

Adam Yin, Brayden Ritter, Henju Duvenhage, Jezryl Austria

December 8th 2025

BME 677 Final Project:

Deep Learning Approaches for MRI-Based Dementia Classification:  
A Comparative Analysis of VGG19, ResNet, and U-Net-Based Image Classification

Professor Maral Aminpour

## **Abstract**

Dementia, particularly Alzheimer's disease (AD), is a progressive neurodegenerative condition for which early detection is essential but remains challenging due to subtle anatomical changes in early stages. This project investigates deep-learning methods for MRI-based classification of dementia severity by comparing three convolutional neural network architectures: VGG19, ResNet50, and a U-Net-based classifier. T1-weighted MRI slices from the OASIS dataset were preprocessed through cropping, resizing, and three-channel conversion. The Moderate Dementia class was removed due to insufficient patient representation, and patient-wise splitting was used to prevent data leakage. VGG19 and ResNet50 were fine-tuned using transfer learning, while U-Net was trained end-to-end. All models were trained with the Adam optimizer, categorical cross-entropy loss, early stopping, and learning rate scheduling. Performance was assessed using accuracy, confusion matrices, and precision-recall-F1 metrics. Results showed modest performance across all models. ResNet50 demonstrated the most balanced classification with a test accuracy of 0.56, while U-Net achieved strong recall for Mild Dementia but displayed unstable validation behavior. VGG19 showed stable convergence but struggled to capture subtle disease-related features. All models showed difficulty distinguishing Very Mild Dementia from Non-Demented cases, reflecting both dataset imbalance and the gradual nature of early AD progression. To conclude, the findings highlight both the potential and the limitations of 2D CNN-based dementia classification. Future improvements may include larger and more balanced datasets, standardized MRI preprocessing, stronger regularization, and 3D CNN architectures.

## **Introduction**

The World Health Organization (WHO) determined that in 2021 more than 57 million people had dementia (formally Major Neurocognitive Disorder) worldwide, with a majority of these individuals living in low to middle income countries [1]. The prevalence of dementia is predicted to rise significantly in the following decades, becoming the second most common cause of death in 2040 worldwide, surpassing diseases such as lung cancer [2]. This is accompanied by more than 10 million new cases per year, costing countries billions of dollars in covering care providers' costs. The main underlying condition causing dementia is Alzheimer's disease, with over 70% of cases being caused by this disease [3]. The condition is marked by a decline in neurological function (synaptic and loss of neural connection) due to atrophy of the cerebral cortex. Key proteins (Amyloid  $\beta$  and tau proteins) that are misfolded deposit themselves close to neural cells, which leads to a series of cascading factors causing cognitive decline [4].

Although there is no cure for Alzheimer's disease at the moment, Disease Modifying Therapies (DMT), symptomatic approaches and clinical trials - such as stem cell treatments and non-pharmacological brain stimulation - are being conducted to address the need for a permanent solution to the condition. Symptomatic treatments also provide a comprehensive approach of care, where tailored support from medical practitioners, clinical caregivers, and pharmacological interventions from FDA approved sources are administered [5]. While these treatments can increase lifespan and ensure a better quality of life, only the symptoms are treated as opposed to managing the origin or understanding the mechanism of the disease. Preventative diagnosis, before patients cross the threshold into dementia, presents an important opportunity for early protective factors as well as intervention with disease-modifying therapy [6].

Current literature points towards a shift from traditional machine learning methods to deep learning methods due to their abilities of higher level image processing [7][8]. Machine learning approaches, such as using neural networks used to analyze MRI images and predict the level of dementia a patient presents with, also called MCI-to-AD conversion, is one of the leading approaches for early diagnosis of AD [9]. This approach focuses on detecting and predicting Alzheimer's disease severity based on differences in affected brain regions, such as atrophy in the hippocampus, entorhinal cortex and amygdala due to widespread cell death. Other work has adapted pretrained models such as VGG and ResNet for MRI-2D-slice-based AD detection, relying on transfer learning to distinguish healthy from diseased states, and shows evidence of high accuracy observed in binary classification problems [10] [11]. More recent approaches have explored segmentation-based architectures such as U-Net, which has the benefit of preserving spatial detail and has shown promising performance for AD diagnosis [12]. However, much of this literature still focuses on coarse disease contrasts and limited combinations of disease stages, with a noticeable drop in performance when adjacent clinical categories on the dementia spectrum are included in a multiclass classification task, as opposed to a binary classification task [13][14]. In addition, many pipelines rely on heavily processed MRI scans that are normalized to a brain atlas [15], which can blur subtle regional atrophy patterns that are critical for early-stage diagnosis. We aim to not use standardized principles, such as normalization, as finding an approach that does not utilize a brain atlas, but that still produces accurate results, could prove to be a significant leap in AD classification using CNN models.

Many machine learning models are well-suited to tasks of this nature, but finding the optimal solution also requires fine tuning the hyperparameters available within each of these models, leaving nearly infinite combinations to be tested. With regards to current work on early dementia classification using MRI images, Shamrat et al (2023) compared five deep learning models (including ResNet50 and VGG16) and modified the highest accuracy model using a CNN classifier called AlzheimerNet [16]. A 98% accuracy was achieved by using 60000 images from the ADNI database, including preprocessing methods such as data augmentation and varying post-processing metrics included. Yang et al. (2018) also trained two main models, 3D-VGGNet and 3D-ResNet, to classify AD from brain MRI scans. [17]

The objective of our study focussed on building, training and comparing different deep learning models to detect early stage Alzheimer's disease by classifying MRI images into three different dementia stages, namely Non-Dementia, Mild-Dementia and Very Mild Dementia. This includes handling class imbalances from the OASIS MRI dataset [18] and implementing the pretrained models to generalize well to unseen data. Based on the literature search, we aim to implement a comparative analysis of three different deep learning models, namely U-Net [19], VGG19 [20] and ResNet50 [21].

## **Materials and Methods**

### **Exploratory Data Analysis**

The dataset originally consisted of four diagnostic categories: Non-Demented, Very Mild Demented, Mild Demented, and Moderate Demented. However, the Moderate Demented class contained imaging data from only two unique patients, making it scientifically unsuitable for reliable model training. Although extensive data augmentation could artificially expand this class, such synthetic enlargement would simply replicate variations of the same two subjects rather than provide meaningful

biological variability. This would lead to overfitting, inflated accuracy, and degraded generalizability. Therefore, the Moderate Demented class was removed, and the final dataset included three classes: Non-Demented, Very Mild Demented, Mild Demented.

This decision prioritizes data reliability and scientific validity over dataset completeness, which is considered best practice for sensitive medical imaging applications. Recent large-scale reviews on medical machine learning emphasize that classes with extremely limited subject representation introduce significant class noise, sampling bias, and instability in deep learning models, ultimately leading to poor real-world generalization despite high apparent accuracy [22]. As emphasized in contemporary medical AI evaluation frameworks, maintaining balanced and biologically representative class distributions is critical for producing trustworthy and clinically meaningful models. Consequently, removing the Moderate Demented class reduced class noise, improved model stability, and established a reliable baseline for future expansion when sufficiently diverse data became available.

#### Dataset Distribution (Removing Moderate Dementia)

Class	Total Patients	Total Images	Notes
Non-Demented	266	67,200	Largest Class
Mild Dementia	21	5,002	
Very Mild Dementia	58	13,700	
Moderate Dementia	2	488	Removed due to insufficient variance

Table 1: Raw data distributions from the Oasis dataset for the 4 classes, including the number of patients and images.

To ensure compatibility with pretrained convolutional neural networks, all MRI slices originally stored as single-channel grayscale images, were converted to three-channel RGB format. Although the RGB channels contain identical information (since they are replicated grayscale values), this conversion allows the use of architectures such as VGG19 and ResNet50, which expect 3-channel inputs.

All images were cropped to remove large black background regions and retain only the relevant brain anatomy. They were then resized to  $240 \times 240$  pixels and saved in JPEG format for consistent input dimensions during training. This preprocessing step improves model focus on anatomical features and reduces unnecessary pixel variation caused by background noise.

#### Train–Validation–Test Split

To prevent information leakage across model development stages, the dataset was partitioned at the patient level rather than the image level. Each MRI scan contains thousands of slices, and for this study, approximately 5,000–5,500 images were randomly selected from each patient. This standardized

and randomized extraction volume ensured a balanced dataset across classes, preventing situations where classes with more MRI slices would dominate training and bias the model.

Equally important, all slices belonging to the same patient were kept within a single subset. No patient contributed images to more than one of the train, validation, or test sets. This is crucial in medical imaging: if slices from the same patient appeared in multiple subsets, the model could learn patient-specific anatomical patterns rather than disease-related features, artificially inflating accuracy.

After removing the Moderately Demented class, the remaining patients were divided into:

- 70% Training patients
- 15% Validation patients
- 15% Testing patients

This strategy directly follows the findings of Yagis, E. (2021) [23], who demonstrated that slice-level splitting of volumetric brain MRI data leads to severe data leakage and artificially inflated classification accuracy by as much as 30–55% across multiple neuroimaging datasets. Their controlled experiments further showed that even randomly labeled MRI data could yield near-perfect accuracy when improper slice-level splitting was used, highlighting the critical risk of identity confounding. By enforcing strict patient-wise separation in our dataset, our study avoids this methodological pitfall and ensures that the reported performance reflects true disease-related generalization rather than subject-specific anatomical memorization.

All extracted slices from each patient (5,000–5,500 images) were saved in separate train/validation/test directories, ensuring strict isolation and eliminating data leakage.

The resulting distribution is summarized below:

Class	Train patients	Validation Patients	Test Patients	Total Images
Non-Demented	14	3	4	5,002
Very Mild Dementia	14	3	4	5,063
Mild Dementia	14	3	4	5,002

Table 2: Final distribution of the train, validation and test splits from the 21 patients per class.

## Architecture of Models

Our study used VGG19, ResNet50, and U-Net-based image classifiers to classify MRI images into different demented groups. VGG19 is a well-established deep CNN architecture that builds depth through repeated 3x3 convolutions and max-pooling layers. Previous work by Antony et al. (2023) showed that VGG19 not only achieves higher accuracy than VGG16 (~84% vs. ~81%) but also performs well with limited medical imaging data due to its pretrained feature extractor [24]. Therefore, our project uses a pretrained VGG19 model to avoid the need for extensive dataset collection, since training deep

CNNs from scratch would be extremely time-consuming and computationally expensive. Structurally, VGG19 consists of 19 weight layers organized into five convolutional blocks, each containing stacked 3x3 convolutions followed by max-pooling, and ends with three fully connected layers [17]. For adaptation to our classification task, the original VGG19 classification head composed of three fully connected layers trained on ImageNet was removed and replaced with a custom fully connected classifier consisting of a global average pooling layer, a dense layer for feature learning, a dropout layer for regularization, and a final softmax output layer matching the number of dementia classes. The convolutional base of VGG19 was retained as a pretrained feature extractor to leverage transfer learning and improve performance under limited data conditions.

ResNet50 is a deep residual network designed to overcome the degradation problem in very deep CNNs by using skip connections, which allow the model to learn residual mappings and improve gradient flow. It contains 50 layers, beginning with a 7x7 convolution, followed by 48 convolutional layers inside bottleneck blocks, and a final fully connected layer [25]. Moreover, ResNet50 has demonstrated strong performance on Alzheimer's MRI data when properly trained [26] [27]. We included ResNet50 in this study because its deeper residual architecture enables the extraction of more abstract and progressively richer feature representations compared with classical CNN architectures. By leveraging identity shortcut connections, ResNet50 mitigates the vanishing gradient problem and enables stable training of very deep networks. Comparing ResNet50 with VGG19 allows us to evaluate whether a modern residual network outperforms a traditional deep CNN for MRI-based dementia classification. A pretrained ResNet50 model (ImageNet) was adopted using a transfer learning strategy to reduce training time and computational cost. The original ImageNet classification head was removed and replaced with a custom classifier consisting of a global average pooling layer, a fully connected dense layer, a dropout layer for regularization, and a final softmax output layer matching the number of dementia classes. The convolutional backbone was retained as a pretrained feature extractor, and MRI images were resized and converted to three-channel format to match the network input requirements.

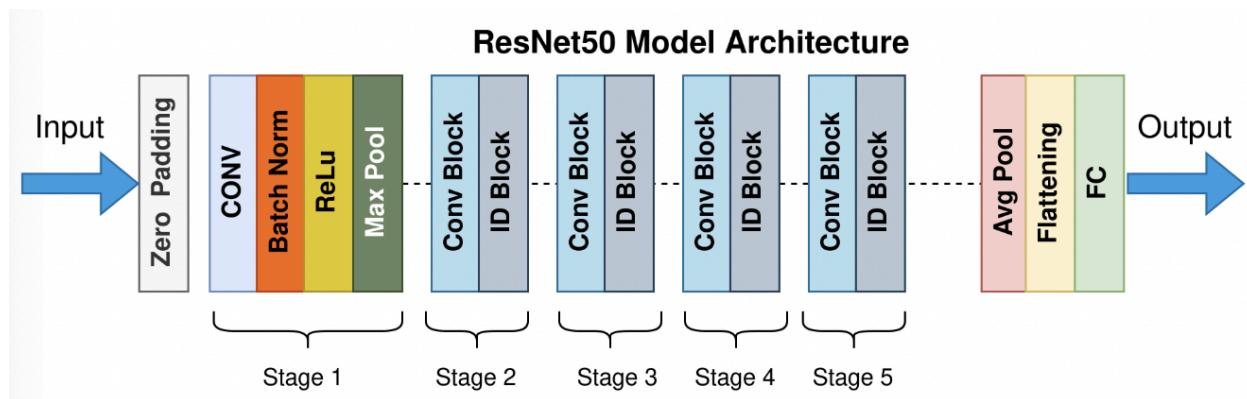


Figure 1. ResNet50 model architecture. [25]

The U-Net-based image classifier is a convolutional encoder-decoder architecture with symmetric downsampling and upsampling paths [28]. The encoder consists of repeated 3x3 convolutions followed by ReLU activation and 2x2 max-pooling, which progressively reduce the spatial resolution while increasing the number of feature channels. This allows the network to capture hierarchical contextual information. At the deepest point of the architecture, the bottleneck layer encodes the most abstract representation of the MRI image with the highest level of feature complexity.

The decoder mirrors the encoder by using upsampling (via transpose convolutions) and 3x3 convolutions to gradually recover spatial resolution while reducing the number of feature channels. Skip connections concatenate feature maps from encoder to the corresponding decoder layers, ensuring that fine-grained spatial information is preserved throughout reconstruction. The network concludes with a final 1x1 convolution in the output layer.

A U-Net-based architecture was implemented as an image-level classifier in this study to leverage its strong spatial feature preservation capability for MRI data. The encoder-decoder structure with skip connections was used to extract both low-level anatomical details and high-level contextual features. The original segmentation output layer was removed and replaced with a custom classification head consisting of a global average pooling layer, a fully connected dense layer for feature learning, a dropout layer for regularization, and a final softmax output layer matching the number of dementia classes. MRI images were resized and normalized prior to training to ensure compatibility with the network input and numerical stability. This configuration enabled U-Net to function as a classifier while preserving its spatial representation advantages for MRI-based dementia classification.

## Training and Optimization

All models in this project were trained using a consistent training pipeline to ensure fair comparison across architectures. A batch size of 16 was used to balance between GPU memory constraints with stable gradient updates, and smaller batches have been shown to improve generalization performance on limited medical imaging datasets [29] [30]. To optimize model weights, categorical cross-entropy was used as the loss function for the three-class dementia classification task (Non-Demented, Very Mild Demented, Mild Demented), and the Adam optimizer was selected due to its adaptive learning rates and stable convergence when fine-tuning pretrained CNNs. Weight decay ( $1\times10^{-4}$ ) was applied as L2 regularization to reduce overfitting.

Both the VGG19 and ResNet50 classifiers followed a two-phase transfer-learning strategy. In Phase 1, most pretrained backbone layers were frozen and only the final classification layers were trained using a learning rate of  $1\times10^{-4}$ . This allowed the classifier to adapt to MRI-specific features without disrupting the pretrained weights. In Phase 2, all layers were unfrozen for fine-tuning using a smaller learning rate ( $1\times10^{-5}$ ), enabling deeper feature refinement relevant to dementia-related structural differences in the MRI images of the brain.

Additionally, A ReduceLROnPlateau scheduler automatically decreased the learning rate when validation loss plateaued, ensuring efficient convergence and preventing performance stagnation. EarlyStopping with a patience of three to five epochs was used to avoid overfitting by terminating training once validation performance stopped improving. The model checkpoint with the lowest

validation loss was retained as the final model for evaluation. This unified training strategy ensures fair comparison across VGG19 and ResNet50 while effectively leveraging transfer learning to adapt pretrained CNNs for MRI-based dementia classification.

Unlike VGG19 and ResNet50, the U-Net-based image classifier was trained end-to-end because it does not rely on a pretrained ImageNet backbone. The model was optimized using Adam with an initial learning rate of  $1 \times 10^{-4}$  and the same categorical cross-entropy loss as the other architectures. ReduceLROnPlateau was applied to adjust the learning rate dynamically when validation loss plateaued, and EarlyStopping with a patience of five epochs prevented overfitting. Additionally, ModelCheckpoint was used to save the best-performing U-Net weights. This training strategy allowed the encoder-decoder architecture to learn both global and fine-grained structural features from MRI imaging data without requiring a transfer-learning pretrained phase.

### **Evaluation of Performance: Accuracy, Confusion Matrix, train-val loss accuracy curves, classification report**

To evaluate the performance of all three models in more detail, we used test accuracy and loss, confusion matrix, training and validation curves, and classification report to have a comprehensive understanding of classification quality across the three dementia categories (Non-Demented, Very Mild Demented, Mild Demented).

#### **Test Accuracy and Loss**

Each model was evaluated on the independent test dataset using its final saved weights. Test accuracy reflects the proportion of correctly classified MRI images, while test loss indicates overall fit of the model to unseen data. All predictions were obtained from the softmax output using the maximum probability class.

#### **Confusion Matrix**

To better understand classification behavior across categories, we computed confusion matrices by comparing predicted labels with actual labels. This is because confusion matrix highlights not only correct predictions but also patterns of misclassification, enabling us to improve explainability to the performance of the three CNN models. Matrices were visualized using heatmaps for all models.

#### **Training and Validation Curves**

Training history was plotted for each model, including training vs. validation loss and accuracy across epochs. These curves help recognize model behavior during optimization, such as overfitting, underfitting, and stable convergence. Moreover, in the U-Net-based image classifier, these curves also indicate the effects of early stopping and learning rate scheduling on convergence stability.

#### **Classification Report**

Precision, recall, and F1 scores were computed for each class in the classification report. Precision indicates the reliability of positive predictions, while recall measures the proportion of actual

samples correctly identified. F1 scores balance both features to have a more comprehensive evaluation. Macro and weighted averaged scores were also examined to account for potential class imbalance in the dataset.

## **Results and Discussion**

### **ResNet50**

Figure 2 below shows the classification report, confusion matrix and losses for the ResNet50 model. The metrics shown in Figure 2a identifies the Non-Demented class having the highest performance across the board, where the recall (0.73) showed that the model identified healthy individuals better than patients with the disease. The Mild Dementia classification had moderate performance, and could only partially separate this specific class from the others, where the Very Mild Dementia class had a poor performance (recall of 0.27) that aligns with struggling to identify early or subtle dementia cases. Overall accuracy of the model stands at 0.56.

The confusion matrix (figure 2b) also verifies that the Non-Demented class was classified correctly most often (715) due to the advantage of the majority class, however the Mild and Very Mild cases had significant confusion with the other classes. An alarming indicator shows a significant part of the images predicted as Non-Dementia while the actual classification is Very Mild Dementia, showing a false negative or Type II error that the model undergoes. Implications such as delayed treatments and undetected risk could take place, which nullifies the objectives of our models.

Figure 2c shows a gradual, rapid decline in the training losses with an increase in the number of epochs, however, the validation losses remain fluctuating at 1.2, which is a prime example of the model overfitting due to the losses diverging. This is indicative of the data being memorized as opposed to the model generalizing to new data. Training and validation accuracy shows similar understanding, where the training accuracy reaches almost perfect accuracy at 0.95, but plateaus at 8 epochs for the validation accuracy around 0.7. Increasing the number of epochs does not improve the performance, and the model capacity exceeds what the small dataset can support. Additional regularization techniques could be explored. Additionally, class weights were modified in attempts to give more balance to the datasets, and while theoretically adjusting the weights towards the less significant classes should increase the overall accuracy, this was not the case for our implemented model.

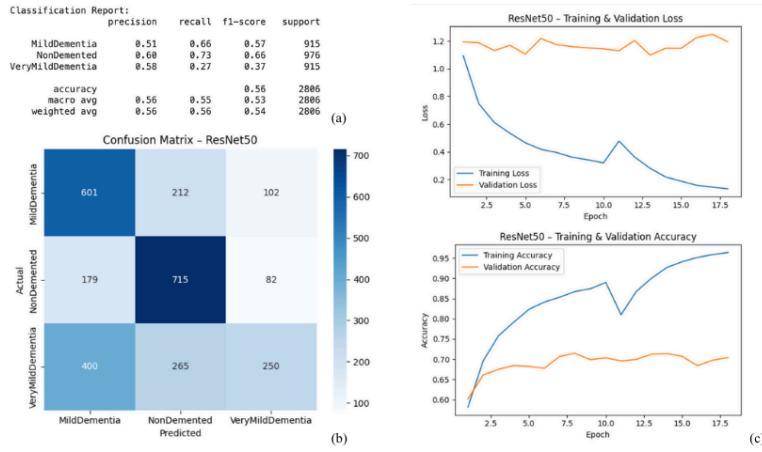


Figure 2. Metrics that were extracted from the ResNet50 model to evaluate its performance. (a) Classification report denoting the Precision, Recall, f1-score and support for all three classes, as well as weighted and macro averages. (b) Confusion matrix defining the performance of classification based on severity. (c) Graphs depicting the training and validation losses (top) and training and validation accuracies (bottom) over a range of epochs.

## U-Net

The U-net classifier achieved an overall accuracy of 0.55, on the test set, where performance substantially varied across class distributions. Figure 3a indicates Mild Dementia achieving a high recall of 0.84, indicating a majority correct classification of the indicated class. In contrast, Very Mild Dementia had considerably lower recall which could again suggest the difficulty in identifying subtle differences in early stage dementia. Non-Demented class performance was moderate (0.48).

Aforementioned trends were confirmed in the confusion matrix (Figure 2b), where correct classifications for the Mild Dementia class were most prominent. Notably, this model decreased the amount of false negatives in comparison to the ResNet50 Model when referring to the Mild Dementia and Non-Demented classes. The Very Mild Dementia samples were still frequently misclassified as Non-Demented or Very Mild Dementia, which reinforces the idea that even with the U-Net models increased sensitivity to precise boundary detections in medical images, the skewed data distributions still had a profound effect on the classification.

Figure 2c highlights the training and validation losses for the U-Net model, where training losses steadily decreased as the number of epochs increased. High instability in the validation losses were observed, fluctuating between 2.5 to 5. This could point to multiple limitations, including an unoptimized learning rate, small batch sizes or noisy, unshuffled data. With similar patterns in the training and validation accuracies, overfitting is a substantial notifier from the model's performance. U-net thrives on segmentation and with a large number of parameters, the smaller, imbalanced dataset could amplify the errors that the model connects to. This also calls for increased regularization, combination of model architectures or refined preprocessing for better metrics in addition to identifying balanced datasets.

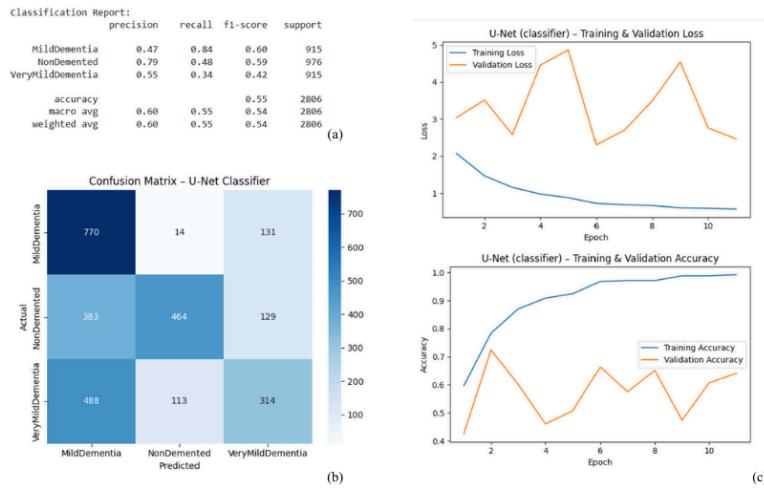


Figure 3. Metrics that were extracted from the U-Net model to evaluate its performance. (a) Classification report denoting the Precision, Recall, f1-score and support for all three classes, as well as weighted and macro averages. (b) Confusion matrix defining the performance of classification based on severity. (c) Graphs depicting the training and validation losses (top) and training and validation accuracies (bottom) over a range of epochs.

## VGG19

VGG19's plain architecture, in comparison to U-Net and ResNet-50 has drawbacks when looking at the extraction of multi-scale features relevant to early Dementia classification. An overall accuracy of 0.5 with varying performance across classes was observed for the VGG19 model. Figure 4a lists a high recall for the Non-Demented class, similarly to the ResNet50 model, the model does better to predict the aforementioned class accurately (716) in Figure 4b. It still struggles with Very Mild and Mild Dementia precision, where only 0.48 and 0.56 were respectively reported. This emphasizes a lack of reliability for this CNN, where a mere 0.35 recall for the Mild Dementia class was observed.

An increased stability in the losses and accuracies with varying epochs was indicated in Figure 4c, when comparing to the previous models, which is indicative of a more gradual, stable learning process. Still, a divergence occurs with the losses, where the validation loss reaches 1. The validation accuracy stabilizes at around 0.6 while the training accuracy increases to around 0.7. Overfitting is moderate in comparison with the other models. Despite the successful optimization on natural images using ImageNet, the VGG19's simple stacked convolutional structure shows decreased adaptability to a medical context. The rigid design makes it less flexible - which is not ideal for grayscale adaptation from its RGB training.

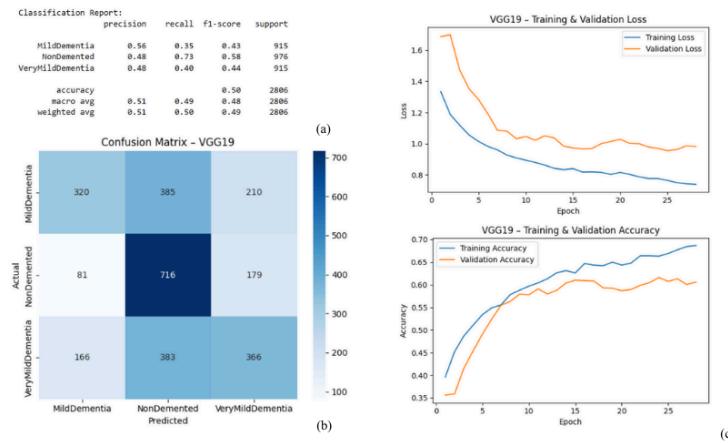


Figure 4. Metrics that were extracted from the VGG19 model to evaluate its performance. (a) Classification report denoting the Precision, Recall, f1-score and support for all three classes, as well as weighted and macro averages. (b) Confusion matrix defining the performance of classification based on severity. (c) Graphs depicting the training and validation losses (top) and training and validation accuracies (bottom) over a range of epochs.

Overall, Resnet50 emerged as the most balanced performer, with U-Net excelling on the Mild Dementia recognition, but lacking in consistency. The VGG19 struggled with subtle pathology. When implementing these 3 models to balance architectural design, computational resource optimization and overall performance, multiscale feature learning for heavily imbalanced, smaller datasets require additional tuning to ensure comparable metrics for early stage dementia detection.

## Conclusion

Misclassification happens primarily between adjacent clinical stages. A reason for the difficulty in classifying adjacent severities could look at early dementia severity as a continuum rather than sharp, categorical boundaries, hence determining classes as non, mild, very mild and moderate dementia imposes rigid labels on what is biologically a gradual spectrum of hippocampal degradation. These subtle differences between boundaries almost certainly imposed challenges in image classification tasks, such as the ones given to the models we selected. This challenge is likely at its peak when attempting to decide between classifying patients into the very mild and mild dementia groups.

Within a broader context, the three CNN models evaluated in this study demonstrated modest performance. ResNet50 achieved the most balanced results across classes indicated overfitting driven by limited patient diversity. The U-Net classifier, while proving more effective for identifying mild dementia, showed unstable validation behaviour likely indicating its high parameter count relative to the datasets size. VGG19 displayed more stable convergence but lacked the multi-scale representational capacity required to detect subtle anatomical changes in the MRI images provided. Collectively these outcomes illustrate the inherent difficulty of using 2D CNN models to classify Alzheimer's disease from individual

MRI slice images. The differences between classifications in Alzheimer's disease are known to be highly-patient dependent and may not always be visually distinct without deeper knowledge in the field.

Several limitations of the dataset and preprocessing pipeline shaped these results. The small number of unique patients given the complexity of the disease, combined with notable class imbalances constrained the models ability to generalize beyond large scale patterns. The inherent challenge with MRI imaging data is that not everyone's brain is built the same, so in MRI whole-brain comparison studies, slices are typically downscaled and warped onto a brain atlas to increase generalization across patients, this is challenging to implement in preprocessing without immense resources, however [31]. Future work could benefit from following standardized preprocessing methods used in MRI whole-brain image classification studies, implementing increased regularization, using more balanced datasets, or even implementing 3D CNN architectures.

Despite the challenges faced, the project provides a valuable foundation for understanding how different CNN architectures interpret MRI data across the Alzheimer's disease spectrum. The insights gained here highlight both the promise and the limitations of deep learning for early dementia detection and suggest clear avenues for future methodological refinement.

## References

- [1] WHO, "Dementia - Fact Sheets," World Health Organization. Accessed: Nov. 24, 2025. [Online]. Available: <https://www.who.int/news-room/fact-sheets/detail/dementia>
- [2] E. Nichols *et al.*, "Estimation of the global prevalence of dementia in 2019 and forecasted prevalence in 2050: an analysis for the Global Burden of Disease Study 2019," *Lancet Public Health*, vol. 7, no. 2, pp. e105–e125, Feb. 2022, doi: 10.1016/S2468-2667(21)00249-8.
- [3] P. D. Emmady, C. Schoo, and P. Tadi, "Major Neurocognitive Disorder (Dementia)," *StatPearls*, Nov. 2022, Accessed: Dec. 07, 2025. [Online]. Available: <https://www.ncbi.nlm.nih.gov/books/NBK557444/>
- [4] D. J. Selkoe and J. Hardy, "The amyloid hypothesis of Alzheimer's disease at 25 years," *EMBO Molecular Medicine* 2016 8:6, vol. 8, no. 6, pp. 595–608, Mar. 2016, doi: 10.15252/EMMM.201606210.
- [5] K. G. Yiannopoulou and S. G. Papageorgiou, "Current and Future Treatments in Alzheimer Disease: An Update," *J Cent Nerv Syst Dis*, vol. 12, p. 1179573520907397, 2020, doi: 10.1177/1179573520907397.
- [6] B. Dubois, A. Padovanib, P. Scheltensc, A. Rossid, and G. D. Agnello, "Timely diagnosis for alzheimer's disease: A literature review on benefits and challenges," *Journal of Alzheimer's Disease*, vol. 49, no. 3, pp. 617–631, 2015, doi: 10.3233/JAD-150692;CTYPE:STRING:JOURNAL.
- [7] S. Modak, E. Abdel-Raheem, and L. Rueda, "Applications of deep learning in disease diagnosis of chest radiographs: A survey on materials and methods," *Biomedical Engineering Advances*, vol. 5, p. 100076, Jun. 2023, doi: 10.1016/J.BEA.2023.100076.
- [8] T. Islam *et al.*, "COMPARATIVE ANALYSIS OF NEURAL NETWORK ARCHITECTURES FOR MEDICAL IMAGE CLASSIFICATION: EVALUATING PERFORMANCE ACROSS DIVERSE MODELS," *American Journal of Advanced Technology and Engineering Solutions*, vol. 4, no. 01, pp. 01–42, Apr. 2024, doi: 10.63125/FEED1X52.
- [9] W. Lin *et al.*, "Convolutional neural networks-based MRI image analysis for the Alzheimer's disease prediction from mild cognitive impairment," *Front Neurosci*, vol. 12, no. NOV, p. 412254, Nov. 2018, doi: 10.3389/FNINS.2018.00777/BIBTEX.

- [10] D. AlSaeed, S. F. Omar, D. AlSaeed, and S. F. Omar, “Brain MRI Analysis for Alzheimer’s Disease Diagnosis Using CNN-Based Feature Extraction and Machine Learning,” *Sensors* 2022, Vol. 22, vol. 22, no. 8, Apr. 2022, doi: 10.3390/S22082911.
- [11] M. U. Ali, S. J. Hussain, M. Khalid, M. Farrash, H. F. M. Lahza, and A. Zafar, “MRI-Driven Alzheimer’s Disease Diagnosis Using Deep Network Fusion and Optimal Selection of Feature,” *Bioengineering (Basel)*, vol. 11, no. 11, Nov. 2024, doi: 10.3390/BIOENGINEERING11111076.
- [12] Z. Fan *et al.*, “U-net based analysis of MRI for Alzheimer’s disease diagnosis,” *Neural Computing and Applications* 2021 33:20, vol. 33, no. 20, pp. 13587–13599, Apr. 2021, doi: 10.1007/S00521-021-05983-Y.
- [13] F. Ramzan *et al.*, “A Deep Learning Approach for Automated Diagnosis and Multi-Class Classification of Alzheimer’s Disease Stages Using Resting-State fMRI and Residual Neural Networks,” *Journal of Medical Systems* 2019 44:2, vol. 44, no. 2, pp. 37-, Dec. 2019, doi: 10.1007/S10916-019-1475-2.
- [14] S. B. Shahid *et al.*, “Novel deep learning for multi-class classification of Alzheimer’s in disability using MRI datasets,” *Frontiers in Bioinformatics*, vol. 5, p. 1567219, Aug. 2025, doi: 10.3389/FBINF.2025.1567219/BIBTEX.
- [15] A. Casamitjana *et al.*, “A probabilistic histological atlas of the human brain for MRI segmentation,” *Nature* 2025, pp. 1–8, Nov. 2025, doi: 10.1038/s41586-025-09708-2.
- [16] F. M. J. M. Shamrat *et al.*, “AlzheimerNet: An Effective Deep Learning Based Proposition for Alzheimer’s Disease Stages Classification From Functional Brain Changes in Magnetic Resonance Images,” *IEEE Access*, vol. 11, pp. 16376–16395, 2023, doi: 10.1109/ACCESS.2023.3244952.
- [17] C. Yang, A. Rangarajan, and S. Ranka, “Visual Explanations From Deep 3D Convolutional Neural Networks for Alzheimer’s Disease Classification,” *AMIA Annu Symp Proc*, vol. 2018, pp. 1571–1580, Mar. 2018, Accessed: Dec. 07, 2025. [Online]. Available: <https://arxiv.org/pdf/1803.02544>
- [18] 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, doi: 10.1162/JOCN.2007.19.9.1498.
- [19] W. Weng and X. Zhu, “INet: Convolutional Networks for Biomedical Image Segmentation,” *IEEE Access*, vol. 9, pp. 16591–16603, 2021, doi: 10.1109/ACCESS.2021.3053408.
- [20] K. Simonyan and A. Zisserman, “Very Deep Convolutional Networks for Large-Scale Image Recognition,” *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, Sep. 2014, Accessed: Dec. 07, 2025. [Online]. Available: <https://arxiv.org/pdf/1409.1556>
- [21] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2016-December, pp. 770–778, Dec. 2016, doi: 10.1109/CVPR.2016.90.
- [22] M. Salmi, D. Atif, D. Oliva, A. Abraham, and S. Ventura, “Handling imbalanced medical datasets: review of a decade of research,” *Artificial Intelligence Review* 2024 57:10, vol. 57, no. 10, pp. 273-, Sep. 2024, doi: 10.1007/S10462-024-10884-2.
- [23] E. Yagis *et al.*, “Effect of data leakage in brain MRI classification using 2D convolutional neural networks,” *Sci Rep*, vol. 11, no. 1, p. 22544, Dec. 2021, doi: 10.1038/S41598-021-01681-W.
- [24] F. Antony, H. B. Anita, and J. A. George, “Classification on Alzheimer’s Disease MRI Images with VGG-16 and VGG-19,” *Smart Innovation, Systems and Technologies*, vol. 312, pp. 199–207, 2023, doi: 10.1007/978-981-19-3575-6\_22/TABLES/3.

- [25] “The Annotated ResNet-50 | Towards Data Science.” Accessed: Dec. 07, 2025. [Online]. Available: <https://towardsdatascience.com/the-annotated-resnet-50-a6c536034758/>
- [26] Y. Zhang, S. Wang, P. Phillips, J. Yang, and T. F. Yuan, “Three-Dimensional Eigenbrain for the Detection of Subjects and Brain Regions Related with Alzheimer’s Disease,” *Journal of Alzheimer’s Disease*, vol. 50, no. 4, pp. 1163–1179, Feb. 2016, doi: 10.3233/JAD-150988;WGROUP:STRING:PUBLICATION.
- [27] D. Bansal, R. Chhikara, K. Khanna, and P. Gupta, “Comparative Analysis of Various Machine Learning Algorithms for Detecting Dementia,” *Procedia Comput Sci*, vol. 132, pp. 1497–1502, Jan. 2018, doi: 10.1016/J.PROCS.2018.05.102.
- [28] Z. Fan *et al.*, “U-net based analysis of MRI for Alzheimer’s disease diagnosis,” *Neural Computing and Applications* 2021 33:20, vol. 33, no. 20, pp. 13587–13599, Apr. 2021, doi: 10.1007/S00521-021-05983-Y.
- [29] D. Masters and C. Luschi, “Revisiting Small Batch Training for Deep Neural Networks,” Apr. 2018, Accessed: Dec. 07, 2025. [Online]. Available: <https://arxiv.org/pdf/1804.07612>
- [30] M. Salmi, D. Atif, D. Oliva, A. Abraham, and S. Ventura, “Handling imbalanced medical datasets: review of a decade of research,” *Artificial Intelligence Review* 2024 57:10, vol. 57, no. 10, pp. 273-, Sep. 2024, doi: 10.1007/S10462-024-10884-2.
- [31] A. Casamitjana *et al.*, “A probabilistic histological atlas of the human brain for MRI segmentation,” *Nature* 2025, pp. 1–8, Nov. 2025, doi: 10.1038/s41586-025-09708-2.

## **Author Contributions**

### **Adam Yin:**

- Conceptualization
- Formal analysis
- Investigation
- Methodology
- Software
- Validation
- Visualization
- Writing

### **Brayden Ritter:**

- Conceptualization
- Investigation
- Methodology
- Project administration
- Software
- Visualization
- Writing

### **Henju Duvenhage:**

- Conceptualization
- Investigation
- Methodology
- Project administration
- Software
- Visualization
- Writing

### **Jezryl Austria:**

- Conceptualization
- Data curation
- Investigation
- Methodology
- Software
- Supervision
- Writing

**Statement of AI use:**

We acknowledge the use of Chat-GPT 5.0 (<https://chatgpt.com/>). The use of generative AI was implemented in aiding the understanding of complex topics and concepts pertaining to the documentation above. For this proposal, feedback was implemented based on spell checking, concept review and general flow of the documentation. All writing, critical analysis and interpretations are our own. The AI tool did not generate or adjust data (interpret) experimental outcomes.