasizes the urgent need for improved care and support systems [4].

Most common stage of Alzheimer's disease is dementia, which primarily affects older adults and is defined by a progressive loss of memory, cognitive function, and everyday functioning. The progression of Alzheimer's dementia usually occurs in stages: early or preclinical, when there may be no symptoms or only very mild memory lapses; mild cognitive impairment (MCI), where memory problems become more noticeable but daily activities are still manageable; mild dementia, when communication, recent memories, and familiar tasks become difficult, often accompanied by mood swings and confusion; moderate dementia, when more help is required as the ability to recognize loved ones diminishes and behavioral changes intensify; and severe dementia, when people become completely dependent on others for care, losing the ability to

recognize and communicate, even close family members [5].

In the below Figure 1 clearly shows the difference between the healthy and Alzheimer's affected brain. In AD, the brain undergoes shrinkage, loses moisture, and develops enlarged ventricles, distinguishing it from a healthy brain and indicating significant neuronal and tissue loss.

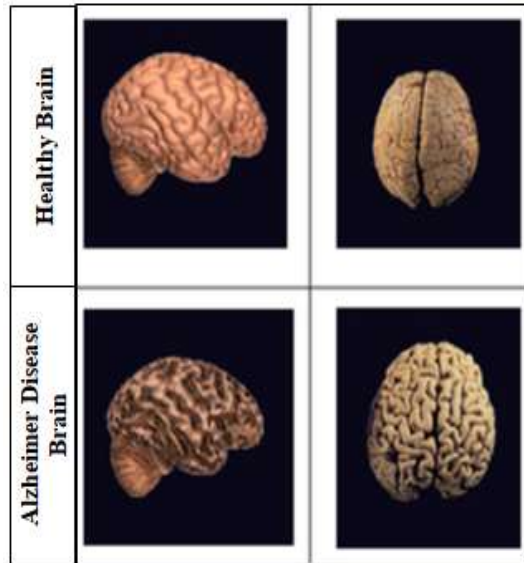


Figure 1: Healthy Brain vs Alzheimer's Brain

Numerous imaging methods are there for the identification and prediction of the AD which includes MRI, PET, and CT scans [6]. Structural MRI detects cortical and hippocampal atrophy and the PET images are used to detect the hypometabolism in temporoparietal cortex and the amyloid accumulation [7]. Accurately identifying brain soft tissues and differentiating them from healthy tissues is crucial for early detection of dementia and AD, but manual analysis of MRI scans or medical records is time-consuming and prone to errors due to the subtle similarities among the affected and healthy tissues [8].

It is challenging to correctly identify the stage of Alzheimer's disease from MRI scans because the early stages are very similar, and the changes to the brain are small. Most current methods, including CNN, transfer learning, and Transformers, fail to capture both the overall brain structure and the minute details needed for accurate prediction, specifically when the dataset is limited or imbalanced. This results in the increasing need for a stronger and more reliable approach. To

address this challenge, our work proposes a dual-backbone model that combines Swin Transformer with EfficientNet-B7 for the learning of both broad and fine features; a balanced pseudo-RGB dataset that helps avoid bias; and improves four-stage Alzheimer's classification supported by clear visual explanations through Grad-CAM.

2. LITERATURE SURVEY

Saratxaga et al. (2021) presented deep learning models to classify individuals with Alzheimer's disease using OASIS MRI data. Their process involved using various preprocessing techniques, including skull stripping and normalization, as well as developing custom 2D/3D convolutional neural networks (CNNs) and architectures for transfer learning, which resulted in obtaining high scores for both binary detection of AD (0.93 balanced accuracy) and three-stage classification of AD (0.88). However, the small size of the dataset, the recurrence of some subjects, and the extreme imbalance among classes created limitations for the overall generalizability of the models. These results show that better feature extraction methods and stronger multidimensional techniques are still needed to improve the accuracy and overall performance of AD diagnosis [9]. Kwok Tai Chui and colleagues devised a strong framework for detecting Alzheimer's disease based on MRI scans using GANs, CNNs, and transfer learning to achieve higher accuracy and alleviate issues of imbalance in class sizes. Initially trained on OASIS-1, OASIS-2, and OASIS-3 datasets, this method takes advantage of augmentation from fake MRI scans created by GANs and learned behaviors from other datasets. This resulted in gains in inaccuracy of 2.85 - 3.88% through augmentation and 2.43 - 2.66% due to knowledge transfer. The method still has some limitations because it relies heavily on synthetic data produced by GANs and differences between the three different OASIS datasets and will not generalize well to other forms of MRs for which no training has been done, therefore indicating that there is a need for improved/generalized feature extraction techniques for multi-stage AD classification [10].

Yagis et al. have proposed a 3D CNN model for the diagnosis of Alzheimer's using T1-weighted structural MRI to maintain the whole volumetric information that is lost in 2D slice-based methods. The performance of their customized 3D VGG-like network, evaluated on ADNI and OASIS datasets with 5-fold cross-validation, reached an

accuracy of 73.4% for ADNI and 69.9% for the OASIS dataset. Although outperforming 2D models with similar architectures, this approach was limited by relatively lower accuracy and high computational cost, with weak cross-dataset generalization, raising the need for efficient and robust feature-learning methodologies for multi-stage AD classification [11]. A deep learning framework for the identification of Alzheimer's disease from MRI images has been published by Pradhan et al. The implementation of their framework was based on the application of pre-trained CNNs such as VGG19 and DenseNet169 to data contained in the Kaggle repository and consists of patients with four different stages of AD. This study produced a model with good accuracy (82-88%) for the classification of AD into multiple classes. Unfortunately, the authors used a small and unbalanced dataset, which makes their findings less generalizable. In order to ensure proper diagnosis of a patient at different stages of AD, the development of better feature extraction techniques is needed. [12].

Natmpakis et al., proposed a framework using deep learning for binary classification of binary dementia diagnosis by using 3D structural MRI scans of elderly individuals and evaluated on OASIS-1/OASIS-2 and ADNI datasets to ensure both performance and generalizability. The methodology starts with slice selection technique and extracts mostly 140 informative slices per subject based on brain-region relevance and edge density, and reduces the noise from irrelevant regions. The regions of the features that the model uses are explained via explainable AI tools. The ensemble method attained 94.12% accuracy on ADNI dataset [13]. Battineni et al., implemented a deep CNN architecture for multiclass classification of AD using structural MRI scans from OASIS-3 dataset which comprises of 2,168 images across stages from very mild cognitive impairment to dementia. The model is trained using the dataset of 70% training data and tested using 10-fold cross validation and achieves 83.3% accuracy and it is outperformed on traditional classifiers like support vector machine and logistic regression. The CNN is utilized the automatic feature learning and eliminates the handcrafted feature engineering and achieves best accuracy [14].

ChakraBortty et al., proposed an ensemble framework emerging Vision transformers (ViT-B16) with three CNN architectures at the cutting edge i.e. ResNet152V2, VGG19, and

EfficientNetV2B3 used for recognizing the AD by using the OASIS dataset's structural MRI images. The ensemble models uses weighted-average strategy optimized via the grasshopper optimization algorithm. The model achieves 97.31% accuracy and outperformed the prior benchmark approaches. The study also highlights the benefit of combination of ViTs and CNN and accompanies both long-range self-attention and convolutional feature extraction and introduces the ensemble optimization [15]. Saracoglu and acilar proposed AD Net, a convolutional neural network which specifically designed for diagnosing AD with the OASIS-1 MRI dataset. The model built three different datasets by extraction slices from MRI scans which are taken from first, middle and third quarters of 128 available slices to identify the plane contained the most diagnostic relevant information. The data is divided in 8:2 ratios and the model is validated by using three iterations of five-fold cross-validation and achieves the accuracy of 97.05%. The study tells that ADNet is particularly applied to the most informative sagittal slices and used for transfer learning [16].

Rajendiran et al., observed different multiple transfer learning models for the effectiveness of those models which are used for the identifying AD from MRI images and it presents the findings in the international journal of health sciences. In this author compared several pretrained CNN models like VGG-16, Inception and AlexNet used for the multiclass setup to distinguish between normal cognition and AD using MRI scans. The workflow includes careful MRI preprocessing and layer wise fine-tuning and incorporates at feature level classifiers which give better results [17]. Hussian et al., proposed a custom 12-layer CNNs for AD classification of patents and healthy ones and controls using T1 weighted MRI scans collected from OASIS-1 dataset. The model is compared with various pretrained architectures like Inceptionv3, MobileNetV2, Xception, and VGG. The proposed architecture achieves 97.5% accuracy and outperforms over many baseline models on the same dataset. The study demonstrates the careful design of the lightweight CNN can also outperform than heavier pre trained networks in MRI-based Binary classification of AD [18].

Basheer et al., proposed a model using deep neural network to recognize dementia stages using MRI images which are collected from OASIS dataset and the framework integrates feature

engineering and Some methods for reducing dimensionality like PCA to improve model performance by decreasing overfitting. The model achieves 94.44% accuracy and outperforms over traditional machine learning classifiers. This research also highlights the effectiveness of combining clinical features with deep learning dementia classification [19]. Salami et al., developed a clinical decision support system that uses the OASIS-3 dataset to diagnose AD which contains longitudinal neuro-imaging and cognitive data. The model used a combination of preprocessing techniques and feature selection process along with the machine learning classifiers to build a robust diagnostic pipeline. The model is explained by using SHAP which is an Explainable AI which highlights the feature importance in decision making. The classifiers tested using Random Forest and support Vector machine which shows superior performance. The model achieves 90% classification accuracy [20].

Alsan and Ozupak compares and gives us an comprehensive review of different machine learning algorithms for the automatic detection of AD using structural neuroimaging data and cognitive scores. The study explains various classifiers Random Forest, Support vector machine (SVM), KNN, gradient boosting machines on benchmark datasets. The random forest emerging top performance by achieving 91.3% accuracy. The paper explains the importance of feature selection and dimensionality reduction techniques. It also highlights the ML models behavior under various typed of data and underscores the potential of ensemble methods in AD prediction task [21]. Garg et al., uses OASIS MRI data to present a deep learning method for AD detection. The model consists of 19 convolutional neural networks which is designed to accurately differentiate between Alzheimer's and non-Alzheimer's cases by extracting hierarchical features from brain MRI images. The architecture contains multiple pooling and convolutional layers which are followed by fully connected layers. This 19-layered architecture achieves high accuracy than the traditional ones. This paper explains the potential of the deep CNN architecture for the automatic AD diagnosis from neuroimaging data [22].

Swain et al., introduces a deep learning model that uses enhanced MRI data from the OASIS-1 dataset to identify AD. The writers approach uses the deep CNN and trained on a dataset and it is expanded through methods for data

augmentation that decrease overfitting and increase the generalization. The architecture contains several max-pooling layers and convolutional layers with Relu activation function and dropout regularization. Model achieves high accuracy when compared to non-augmented baseline models [23]. Puente-castro et al., proposed an automatic system using deep learning techniques based on structural MRI data for diagnosis of AD by. The author implements and compares various convolutional neural network architectures which include 3D and 2D variants to identify the relation between the changes in brain scans. The feature extraction method is accompanied to increase the capability of the CNN architectures to capture the spatial patterns associated with disease progression. The model is evaluated using different publicly available datasets and demonstrates strong performance. The paper highlights the potential for the clinical adoption, especially when it paired with Explainable AI components for result interpretation [24].

Zhou et al. (2025) created a deep learning model to detect Alzheimer's disease at an early stage using 3D MRI scans. This study utilized a 3D Convolutional Neural Network (3D-CNN) to analyze the complete 3D brain structure, allowing the model to automatically detect Alzheimer's indicators with high accuracy, which demonstrated the strong potential of deep learning for AD diagnosis. However, this study faced several limitations, including uneven image distribution across the different Alzheimer's disease (AD) stages, significant difficulty in interpreting the model's decisions (a 'black box' problem), and the potential for the model to fail when applied to data from different patient populations or various MRI equipment [25].

3. RESEARCH METHOD AND EXECUTION

3.1 Experimental Setup

The proposed model was developed using the Python, with PyTorch used as the main deep learning framework and additional support from libraries such as timm, torchvision, NumPy, and scikit-learn for model building, preprocessing, and testing. The main modules implemented for this study include the data preprocessing module, the dual-backbone feature extraction module integrating Swin Transformer and EfficientNet-B7, the classification module, and the training-validation loop with mixed precision support. All experiments are run on the Kaggle Notebook

platform, Python 3.10, and GPU support. The system used an NVIDIA Tesla T4 GPU with 16 GB VRAM, along with powerful cloud CPUs and more than 13 GB of RAM. This setup offered the required computational power to train both transformer and CNN models effectively and with ease.

3.2 Dataset Description

For this work, the Alzheimer's Synthesized Dataset was used, which was downloaded from Kaggle. The dataset contains 8,000 brain images in pseudo-RGB form, which were derived from MRI scans. There are four folders in the dataset, and each folder contains 2,000 images for a different stages of Alzheimer's disease. These images are converted into a three-channel format to mimic color images, making them suitable for CNN and Transformer models. This is done by turning grayscale MRI slices into three-channel images, possibly by copying or adjusting intensity values across channels, allowing the use of models designed for regular color images. To fix the common problem of uneven class sizes in Alzheimer's datasets, this version uses techniques like rotating, flipping, and adding noise to the images to create a balanced set. This balance helps ensure fair training and testing of the model, avoiding bias toward any class. The images are in PNG format and have different original sizes but are made uniform through preprocessing for model use. Pseudo-RGB images change single-channel data into three channels, making the features clearer and easier for deep learning models to understand [26].

3.3 Preprocessing

Properly preparing the dataset is vital for keeping it organized and improving the model's performance. The dataset is splitted into two parts: 85% for training the model and 15% for checking how well it works (validation). This made sure that each class was evenly represented in both parts. To help the model learn better and make accurate predictions on new data, several changes applied to the training images. First, all images were resized to 224×224 pixels for consistency. Then, some images were randomly flipped horizontally and rotated slightly—up to 10 degrees—to increase diversity. Also made small tweaks to brightness and contrast so the model could handle different lighting conditions. For the validation set, only resizing, tensor conversion, and the same

normalization were applied to test the model on unchanged images. The training images were randomly shuffled to avoid the model picking up on unhelpful patterns. This process ensured the images were properly formatted, varied, and ready for neural network training.

4. METHODOLOGY

4.1 Proposed Models

The proposed model uses two advanced backbone models: the Swin Transformer (Small version) and EfficientNet-B7. Among the EfficientNet family (B0–B7), EfficientNet-B7 achieved the best performance with less number of training epochs. These were chosen because they work well together—the Swin Transformer is great at understanding large-scale patterns across the image using its layered attention system, while EfficientNet-B7 is highly effective at picking up fine details thanks to its smart convolutional structure.

4.1.1 Swin Transformers

The below Figure 2 shows the Swin Transformer, developed by Liu and colleagues in 2021, marks a major change from traditional CNNs by using a layered Transformer design with shifted windows, enabling efficient computation on high-resolution images. Unlike standard Vision Transformers (ViTs), which apply global self-attention, the Swin Transformer computes self-attention locally within non-overlapping windows, using a shifted window approach to allow interactions across adjacent windows while reducing computational complexity from quadratic($O(N^2)$) to linear ($O(N)$) relative to the image size N . The Swin-Small variant is designed for 224×224 input images. Initially, the RGB image of size (3 X 224 X 224) is divided into non-overlapping 4×4 patches, resulting in $(224/4)^2 = 3,136$ patches, each of 48 dimensions ($4 \times 4 \times 3$), which are linearly projected to an embedding dimension of 96, forming a feature map of $56 \times 56 \times 96$.

The architecture consists of four hierarchical stages. Stage 1 has 2 Swin Transformer blocks with embedding dimension 96 and 3 attention heads, producing an output of $56 \times 56 \times 96$. Stage 2 has 2 layers, each using 192 features and 6 attention heads. It begins by combining image patches, cutting the size in half, which gives a feature map of $28 \times 28 \times 192$. Stage 3 includes 18 layers with

384 features and 12 heads, and again reduces the size, resulting in a $14 \times 14 \times 384$ output. Stage 4 has 2 layers with 768 features and 24 heads, and after merging patches, it produces a final feature map of $7 \times 7 \times 768$. Within each block, multi-head self-attention (MSA) is computed in local windows of size 7×7 , and shifted window MSA (SW-MSA) alternates by shifting windows by $(\text{window_size}/2)$ in both dimensions to facilitate cross-window information flow. Each attention block is followed by layer normalization, residual connections, and a 2-layer MLP with GELU activation. After the final stage, average pooling reduces the feature map to a 768-dimensional vector. The Swin-Small model contains approximately 28 million parameters and requires about 4.5G FLOPs for 224×224 input images. This design works well for medical image classification because it can effectively understand both detailed and wide-ranging patterns. That makes it especially useful for spotting small changes in brain structure that may signal the development of Alzheimer's disease.

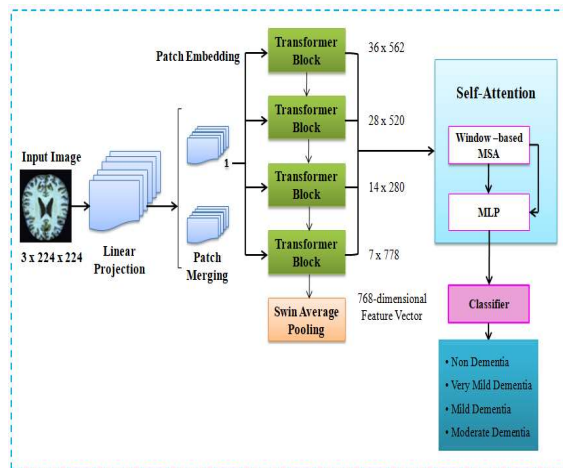


Figure 2: Architecture of Swin Transformer

4.1.2 Efficient Net B7

The Figure 3 represents the EfficientNet, introduced by Tan and Le in 2019, is a group of convolutional neural networks designed to balance accuracy and efficiency by scaling depth, width, and image resolution all at once[27]. EfficientNet-B7 is the biggest and most powerful version in this group, focusing on deeper layers and higher image quality for demanding tasks. Its structure is based on MBConv blocks, which are lightweight and effective, and it includes special components like squeeze-and-excitation modules and Swish activation functions to boost performance. The network starts with a stem layer that applies a 3×3

convolution using 64 filters with a stride of 2 on the input image sized $3 \times 224 \times 224$. This is followed by batch normalization and a Swish activation function, which reduces the output feature map to $112 \times 112 \times 64$. After this, the network passes through seven blocks, each increasing in complexity and depth. Block 1 includes 3 MBConv1 layers with 32 channels and a 3×3 kernel. This network structure details a progression across four blocks, all utilizing the MBConv6 layer type. Block 2 uses four layers with a 3×3 kernel, a stride of 2, and outputs 48 channels. Block 4 has 6 MBConv6 layers with 112 channels, using a 3×3 filter and stride 2 to downsample. Block 5 also has 6 MBConv6 layers, with 192 channels, a 5×5 filter, and stride 1 to keep spatial size. Block 6 includes 5 MBConv6 layers with 320 channels, a 5×5 filter, and stride 2 for further downsampling. Block 7 concludes with a single MBConv6 layer of 1280 channels, employing a 3×3 filter and a stride of 1.

Through these stages, the network progressively captures increasingly complex and detailed features. After that, another 1×1 convolution reduces the channels back to the original size, and if the input and output shapes match, a shortcut connection is added to preserve information. At the end of the network, a global average pooling layer creates a 2560-value feature vector (in the case of EfficientNet-B7), followed by dropout and a final linear layer for classification—though this last layer is usually removed when using the model as a feature extractor. EfficientNet-B7 is a powerful and computationally efficient model, boasting roughly 66 million parameters and requiring 37 billion operations. EfficientNet-B7 is especially good at detecting small details in images, complementing models like the Swin Transformer that focus on broader patterns. This makes it especially useful for identifying Alzheimer's-related changes in brain scans, such as tissue loss or enlarged ventricles.

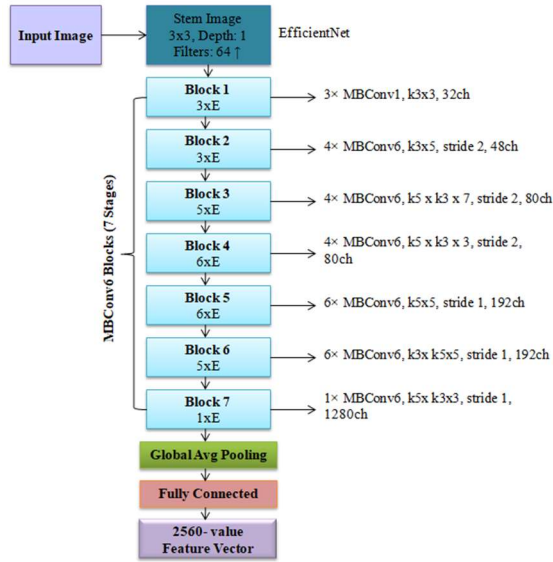


Figure 3: Architecture of EfficientNetB7

4.1.3 Proposed Model Swin-Efficient:

The proposed model that shown in Figure 4 explains the workflow which includes input to output complete working. This method classifies Alzheimer's disease by integrating features from two powerful neural networks: the Swin Transformer (Small) and EfficientNet-B7. Input images are processed by both backbones simultaneously to extract distinct feature representations. These individual features are then concatenated to form a combined vector, which is finally fed into a shared classifier head for prediction. By blending the Swin Transformer's hierarchical attention, which is effective at capturing broad, structural patterns, with the EfficientNet-B7's convolutional strength for identifying fine local details, this approach significantly enhances the model's ability to discriminate accurately between the Alzheimer's classes.

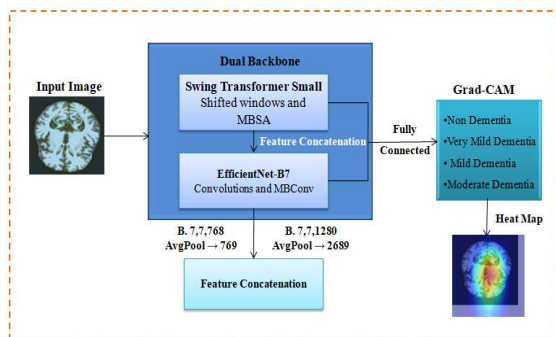


Figure 4: Architecture of Swin-Efficient

The model takes the input images and sends them to both the Swin Transformer and EfficientNet-B7 at the same time to get different kinds of information (features). These two sets of information are then merged into one combined data piece. This single piece of data is then used by a shared decision-maker to make the final prediction. This method achieves success by combining the Swin Transformer's ability to see large-scale, overall patterns with EfficientNet-B7's strength in detecting small, localized details, which leads directly to superior and more accurate final results. The image is first patch-embedded and then processed through four hierarchical stages, producing a feature map of '(B, 7, 7, 768)'. Global average pooling reduces this to a 768-dimensional vector per image. At the same time, the EfficientNet-B7 branch—pre-trained and stripped of its classification layer—processes the input image through its initial stem and seven MBConv blocks. This results in a feature map with dimensions '(B, 7, 7, 1280)', which is subsequently transformed into a 2560-dimensional vector via global pooling.

The feature vectors from the Swin Transformer and EfficientNet-B7 branches are joined along the feature axis, resulting in a unified representation of shape '(B, 3328)'. The model takes the combined feature vector (which is 3,328 dimensions long) and sends it to a simple, fully connected classifier. First, a linear layer shrinks this vector down to 512 dimensions. After the data's dimension is reduced, then immediately passes through a ReLU layer to introduce the necessary non-linear complexity for the model to learn effectively. This is instantly followed by a dropout layer, which randomly disables 50% of the neurons to prevent the model from overfitting (simply memorizing the training examples). In the final step, a second linear layer transforms the 512-dimensional output into 4 final logit values, where each value represents the model's score for one of the four possible Alzheimer's disease classes. Softmax can be applied post-hoc to obtain class probabilities, though raw logits are used during training with Cross-Entropy Loss.

The model is trained end-to-end, with both backbones unfrozen, using the AdamW optimizer (learning rate 2e-4, weight decay 1e-5) and a ReduceLROnPlateau scheduler (patience 3, factor 0.5) monitoring validation accuracy. Processing in parallel allows the Swin branch to capture shifted-window attention for multi-scale brain features,

such as cortical thinning, while EfficientNet captures fine-grained convolutional patterns, such as hippocampal volume changes. Feature maps evolve from the input $(3 \times 224 \times 224)$ through the Swin stages $(56 \times 56 \times 96 \rightarrow 28 \times 28 \times 192 \rightarrow 14 \times 14 \times 384 \rightarrow 7 \times 7 \times 768 \rightarrow \text{avg pool to } 768)$, which is then concatenated with EfficientNet outputs $(112 \times 112 \times 64 \rightarrow \dots \rightarrow 7 \times 7 \times 1280 \rightarrow \text{avg pool to } 2560)$, resulting in the final fused vector (3328) that flows through the classifier head.

To ensure the model's predictions are clear and easy to interpret, the Grad-CAM technique is used specifically on the EfficientNet part of the network. This tool creates a visual map showing which parts of the image the model focuses on when making its final classification. This technique highlights class-discriminative areas in the input image, assisting clinicians in understanding model decisions, for instance, by focusing on atrophy-prone regions of the brain. The custom `EffGradCAM` class registers forward and backward hooks on the target convolutional layer (the last Conv2d in EfficientNet-B7's final MBConv block) to capture high-level features. During inference, the Swin features are detached, and the input is forwarded through EfficientNet to extract activations and compute logits. Gradients of the predicted class with respect to the target layer are used to obtain channel-wise weights, which are multiplied with the activations, passed through ReLU, upsampled to the input size (224×224) , and normalized. The final step is to overlay the resulting heatmap (which shows the model's focus) onto the original image. The resulting heatmap, which indicates the areas the model focused on, is visually layered onto the original image. This is achieved using a 'jet' color gradient and set to be 50% transparent. This method clearly shows the critical areas, such as the ventricles, that led to the Alzheimer's classification. This explainability feature is necessary because that makes the model's decision process easy to understand.

4.2 Algorithm

Figure 5 presents the step-by-step procedure of the algorithm

- **Input Image:** The system takes input image which is a pseudo-RGB MRI scan sized $224 \times 224 \times 3$.
- **Preprocessing:** The image is resized to 224×224 , randomly flipped and rotated

while training, converted into a tensor, and normalized.

- **Swin Transformer Path:** The preprocessed image is split into small patches, processed with attention blocks, and turned into a global feature vector using average pooling.
- **EfficientNet-B7 Path:** The same image went by EfficientNet-B7 model, where convolution and SE modules extract deep features and create another pooled vector.
- **Feature Fusion:** Both feature vectors (from Swin and EfficientNet) are joined together into one combined representation. **Linear Layer:** The combined feature vector goes by a fully connected layer to shrink its size and get it ready for classification.
- **ReLU & Dropout:** The ReLU include non-linearity so the model can learn better, and Dropout is used to reduce overfitting and makes the model more general. **Classification Layer:** A final linear layer outputs four values, one for each Alzheimer's type.
- **Softmax:** These outputs are converted into probabilities, and the highest probability class is taken as the final result.
- **Grad-CAM:** After training, Grad-CAM is worked on EfficientNet-B7's last convolutional layer to show which areas are mostly the model was focused for prediction. Then highlight those area with heatmaps.
- **Final Output:** The system provides the predicted Alzheimer's class, the probability scores, and the Grad-CAM heatmap placed on the MRI image for easy understanding.

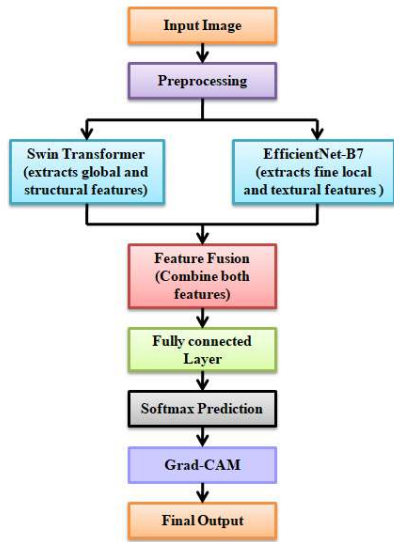


Figure 5: Flowchart of the proposed Swin Transformer + EfficientNet-B7 classification framework.

5. RESULTS& ANALYSIS

The synthetic Alzheimer's pseudo-RGB dataset was used for training and evaluating the proposed Swin Transformer model, which combines Swin Transformer and EfficientNet-B7. The validation set consists of 300 images for each of the four balanced classes. The balanced class distribution decreased bias in both model training and model evaluation by ensuring that no group was under-represented. This constant representation made possible to perform a thorough and fair evaluation across all stages of dementia, ensuring that no class had an excessive impact on the model's performance.

Figure 5 depicts the training behavior of the proposed Swin-Transformer model. Figure 5(a) presents the loss values for both training and validation sets. During the initial few epochs, the curves showed a steep decline before flattening and stabilizing. The validation loss stabilized at about 0.05 by the 18th epoch, remaining around the training loss. This suggests that the model achieved strong generalization with minimal overfitting. Figure 5(b) shows the accuracy curves. In the early epochs, both training and validation accuracy increased significantly, by the seventh epoch reached over 95%. The model constantly learnt to classify the MRI scans with high reliability, as shown by the accuracy values for both sets approaching 99% by its final epochs. Together, these results highlight the model's strong learning dynamics and its ability to extract discriminative

features without significant performance degradation.

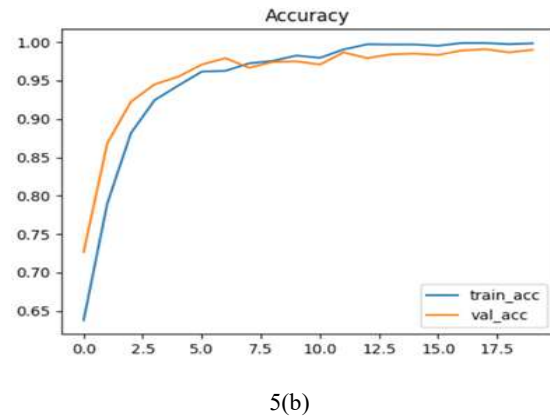
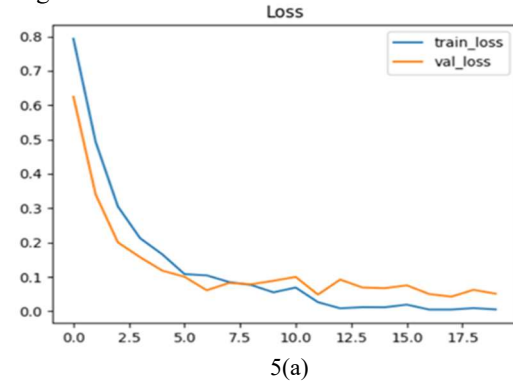


Figure 5: Training and Validation Loss and Accuracy of the Dual-Backbone Model

The confusion matrix in Figure 6 displays the model's predictions by class. The model achieved overall 99% accuracy in both the Mild Demented and Moderate Demented categories, properly classifying all samples. This indicates high ability to identify dementia in its intermediate and severe stages, when anatomical alterations in the brain are more significant. Few misclassifications were observed between Non Demented and Very Mild Demented classes, with 6 misclassified Non Demented cases and 5 misclassified Very Mild Demented cases. This overlap reflects the difficulty of early-stage Alzheimer's, when MRI differences are subtle and clinical diagnosis remains challenging even for human experts. Additional information about precision, recall, and F1-scores across classes can be found in the validation classification report. The model demonstrated high efficiency in classifying all dementia categories, achieving precision, recall, and F1-score values of 0.99.

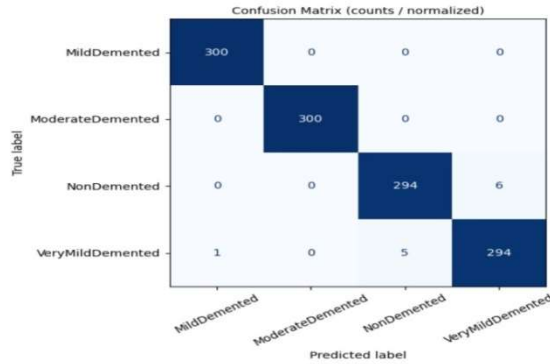
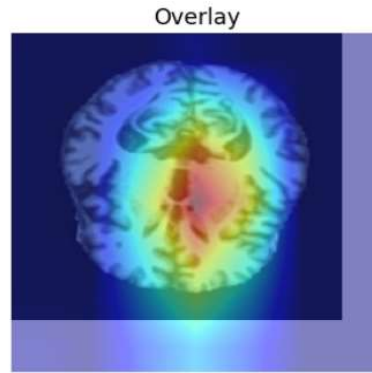


Figure 6: Confusion Matrix of the Proposed Model on Validation Data



7(b)

Figure 7: Grad-CAM Heatmaps for Alzheimer's Stages

Gradient-weighted Class Activation Mapping (Grad-CAM) was used to visualize class-specific regions in the input MRI scans in order to examine the Swin-transformer model's decision-making process. Figure 7(a) shows the original MRI brain scan without any model-based highlighting. This image shows the anatomical baseline that can be used to compare the model's attention. Figure 7(b) presents the Grad-CAM heatmap overlaid on the same scan. The red and yellow regions correspond to areas of high model activation. The overlay shows that the model focuses mostly on regions of the central brain, including the hippocampus and surrounding medial temporal lobes, which are considered to be indicators of the neurological causes of Alzheimer's disease. Red and yellow patches represent high activations, indicating that the model's predictions are focused on clinically relevant regions rather than irrelevant background features. The model's predictions are easier to understand and have more clinical value when they are in line with recognized biomarkers.

The model's strongest activations match anatomical regions known to be affected in the course of Alzheimer's disease, as confirmed by the overlay image. These visualizations enhance interpretability and strengthen the model's predictions' clinical credibility. The Swin-Transformer model showed excellent overall classification performance, achieving high accuracy across all phases of dementia. Although a minor confusion was observed between non-demented and very mildly demented cases, revealing the underlying difficulty of early-stage Alzheimer's detection, the flawless performance in the mildly and moderately demented groups demonstrates the model's effectiveness in capturing advanced neurodegenerative patterns. Importantly, the Swin-Transformer framework's high precision and recall scores across all classes confirm its effectiveness in detecting important clinical patterns, indicating its great potential as a decision-support tool for neuroimaging-based Alzheimer's diagnosis.



7(a)

Table 1: Accuracy comparison of different models for Alzheimer's disease detection.

Model	Accuracy
GoogleNet [28]	96.39%
AlexNet [28]	94.08%
Pre-trained VGG [29]	98.73%
3D-CNN [30]	98%
Proposed Swin-Transformer + EfficientNet-B7	99%

Table 1 compares the accuracy of some of the models. While models such as GoogleNet, AlexNet, VGG, and 3D-CNN present good results, within the range of 94% to 98.73%, none have reached the performance achieved with the proposed Swin Transformer + EfficientNet-B7 framework. Combining the global features from the Swin Transformer with fine-grained details extracted from EfficientNet-B7 enables our model to learn about both large structural changes and subtle local patterns in MRI scans. It thus becomes more capable of distinguishing stages of Alzheimer's. This leads it to achieve the highest accuracy of 99%, becoming the best-performing model in comparison with the others.

6. CONCLUSION

This study presents the dual-backbone deep learning framework that couples Swin Transformer with EfficientNet-B7, aiming to classify AD stages from pseudo-RGB MRI scans with high accuracy. Further, by combining the strengths of the Swin Transformer for capturing large-scale structural patterns with those of EfficientNet-B7 for extracting fine local details, the proposed model yielded 99% accuracy, while consistently high F1-scores confirmed the model's strong effectiveness in all AD stages, especially for mild and moderate demented cases. The integration of Grad-CAM further enhanced model interpretability and emphasized clinically important regions, including the hippocampus, validating the fact that the model is using meaningful anatomical features when making predictions. All these results together mean that the proposed system is a reliable, interpretable, and effective tool that could assist in diagnosing Alzheimer's disease; thus, the future direction would be to validate the model on real clinical datasets, perform multimodal imaging, improve early-stage detection, and enhance explainability using advanced XAI techniques.

REFERENCES:

- [1] P. Murala and K. N. Rao, "Multi-Class Brain Tumor Diagnosis Using a Vision Transformer with MRI Image Segmentation", *Eng. Technol. Appl. Sci. Res.*, vol. 15, no. 4, pp. 26120–26127, Aug. 2025.
- [2] Ghahnavieh, A.E., Luo, S., &Chiong, R. (2019). Deep learning to detect Alzheimer's disease from neuroimaging: A systematic literature review. *Computer methods and programs in biomedicine*, 187,105242 .
- [3] Shehri, W.A. (2022). ADdiagnosis and classification using deep learning techniques. *PeerJ Computer Science*, 8.
- [4] 2024 Alzheimer's disease facts and figures. *Alzheimers Dement.* 2024 May;20(5):3708-3821. doi: 10.1002/alz.13809. Epub 2024 Apr 30. PMID: 38689398; PMCID: PMC11095490.
- [5] El-latif, A.A., Chelloug, S.A., Alabdulhafith, M., &Hammad, M. (2023). Accurate Detection of ADUsing Lightweight Deep Learning Model on MRI Data. *Diagnostics*, 13.
- [6] B, P., Balaji, P., Chaurasia, M.A., Bilfaqih, S.M., Muniasamy, A., &Alsid, L.E. (2023). Hybridized Deep Learning Approach for Detecting Alzheimer's Disease. *Biomedicines*, 11.
- [7] S. G. Mueller, M. W. Weiner, L. J. Thal, R. C. Petersen, C. R. Jack, W. Jagust, J. Q. Trojanowski, A. W. Toga, and L. Beckett, "Ways toward an early diagnosis in Alzheimer's disease: The ADneuroimaging initiative (ADNI)," *Alzheimer's Dementia*, vol. 1, no. 1, pp. 55–66, Jul. 2005.
- [8] Mohammed, B.A., Senan, E.M., Rassem, T.H., Makbol, N.M., Alanazi, A.A., Al-Mekhlafi, Z.G., Almurayziq, T.S., &Ghaleb, F.A. (2021). Multi-Method Analysis of Medical Records and MRI Images for Early Diagnosis of Dementia and ADBased on Deep Learning and Hybrid Methods. *Electronics*.
- [9] Saratxaga, C.L., Moya, I., Picón, A., Acosta, M., Moreno-Fernandez-de-Leceta, A., Garrote, E., &Bereciartúa-Pérez, A. (2021). MRI Deep Learning-Based Solution for ADPrediction. *Journal of Personalized Medicine*, 11.
- [10] Chui, K.T., Gupta, B.B., Alhalabi, W.S., &Alzahrani, F.S. (2022). An MRI Scans-Based ADDetection via Convolutional Neural Network and Transfer Learning. *Diagnostics*, 12.
- [11] Yagis, E., Citi, L., Diciotti, S., Marzi, C., Atnafu, S.W., & Herrera, A.G. (2020). 3D Convolutional Neural Networks for Diagnosis of Alzheimer's Disease via Structural MRI. 2020 IEEE 33rd International Symposium on Computer-Based Medical Systems (CBMS), 65-70.
- [12] Pradhan, A., Gige, J., &Eliazzer, M. (2021). Detection of Alzheimer's Disease (AD) in MRI Images using Deep Learning.
- [13] Ntampakis, N., Diamantaras, K., Chouvarda, I., Argyriou, V., &Sarigiannidis, P. (2024). Enhanced Deep Learning Methodologies and MRI Selection Techniques for Dementia

- Diagnosis in the Elderly Population. ArXiv, abs/2407.17324.
- [14] Battineni, G., Chintalapudi, N., Amenta, F., & Traini, E. (2021). Deep Learning Type Convolution Neural Network Architecture for Multiclass Classification of Alzheimer's Disease. *Bioimaging* (Bristol. Print).
- [15] Chakra Bortty, J., Chakraborty, G.S., Noman, I.R., Batra, S., Das, J., Bishnu, K.K., Tarafder, M.T., & Islam, A. (2025). A Novel Diagnostic Framework with an Optimized Ensemble of Vision Transformers and Convolutional Neural Networks for Enhanced ADDetection in Medical Imaging. *Diagnostics*, 15.
- [16] Saraçoğlu, A.S., Acılar, A.M., & ErdaşÇiçek, Ö. (2025). ADNet: A CNN MODEL FOR ALZHEIMER'S DISEASE DIAGNOSIS ON OASIS-1 DATASET. *KahramanmaraşSütçü İmam ÜniversitesiMühendislikBilimleriDergisi*.
- [17] Rajendiran, M., Kumar, D.K., Anu, D.S., & Nair, H. (2022). Detection of AD in MRI images using different transfer learning models and improving the classification accuracy. *International journal of health sciences*.
- [18] E. Hussain, M. Hasan, S. Z. Hassan, T. Hassan Azmi, M. A. Rahman and M. ZavidParvez, "Deep Learning Based Binary Classification for ADDetection using Brain MRI Images," 2020 15th IEEE Conference on Industrial Electronics and Applications (ICIEA), Kristiansand, Norway, 2020, pp. 1115-1120, doi: 10.1109/ICIEA48937.2020.9248213.
- [19] S. Basheer, S. Bhatia and S. B. Sakri, "Computational Modeling of Dementia Prediction Using Deep Neural Network: Analysis on OASIS Dataset," in *IEEE Access*, vol. 9, pp. 42449-42462, 2021, doi: 10.1109/ACCESS.2021.3066213.
- [20] Salami, F., Bozorgi-Amiri, A., Hassan, G. M., Tavakkoli-Moghaddam, R., & Datta, A. (2022). Designing a clinical decision support system for Alzheimer's diagnosis on OASIS-3 data set. *Biomedical Signal Processing and Control*, 74, 103527.
- [21] Aslan, E., & Özüpak, Y. (2025). Comparison of machine learning algorithms for automatic prediction of Alzheimer disease. *Journal of the Chinese Medical Association*, 88(2), 98-107.
- [22] G. Garg, C. Prabha, B. Ahuja, R. Singh and A. Agarwal, "An In-Depth Study of Alzheimer's Detection: Leveraging OASIS MRI with a 19-Layer CNN," 2024 International Conference on Expert Clouds and Applications (ICOECA), Bengaluru, India, 2024, pp. 758-763, doi: 10.1109/ICOECA62351.2024.00136.
- [23] B. K. Swain, R. Abhisika, S. K. Rout, S. Mohapatra, S. Hasan and M. Mishra, "Enhancing Alzheimer's disease Diagnosis with Augmented OASIS-1 MRI Data: A Deep Convolutional Neural Network Approach," 2024 11th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), Noida, India, 2024, pp. 1-5, doi: 10.1109/ICRITO61523.2024.10522302.
- [24] Puente-Castro, A., Fernandez-Blanco, E., Pazos, A., & Munteanu, C. R. (2020). Automatic assessment of AD diagnosis based on deep learning techniques. *Computers in biology and medicine*, 120, 103764.
- [25] Zhou, J., et al. (2025). A deep learning model for early diagnosis of Alzheimer's disease using 3D MRI. *Scientific Reports*, 15(1), 1-11.
- [26] Ayon, A. M. (2024). Alzheimer's Synthesized Dataset. *Kaggle*. <https://www.kaggle.com/datasets/masud1901/alzheimers-synthesized-dataset>.
- [27] M. Tan and Q. V. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," arXiv preprint arXiv:1905.11946, 2020.
- [28] 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, 103217.
- [29] 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.2021.01.008>.
- [30] Basaia, S., Agosta, F., Wagner, L., Canu, E., Magnani, G., Santangelo, R., & Filippi, M. (2019). Automated classification of Alzheimer's disease and mild cognitive impairment using a single MRI and deep neural networks. *NeuroImage: Clinical*, 21, 101645. <https://doi.org/10.1016/j.nicl.2018.101645>.
- [31]