# Diagnosis of Alzheimer's Disease Using 3D and 2D Convolutional Neural Networks

Siya Goel
Stanford University
siyagoel@stanford.edu

Nikhil Sharma
Stanford University
sharmnik@stanford.edu

Zhi Zheng
Stanford University
zzheng88@stanford.edu

## Abstract

*Alzheimer's Disease (AD), the most common type of dementia, significantly impacts millions and complicates early diagnosis. This project employs advanced 2D and 3D convolutional neural networks (CNNs) to enhance early AD detection by analyzing MRI images from the OASIS-1 dataset. We evaluate these models alongside benchmarks such as BrainNet2D and BrainNet3D, focusing on accurate AD stage detection. Initial findings indicate that models pretrained on ImageNet, like InceptionV3 and VGG16, surpass custom models including BrainNet2D. Hybrid CNN-LSTM models and innovative 3D-CNNs like 3D-DenseNet and modified BrainNet3D demonstrate strong potential in detecting all AD stages. Future efforts will concentrate on refining models and expanding datasets to improve diagnostic precision and address MRI data analysis challenges.*

## 1. Introduction

Alzheimer's Disease (AD) affects over 6.7 million people in the US and 35 million people worldwide. Effective treatments are scarce, partly because interventions often start too late in the disease's progression. This delay leads to the failure of many clinical trials and underscores the urgent need for early diagnosis to improve treatment outcomes [6]. Currently, only 5% of early-stage AD cases are diagnosed promptly ([10]).

This project aims to develop and evaluate deep learning models for early AD detection using MRI images from the OASIS-1 dataset as inputs. We employ a range of models, including custom and standard 2D and 3D CNNs, to output the AD stage of each image—very mild, moderate/mild, or healthy. Our goal is to assess these models' effectiveness against benchmarks like BrainNet2D and BrainNet3D by measuring accuracy, sensitivity, specificity, balanced accuracy, and ROC, therefore enhancing early-stage diagnostic accuracy and addressing challenges in MRI data analysis.

## 2. Related Work

AD is primarily diagnosed manually by doctors using CT and MRI scans, with concerns about false positives limiting the use of machine learning (ML). Although ML is underutilized in AD diagnosis due to accuracy issues, ongoing research aims to enhance its efficacy [16]. Studies using the OASIS dataset have mainly employed traditional ML techniques like SVM, decision trees, and logistic regression, achieving 78%-86% accuracy [3]. These methods often fail to fully capture the complex patterns and spatial diversity of the images.

Recent studies on the OASIS dataset have primarily utilized ResNet and Inception-ResNet architectures, achieving 95% accuracy with ResNet-50 and Inception-ResNet-v2 on the OASIS-3 dataset [14] [7]. These models leverage ResNet's skip connections for simpler training and Inception-ResNet's design for enhanced efficiency and complexity management. Despite their success, ResNet faces issues with complexity and generalization, and the Inception architecture's performance on OASIS-1 is untested, although it's used widely in medical contexts [5]. Additionally, DenseNet-121 and VGG-16 have been explored, reaching 91% and 96% accuracy respectively on OASIS-3, but their efficacy on OASIS-1 remains unknown [12] [8].

Custom 2D CNN models like BrainNet2D have successfully diagnosed AD using OASIS datasets, with 93% accuracy in AD prediction and 88% in stage differentiation across OASIS-1 and OASIS-2 [13].. These models are efficient, using only five convolutional blocks. Meanwhile, larger, more computationally intensive hybrid models combining EfficientNetB5 and Inception-ResNet-v2 achieved similar precision and recall scores of around 0.96 on OASIS-1 [11]. Additionally, RNNs/LSTMs, which excel at sequential data processing, reached 96% accuracy on OASIS-3 [4]. However, no CNN-RNN hybrids have been tested on OASIS-1, which could potentially enhance feature extraction and sequential analysis for better results.

3D architectures have seen limited exploration on the OASIS datasets compared to 2D models. For example, a 3D VGG model only achieved 69.9% accuracy on OASIS-1 [17]. While 3D ResNet and 3D DenseNet have shown promise in extracting spatial-temporal information and offering parameter efficiency in other medical contexts like CT colonographies, they remain untested on OASIS [15]. Moreover, custom models like BrainNet3D, a 3D adaptation of BrainNet2D, exist but require further investigation to enhance diagnostic accuracy [13].

# 3. Dataset

## 3.1. Overview

The project utilizes the OASIS dataset from OASIS Brains (wustl.edu), featuring MRI brain scans from 416 individuals aged 18-96. This dataset includes 100 subjects with AD and 316 without, grouped under OASIS-1. It provides demographic and clinical data such as age, gender, education, and mini-mental state examination scores, stored across 12 TAR files containing 3-D data. A primary challenge in data pre-processing is the imbalance in the dataset, with 316 cognitive normal cases and fewer cases of very-mild (70 instances), mild (28 instances), and moderate dementia (2 instances). We merged the mild and moderate categories to address this.

## 3.2. 2D CNN Dataset Modifications

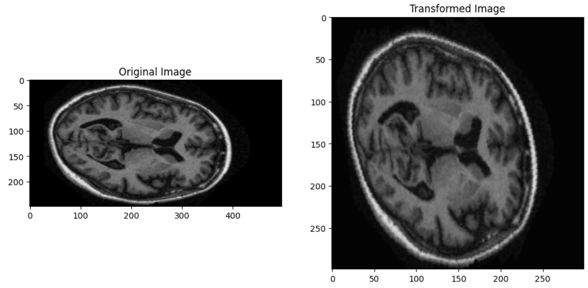


Figure 1: Original Image vs Data Augmented Image

We extracted 2-D slices from indexes 60-80 and 100-120 of each TAR file in the OASIS1 dataset, converting them to JPG format for analysis. To avoid data leakage, we used a subject-level split, ensuring slices from each patient were only in the training/validation or test set, crucial for preventing overinflated accuracy metrics due to the high correlation between adjacent MRI slices [18].

From these slices, we generated 86,437 JPG images, resized to 224x224 pixels for compatibility with most 2D CNN models, except InceptionV3, which required 299x299 pixels (Figure 1). To address class imbalance, we enhanced underrepresented categories through augmentation techniques such as random rotations, horizontal flips, and resized crops.

After data augmentation, the training dataset included approximately 47,336 healthy, 12,000 very mild, and 12,000 mild/moderate samples. The validation set comprised 13,298 healthy, 3,050 very mild, and 1,159 mild/moderate samples. The testing dataset contained about 6,588 healthy, 1,220 very mild, and 488 mild/moderate samples.

## 3.3. 2D Custom CNN Dataset Modifications

The dataset for the 2D Custom models was sourced from the 2D-processed OASIS-1 Collection [1]. It originally transformed from 3D to 2D images along the coronal plane. To combat data imbalance in multi-classification, the mild and moderate classes were combined, resulting in a total of 15,057 non-demented, 13,725 very mild, and 5,429 combined mild and moderate samples. These were divided into training, validation, and test sets in a 0.6, 0.2, and 0.2 ratio, leading to sizes of 20,564, 6,854, and 6,864, respectively. Measures were taken to maintain consistent class distributions across these sets to address data imbalance. Additionally, each image slice was resized to 128x128 and reordered to (channel, height, width) format to streamline processing and enhance model compatibility.

## 3.4. 3D CNN Dataset Modifications

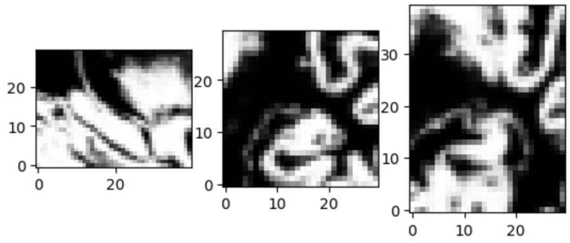


Figure 2: Images of the Right Hippocampus for OASIS Sample 1

The 3D data was extracted from TAR files using Clinica, a widely-used platform in neuroimaging studies.

Due to the dataset's small size, Clinica applied non-linear regression, segmented grey matter, and converted the data into tensor format [2]. We trained and tested the 3D architectures exclusively using the right hippocampus to mitigate overfitting. This choice is informed by research indicating that the right hippocampus, which plays a crucial role in processing spatial information, is frequently affected in the early stages of AD [9]. Thus, the resulting image size was reduced to $30 \times 40 \times 30$ from the original $121 \times 45 \times 121$. A sample image is shown in Figure 2.

We allocated 80% of the data for training and validation, and 20% for testing, using 300 samples (242 healthy, 37 very mild, 21 mild/moderate) for the former and 77 samples (62 healthy, 8 very mild, 7 mild/moderate) for the latter. We ensured an even distribution of age, gender, and diagnosis between groups.

# 4. Methods

## 4.1. 2D CNN Models

### 4.1.1 BrainNet2D

Our baseline models, binary and multiclass BrainNet2D architectures referenced from Saratxaga et al. [13], are re-implemented with our preprocessing methods for direct comparison on OASIS datasets. BrainNet2D consists of five Convolutional Blocks, each with a layer sequence of convolution, batch normalization, ReLU activation, and max pooling, with channel outputs increasing from 8 to 128. The architecture concludes with a global average pooling layer and a dense layer, comprising 522 thousand trainable parameters and a total size of 2.09 MB.

BrainNet2D's block structure enhances its efficiency in feature extraction and training. The early blocks focus on simple features such as edges, with deeper blocks handling complex patterns crucial for specific medical diagnoses. This modular design facilitates easy modifications, promotes faster convergence with stable gradients, and improves generalization to new data, thus boosting diagnostic accuracy. However, BrainNet2D has limitations. The Global Average Pooling (GAP) and Dense Layers may cause loss of crucial spatial details and reduce feature distinction, essential for accurate neurological assessments. Dense layers also risk overfitting and pose scalability issues with larger datasets. Insufficient convolutional blocks might limit the model's ability to learn detailed patterns, affecting performance and generalization. These factors highlight the need for a balanced network design in critical applications like medical imaging.

### 4.1.2 Pre-Trained 2D CNN Models

After benchmarking BrainNet2D, we assessed the capabilities of popular 2D CNN frameworks—DenseNet121, InceptionV3, and VGG16—using the OASIS-1 dataset for AD diagnosis. These models were selected for their diverse layer depths, kernel sizes, and connectivity, offering thorough evaluations for MRI brain scan processing. We adapted each model by replacing the original final fully connected layer with a linear layer designed to fit the required number of classes. Our approach included both binary models, which determined the presence or absence of AD, and multi-class models, which categorized stages to enhance early-stage detection accuracy.

We modified the DenseNet121 architecture for AD classification by implementing a two-stage linear transformation in its classifier, which first halves feature dimensions using a ReLU activation and then maps these to disease classes. This adaptation maintains dense connectivity and deep supervision, improves gradient flow and feature reuse, and keeps a low parameter count at 526 thousand trainable of 7.5 million total, with a model size of about 30 MB.

InceptionV3 was optimized for medical imaging with a dual-stage linear transformation in its classifier, reducing dimensions by a third for better feature discrimination and adjusting classifiers to enhance gradient flow and stability. Designed for a 299x299 pixel input, it includes 25.9 million total parameters, 1.6 million trainable, and is 103.761 MB in size.

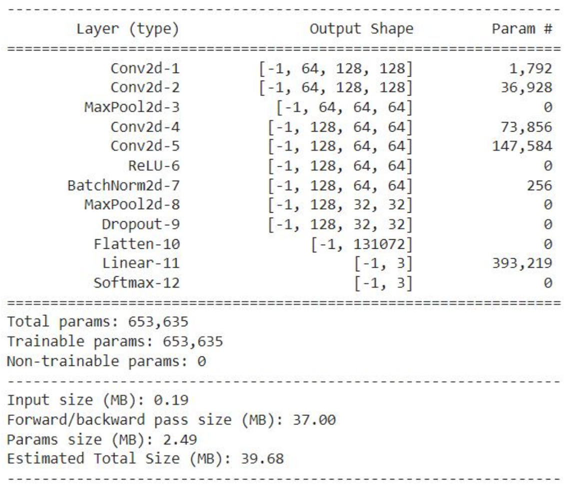
VGG16 was optimized for AD detection by adding two linear layers in the final classification stage, halving and then mapping output dimensions to enhance feature extraction. The final model includes 142 million total parameters, 8.4 million trainable, and is approximately 570.629 MB in size.

## 4.2. Custom Model of 2D CNN and RNN

CNNs excel in image data processing and multiclass classification due to their ability to capture spatial features, while RNNs are preferred for sequential tasks like time-series analysis. Although 3D CNNs are suited for 3D medical images, their complexity and high parameter count pose training challenges. In our study, we integrate 2D CNNs with RNNs by segmenting 3D MRI data along the coronal plane, using 2D CNNs to extract features from each slice, which are then sequentially analyzed by an RNN. This approach is compared to a custom 2D CNN model.

#### 4.2.1 Custom 2D CNN Architectures

The custom 12-layer CNN (Figure 3) is tailored for extracting features from 2D MRI slices for AD analysis. It begins with convolutional layers using 3-sized kernels for initial feature extraction, followed by pooling layers that downsample feature maps to reduce spatial dimensions while retaining critical features. ReLU activation enhances non-linearity and is followed by batch normalization to stabilize activations and expedite training. To mitigate overfitting, a dropout layer with a 0.25 probability is applied after normalization. The data is then flattened and directed to a fully connected layer for multiclass classification, enabling the model to learn and capture complex AD-related patterns.



```
----------------------------------------------------------------
        Layer (type)               Output Shape         Param #
================================================================
            Conv2d-1        [-1, 64, 128, 128]           1,792
            Conv2d-2        [-1, 64, 128, 128]          36,928
         MaxPool2d-3          [-1, 64, 64, 64]               0
            Conv2d-4         [-1, 128, 64, 64]          73,856
            Conv2d-5         [-1, 128, 64, 64]         147,584
              ReLU-6         [-1, 128, 64, 64]               0
       BatchNorm2d-7         [-1, 128, 64, 64]             256
         MaxPool2d-8         [-1, 128, 32, 32]               0
           Dropout-9         [-1, 128, 32, 32]               0
          Flatten-10              [-1, 131072]               0
           Linear-11                   [-1, 3]         393,219
          Softmax-12                   [-1, 3]               0
================================================================
Total params: 653,635
Trainable params: 653,635
Non-trainable params: 0
----------------------------------------------------------------
Input size (MB): 0.19
Forward/backward pass size (MB): 37.00
Params size (MB): 2.49
Estimated Total Size (MB): 39.68
----------------------------------------------------------------
```

Figure 3: Custom 2D CNN

#### 4.2.2 Custom 2D CNN and RNN Architecture

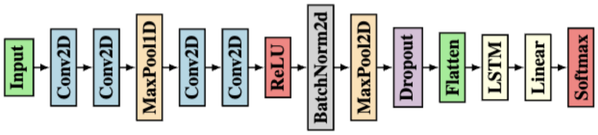


Figure 4: Custom 2D CNN and LSTM

Building on the 2D CNN architecture, we aim to improve the model's capability to interpret temporal dynamics in MRI slices by incorporating Long Short-Term Memory (LSTM) units. After the CNN layers extract features from the MRI data, these feature maps are reshaped and fed into the LSTM layer. We treat individual slices from segmented 3D MRI volumes as sequential data instances, which the LSTM processes to create a consolidated output. This output is then refined through a fully connected layer before moving to the final multiclass classification stage (Figure 4).

### 4.3. 3D CNN Architectures

#### 4.3.1 BrainNet3D Architecture

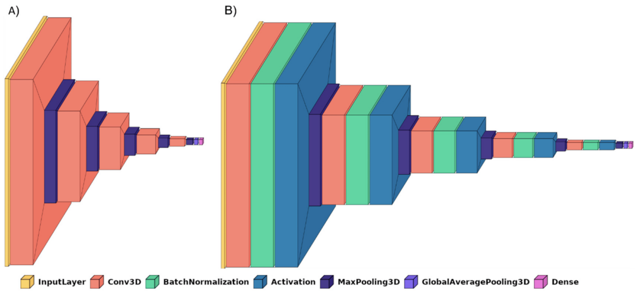


Figure 5: Architecture for BrainNet3D

The baseline for the 3D-CNN models is a binary class BrainNet3D, as shown in Figure 5. Note that this architecture is exactly the same as the BrainNet2D architecture but extends to 3D. Thus, BrainNet3D also benefits from the Convolutional Blocks because of their complexity and faces drawbacks due to choices with model architecture.

#### 4.3.2 Modified BrainNet3D Architecture

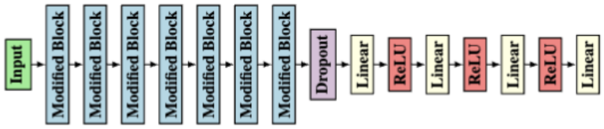


Figure 6: Architecture of Modified Neural Network (More Detailed Architecture in Appendix)

The BrainNet3D architecture was refined to the Modified BrainNet3D to address performance and robustness, selecting binary over multiclass models due to class imbalances (Figure 6). Modifications included replacing GAP and Dense Layers with a Dropout Layer, adding three pairs of Linear and ReLU Layers, and an extra Linear Layer to enhance complexity and nonlinearity, aligning with standard neural network practices for improved efficacy.

Conv3D
BatchNorm
ReLU
MaxPool3D
Conv3D
BatchNorm
CustomPool3D

(a) Convolutional Block (b) Modified Convolutional Block

Figure 7: Changes Between the Two Blocks

Further enhancements to the Convolutional Block aimed at optimizing performance and reducing overfitting involved removing ReLU layers, simplifying the block, and substituting the standard MaxPool Layer with a CustomMaxPool Layer for 3D max pooling. This allowed for dynamic padding to better process varying input dimensions (Figure 7).

The final key enhancement was increasing the number of Convolutional Blocks in the model from five to seven, enhancing feature extraction and learning capacity through more parameters and deeper architecture. The optimal setup now includes seven layers with progressively increasing channel sizes: 8, 16, 32, 64, 128, 256, and 512, each using a convolution filter size of 3 for refined feature extraction at each stage.

## 4.4. 3D ResNet and 3D DenseNet

Binary class 3D ResNet and 3D DenseNet architectures were explored due to their common use in MRI scan diagnosis studies.

3D ResNets build on 2D models like ResNet-18, 34, 50, 101, and 152, adapted for volumetric data. These architectures employ residual blocks with batch normalization and LeakyReLU activation, enhancing gradient flow and preventing the vanishing gradient problem. This allows for deeper networks, making them effective for MRI analysis. The modification to LeakyReLU from standard ReLU further improves training performance.

3D DenseNets extend from 2D models like DenseNet-121, 169, 201, and 264, suited for processing volumetric data. They incorporate dense connections that enhance feature propagation, crucial for detecting subtle medical imaging features. The structure features densely connected blocks interspersed with transition layers that manage dimensions through convolution and pooling. Each block uses Batch Normalization, ReLU activation, and 3x3 Convolution, optimizing parameter use by concatenating features. Adjustments like dropout in each DenseBlock layer increase robustness against overfitting in complex 3D environments.

## 4.5. Loss Functions

### 4.5.1 2D CNN and 3D CNN

In our study, 2D CNN pre-trained models employed a cross-entropy loss function with class weights to counteract data class imbalances. These weights were derived from the inverse frequency of classes in the training data to prevent bias toward more frequent classes. For 3D-CNN architectures and BrainNet3D, we used a non-weighted cross-entropy loss function, which showed significant improvement in validation loss curves over time, validating our choice.

The loss function is defined as:

$$L(y, \hat{y}) = -\sum_{i=1}^{C} w_i \cdot y_i \cdot \log(\hat{y_i}),$$

where $w_i$ is the class weight for class $i$, used only for weighted models.

### 4.5.2 Custom 2D CNN and RNN

We initially used the Adam optimizer for training our Custom 2D CNN and RNN models but observed limited improvement in validation accuracy, despite various learning rate adjustments. This indicated a possible mismatch between the optimizer and our model architecture, as Adam may struggle in high-dimensional parameter spaces. Consequently, we switched to the simpler and more robust SGD with momentum set to 0.9, which significantly improved our models' performance.

SGD is represented by the equations:

$$v_{t+1} = \beta \cdot v_t + (1-\beta) \cdot \nabla L(\theta_t), \quad \theta_{t+1} = \theta_t - \alpha \cdot v_{t+1}$$

where $v_{t+1}$ is the updated momentum vector, $\beta$ is the momentum coefficient, $\nabla L(\theta_t)$ is the gradient of the loss function $L$ with respect to the parameters $\theta_t$ at time $t$, $\theta_{t+1}$ is the updated parameter vector, and $\alpha$ is the learning rate.

# 5. Experiments and Results

## 5.1. Metrics

We used several metrics to evaluate our multiclass 2D CNN models on validation and test datasets: accuracy, balanced accuracy, ROC with OVR Accuracy, sensitivity, and specificity. Balanced accuracy was crucial for handling varying class sizes, while ROC scores assessed class discrimination effectiveness. Sensitivity was key for accurately identifying AD patients, and specificity ensured the correct identification of healthy individuals.

We used the same metrics for binary 3D CNN models. Additionally, we assessed the accuracy of specific AD stages by calculating the ratio of correctly identified cases within the mild-moderate and very-mild stages, deriving overall AD accuracy from these stage-specific accuracies.

## 5.2. 2D CNN Architectures

### 5.2.1 Hyperparameter Tuning

| Model | Learning rate | Batch size | Beta1 | Beta2 | Epsilon |
|---|---|---|---|---|---|
| VGG16 | 0.0016 | 128 | 0.92 | 0.999 | 1e-8 |
| DenseNet121 | 0.00012 | 64 | 0.90 | 0.992 | 1e-7 |
| InceptionV3 | 0.0003 | 256 | 0.90 | 0.999 | 1e-8 |

Table 1: Summary of Best Hyperparameter Values for Pre-Trained 2D CNN Models

We optimized the hyperparameters of pretrained 2D CNN models using the Optuna framework, specifically focusing on the Adam optimizer for the VGG16 model. We tested learning rates from 1e-5 to 1e-2, finding 0.0016 optimal for fast yet stable learning. The best results came with a batch size of 128 and moment decay rates (beta1 and beta2) fine-tuned to 0.92 and 0.999, respectively. For details on other models, see Table 1.

### 5.2.2 Model Performance

In the binary classification setting as shown in Table 2, the BrainNet2D model achieves high accuracy (0.809) but low sensitivity (0.370) for the demented class, suggesting many mild and very mild cases are misclassified as healthy. This illustrates the limits of relying solely on accuracy in cases of class imbalance. In contrast, the VGG16 model, despite a lower accuracy (0.607), shows the highest balanced accuracy (0.759) among the models, emphasizing its reliability.

| Model | Acc | Balanced Acc | Area Under ROC | Sensitivity | Specificity |
|---|---|---|---|---|---|
| BrainNet2D | 0.809 | 0.647 | 0.856 | 0.370 | 0.923 |
| DenseNet121 | 0.812 | 0.770 | 0.884 | 0.697 | 0.842 |
| InceptionV3 | 0.808 | 0.789 | 0.887 | 0.579 | 0.868 |
| VGG16 | 0.607 | 0.759 | 0.777 | 0.915 | 0.527 |

Table 2: Performance of baseline and 2D CNN pretrained models in binary classification

| Model | Acc | Balanced Acc | Area Under ROC |
|---|---|---|---|
| BrainNet2D | 0.739 | 0.447 | 0.834 |
| DenseNet121 | 0.6766 | 0.549 | 0.808 |
| InceptionV3 | 0.534 | 0.605 | 0.805 |
| VGG16 | 0.604 | 0.723 | 0.802 |

Table 3: Performance of Baseline and 2D CNN Pre-Trained Models in Multiclass Classification

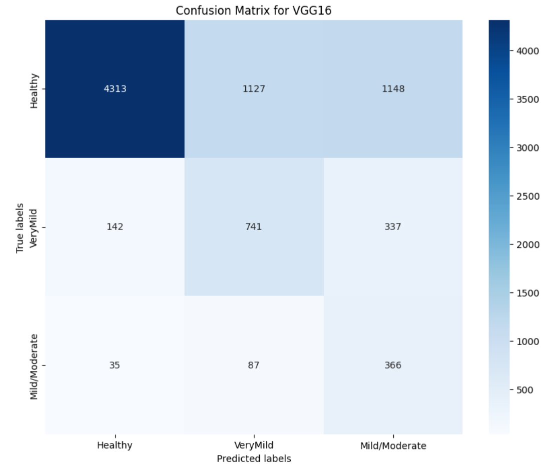


Figure 8: Confusion Matrix of VGG16 Model for Multi-Class

In multi-class classification, shown in Table 3, VGG16 achieves the highest balanced accuracy (0.723), effectively managing inter-class variance and proving suitable for evenly representing all classes, as further illustrated by its confusion matrix (Figure 8). Meanwhile, BrainNet2D, with moderate accuracy (0.739), has the lowest balanced accuracy (0.447), indicating significant performance disparities across classes.

### 5.2.3 Qualitative Analysis

(a) Healthy patient

(b) Mild/Moderate patient

Figure 9: Comparison of MRI for Healthy and Demented Patient

The first image (Figure 9a) shows an MRI slice from a healthy individual, and the second (Figure 9b) from a patient with mild/moderate dementia. The VGG16 model correctly identified both, demonstrating its capability to distinguish health stages. In contrast, the BrainNet2D model misclassified both as healthy, reflecting its frequent inability to detect mild/moderate conditions. This indicates that BrainNet2D may lack the necessary computational sophistication to identify subtle pathological markers in MRI images. VGG16's superior performance is also likely enhanced by its extensive pre-training on the diverse ImageNet dataset, improving its feature extraction capabilities for better medical image analysis.

## 5.3. Custom 2D CNN and RNN

### 5.3.1 Hyperparamters Tuning

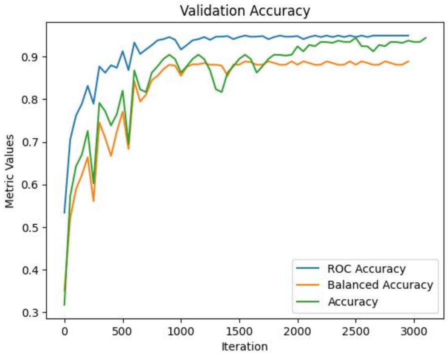


Figure 10: Accuracy of Custom 2D CNN and LSTM on Validation Dataset

In our study, we tested learning rates within the range [1e-2, 1e-4], finding that approximately 1e-3 produced optimal results. At this rate, training loss consistently decreased while validation accuracy steadily increased, indicating no overfitting on the training set. Additionally, we determined that a momentum parameter of 0.9 in SGD optimizes model performance by enabling faster convergence and balancing rapid progress with stability. This value effectively averages gradients over multiple iterations, enhancing the optimization process (Figure 10).

### 5.3.2 Model Performance

| Metric | 2D CNN | 2D CNN + LSTM |
|---|---|---|
| Accuracy | 0.8972 | 0.9833 |
| ROC OvR Accuracy | 0.95543 | 0.9907 |
| Balanced Accuracy | 0.8775 | 0.9833 |

Table 4: Performance metrics of Custom 2D CNN vs 2D CNN and LSTM on Test Dataset

After training for five epochs, we assessed each model's performance on the test dataset (Table 4). Adding an LSTM layer to the 2D CNN significantly improved classification accuracy, suggesting that the LSTM effectively captured temporal dependencies in the 2D MRI slices, enhancing predictive accuracy. This enhancement also extended to balanced accuracy, demonstrating that the combined 2D CNN and LSTM model performs well in terms of sensitivity and specificity, effectively handling the imbalanced class distributions in our data. Both models showed strong ROC accuracy, with the 2D CNN + LSTM model outperforming the standalone 2D CNN, indicating robust discrimination capabilities between the positive class and all other classes across various thresholds.

### 5.3.3 Qualitative Analysis

The combined CNN and LSTM model outperforms standalone CNNs because LSTMs excel in capturing hierarchical features. In our setup, where 2D images are derived from 3D MRI brain scans, LSTM effectively recognizes spatial dependencies across slices from the same original 3D image. By complementing CNN's ability to extract spatial features, LSTM enhances the model's capability to detect dementia areas in 2D MRI slices by integrating both spatial and temporal information effectively.

## 5.4. 3D CNN Architectures

### 5.4.1 Hyperparamter Tuning

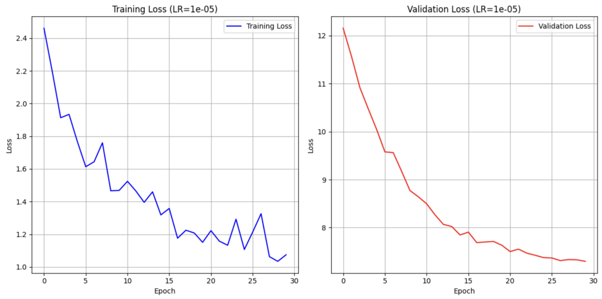


Figure 11: Training Loss vs Validation Loss for Baseline, BrainNet3D, at Optimal Learning Rate

The mini-batch size was set at 4 to manage class imbalance, improve generalization, and increase model stability through more frequent weight updates. We chose the Adam optimizer for its efficiency and robustness, which calculates exponentially decaying averages of past gradients and squared gradients. The primary hyperparameter tuned was the learning rate, ranging from 1e-3 to 1e-6. For BrainNet3D, ModifiedBrainNet3D, 3D ResNet-101 and 3D DenseNet-121, 201, and 264, the optimal rate was 1e-5. For ResNet-18, 34, 50, and 152 and DenseNet-169, it was 5e-6, reflecting differences in network size and architecture. At these rates, training loss decreased while validation loss stabilized, indicating optimal performance (Figure 11). The number of epochs was adjusted based on the validation loss curve to prevent overfitting.

### 5.4.2 Model Performance

| Model | Acc. | Balanced Acc. | Sensitivity | Specificity |
|---|---|---|---|---|
| BrainNet3D | 0.883 | 0.700 | 0.400 | 1.000 |
| Modified BrainNet3D 7 Blocks | 0.909 | 0.842 | 0.733 | 0.952 |
| Modified BrainNet3D 10 Blocks | 0.883 | 0.776 | 0.600 | 0.952 |
| 3D ResNet-18 | 0.857 | 0.734 | 0.533 | 0.935 |
| 3D ResNet-34 | 0.896 | 0.809 | 0.666 | 0.952 |
| 3D ResNet-50 | 0.870 | 0.869 | 0.867 | 0.871 |
| 3D ResNet-101 | 0.896 | 0.860 | 0.919 | 0.800 |
| 3D ResNet-152 | 0.857 | 0.785 | 0.667 | 0.903 |
| 3D DenseNet-121 | 0.909 | 0.767 | 0.533 | 1.000 |
| 3D DenseNet-169 | 0.883 | 0.826 | 0.733 | 0.919 |
| 3D DenseNet-201 | 0.857 | 0.760 | 0.600 | 0.919 |
| 3D DenseNet-264 | 0.883 | 0.725 | 0.467 | 0.984 |

Table 5: General Performance Metrics of the Models

| Model | Very-Mild Accuracy | Mild-Moderate Accuracy | AD Accuracy |
|---|---|---|---|
| BrainNet3D | 0.250 | 0.571 | 0.411 |
| Modified BrainNet3D 7 Blocks | 0.625 | 0.857 | 0.741 |
| Modified BrainNet3D 10 Blocks | 0.375 | 0.857 | 0.616 |
| 3D ResNet-18 | 0.250 | 0.857 | 0.554 |
| 3D ResNet-34 | 0.500 | 0.857 | 0.679 |
| 3D ResNet-50 | 0.750 | 1.000 | 0.875 |
| 3D ResNet-101 | 0.750 | 0.857 | 0.804 |
| 3D ResNet-152 | 0.625 | 0.714 | 0.670 |
| 3D DenseNet-121 | 0.500 | 0.857 | 0.679 |
| 3D DenseNet-169 | 0.625 | 0.857 | 0.741 |
| 3D DenseNet-201 | 0.375 | 0.857 | 0.616 |
| 3D DenseNet-264 | 0.500 | 0.857 | 0.679 |

Table 6: Specific Stage Accuracies of the Models

The baseline model, BrainNet3D, achieved an 88.30% accuracy but only a 70% balanced accuracy, indicating low sensitivity and high specificity, which hindered its ability to diagnose early stages of AD accurately (Table 5). Despite a slight drop in accuracy, models like some ResNets and DenseNets showed improvements in sensitivity, balanced accuracy, and stage-specific accuracies due to their ability to detect subtle changes, particularly in critical AD areas like the hippocampus (Table 6). However, these gains in sensitivity typically resulted in a significant reduction in specificity, a common tradeoff for higher diagnostic sensitivity.

For the Modified BrainNet3D model, reducing the number of blocks from 10 to 7 improved performance, likely due to fewer layers minimizing issues like gradient problems and diminishing returns. Similarly, ResNet-50, ResNet-101, and DenseNet-169 exhibited better performance in AD detection compared to other ResNet and DenseNet models, balancing complexity effectively (Table 6 and Table 5). Among the models, optimal ResNet configurations outperformed the best DenseNet and Modified BrainNet3D 7-block models in stage accuracies, balanced accuracies, and sensitivities, with similar overall accuracies. ResNet's skip connections, which allow bypassing less effective layers, contributed to more efficient training and better generalization compared to DenseNet.

### 5.4.3 Qualitative Analysis

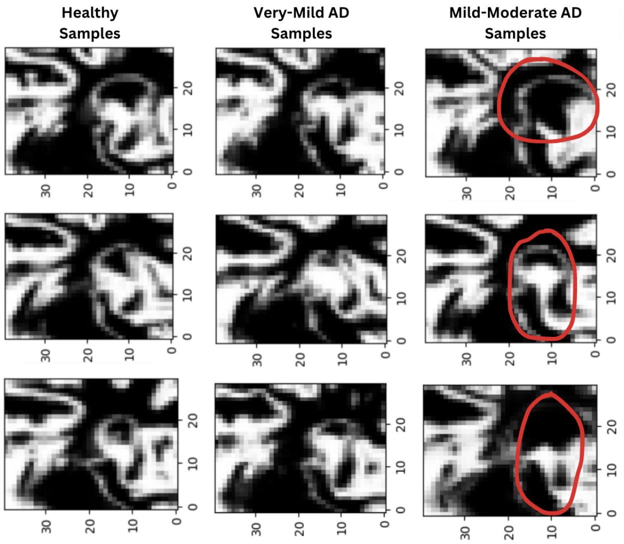


Figure 12: Examples of Healthy Samples, Very-Mild AD, and Mile-Moderate AD Samples

The models improved AD sample identification over the baseline, particularly in differentiating disease stages, yet they showed lower accuracy for very-mild samples compared to mild-moderate samples (Table 6). This lower performance for very-mild AD can be linked to their subtle visual differences from healthy samples, which often only show minor deformations. In contrast, mild-moderate AD samples are easier to detect due to significant hippocampal loss as shown by the indicated regions in Figure 12. Additionally, the models' reduced specificity and accuracy in recognizing healthy samples suggests that the visual similarities between healthy and very-mild AD samples contribute to misclassification.

The models accurately identified about 85.7% of mild-moderate samples, which display more severe deformations than very-mild AD. However, their resemblance to healthy samples in features and shape made differentiation challenging, impacting diagnosis accuracy.

## 6. Conclusion and Future Work

The study found that pretrained 2D CNNs like the BrainNet2D model tended to overfit to the 'Healthy' label, whereas InceptionV3 and VGG16, both pretrained on ImageNet, showed better performance in classifying multiple stages of AD. Future efforts will focus on fine-tuning more layers of VGG16 and InceptionV3, utilizing the OASIS-2 dataset to better represent underrepresented stages.

Additionally, a hybrid model combining CNN and LSTM outperformed standalone CNNs by utilizing CNNs for spatial feature extraction and LSTMs for capturing temporal dependencies, improving the detection of complex dementia patterns in 2D MRI slices. Future work will investigate lighter versions of both CNN and LSTM to enhance model efficiency and reduce complexity. In 3D modeling, ResNet models with skip connections, specifically ResNet-50 and ResNet-101, surpassed other models including the Modified BrainNet3D and DenseNet in detecting AD. Future research should focus on optimizing these models for improved specificity and detection of very-mild AD stages.

# 7. Appendices

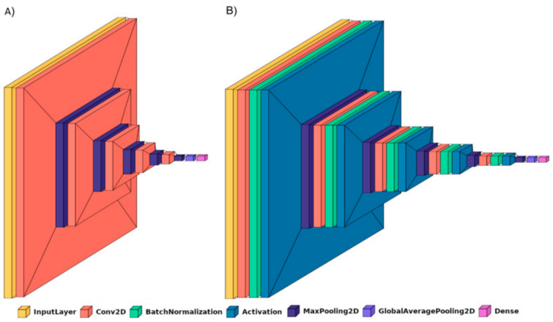


Figure 13: Architecture for BrainNet2D

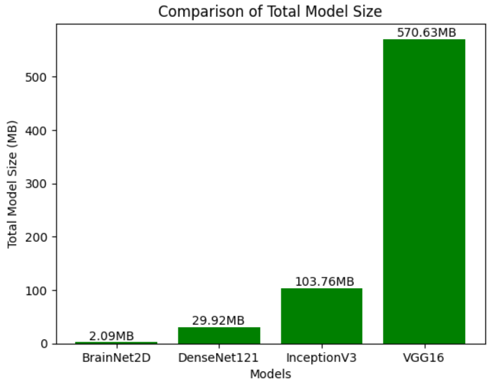


Figure 14: Comparison of Total Model Size for 2D CNN models

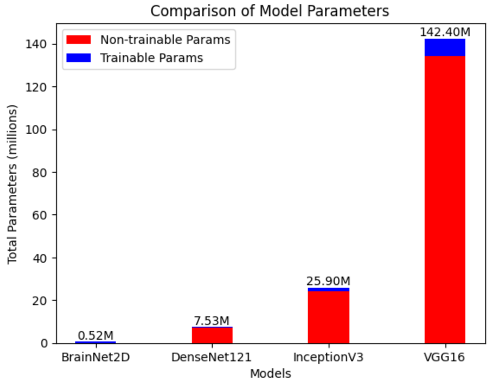


Figure 15: Comparison of Total Model Parameter size for 2D CNN models

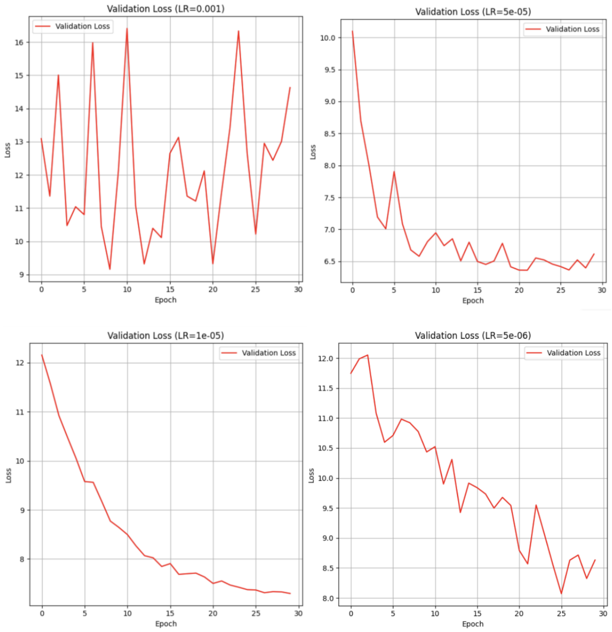


Figure 16: Example of Different Learning Rates Effecting the Shape of the Loss Curve

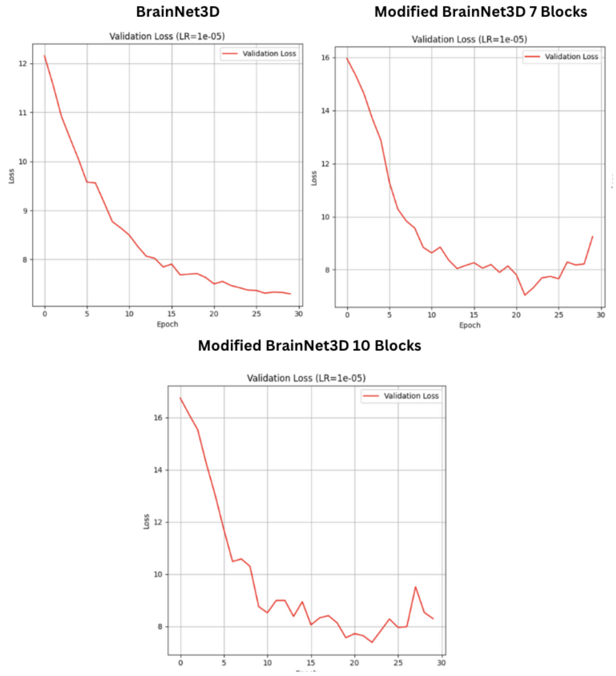


Figure 17: Validation Loss Curves of BrainNet3D and Modified BrainNet3D Models

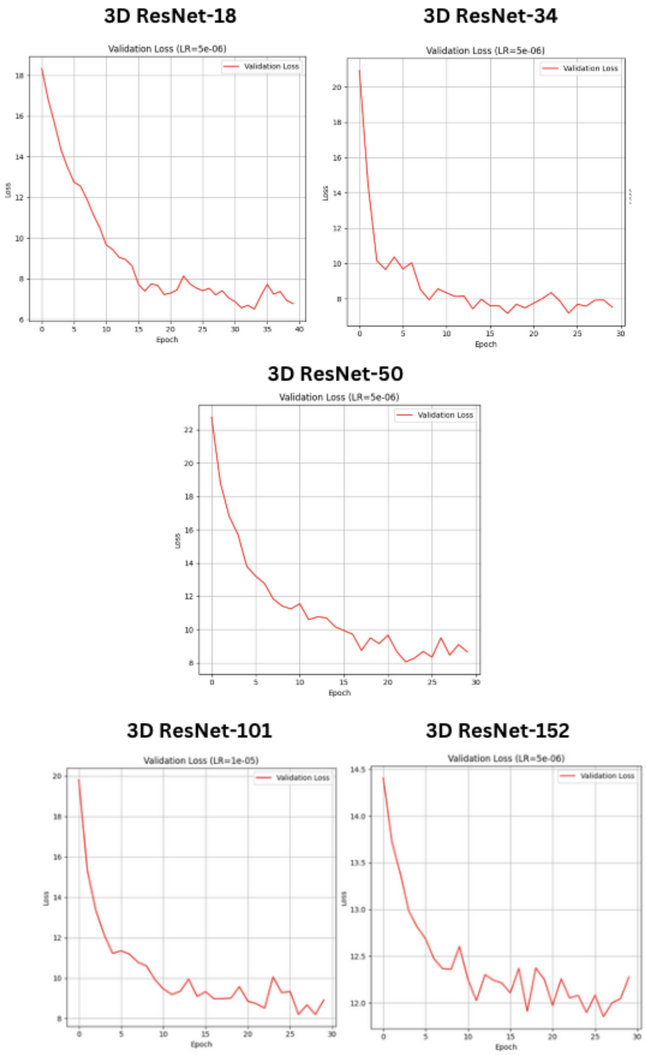


Figure 18: Validation Loss Curves of ResNet Models

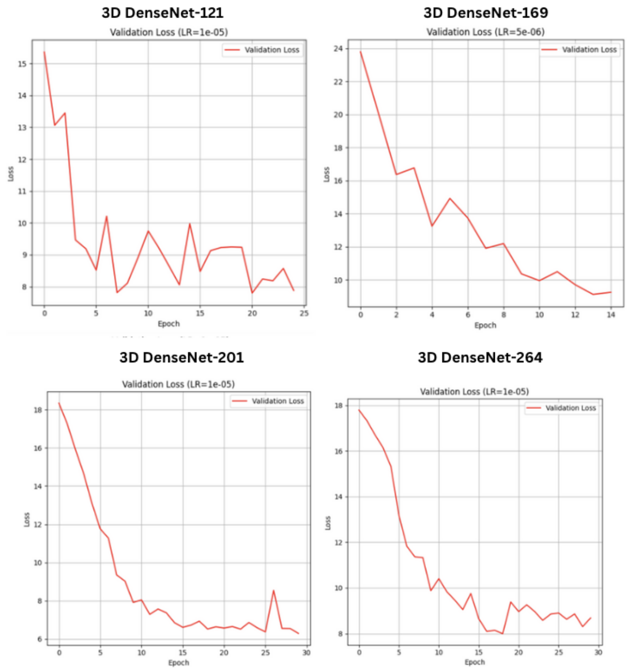


Figure 19: Validation Loss Curves of DenseNet Models

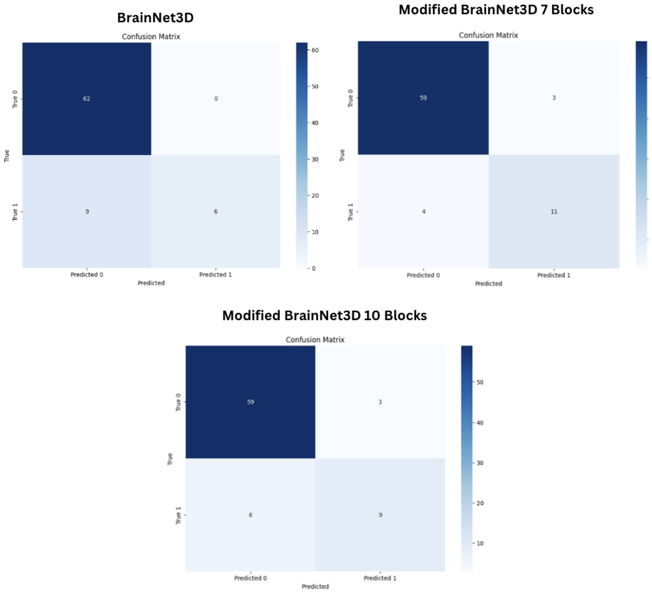


Figure 20: Confusion Matrices of BrainNet3D and Modified BrainNet3D Models

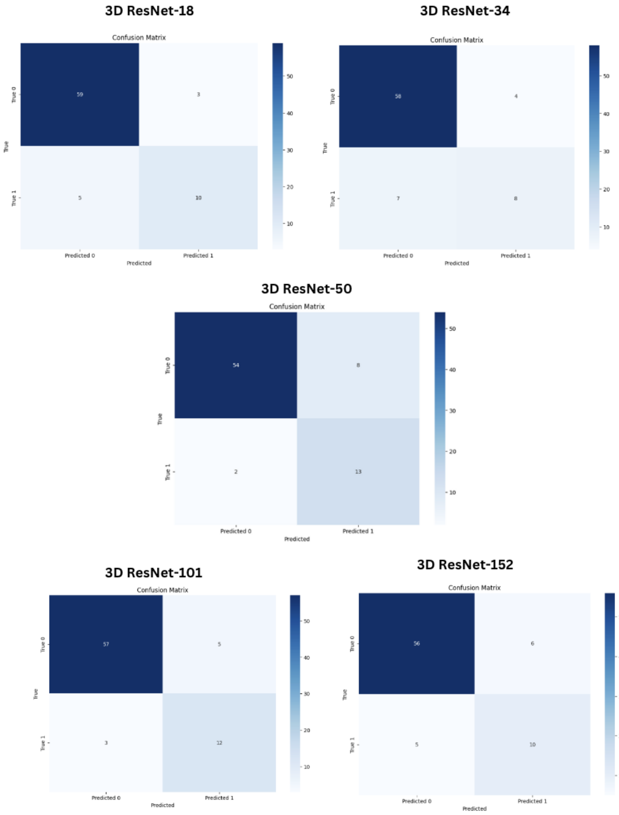


Figure 21: Confusion Matrices of ResNet Models

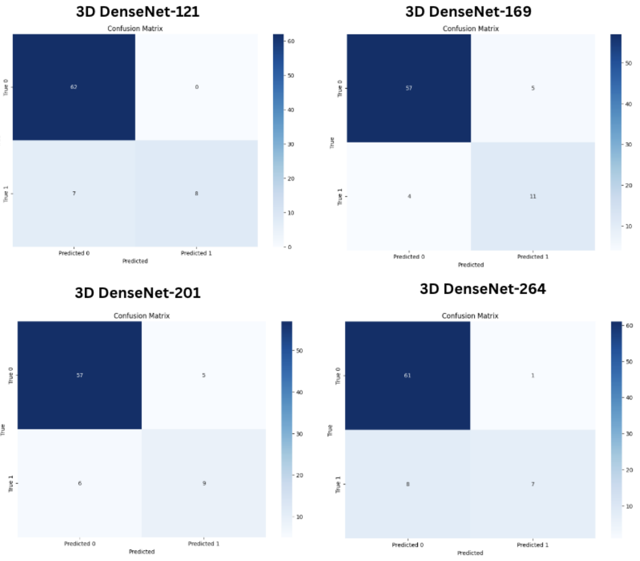


Figure 22: Confusion Matrices of 3D DenseNet Models

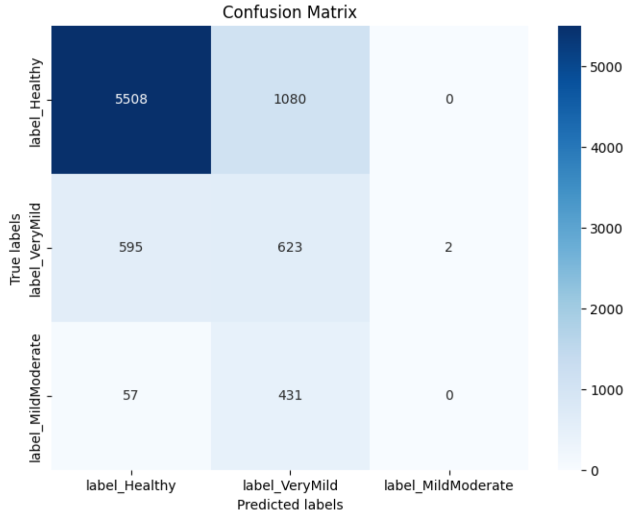


Figure 23: Confusion Matrix of BrainNet2D Model for Multi-Class

## 8. Contributions & Acknowledgements

In this project, Nikhil, Celia, and Siya each played pivotal roles in developing and advancing our models. Nikhil focused on crafting the 2D CNN models, Celia expanded our scope by developing both 2D CNN and RNN models, and Siya specialized in the 3D CNN models. Collectively, we implemented these models, gathered results, and meticulously documented our findings.

## References

[1] https://www.kaggle.com/code/adisongoh/alzhemier-classification-with-pretrained-densenet.

[2] https://aramislab.paris.inria.fr/files/data/databases/DL4MI/OASIS-1-dataset_pt_new.tar.gz.

[3] P. Baglat, A. W. Salehi, A. Gupta, and G. Gupta. Multiple machine learning models for detection of alzheimer's disease using oasis dataset. In *Re-imagining diffusion and adoption of information technology and systems: A continuing conversation: IFIP WG 8.6 international conference on transfer and diffusion of IT, TDIT 2020, Tiruchirappalli, India, December 18–19, 2020, proceedings, part I*, pages 614–622. Springer, 2020.

[4] S. Gnanasegar, B. Bhasuran, and J. Natarajan. A long short-term memory deep learning network for mri based alzheimer's disease dementia classification. *J Appl Bioinforma Comput Biol 9: 6. doi: 10.37532/jabcb. 2020.9 (6)*, 187:2, 2020.

[5] S. Kollem, L. L. Joseph, U. Desai, S. Tayal, J. Ajayan, S. Bhattacharya, P. Rajasree, et al. Incnet: Brain tumor detection using inception and optimization techniques. In *2022 International Conference on Emerging Techniques in Computational Intelligence (ICETCI)*, pages 71–75. IEEE, 2022.

[6] C. Laske, H. R. Sohrabi, S. M. Frost, K. López-de Ipiña, P. Garrard, M. Buscema, J. Dauwels, S. R. Soekadar, S. Mueller, C. Linnemann, et al. Innovative diagnostic tools for early detection of alzheimer's disease. *Alzheimer's & Dementia*, 11(5):561–578, 2015.

[7] A. Loddo, S. Buttau, and C. Di Ruberto. Deep learning based pipelines for alzheimer's disease diagnosis: a comparative study and a novel deep-ensemble method. *Computers in biology and medicine*, 141:105032, 2022.

[8] E.-G. Marwa, H. E.-D. Moustafa, F. Khalifa, H. Khater, and E. AbdElhalim. An mri-based deep learning approach for accurate detection of alzheimer's disease. *Alexandria Engineering Journal*, 63:211–221, 2023.

[9] M. Parizkova, O. Lerch, R. Andel, J. Kalinova, H. Markova, M. Vyhnalek, J. Hort, and J. Laczó. Spatial pattern separation in early alzheimer's disease. *Journal of Alzheimer's Disease*, 76(1):121–138, 2020.

[10] C. Reitz, E. Rogaeva, and G. W. Beecham. Late-onset vs nonmendelian early-onset alzheimer disease: A distinction without a difference? *Neurology: Genetics*, 6(5):e512, 2020.

[11] S. U. Sadat, H. H. Shomee, A. Awwal, S. N. Amin, M. T. Reza, and M. Z. Parvez. Alzheimer's disease detection and classification using transfer learning technique and ensemble on convolutional neural networks. In *2021 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 1478–1481. IEEE, 2021.

[12] F. Salami, A. Bozorgi-Amiri, G. M. Hassan, R. Tavakkoli-Moghaddam, and A. Datta. Designing a clinical decision support system for alzheimer's diagnosis on oasis-3 data set. *Biomedical Signal Processing and Control*, 74:103527, 2022.

[13] C. L. Saratxaga, I. Moya, A. Picón, M. Acosta, A. Moreno-Fernandez-de Leceta, E. Garrote, and A. Bereciartua-Perez. Mri deep learning-based solution for alzheimer's disease prediction. *Journal of personalized medicine*, 11(9):902, 2021.

[14] M. Talo, O. Yildirim, U. B. Baloglu, G. Aydin, and U. R. Acharya. Convolutional neural networks for multi-class brain disease detection using mri images. *Computerized Medical Imaging and Graphics*, 78:101673, 2019.

[15] T. Uemura, J. J. Näppi, T. Hironaka, H. Kim, and H. Yoshida. Comparative performance of 3d-densenet, 3d-resnet, and 3d-vgg models in polyp detection for ct colonography. In *Medical Imaging 2020: Computer-Aided Diagnosis*, volume 11314, pages 736–741. SPIE, 2020.

[16] J. Weller and A. Budson. Current understanding of alzheimer's disease diagnosis and treatment. *F1000Research*, 7, 2018.

[17] E. Yagis, L. Citi, S. Diciotti, C. Marzi, S. W. Atnafu, and A. G. S. De Herrera. 3d convolutional neural networks for diagnosis of alzheimer's disease via structural

mri. In *2020 IEEE 33rd International Symposium on Computer-Based Medical Systems (CBMS)*, pages 65–70. IEEE, 2020.

[18] E. Yagis, A. G. S. De Herrera, and L. Citi. Generalization performance of deep learning models in neurodegenerative disease classification. In *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1692–1698, 2019.