Predicting Alzheimer's Disease Using Structural MRIs: Activation Map Analysis of Memory Related Brain Regions

Rajdeep Singh, John Salloum, Atul Aravind Das

May 3, 2025

1 Abstract

Alzheimer's Disease (AD) is a progressive neurodegenerative disorder and the primary cause of dementia, characterized by memory loss and cognitive decline. Due to the lack of a cure, early detection is necessary. This project develops a deep learning classification model trained on 3D T1-weighted structural MRI scans that analyzes structural patterns in memory related regions of the brain in order to predict patient AD onset. Four models — Custom CNN, VGG-16, ResNet18 and ResNet50 — were evaluated using a dataset curated from the ADNI database. VGG-16 suffered from overfitting, and ResNet50 performed poorly on 3D data despite having solid results on 2D data. ResNet18 and the Custom CNN performed the best on the 3D MRI data, with validation accuracies near 93%.

Impact: By showing that deep learning models can reliably identify Alzheimer's stages from MRI scans, this study may help patients with this progressive neurodegenerative disease receive an earlier diagnosis and treatment.

2 Introduction

Alzheimer's Disease (AD) is a progressive neurodegenerative disorder which impairs cognitive functions – especially memory. [1] Globally, this disease affects millions, causing loss of dependence, diminished quality of life, and emotional and financial strain on caregivers and health systems. Currently there is no cure for this disorder, making early detection and preventive care the only possible approaches for health care providers.

Deep learning models, particularly convolutional neural networks (CNN), have shown promise in analyzing medical imaging data to detect subtle structural changes in the brain.[2] Specifically, in AD these models have shown success in using MRI scans to identify degradation in key memory related structures like the hippocampus and amygdala.

Our project seeks to develop a custom CNN which is capable of classifying MRI scans into three categories: cognitively normal (CN), mild cognitive impairment (MCI – the precursor to Alzheimer's) and Alzheimer's Disease (AD). Our project was carried out in the following order. We began by training CNNs (Custom CNN, ResNet50, VGG16) on 2D data to establish baseline performance. Following that, we conducted extensive dataset cleaning to only contain MRI slices that display key memory regions (hippocampus and amygdala). Finally the CNNs were adapted to process 3D-volumetric data to capture spatial information across slices and improve classification accuracy. Activation maps were generated for ResNet50 and VGG16 to visualize what regions of the structural

[1] Alzheimer's Association (n.d.)
[2] Li et al. (2019)

MRI's were impacting their predictions. The insights from these maps were used to inform potential improvements to the custom CNN architecture.

This project is just the starting point of a more comprehensive implementation of deep learning in AD Detection. Researchers are trying to create models that can analyze a cognitively normal patient's MRI scan and biomarker data to determine their risk of developing AD and predict the potential onset of the disease they may face.[3]

3 Related Works

Recent studies have demonstrated that the application of deep learning models on analyzing structural MRI (sMRI) data for Alzheimer Disease (AD) classification hold very promising results. In "Deep Learning Model for Prediction of Progressive Mild Cognitive Impairment to Alzheimer's Disease", the researchers compared the performance of two popular CNN architectures, ResNet-50 and VGG-16, to their custom CNN.[4] Their results showed that VGG-16 was able to achieve the highest classification performance with an accuracy of 78.57%, while ResNet-50 also achieved comparable levels of performance. The classification task performance of these two architectures is what prompted us to use these models as the baselines in our approach. Additionally, in "A Novel Method for Diagnosing Alzheimer's Disease from MRI Scans Using the ResNet50 Feature Extractor and the SVM Classifier", researchers proposed combining ResNet-50 for feature extraction with Support Vector Machine (SVM) for classification.[5] This method achieved some of the best results in literature yielding a validation accuracy of 99.52%. These papers confirmed that our approach is valid and that we should focus on using VGG-16 and ResNet-50 on both 2D and 3D data to establish baseline performance.

In AD research, multimodal learning is used to refer to the combining of structural imaging (like sMRI) and other diagnostic data such as clinical assessments or genetic data. Some research has been conducted and successfully classified patient onset by integrating sMRI with other biomarkers. However, multimodal learning drastically increases the task of data gathering and preprocessing, which is why it was ruled out for our purposes. Despite not being able to implement the multimodal approach, it provides insights into where deep learning in AD detection is heading.[6]

4 Approach/Methodology

4.1 Dataset Preparation

Initially, we implemented our CNN models using 2D data extracted from the midline slice of each patient's scan. This decision served two purposes. First, it allowed us to verify whether we correctly implemented the CNN architecture. Second, our dataset was not yet fully cleaned to contain only memory-related slices. Our dataset contained complete structural MRI scans, ranging between 160–200 scans per patient. Since the brain is not a uniform organ, the position of the memory-related structures varies across individuals. The hippocampus and amygdala appear roughly three-quarters of the way through each of the brain's hemispheres. The midline slice used for the baseline testing does not contain the memory related structures. Software like FreeSurfer could be used to precisely locate the slices that contain these specific regions, but it would take anywhere from 12–24 hours to process each scan. Since our dataset contains upwards of 1,500 scans, this approach was simply infeasible.

[3] Lew et al. (2023)
[4] Lim et al. (2022)
[5] Islam et al. (2023)
[6] Qiu et al. (2022)

Instead, we developed a more generalized and scalable approach for identifying slices containing the desired memory structures. Our approach was verified as reasonable by Professor Arash Yazdanbaskh, who heads the Computational Neuroscience and Vision Lab at Boston University. Researchers have demonstrated that the distance from the tip of the temporal lobe to the head of the hippocampus ranges from 3.5 to 4.2 cm, with a mean of 3.8 cm.[7] This indicates that generally, a patient's hippocampus structure falls within a certain set of slices per patient. Then, we manually examined 10 patients from each group (5 male, 5 female) and identified the slices where the left and right hippocampus and amygdala were most visible. Using this data, we calculated the average proportion (relative slice depth) and standard deviation for these structures for each group. We tested these values on a second set of 10 patients per group and found that slices within +/- 2 standard deviations from the group mean captured the desired memory structures. This approach took approximately 8 hours to test and implement across the entire dataset.

4.2 VGG-16 and ResNet50

VGG-16 and ResNet-50 were benchmark models that we saw in the papers, thought of as the golden standard for this kind of problem. VGG-16 and ResNet-50 are effective for image classification because they can extract deep hierarchical features: VGG-16 uses a simple and uniform structure of stacked convolutional layers, while ResNet-50 introduces skip connections that allow for training much deeper networks without vanishing gradients. To have a baseline to compare our custom model with, we decided to initially run both models on the 2D slices from our data, seeing if we could recreate the accuracy of the papers.

To test the models, we decided to make use of Boston University's Shared Computing Clusters (SCC), which allowed us to use higher GPUs and computing power to speed up our runs. We also made use of sweeps, a feature in Weights and Biases that allows us to run multiple instances of our model with varying parameters. Our `sweep.yaml` file had ranges for epochs, batch size, and learning rate that were randomly chosen so we could determine the best parameters for our runs. After we found new optimal parameters, we edited the `sweep.yaml` file to match them and repeated the process accordingly. This way, we were able to run both VGG-16 and ResNet-50 for several runs and maximize their accuracy, both boasting about 50% to 70% accuracy.

4.3 ResNet18

When we switched from 2D slices to 3D slices, surprisingly we saw performance of both of our previous models either not improve or drop slightly. This may be due to increased model complexity relative to dataset size, or the fact that 3D models are more sensitive to noise and misaligned slices. In order to handle this, we implemented a min-max transformation to try and allow the model to handle data better. This helped standardize input distributions across patients and reduced the impact of outliers and scanner variability. Unlike 2D models, which only observe a single cross-sectional slice, the 3D model has access to the full volume of memory-related slices, allowing it to learn richer spatial patterns and structural context across all regions of the brain. Additionally, the smaller size of ResNet-18 likely helped avoid overfitting versus a deeper model like ResNet-50.

During the run, we also made sure to output activation maps in order to understand where the model was learning the most. As we were building and tuning our custom CNN model, we wanted to compare its activation maps to the three models before, seeing what regions of the brain were most important to learning.

[7] Tubbs et al. (2018)

4.4 Custom CNN

We developed our custom CNN architecture to be as lightweight and flexible as possible when making predictions. The goal was to design a model simple enough to test our data processing pipeline while still being expressive enough to capture distinguishing patterns across clinical groups similar to the standard models. Our custom CNN was trained on both 2D and 3D data, with edits being made during the switch from 2D to 3D.

The 2D version consists of three convolutional layers with batch normalization, ReLU activation, and pooling, followed by a fully connected layer for classification. Each image is normalized with min-max scaling, resized to 64×64, and paired with its corresponding label from our patient CSV data (CN, MCI, or AD) based on patient ID. Similar to the code used for running ResNet and VGG-16, training and validation are handled by `train()` and `validate()` functions, which compute loss and accuracy each epoch. The `main()` function sets up data loaders, optimizer (Adam or SGD), learning rate scheduler, and logs results using Weights & Biases.

For the 3D model, we generally used a similar architecture with a few tweaks in order to handle 3D data. Most importantly, the convolutional and pooling layers were changed from 2D to 3D operations, allowing the model to learn spatial features not only within each slice (height and width) but also across slices (depth). We wanted to better capture areas like the hippocampus in order to make better predictions. We also implemented code to create and return activation maps after every run in order to compare with the activation maps of our other models. Once we figured out what areas of the brain were most important based on these maps, we edited the images we were sending in to match these and compare the results.

5 Dataset

A custom dataset was created for this project. We gained access to the Alzheimer's Disease Neuroimaging Initiative (ADNI) database. ADNI is the gold-standard resource used in AD-related research and provides standardized data including various brain scans (MRI, PET, fMRIs, etc.), clinical assessments, and biomarker data.

Based on our literature review, 3D T1-weighted MRI scans are the most commonly used imaging modality for Alzheimer's detection. These scans are preferred due to their high-resolution structural details and ability to capture subtle patterns of brain atrophy. We specifically focused on the sagittal view — vertical slices that divide the brain from left to right — spaced at increments of 1.2 micrometers. This orientation and unified slice increment value provides the clearest and most accurate view of the memory-related brain regions we aimed to capture.

Our dataset consists of 1,574 MRI scans from 637 unique patients. Many patients have multiple scans, which is expected, as clinicians often perform follow-up imaging to monitor progression from MCI or AD. The 1,574 scans distribution consists of 497 CN scans, 658 MCI scans, and 419 AD scans. At the patient-level distribution, the distribution is more balanced: 216 CN patients, 219 MCI patients, and 202 AD patients.

Sex distribution was another important consideration. Research has shown that females tend to have a higher risk of developing AD.[8] The dataset consists of 723 female and 851 male scans, suggesting that the sex distribution is relatively balanced. Given the specific MRI scan types we were aiming for, this dataset represents the most unbiased and representative sample we could achieve from the ADNI database.

As for the size of the dataset, each scan consists of anywhere between 160–200 slices per scan. Since we are using 3D-weighted images for our model's input, the dataset preprocessing is quite large, coming in at around 39 GB of disk space. While we were establishing our 2D baseline results, we used single midline slices per scan. This approach resulted in the dataset being only 354.7 MB.

[8] Moutinho (2025)

For final modeling, we used the approach mentioned in the dataset preparation section, and this resulted in the final dataset for training being just above 5 GB.

6 Evaluation/Results

6.1 ResNet

ResNet-50 initially ran on 2D data, boasting an accuracy of around 60%. We found that an epoch setting of around 20 is where the model stopped increasing in accuracy. We also saw that a low learning rate in the ten-thousandths place led to the highest validation accuracy.

When using 3D data, ResNet-50 performed poorly. This is likely due to how we were sending data to the models. When we switched to ResNet-18 instead, we saw a large increase in performance compared to the 2D data. Validation accuracy increased to as high as 93%, likely due to the model learning more complex structures not previously captured in the 2D model. We also saw from the activation maps that the hippocampus and parts of the cerebrum from the scans were the most important, giving us further insight to use for our custom CNN model.

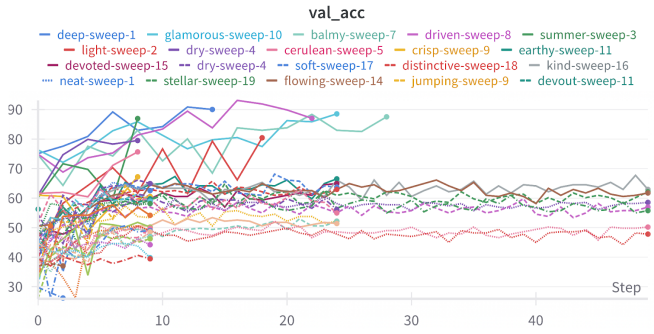

Figure 1: Validation Accuracies of all ResNet runs

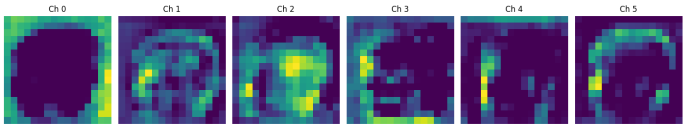

Figure 2: Activation Map generated from ResNet on 3D data

6.2 VGG-16

VGG-16 was initially implemented on the 2D dataset. With regards to the 2D dataset, the validation accuracy was 50%. Since the accuracy was 50% in case of validation and had consistently touched 90% in case of training accuracy, it had suffered an overfitting problem. Thus, VGG-16 was not the ideal neural network for our dataset. To confirm this result, the 3D dataset accuracies were also considered. Despite the training accuracy being 94%, the validation accuracy was only 66%. This was a severe case of overfitting. Therefore, we concluded that VGG-16 is not suitable for the analysis.

The training losses were also inconsistent and not consistently reducing. The training loss also showed a presence of high noise and high variability, thus indicating poor convergence. The validation accuracy was also fluctuating and it was not converging in this case. Also, the validation loss displayed significant spikes in several experiments, while the test accuracy varied drastically between models. These patterns confirm unstable learning and poor generalization, reinforcing that VGG-16 was not well-suited for the dataset.

Figure 3: Train Loss for VGG-16

6.3 Custom CNN

Prior to incorporating the activation map insights from ResNet-18, the best performing version of our custom CNN achieved a validation accuracy of 93.27%, followed by runs reaching 88.9% and 84.8%. The validation accuracy scores themselves are promising, showing generalization in some runs, but the overall variability in the sweeps suggests some form of instability or sensitivity to hyperparameters. Another major contributing factor could be the small batch size values we chose to stick with, as smaller batch sizes yielded better results. The largest batch size in our hyperparameter tuning was 8.

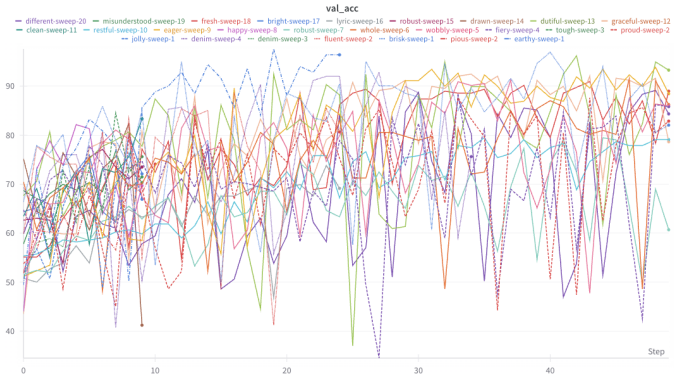

Figure 4: Validation for CNN prior to Activation Map Analysis modifications

The training curves demonstrated training loss consistently decreasing across all epochs, with the majority of the models converging around 0.2 or lower. Similarly, the training accuracy steadily increased, with the higher performing runs achieving over 95% training accuracy. These values confirm that the network was able to fit well to the training data. Activation maps were also generated to better understand which regions the model was relying on.

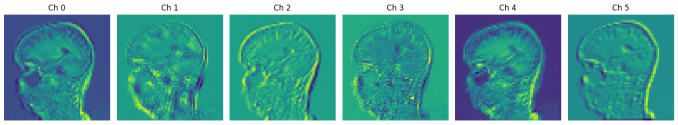

Figure 5: Activation Maps for Unmodified Custom CNN

The activation maps for the custom CNN show broad feature extraction across the entire slice. Channels 0 to 5 indicate that the model is focusing on the edges of the brain rather than focusing on any specific structures. The activations are not concentrated on the specific memory-related structures, and this suggests that the network has not yet associated the connection between these structures and the classification. We analyzed the outputs of ResNet-18 to better understand how we should adapt our preprocessing pipeline and modeling approach.

The activation maps from ResNet-18 showed more focus on brain textures in regions that were more consistent with the memory-related structures. We refined our data pipeline to better align with the ResNet-18 results. The new activation maps demonstrate more focused attention on the brain regions, specifically around the temporal lobe. The model emphasizes fine-grained texture and features that are more consistent with the memory-related areas.

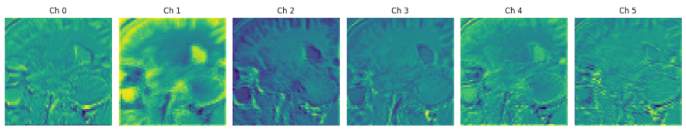

Figure 6: Activation Maps for Modified Custom CNN

Based on the refinements of the activation map analysis, the custom CNN was now demonstrating better performance. The top three runs achieved validation accuracies of 93.2%, 90.8%, and

90.4%. The training losses continued to show steady convergence and the training accuracy also improved, with multiple models achieving higher than 90%. The reduced gap between the training and validation suggests that the modifications made using the activation maps helped the model learn representations that were more generalizable and related to the memory regions of the brain.

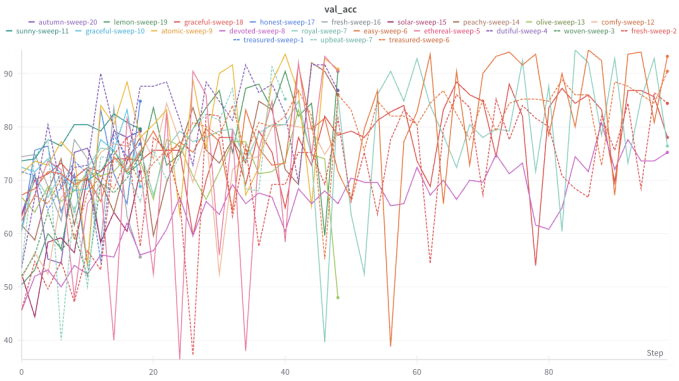

Figure 7: Validation Accuracies for Modified Custom CNN

7 Conclusion

In order to successfully classify brain MRI scans into three clinical categories—Cognitively Normal (CN), Mild Cognitive Impairment (MCI), and Alzheimer's Disease (AD)—deep learning models were developed and evaluated. Our top-performing architecture was the custom 3D CNN, which outperformed both ResNet-50 and VGG-16 with a validation accuracy of 93.2%. Although VGG-16 performed well during training, overfitting and poor generalization were constant problems. Activation maps were crucial in directing model improvements by emphasizing the significance of spatial consistency and region-specific features. Consequently, the final custom CNN improved generalization and robustness by concentrating more precisely on memory-related regions. Predictive power may be further increased in the future by investigating multimodal models that combine clinical and imaging data and implementing more sophisticated slice-targeting techniques. In the end, this study reaffirms the importance of comprehensible and trustworthy early detection methods for reducing the burden of Alzheimer's disease on society.

References

[1] Alzheimer's Association. (n.d.). *What is Alzheimer's disease?* Retrieved from https://www.alz.org/alzheimers-dementia/what-is-alzheimers

[2] Li, H., Habes, M., Wolk, D. A., & Fan, Y. (2019). A deep learning model for early prediction of Alzheimer's disease dementia based on hippocampal magnetic resonance imaging data. *Alzheimer's & Dementia*, 15(8), 1059–1070. https://doi.org/10.1016/j.jalz.2019.02.007

[3] Lew, C. O., Zhou, L., Mazurowski, M. A., Doraiswamy, P. M., Petrella, J. R., & Alzheimer's Disease Neuroimaging Initiative. (2023). MRI-based deep learning assessment of amyloid, tau, and neurodegeneration biomarker status across the Alzheimer disease spectrum. *Radiology*, 309(1), e222441. https://doi.org/10.1148/radiol.222441

[4] Lim, B. Y., Lai, K. W., Haiskin, K., Kulathilake, K. A. S. H., Ong, Z. C., Hum, Y. C., Dhanalakshmi, S., Wu, X., & Zuo, X. (2022). Deep learning model for prediction of progressive mild cognitive impairment to Alzheimer's disease using structural MRI. *Frontiers in Aging Neuroscience*, 14, 876202. https://doi.org/10.3389/fnagi.2022.876202

[5] Islam, F., Rahman, M. H., Nurjahan, N., Hossain, M. S., & Ahmed, S. (2023). A novel method for diagnosing Alzheimer's disease from MRI scans using the ResNet50 feature extractor and the SVM classifier. *International Journal of Advanced Computer Science and Applications (IJACSA)*, 14(6), 103–111. https://doi.org/10.14569/IJACSA.2023.01406131

[6] Qiu, S., Miller, M. I., Joshi, P. S., Lee, J. C., Xue, C., Ni, Y., ... & Kolachalama, V. B. (2022). Multimodal deep learning for Alzheimer's disease dementia assessment. *Nature Communications*, 13, Article 3404. https://doi.org/10.1038/s41467-022-31037-5

[7] Tubbs, R. S., Loukas, M., Barbaro, N. M., Shah, K. J., & Cohen-Gadol, A. A. (2018). External cortical landmarks for localization of the hippocampus: Application for temporal lobectomy and amygdalohippocampectomy. *Surgical Neurology International*, 9, 171. https://doi.org/10.4103/sni.sni$_4$46$_1$7

[8] Moutinho, S. (2025). Women twice as likely to develop Alzheimer's disease as men – but scientists do not know why. *Nature Medicine*, 31(5), 704–707. https://doi.org/10.1038/s41591-025-03491-3