Fine-Tuning Multimodal Foundation Models for Dementia Diagnosis

Maya Ha
mayaha@usc.edu
Ogheneyoma Akoni
akoni@usc.edu
Gustavo Adolpho Lucas De Carvalho
lucasdec@usc.edu

Abstract

Machine learning (ML) methods such as Deep Learning hold great promise for healthcare. However, despite ML becoming ubiquitous in many fields, developing medical ML applications remains very challenging. Two of the main issues with applying Deep Learning in the medical field stem from (1) the lack of ML expertise required to develop such models in the area and (2) Deep Learning requiring significant amounts of data collection, something that is especially tricky in the medical field due to the sensitive nature of the data. One promising paradigm for applications with small amounts of data is fine-tuning Foundation Models. Furthermore, with the advent of open-source Multimodal Foundation Models like LLaVa and user-friendly libraries like HuggingFace, fine-tuning has become relatively beginner-friendly. Unfortunately, another significant problems for ML applications in the medical field is a lack of large compute resources, which Foundation Models usually require. That being said, recent advancements in quantization have made fine-tuning much more cost-effective. This study aims to explore the utilization of foundation models for healthcare applications by fine-tuning LLaVA 7b on a dementia classification task using various neuroimaging data. We demonstrate that we can achieve a remarkably high accuracy, matching the performance of prior work, while using significantly less data and doing so in a cost-effective manner.

1 Introduction

Dementia is a broad term encompassing memory loss and other cognitive deficits that significantly disrupt an individual's daily functioning. Diseases classified as "dementia" result from atypical alterations in the brain. More than 55 million individuals are impacted by dementia, with approximately 10 million new cases emerging each year. Despite the widespread prevalence of dementia, there is still no effective treatment even after years of research in this area. This could be a result of diagnosing dementia in its later stages. However, with rapid improvements in technology, researchers have been utilizing ML models to assist in dementia research. These models have demonstrated performance that is comparable to, or even surpasses, the predictive accuracy of radiologists' imaging assessments.

Diagnosing dementia is a multifaceted process that entails analyzing a wide range of patient data, including medical history, plasma and cerebrospinal fluid biomarkers, cognitive evaluations, neuroimaging, and more factors. Studies indicate that neuronal loss often precedes observable cognitive decline [8], suggesting that early detection methods could help patients prepare for the future and assist researchers' and practitioners' ability to identify early-stage symptoms more efficiently. Recent dementia research has shown promising results in the use of ML models on neuroimaging data like magnetic resonance imaging (MRI) scans for dementia stage prediction [2], [6]. Our research seeks

Preprint. Under review.

to advance this domain by utilizing foundation models for the categorization of MRI scans on a spectrum of demented to non-demented.

Developing successful ML models from the ground up may be inaccessible to those who are not AI/ML engineers. Many researchers in the medical domain possess extensive biological or clinical expertise but may lack expertise in programming, resulting in unexploited research prospects. We aim to bridge that gap by utilizing open source frameworks and exploring pre-trained foundation models. Developing a model for such a complex task like dementia diagnosis requires a substantial amount of resources ranging from large amounts of labeled datasets to significant computational power for training. Due to privacy concerns, limitations on data collection, and other issues, there are not enough comprehensive datasets available in the healthcare industry. To address the data constraint issue, we evaluate foundation models as a potential solution. Foundation models are trained on large, diverse datasets and can be adapted for specific tasks with significantly less data than would be required to train a model from scratch [3]. While fine-tuning foundation models necessitates less data than task-specific models, fine-tuning them still demands considerable computational resources. Because of this, our research also seeks to investigate the effective utilization of constrained computational resources for the fine-tuning of foundation models. This aspect of our research focuses on the practicality of applying foundation models in real-world contexts when access to large-scale computer equipment is limited, such as in smaller medical research labs or low-resource settings. In our tests, we used techniques including LoRA, paged optimizer, bfloat16, and NF4 - all components of the QLoRA optimization strategy - to fine-tune models efficiently on limited GPU resources.

In this study, we investigate the potential of foundation models by evaluating their performance in comparison to specialized dementia classification models. Specifically, we aim to ascertain if foundation models can leverage their pre-trained knowledge to achieve comparable results to specialized models while requiring substantially less training data by fine-tuning the LLaVA foundation model using MRI scans for the classification of dementia diagnoses. Following fine-tuning, we examine the performance of the LLaVA model in comparison to the baseline model, EfficientNet, to determine its effectiveness. Additionally, we examine the minimum amount of training data required to achieve performance levels comparable to those achieved with the complete dataset. To evaluate the model's robustness, we will validate our findings on an additional unseen dataset, ADNI, and assess the model's generalizability by evaluating the model trained with OASIS on ADNI and vice versa. This comprehensive examination seeks to evaluate the feasibility of foundation models as an effective and resource-efficient option for dementia classification tasks.

2 Background and Related Work

2.1 Prior Work

Advancements in deep learning and artificial intelligence (AI) have significantly transformed research in dementia diagnosis. A notable study employed deep learning models to integrate multimodal data, including imaging data, electronic health records (EHR), and genomic single nucleotide polymorphisms (SNPs), for classification of patients into the following dementia categories: cognitively normal (CN), mild cognitive impairment (MCI), and advanced dementia (AD). Distinct networks are independently trained for each data modality- stacked denoising autoencoders for EHR and SNP, and three-dimensional convolutional neural networks (CNNs) for MRI data. Once the networks for each data modality are independently trained, they combine them using various methods, including decision trees, random forests, support vector machines (SVM), and k-nearest neighbors (kNN) for dementia stage prediction.

The study found that deep learning models outperform standard shallow models in single-modality tasks, which frequently rely on handcrafted features created by domain experts. In contrast, deep models learn optimal feature representations on their own during training. However, the study also identified that deep models' performance deteriorates with extremely limited datasets [21]. Given the difficulties related to data scarcity in dementia diagnosis that arise from patient confidentiality and restricted data access, foundation models offer a possible solution by enabling dementia classification with limited training data, overcoming a significant constraint in existing machine learning methodologies for this field. This study aims to investigate the capacity of foundation models to surmount these obstacles while still achieving comparable results to specialized models.

Convolutional neural network (CNN)-based models are commonly used to classify stages of dementia. A study used CNN architectures, EfficientNet and ResNet, as well as a post-processing ensemble learning technique that merged both models' predictions to increase classification performance. The study achieved impressive accuracy levels on the OASIS dataset, with EfficientNet obtaining 98.59% and ResNet reaching 94.59% [7]. We utilize EfficientNet as the baseline model for our research. EfficientNet is a CNN architecture and scaling method that uniformly scales all dimensions of depth/width/resolution using a compound coefficient. It differs from conventional CNN's as it uniformly scales the network dimensions with a fixed set of scaling coefficients [19]. In this study, we will assess whether we can achieve outcomes that are comparable to those of the Efficient Net model and how much less data is required to maintain similar performance levels.

Prior research has concentrated on developing specialized models for dementia diagnosis from the ground up. While this has yielded impressive results, a discernible trend in the previous studies is the challenge posed by insufficient datasets. An evaluation of current state-of-the-art multimodal machine learning models for dementia revealed that 67 out of 92 studies utilized the ADNI dataset [15]. Due to an excessive dependence on the limited available data, there is a need for these pre-trained foundation models, which require a smaller amount of supplementary data to execute a specified task effectively.

Foundation models have demonstrated promising capabilities in medical imaging tasks. A study evaluating large visual-language models (VLMs) compared five VLMs (BiomedCLIP, OpenCLIP, OpenFlamingo, LLaVA, ChatGPT-4) and two CNN-based models (CNN, ResNet-18) across three medical imaging datasets (BTD, ALLIDB2, CX-Ray). For CNN-based methods, the models were trained on the dataset before testing, whereas VLMs were evaluated directly in zero-shot (Biomed-CLIP, OpenCLIP, LLaVA, ChatGPT-4) and few-shot (OpenFlamingo) scenarios. As expected, CNN-based methods performed better on all datasets considering they were pre-trained while the foundation models were not. However, the VLMs still achieved good results without any training, demonstrating their adaptability and efficiency when utilizing restricted datasets [20]. In our study, we expand upon this methodology by fine-tuning a foundation model on the same dataset used for training the CNN-based models, enabling direct performance comparison. Additionally, we will progressively reduce the amount of data used for fine-tuning the foundation model to investigate its ability to maintain comparable performance with minimal data. By doing so, we hope to address the data scarcity issue in dementia diagnosis while leveraging the strengths of foundation models.

2.2 Foundation Models

Foundation models are large deep learning neural networks trained on a diverse array of generalized and unlabeled data requiring minimum fine-tuning for specialized tasks. They are proficient in executing a multitude of general tasks, including language comprehension, text and image generation, and natural language conversation. They generate output from one or more inputs (prompts) in the format of human language instructions. Foundation models are based on complex neural networks including generative adversarial networks (GANs), transformers, and variational encoders. The phrase "foundation model" was initially popularized by the Stanford Institute for Human-Centered Artificial Intelligence. They explain that foundation models are inherently incomplete but function as the fundamental framework from which other task-specific models are developed through adaptation, which is the source of their allure [3]. Multimodal foundation models inherit all the properties of foundation models but with the capability to deal with vision and vision-language modalities [11]. Multimodal foundation models are often able to process multiple modalities (e.g. text, image, video, audio, etc.) at the same time. They can take in modal pairs as input and model the correlation between two different modalities in their pre-training data. This study will adapt the LLaVA multimodal foundation model for the task of dementia classification.

3 Methods

To test the viability of fine-tuning foundation models for healthcare applications, we wanted to ensure that our experiments fit within the constraints detailed earlier. To fit within our expertise constraint, we decided to use LLaVA 1.5 7b, a multimodal foundation model that is readily accessible in HuggingFace. We believe that this fits our expertise constraint as the model and fine-tuning code fit comfortably within the HuggingFace ecosystem, which is relatively beginner-friendly. See Appendix A for details on LLaVA.

To fit within our low-resource constraint, we decided to conduct all of our experiments only using Google Colab, which can provide free (albeit limited) access to a single A100 GPU. However, due to the size of LLaVA, we cannot fine-tune the default model in a single GPU. To achieve this, we leveraged several different methods.

We use Low Rank Adaptors on the text encoder layers of the model instead of fine-tuning the entire model. Introduced by Hu et al. in the paper 'LoRA: Low-Rank Adaptation of Large Language Models' [10], LoRA consists of training low rank weights on top of the original model as adaptor layers. The model layers are frozen and instead, for every weight matrix we want to fine-tune, we train two rectangular matrices that together make a single low rank weight matrix of the same dimensionality as the frozen matrix. Through LoRA we save a significant amount of compute since it allows us to fine-tune the model effectively by only changing a small number of weights.

It is well known that models perform worse if their hidden states and weights are represented with less precision. Brain Float 16, introduced by Google Brain [22], is identical to the float16 data structure except for the fact that it uses 8 bits for the exponent (instead of the usual 5). Empirically it was found that although the data structure only uses 16 bits, models whose hidden states are expressed with Brain Float 16 match the performance of float32 models. Hence, we are able to retain float32 performance while only needing half as much memory for the hidden states.

Furthermore, we represent the weights of our model with NormalFloat4, a 4-bit float data structure designed to keep weights normally distributed even after being quantized down from float32; this has been empirically shown to retain model performance. Finally we use paged optimizers, breaking the optimizer state into chunks and offloading most of it to the CPU, only loading pieces at a time.

All of these techniques together comprise what is known as QLoRA [5], a quantized fine-tuning strategy that also has dedicated libraries such as BitesAndBytes [1] and PEFT (Parameter-Efficient Fine-Tuning Methods) [13] , making it relatively beginner-friendly to implement and hence aligned with our expertise constraint.

4 Experiments

Using the methods outlined above, we fine-tune the LLaVA 1.5 7b model to classify an MRI image included in the prompt. The generated output should be the label, spelled exactly the same as the ground truth to be considered a match. For more details on the implementation, refer to Appendix B.

4.1 Datasets

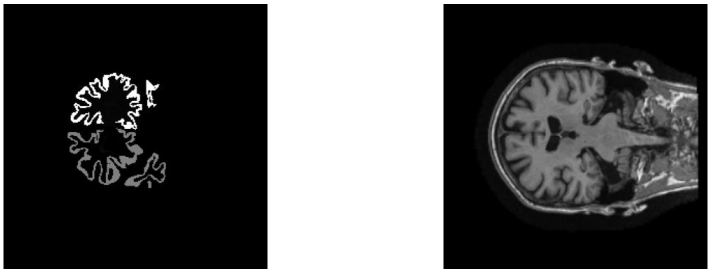
This project utilized the Open Access Series of Imaging Studies (OASIS-1) dataset [14]. We used the Kaggle version of this dataset[1]. The set consists of a cross-sectional collection of 416 subjects aged 18 to 96. For each subject, 3 or 4 individual MRI scans obtained in single scan sessions are included. The subjects are all right-handed and include both men and women. 100 of the included subjects over the age of 60 have been clinically diagnosed with very mild to moderate Alzheimer's disease (AD). The images have been divided into four classes based on dementia progression - mild dementia, moderate dementia, non-demented, or very mild dementia. A limitation of this dataset is that the distribution of labels is very skewed to the "Non Demented" label 3. To obtain 2D images for training, the brain images were sliced along the z-axis into 256 pieces, and slices ranging from 100 to 160 were selected from each patient.

To ensure that any results we obtain are not unique to the OASIS dataset, we utilized a second dataset, derived from the Alzheimer Disease Neuroimaging Initiative Dataset (ADNI) [16]. The ADNI study tracked the progression of the disease using changes to the brain observed in MRI scans alongside other biomarkers. We utilized a Kaggle ADNI dataset, which contains the 1.5T MRI scans from 645 different subjects[2]. Our third dataset, also a Kaggle dataset from ADNI, was segmentation of cortical and subcortical structures across 982 subjects, processed using the FastSurfer pipeline [9] [3]. Both ADNI datasets have the same classes: Alzheimer's Disease (AD), Cognitively Normal (CN),

[1] `https://www.kaggle.com/datasets/ninadaithal/imagesoasis/data`

[2] `https://www.kaggle.com/datasets/mdfahimbinamin/adni-1-5t-filtered-preprocessed-quickseg-dataset/data`

[3] `https://www.kaggle.com/datasets/isratjahankhan/adni-screening-1-5t-segmented-complete-dataset/data`

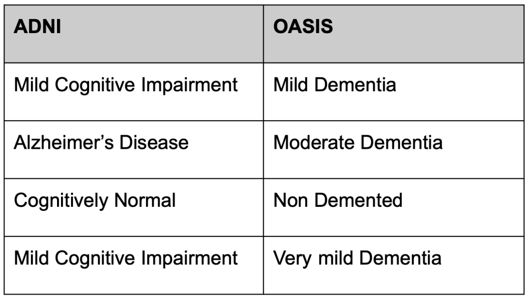


(a) Segmented MRI (b) Non-segmented MRI

Figure 1: Comparison between ADNI segmented and non-segmented MRI's

and Moderate Cognitive Impairment (MCI) groups. We performed the same processing as with the OASIS dataset, where we take slices 100 to 160 from the 3D MRI.

Both OASIS and ADNI employed 1.5T scanners to acquire T1-weighted images, but we also trained a model on segmented ADNI data, and the level of information in these scans is significantly less 1. Part of our interest is to test how this fine-tuning technique performs on different types of imaging data and comparing their effectiveness. For all of these datasets, we utilized Kaggle datasets instead of the original due to simplicity. To the best of the authors' knowledge, these datasets represent the original.



ADNI	OASIS
Mild Cognitive Impairment	Mild Dementia
Alzheimer's Disease	Moderate Dementia
Cognitively Normal	Non Demented
Mild Cognitive Impairment	Very mild Dementia

Figure 2: Label Mapping between OASIS and ADNI datasets.

In addition, to assess the generalizability of these models, we tested the OASIS-trained model on ADNI data, and vice versa. For both of these experiments we used a label mapping indicated in Figure 2. For instance, if a model that was trained on ADNI generates the text "Cognitively Normal" to an image from the OASIS dataset, we interpret that as if the model had output "Non Demented".

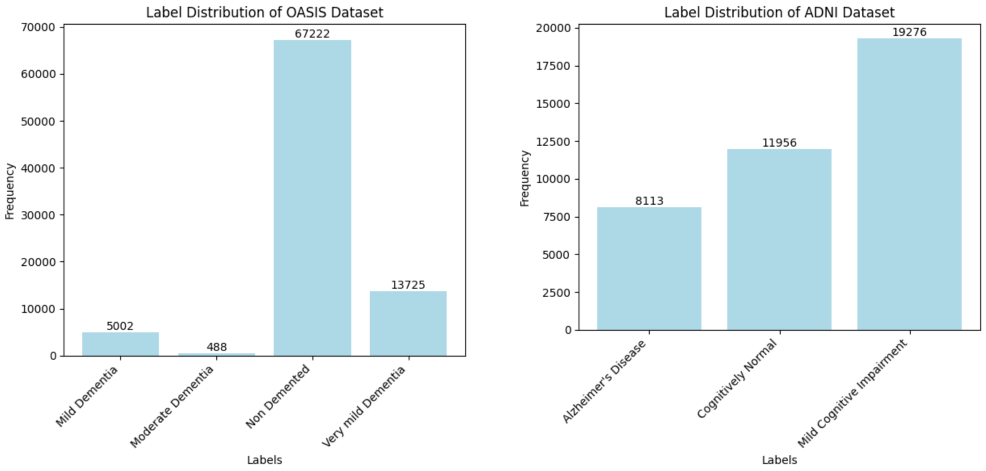


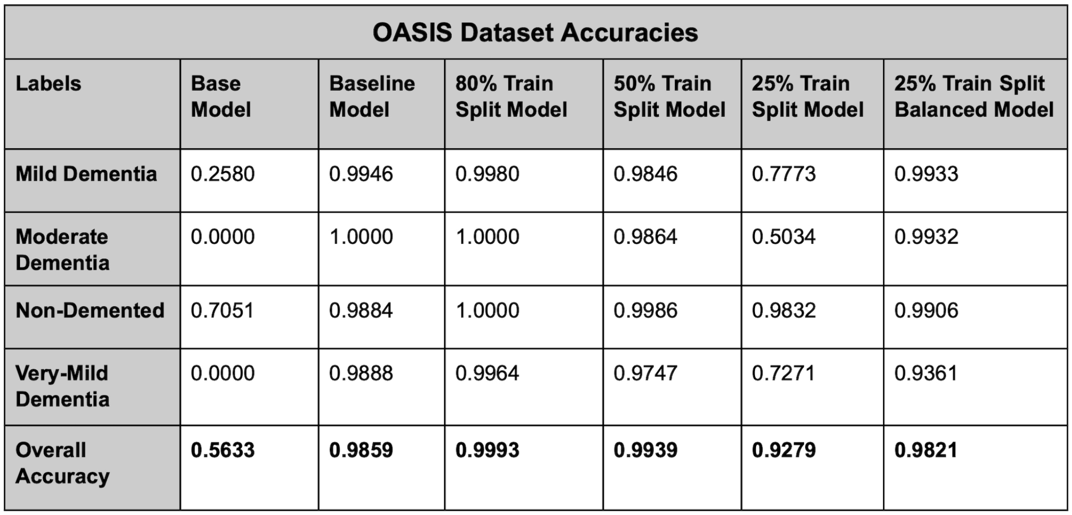
Figure 3: Comparison of Label Distributions Between OASIS (left) and ADNI (right) Datasets

4.2 Training Splits

We applied an 80% training, 10% testing, 10% validation split on OASIS. To test how little data we needed to accomplish our dementia diagnosis task, the training set was progressively reduced to 50%/10%/10% and 25%/10%/5% train/test/validation splits, with results averaged over three shuffles for each split. Stratified sampling was utilized to ensure proportional label distributions across all sets.

For our experiments with the two ADNI datasets, we utilized an 80% training, 10% testing, 10% validation split, with results averaged over three shuffles.

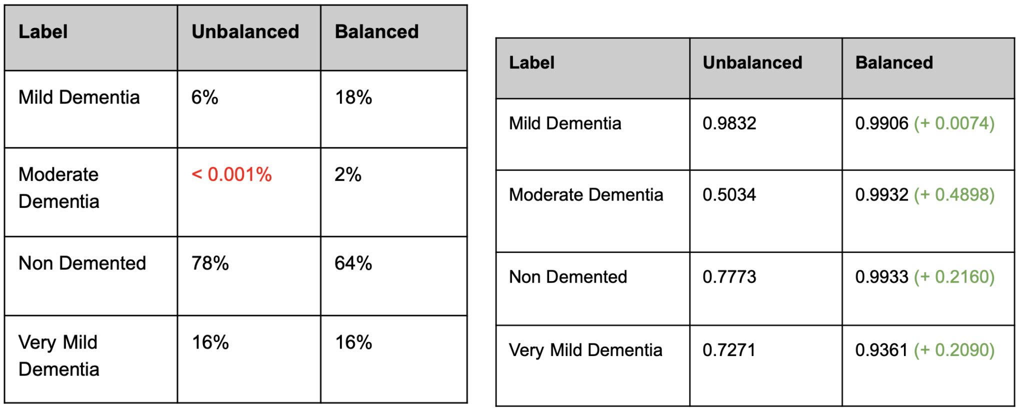
5 Results



OASIS Dataset Accuracies						
Labels	**Base Model**	**Baseline Model**	**80% Train Split Model**	**50% Train Split Model**	**25% Train Split Model**	**25% Train Split Balanced Model**
Mild Dementia	0.2580	0.9946	0.9980	0.9846	0.7773	0.9933
Moderate Dementia	0.0000	1.0000	1.0000	0.9864	0.5034	0.9932
Non-Demented	0.7051	0.9884	1.0000	0.9986	0.9832	0.9906
Very-Mild Dementia	0.0000	0.9888	0.9964	0.9747	0.7271	0.9361
Overall Accuracy	**0.5633**	**0.9859**	**0.9993**	**0.9939**	**0.9279**	**0.9821**

Figure 4: Results on OASIS dataset, averaged across 3 shuffles. "Base Model" are the results of the LLaVa 1.5 7b without fine-tuning, "Baseline Model" are the results of EfficientNet model, and the other columns are the results with different fine-tuning data splits/amounts.

After fine-tuning LLaVA on the 80% split of the OASIS training data, accuracy improved significantly from the base LLaVA model, with an overall accuracy improvement of 42.99% 4. This model was also able to match and even surpass the performance of our baseline slightly. We found that the model fine-tuned on 50% of the OASIS dataset performed very similarly to the model fine-tuned on 80% of the dataset. However, performance started to degrade when we used 25% of the dataset for training, which is why we conducted our label distribution re-balancing experiment at this point.



Label	Unbalanced	Balanced
Mild Dementia	6%	18%
Moderate Dementia	< 0.001%	2%
Non Demented	78%	64%
Very Mild Dementia	16%	16%

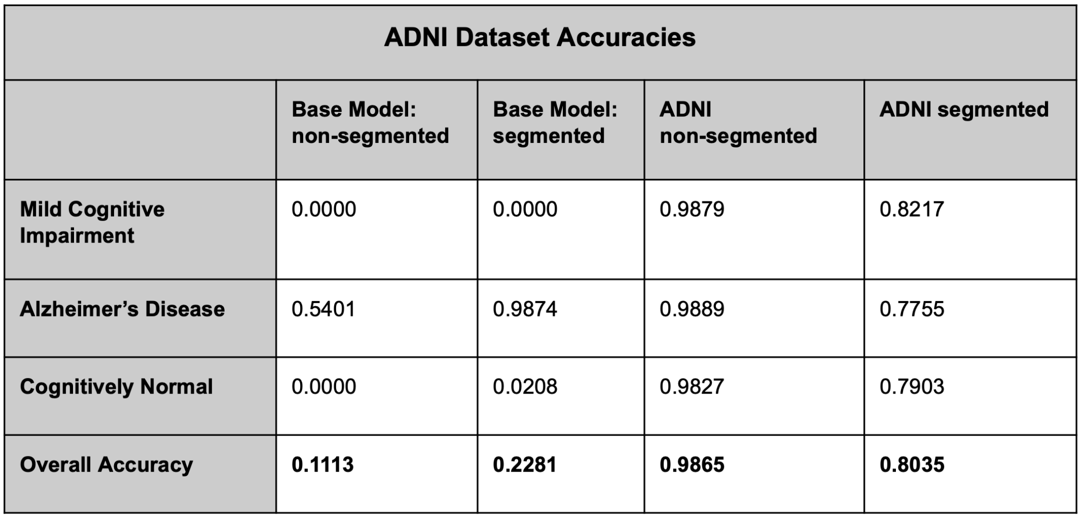
(a) Label Proportion of 25% Training Split

Label	Unbalanced	Balanced
Mild Dementia	0.9832	0.9906 (+ 0.0074)
Moderate Dementia	0.5034	0.9932 (+ 0.4898)
Non Demented	0.7773	0.9933 (+ 0.2160)
Very Mild Dementia	0.7271	0.9361 (+ 0.2090)

(b) Average accuracy over 3 splits

Figure 5: Rebalancing Data Experiment

We re-balanced the label distribution of the 25% training split, as seen in Figure 5a, where we removed samples from the majority class, "Non Demented", to increase the number of samples for our minority classes, "Mild Dementia" and "Moderate Dementia," while maintaining the same number of samples for "Very mild Dementia." This re-balancing shot our accuracy back up almost to that of the baseline, at 98.21%. As outlined in Figure 5b, it had a lower accuracy on the "Very mild Dementia" label of 93.61% but had a ∼99% accuracy on the other 3 labels. Interestingly, the accuracy on "Non Demented" samples still increased, even though the number of samples of that label in the training set actually decreased.

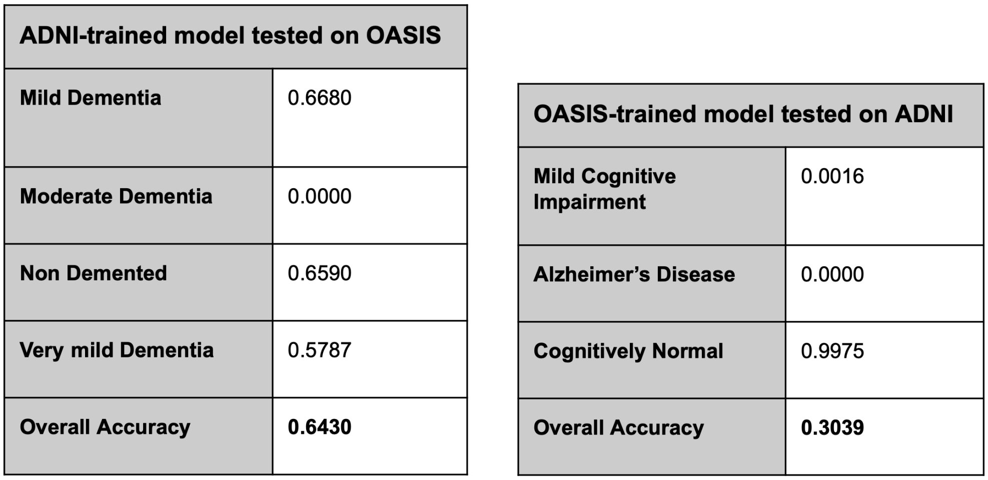


ADNI Dataset Accuracies				
	Base Model: non-segmented	**Base Model: segmented**	**ADNI non-segmented**	**ADNI segmented**
Mild Cognitive Impairment	0.0000	0.0000	0.9879	0.8217
Alzheimer's Disease	0.5401	0.9874	0.9889	0.7755
Cognitively Normal	0.0000	0.0208	0.9827	0.7903
Overall Accuracy	**0.1113**	**0.2281**	**0.9865**	**0.8035**

Figure 6: Results on ADNI dataset, averaged across 3 shuffles. "Base Model" are the results of the model without fine-tuning, while "ADNI" are the results where the model was fine-tuned with "segmented" / "non-segmented" data. The EfficientNet performance (our baseline) on ADNI was trained only on non-segmented data and was not broken down by label so we did not include it in this table. That being said, the overall accuracy for the baseline was 97.25%.

The LLaVA model fine-tuned on non-segmented ADNI produced impressive results as well, with an overall accuracy improvement of 11.13% to 98.65% 6.

The segmented data still performed surprisingly well, given the discrepancy in the amount of detail between the two data types, with an accuracy of 80.35%. However, our findings demonstrate that non-segmented data might be a better input for dementia diagnosis tasks when using a fine-tuned foundation model.



ADNI-trained model tested on OASIS	
Mild Dementia	0.6680
Moderate Dementia	0.0000
Non Demented	0.6590
Very mild Dementia	0.5787
Overall Accuracy	**0.6430**

OASIS-trained model tested on ADNI	
Mild Cognitive Impairment	0.0016
Alzheimer’s Disease	0.0000
Cognitively Normal	0.9975
Overall Accuracy	**0.3039**

Figure 7: Accuracy of ADNI-trained model tested on OASIS data (left) and OASIS-trained model tested on ADNI data (right)

For the two experiments testing the OASIS-trained model on ADNI and vice versa, we found that the generalizability of these models were poor, even though the MRI scans across datasets were generally acquired through very similar means (strength, device, etc.). The ADNI-trained model tested on OASIS obtained an overall accuracy of 64.30%, and the OASIS-trained model tested on ADNI obtained an overall accuracy of 30.39%.

6 Discussion

The most significant weakness of the model trained on the OASIS 50% split is the misclassification of "Very mild Dementia" as "Non Demented" 8. We hypothesize that this is due to the two labels being inherently more difficult to discriminate between than the others, and there being significantly more "Non Demented" samples.

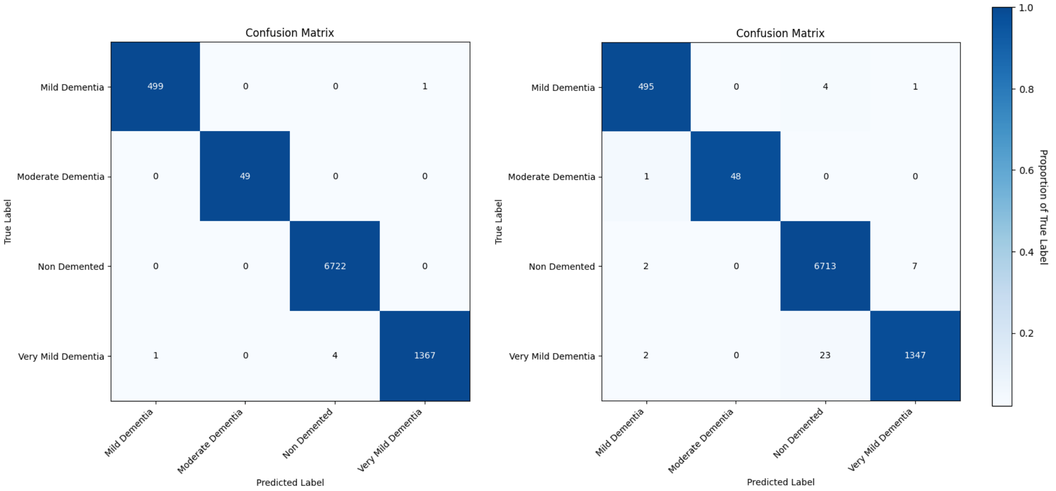


Figure 8: Confusion matrices for 80% (left) and 50% (right) OASIS train splits, where the color intensity of each cell represents the proportion of samples for the true label represented in that cell (each row adds up to 1).

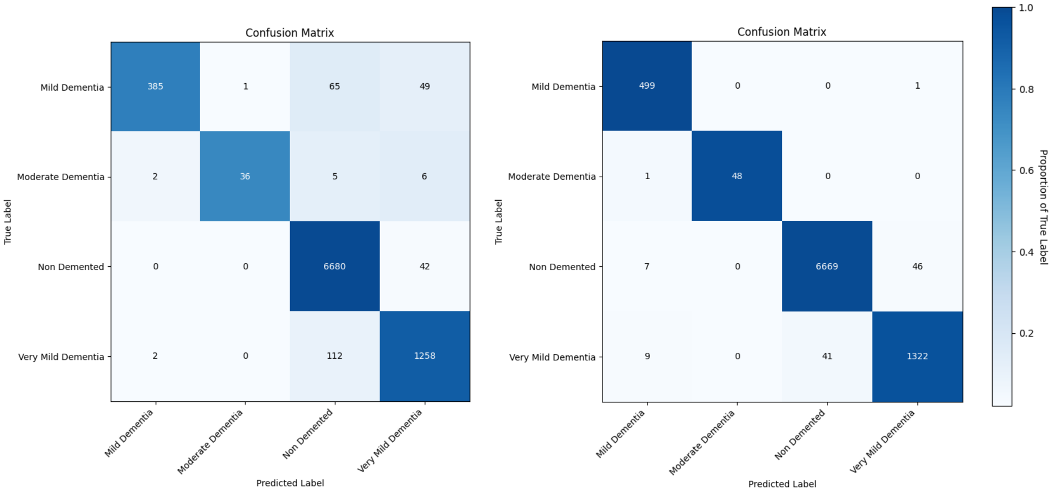


Figure 9: Confusion matrices for 25% (left) and 25% balanced (right) OASIS train splits, where the color intensity of each cell represents the proportion of samples for the true label represented in that cell (each row adds up to 1).

As shown in Figure 9, the model using a 25% training split has a lot of "Very mild Dementia" that is classified as "Non Demented," similarly to the model trained on the 50% split. In addition, we now see "Mild Dementia" being classified as "Non Demented" and "Very mild Dementia" much more frequently. When we balance the distribution of labels, we can see that the confusion matrix drastically improves, and the number of "Very mild Dementia" samples falsely classified as "Non Demented" drops from 112 to 41. The number of "Mild Dementia" falsely classified as "Non Demented" drops from 65 to 0, and the number of "Mild Dementia" falsely classified as "Very mild Dementia" drops from 49 to 1. Even though the number of "Non Demented" samples in the re-balanced training set is lower, the model still performs well on that label.

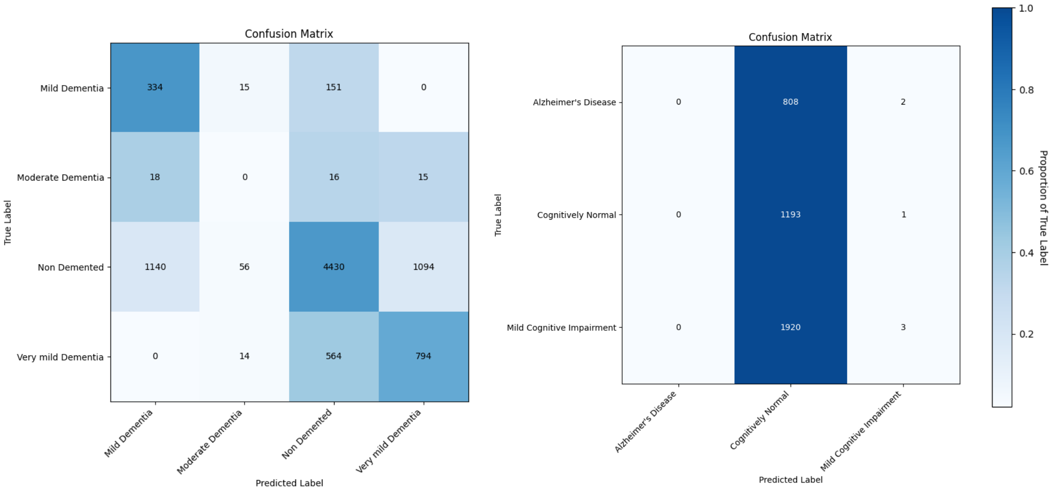


Figure 10: Confusion matrices for the ADNI-trained model tested on OASIS data (left) and the OASIS-trained model tested on ADNI data (right), where the color intensity of each cell represents the proportion of samples for the true label represented in that cell.

In Figure 10, we have the confusion matrices for the two experiments training LLaVA on ADNI and testing on OASIS, and vice versa. The confusion matrix for the ADNI-trained model tested on OASIS is more spread out. There are barely any samples classified as "Moderate Dementia," and there were actually 0 "Moderate Dementia" samples classified correctly. In the confusion matrix for the OASIS-trained model tested on ADNI, we can clearly that it classifies almost all the samples in the ADNI test set as "Cognitively Normal."

One notable difference between these two datasets is the label distribution 3. While OASIS has an overwhelming majority of “Non Demented” samples, ADNI is much more balanced and has a majority of “Moderate Cognitive Impairment” samples. This may explain why the OASIS-trained model classified so many ADNI samples as “Cognitively Normal,” because that is our mapped label equivalent for "Non Demented."

Another hypothesis is that due to OASIS having 1 more class than ADNI, this might make testing the ADNI-trained model on OASIS data an easier task, because it only has to predict “Mild Cognitive Impairment” to be considered correct for either “Very mild Dementia” or “Mild Dementia” samples. Since MCI covers both of these labels, the challenge of discriminating between the two is removed. This is reflected in the results, where the OASIS-trained model performed very poorly on ADNI, and the ADNI-trained model performed better on OASIS 7.

Our findings show that multimodal foundation models have the ability to produce outcomes comparable to specialized ML models on domain-specific tasks. However, we discovered a significant limitation in the generalizability of these models. This difficulty could be due to overfitting to the specific data distribution used during training, as each model was only trained on data from a singular study. This issue reflects a larger problem in the medical field which is a lack of a diverse representative of datasets. Current datasets frequently fail to represent the full range of variability among populations, locations, and clinical circumstances. This limits the model’s capacity to effectively generalize to new data. To address this, future research could concentrate on training models with data obtained from different studies through data-mixing strategies, or perhaps leveraging in-context learning to improve performance. By combining data from many sources, one might be able to produce a more comprehensive dataset that better depicts the heterogeneity found in real-world clinical situations.

7 Future Work and Conclusion

Dementia diagnosis is a difficult endeavor that encompasses multiple elements to enhance diagnostic precision. While our findings show that foundation models can perform well with substantially less data, future research can expand on this by including more diverse and representative datasets through data-mixing. This would aid in exploring foundations model generalizability and would be a step toward bridging the gap between experimental performance and real-world applications in healthcare AI. Another way to further improve classification is by exploring various modalities of the neuroimaging data, as this study focused on classification with T1-weighted MRI neuroimaging data. Future studies could extend this work by incorporating other neuroimaging modalities, such as fluorodeoxyglucose-positron emission tomography (FDG-PET), functional near-infrared spectroscopy (fNIRS), functional MRI (fMRI), and electroencephalography (EEG). Research indicates that MRI and PET scans provide complementary information [17]. Future research could aim to explore additional neuroimaging modalities to see how effectively a combination of these modalities contributes to data reduction for fine-tuning foundational models. One could also include clinical data (demographics, symptoms, medications) in the fine-tuning prompt in addition to imaging data. Using these several modalities may provide a more complete understanding of dementia and its progression. Furthermore, other techniques could shift from categorizing patients into distinct diagnostic groups to assessing Alzheimer's disease patients along a functional continuum. Applying ML in the medical realm is extremely promising, with many unexplored avenues, and our project is a step towards making these tools more accessible and useful.

References

[1] Younes Belkada, Tim Dettmers, Artidoro Pagnoni, Sylvain Gugger, and Sourab Mangrulkar. Making llms even more accessible with bitsandbytes, 4-bit quantization and qlora, 2023.

[2] Bahare Bigham, Seyed Amir Zamanpour, and Hoda Zare. Features of the superficial white matter as biomarkers for the detection of alzheimer's disease and mild cognitive impairment: A diffusion tensor imaging study. *Heliyon*, 8(1), 2022.

[3] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

[4] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, March 2023.

[5] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: efficient finetuning of quantized llms. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2024. Curran Associates Inc.

[6] Sina Fathi, Ali Ahmadi, Afsaneh Dehnad, Mostafa Almasi-Dooghaee, Melika Sadegh, and Alzheimer's Disease Neuroimaging Initiative. A deep learning-based ensemble method for early diagnosis of alzheimer's disease using mri images. *Neuroinformatics*, 22(1):89–105, 2024.

[7] Rani Ghassan Al Rahbani, Anastasia Ioannou, and Tao Wang. Alzheimer's disease multiclass detection through deep learning models and post-processing heuristics. *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, 12(1):2383219, 2024.

[8] Sven Haller, Hans Rolf Jäger, Meike W Vernooij, and Frederik Barkhof. Neuroimaging in dementia: more than typical alzheimer disease. *Radiology*, 308(3):e230173, 2023.

[9] Leonie Henschel, Sailesh Conjeti, Santiago Estrada, Kersten Diers, Bruce Fischl, and Martin Reuter. Fastsurfer - a fast and accurate deep learning based neuroimaging pipeline. *NeuroImage*, 219:117012, October 2020.

[10] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021.

[11] Chunyuan Li, Zhe Gan, Zhengyuan Yang, Jianwei Yang, Linjie Li, Lijuan Wang, Jianfeng Gao, et al. Multimodal foundation models: From specialists to general-purpose assistants. *Foundations and Trends® in Computer Graphics and Vision*, 16(1-2):1–214, 2024.

[12] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023.

[13] Sourab Mangrulkar and Sayak Paul. Peft: Parameter-efficient fine-tuning of billion-scale models on low-resource hardware, 2023.

[14] Daniel S. Marcus, Tracy H. Wang, Jamie Parker, John G. Csernansky, John C. Morris, and Randy L. Buckner. Open access series of imaging studies (oasis): Cross-sectional mri data in young, middle aged, nondemented, and demented older adults. *Journal of Cognitive Neuroscience*, 19(9):1498–1507, 09 2007.

[15] Sophie A Martin et al. Interpretable machine learning for dementia: A systematic review. *Alzheimer's & Dementia: The Journal of the Alzheimer's Association*, 19(5):2135–2149, 2023.

[16] R. C. Petersen, P. S. Aisen, L. A. Beckett, M. C. Donohue, A. C. Gamst, D. J. Harvey, Jr Jack, C. R., W. J. Jagust, L. M. Shaw, A. W. Toga, J. Q. Trojanowski, and M. W. Weiner. Alzheimer's disease neuroimaging initiative (adni): clinical characterization. *Neurology*, 74(3):201–209, 01 2010.

[17] Arjun Punjabi, Adam Martersteck, Yanran Wang, Todd B Parrish, Aggelos K Katsaggelos, and Alzheimer's Disease Neuroimaging Initiative. Neuroimaging modality fusion in alzheimer's classification using convolutional neural networks. *PloS one*, 14(12):e0225759, 2019.

[18] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. *CoRR*, abs/2103.00020, 2021.

[19] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019.

[20] Minh-Hao Van, Prateek Verma, and Xintao Wu. On large visual language models for medical imaging analysis: An empirical study. In *2024 IEEE/ACM Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE)*, pages 172–176. IEEE, 2024.

[21] Janani Venugopalan et al. Multimodal deep learning models for early detection of alzheimer's disease stage. *Scientific reports*, 11(1):3254, 2021.

[22] Shibo Wang and Pankaj Kanwar. Bfloat16: The secret to high performance on cloud tpus, 2019.

Appendix A LLaVA Model Details

LLaVA: Large Language and Vision Assistant is an end-to-end trained large multimodal model that connects a vision encoder and an LLM for general purpose visual and language understanding [12]. Expanding on the notion that machine-generated instruction-following data has improved zero-shot capabilities on unseen tasks in the language domain, LLaVA applies this to the multimodal sphere by integrating both visual and textual input. LLaVA is fine-tuned on multimodal language-image instruction-following data generated by language-only GPT-4. It serves as a visual conversational language assistant that exhibits human-like interactions with users. LLaVA is developed by connecting the open-set visual encoder of CLIP ViT-L/14 [18] with the language decoder Vicuna [4], and fine-tuning on generated instructional vision-language data 11. It first uses a simple projection layer to map image features into the word embedding space, effectively training a compatible visual tokenizer for the frozen LLM. The second stage is end-to-end fine-tuning, in which they keep the visual encoder weights frozen and continue to update both the pre-trained weights of the projection layer and the LLM. Through the incorporation of a vision encoder, LLaVA can analyze image inputs and provide responses for diverse vision-language tasks, showcasing robust capabilities in visio-linguistic comprehension.

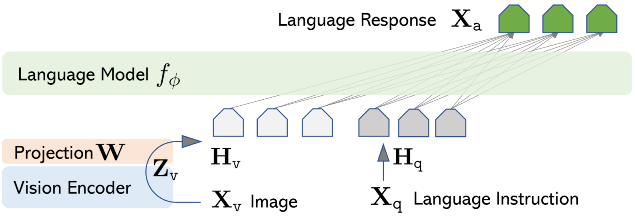


Figure 11: LLaVA Model Architecture

Appendix B Fine-Tuning Details

We use a batch size of 5, learning rate of $1 * 10^{-4}$ and cross-entropy loss as the loss function. We accumulate the gradient of the model over three batches before stepping the optimizer, resulting in an effective batch size of $batch\_size * 3$. The AdamW optimizer was used to address overfitting issues. Validation accuracy was monitored to implement early stopping with a patience of 3 epochs and a maximum of 7 epochs to prevent overfitting. During testing, batch inference was performed with a batch size of 20 to evaluate the model's performance.

The model was fine-tuned using the following prompt, and evaluated using the same prompt without the ground truth:

"USER: [image]

Classify this image as [comma separated labels].

ASSISTANT: [ground truth]"