Interpretable and Inclusive MRI-Based Neurodegenerative Disease Classification

Christina Hahn
University of Washington
chahn317@cs.washington.edu

Di Mao
University of Washington
dimao@cs.washington.edu

Molly Park
University of Washington
parkmoll@cs.washington.edu

Abstract

Automated solutions for medical imaging interpretation offer a promising way to improve the accessibility, accuracy, and consistency of diagnosis. In medical imaging diagnostic tasks, modern deep learning models have shown predictive performance competitive with that of board-certified radiologists [14]. However, despite the performance gains associated with increasingly complex models, key interpretability challenges limit wide scale clinical adoption. In this project, we implement an end-to-end deep learning system that identifies the presence of dementia from a patient's brain MRI (magnetic resonance imaging) scan and provides an explanation of salient regions used for prediction through gradient-weighted class activation maps. We investigate the predictive capabilities of ViT, ResNet, and Inception models, and we also evaluate the impact of various data augmentation techniques on performance. We then do a qualitative evaluation of a select subset of the most influential training images by visualizing regions of the image that were used for prediction, and determining whether or not they align with known markers of neurological disease. Our codebase is available on GitHub.

1. Introduction

Neurological diseases such as Alzheimer's disease (AD), Parkinson's disease, and frontotemporal dementia are leading causes of disability and mortality worldwide, impacting the quality of life for millions [4]. According to the World Health Organization, about 57 million people in the world have dementia, and the number increases by 10 million every year [10]. Thus, early and accurate diagnostic tools are crucial.

The advancement of deep learning models has led to the development of neurology diagnostic tools that assist physicians with brain imaging interpretation. Though promising, the adoption of deep learning in clinical settings is hindered by the fact that deep learning models function as "black boxes," offering little insight into how its decisions are made [16, 17]. This lack of transparency limits their acceptance in medical practice. Additionally, the underrepresentation of diverse demographic groups in brain imaging datasets introduces biases that can negatively affect model performance, especially for underrepresented populations. These biases raise concerns about the fairness and generalizability of AI-driven diagnostic tools.

To address these challenges, this project focuses on two key aspects: the interpretability of deep learning models and the inclusion of underrepresented groups in the training data. We use the BrainLat dataset, which provides brain imaging data from 780 individuals in Latin America, to examine the potential of deep learning models for equitable and interpretable diagnostics [11]. However, we found no existing literature using the BrainLat MRI dataset for this type of task. We chose this dataset for its focus on an underrepresented population, since most literature about neurodegeneration focuses on Europe, the US, and other high-income regions [9]. We also curated a hybrid dataset combining BrainLat with the OASIS dataset, which is a cross-sectional collection of 416 subjects, including individuals with early-stage AD [3]. We then used state-of-the-art deep learning-based architectures, such as Vision Transformers (ViT), ResNet, and InceptionV3, to analyze this brain imaging data. Finally, we performed several interpretability analyses to identify the most important features contributing to model decisions and assess their alignment with established neuroscience literature.

Our findings will demonstrate the feasibility of creating trustworthy and interpretable AI-driven diagnostic tools, addressing both technical and ethical challenges in deploying these systems in clinical practice.

2. Related Work

Advances in deep learning over the recent decades, particularly in deep convolutional networks and self-supervised learning, have resulted in major progress in the field of detecting neurological diseases through neuroimaging. Many prominent studies have supported the effectiveness of CNNs in detecting various diseases from different forms of scans. For instance, Helaly et al. developed a

deep learning framework using CNNs (and transfer learning) trained on the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset to classify MRI scans into stages of AD, achieving up to 93% accuracy [5].

While the performance and accuracy of these models are often high, result interpretability remains a constant challenge. Previous studies have explored post-hoc explainability methods to visualize the specific brain regions that influence model decisions, though not many are used to detect AD specifically. Wang et. al used a combination of Guided GradCAM (GGC), Layer-wise Relevance Propagation (LRP), and Integrated Gradients (IG) to enhance transparency in AD prediction models, identifying brain regions consistent with known pathological findings [18]. This study demonstrated that heatmaps generated by interpretability methods like GradCAM, Integrated Gradients, and Layer-wise Relevance Propagation align with AD disease patterns, performing better than SVM coefficients [18]. However, the availability of these techniques in the clinical setting is limited and their reliability across diverse populations have not been studied.

Due to legal and privacy issues, it is difficult to curate a large amount of brain data [13]. As such, many studies train their models with datasets like ADNI, which consists primarily of North American and European populations. However, the lack of diversity in these datasets raises growing concerns about bias in deep learning models. Not only are underrepresented groups at greater risk for AD, but models trained on non-diverse data may also result in reduced diagnostic performance when applied to underrepresented populations [19]. Previous efforts to address these issues include the collection of demographically inclusive datasets. Notably, the BrainLat project represents one of the first large-scale initiatives that focus on collecting neuroimaging data from Latin American populations, with a goal of supporting culturally and regionally relevant neurological research.

We build directly on this context by exploring the application of interpretability methods within diverse population datasets.

3. Methods

At a high level, our experimental setup can be divided into three steps: data preprocessing, transfer learning, and interpretability analysis. After normalizing our image data, we fine-tuned three popular deep learning-based image classification models (ViT-B/16, ResNet50, and Inception-v3), then finally conducted an interpretability analysis using GradCAM.

The goal of our experiments were threefold: (1) evaluate the impact of various data preprocessing and augmentation techniques on the model's ability to correctly classify neurological disease state, (2) compare the performance of competitive modern vision model architectures, and (3) investigate if parts of the image used by top-performing models for inference aligned with known anatomical predictors of Alzheimer's disease. The objective of these experiments was to not only develop a model with strong predictive capability, but also to assess the trustworthiness of those predictions.

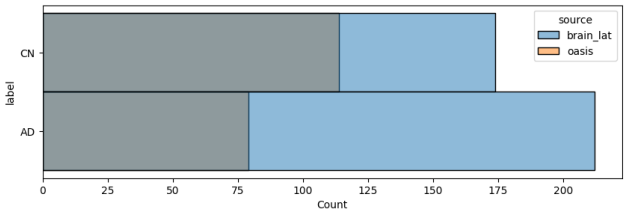


Figure 1. Distribution of Alzheimer's Disease (AD) and healthy control (CN) samples in the BrainLat and OASIS datasets.

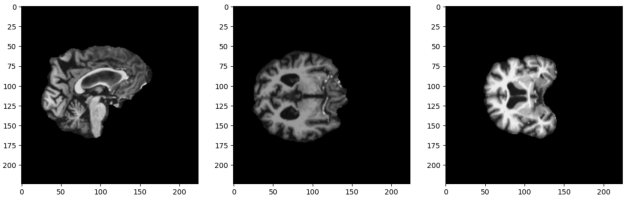


Figure 2. Example sagittal (left), axial (middle), and coronal (right) skull-stripped image slices.

3.1. Dataset and Preprocessing

We used the BrainLat dataset, which is comprised of brain imaging data from Latin American individuals in multiple medical centers across Argentina, Chile, Colombia, Mexico, and Peru. The dataset was missing a substantial number of MRI images (513/780 present total, 389/780 for the AD and control classes), so we experimented with mixing in a subset of the OASIS-1 dataset (N=235) to increase the available number of samples. For consistency, we only used the first MRI per patient from OASIS-1, and omitted non-AD classes from BrainLat.

The dataset of AD and CN images were divided into an 80%-10%-10% train-validation-test split. We normalized each image's pixel value by subtract the dataset's minimum values and dividing by the range. To ensure compatibility with existing 2D deep learning frameworks, we only used the middle slice of each MRI volume along each dimension. To account for size heterogenity in our multi-medical center data, we resized images to a resolution of 224×224 pixels (and 299×299 for training Inception-V3) by scaling and padding. We did not change the aspect ratio to maintain the integrity of the original image.

3.2. Dataset Variations

Here, we evaluate how different data preprocessing and augmentation techniques impact predictive performance. We crafted the following variations of our input dataset:

- BrainLat-Standard (N=389): Middle slices are taken from the axial perspective, using normalized volumes in the original BrainLat dataset alone.
- BrainLat-skullstrip: BrainLat MRI volumes initially processed using SynthStrip [6], a segmentation software that strips out non-brain tissue and background artifacts. Volumes were then sliced in the same manner as BrainLat-standard. We hypothesized that this preprocessing step would reduce variability attributed to differences in acquisition methods across medical centers.
- BrainLat-OASIS (N=624): Mix of the BrainLat (N=389) dataset with images from OASIS-1 (N=235) that were also labeled with AD diagnoses. Since these were taken from different studies in different countries, all images were skullstripped.

To investigate the impact of regularization on performance, we also crafted an alternative version of BrainLat-standard where for each image slice, a randomly sized patch had a 25% chance of being blacked out.

3.3. Model Architecture and Training

Three deep learning architectures were selected for this study given their strong performance on general image tasks: Vision Transformer (ViT-B/16), ResNet-50, and Inception-V3. Models were pretrained on ImageNet-1k [2]. We hypothesized that leveraging the general image representations learned during pretraining would be beneficial given our small dataset. Ideally we would use pretrained weights from other MRI tasks, though we were unable to find these open source. We finetuned these models further on our processed BrainLat dataset for a maximum of 20 epochs using cross entropy loss. We used PyTorch's StepLR learning rate policy, and we also implemented early stopping to prevent overfitting. Hyperparameters were determined through random search over the same ranges of optimizers and dropout rates. After tuning each version over the same range of hyperparameters on each dataset, we compared the accuracy and F1 score of the generated predictions.

3.4. Interpretability

For our interpretability analysis, we used an implementation of TracIn [12] from the library Captum [7] to select the top 10 most influential training examples and identify general trends. TracIn is a method to calculate the influence of a given training example on a given test example. This influence score roughly represents how much higher the loss for the given test example would be if the given training example were removed from the training dataset and the model were re-trained [12].

We also used GradCAM [15] to generate heat maps visualizing the most influential regions of each image. For select images, we conducted qualitative case studies to determine if the concentrated regions in the heat map matched brain regions known to be affected by AD. Early stages of AD associated with hippocampal shrinkage, while later stages are associated with cortical shrinkage [8].

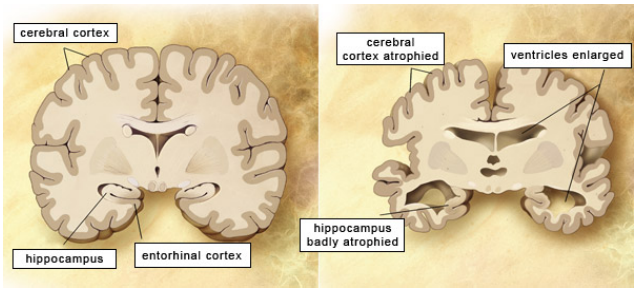


Figure 3. The effect of AD on the brain. Reproduced from [1] under the CC BY-NC-SA 3.0 license.

4. Results

4.1. Model Evaluation

We trained the ViT, ResNet, and Inception-V3 pretrained base models on the three dataset variants mentioned in Section 3.2 (results in Table 1). Here, ResNet-50 showed the strongest performance with the top score in 2/3 of the dataset variants. We hypothesize that ViT's lagging performance despite state of the art results on general benchmarks is likely due to the small size of our dataset. Moreover, all three models performed worst on BrainLat-OASIS. We believe this may be attributed to the fact that the BrainLat and OASIS datasets were collected in different continents, and structural markers of disease can vary across patients from different demographics. OASIS-1 also incorporates patients with early-stage AD, who may have more subtle structural indicators of disease that models may struggle to pick up.

When investigating the impact of data augmentation (random cut outs), we found that models trained with augmentation performed slightly better than the un-augmented models when evaluated on test examples. Though further trials and a larger test set would be necessary to evaluate the significance of this, this result suggested that random augmentation was helpful for reducing overfitting.

Dataset	Model	Accuracy	F1
BrainLat-Standard	ViT-B/16	0.775	0.774
	ResNet-50	**0.800**	**0.799**
	InceptionV3	0.750	0.747
BrainLat-Skullstrip	ViT-B/16	0.775	0.768
	ResNet-50	**0.825**	**0.816**
	InceptionV3	0.800	0.795
BrainLat-OASIS	ViT-B/16	0.661	0.661
	ResNet-50	0.746	0.741
	InceptionV3	**0.780**	**0.779**

Table 1. Accuracy (proportions) and F1 scores for predictions on the test set across model types.

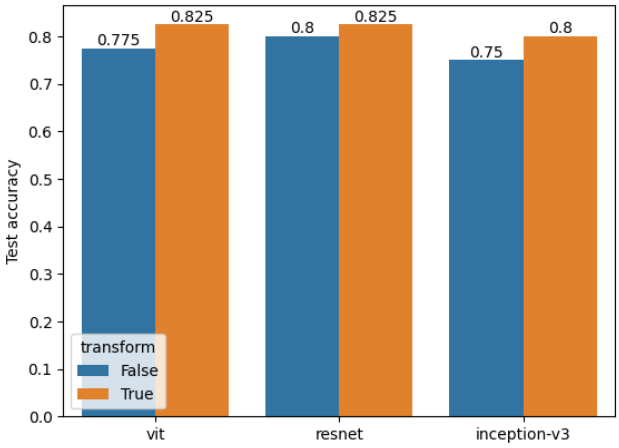


Figure 4. Comparison of model performance when trained with and without data augmentation (transform = with augmentation)

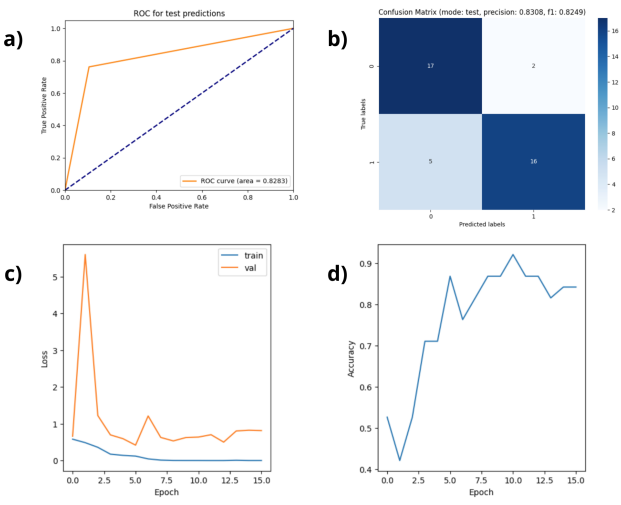


Figure 5. Evaluation and training metrics for the top performing model (ResNet-50 on BrainLat-Standard with augmentation). (a) Receiver Operating Characteristic (ROC) curve, (b) confusion matrix for test examples, (c) train and validation loss curves during training, (d) validation accuracy during training

Moreover, we validated our decision to use the axial slice by training one of our strongest model, ResNet-50, on the axial/coronal/sagittal slices. Since each provides a different perspective of the brain structures of interest, we figured they may have different predictive abilities. Our results indicate that the axial slice is fitting for our task.

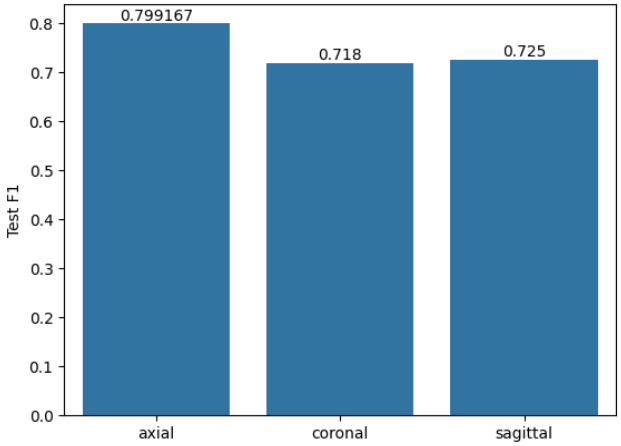
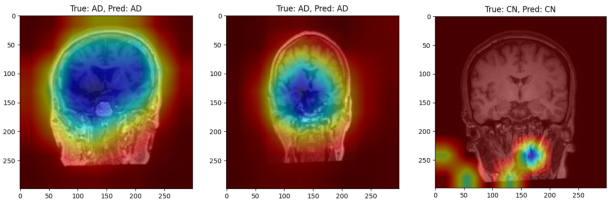


Figure 6. Comparison of performance across image slices using the ResNet-50 model. Here, we use the same patients as test set but used alternative perspectives of their brain volumes.

4.2. Interpretability Evaluation

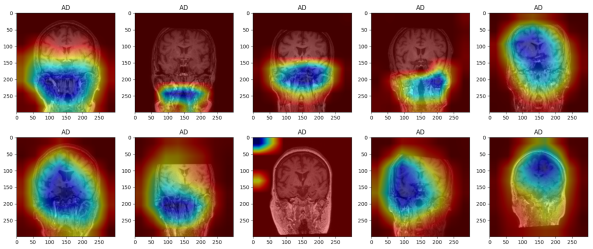
AD is associated with hippocampal shrinkage in earlier stages and cortical shrinkage in later stages [8]. From our interpretability analysis using TracIn and GradCAM, we observed results consistent with these findings.



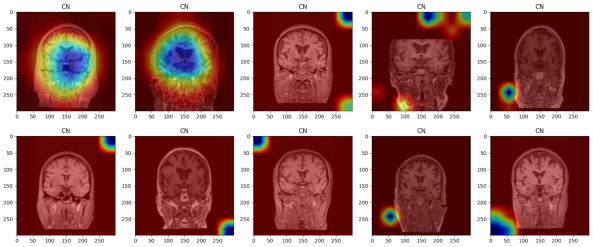
(a) (b) (c)

Figure 7. GradCAM heatmaps of selected test examples for the class target AD. We show (a) a brain with severe AD, (b) moderate AD, and (c) no AD. The coronal slice is used for clearer visualization of brain structures.

The model determines the entire brain to be influential for cases of severe AD, possibly due to the severe cortical shrinkage as discussed in neuroscience literature (Figure 7). For more moderate cases, the model focused more on the hippocampal and lower-cortical regions, which matches the area of atrophy in earlier stages of AD (Figure 7). Finally, the model tended to not focus on any particular area for healthy control brains.



(a)



(b)

Figure 8. The top proponents and opponents of the test example from Figure 7a. (a) shows the top 10 proponents, while (b) shows the top 10 opponents.

To select case studies for further analysis, we identified influential training examples selected by TracIn [12]. More formally TracIn defines the influence of training example z on test example z' as the similarity between their loss gradient vectors (averaged across training checkpoints):

$$\text{TracIn}(z, z') = \sum_{t:z_t=z} \eta_t \nabla \ell(w_t, z') \cdot \nabla \ell(w_t, z)$$

With the intuition that a test and train example are similar if they would change model parameters in a similar direction, TracIn defines "proponents" and "opponents" as training examples the most similar and most different gradient loss vectors respectively. Theoretically, strong proponents made it "easier" to classify the test example, while strong opponents made it more difficult. Thus, strong proponents are typically similar images of the same class as the test image, while strong opponents are ambiguous images of the wrong class.

From Figure 8, several of the top proponents of Figure 7a are similar images of the correct class, showing high degrees of cortical shrinkage. Meanwhile, the top two opponents are similar images of the incorrect class - essentially outliers in the control dataset. Since the differences are difficult to see with raw images, we overlaid GradCAM heatmaps on the training examples shown in Figure 8. Similar trends can be observed across different test examples, indicating that the model has learned to differentiate between healthy brains and brains showing signs of shrinkage, a key symptom of AD.

However, we also observe that there are many influential training examples with heatmaps that do not focus on expected regions. For example, our top performing model (ResNet-50 trained on the BrainLat-Standard dataset with augmentation) obtained 82.5% accuracy, but manual inspection revealed that only 65% of test images had GradCAM patches localized on a region in the brain rather than the skull or background. This indicates that there were correct predictions that utilized irrelevant parts of the image. So, although we can observe general trends, our model should not be expected to reliably identify brain regions of interest. Moreover, applying the GradCAM technique to the vision transformer produced several nonlocalized circular patches tiling the image, resulting in uninterpretable explanations compared to its CNN-based counterparts (9). This is likely because ViTs use flatten patches and attention rather than convolutional feature maps. We consider these inconsistencies to be a limitation of this work, highlighting the need for the development of more interpretable prediction frameworks in the medical domain.

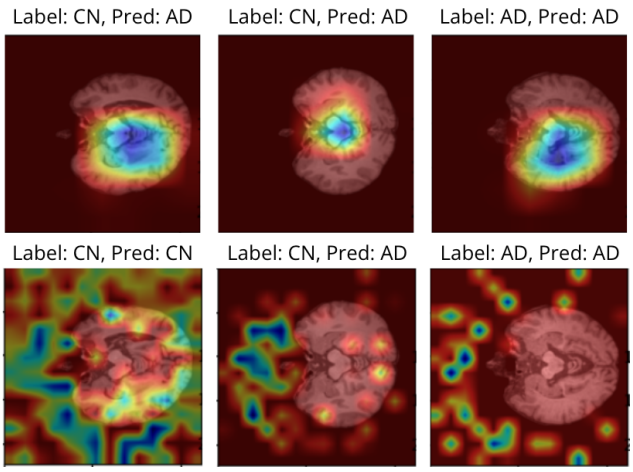


Figure 9. Example of ViT (top) vs. ResNet-50 (bottom) GradCAMs for a random selection of 3 skullstripped axial slices

5. Discussion

The results show the importance of deep-learning neurological disease diagnosis that are both interpretable and inclusive without sacrificing performance. By using imaging data from underrepresented Latin American populations, we address a lack of demographic representation in current neuroimaging research. Furthermore, we show that performance can be maintained while improving generalizability to include underrepresented populations. However, some models, like the ViT-B/16 base model, had fluctuating accuracies, possibly due to the smaller sample size in the BrainLat dataset. This suggests that the Vit-B/16 model offers limited advantages over the other explored CNN architectures.

Our interpretability analysis highlights that task-agnostic evaluation metrics (e.g. accuracy, F1) on a medical imaging tasks does not correlate or imply that a model has learned to use clinically-relevant features for predictions. Interpretability tools, like TracIn and GradCAM, provide insight into model decision-making. The models generally, but not all the time, focus on regions consistent with findings in neuroscience literature. Further research is needed to understand the factors that contribute to the observed variability in model learning with clinically relevant regions.

Further work in developing inherently interpretable models (e.g. via integration with natural language explanations) may improve the trustworthiness of medical AI diagnostic tools, especially if they were to eventually be adopted in clinical practice. As such, advancing AI use in medicine should not only focus on maintaining high performance but also address ethical and practical concerns like transparency and equity.

Future work may include exploring more model architectures and incorporating multimodal components. For this report, we focused on only one data augmentation technique. Many other methods can be explored to increase validation scores. Currently, our models train on MRI slices, but the BrainLat dataset also includes EEG data and cognitive assessment scores. By combining multimodal models or using them in conjunction, clinicians could gain a more holistic view of the patient's situation. Overall, there are many possible directions to improve the interpretability and inclusivity of neurodegenerative disease diagnostic tools.

References

[1] Kuljeet Singh Anand and Vikas Dhikav. Hippocampus in health and disease: An overview. *Annals of Indian Academy of Neurology*, 15(4):239–246, 2012. 3

[2] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 3

[3] Marcus DS, Fotenos AF, Morris JC Csernansky JG, and Buckner RL. Open access series of imaging studies: longitudinal mri data in nondemented and demented older adults. *Journal of cognitive neuroscience*, 22, 2010. 1

[4] Jaimie D Steinmetz et al. Global, regional, and national burden of disorders affecting the nervous system, 1990–2021: a systematic analysis for the global burden of disease study 2021. *The Lancet Neurology*, 23(4):344–381, April 2024. 1

[5] Hadeer A. Helaly, Mahmoud Badawy, and Amira Y. Haikal. Deep learning approach for early detection of alzheimer's disease. *Cognitive Computation*, 14:1711–1727, 2022. 2

[6] Andrew Hoopes, Jocelyn S. Mora, Adrian V. Dalca, Bruce Fischl, and Malte Hoffmann. Synthstrip: skull-stripping for any brain image. *NeuroImage*, 260:119474, 2022. 3

[7] Narine Kokhlikyan, Vivek Miglani, Miguel Martin, Edward Wang, Bilal Alsallakh, Jonathan Reynolds, Alexander Melnikov, Natalia Kliushkina, Carlos Araya, Siqi Yan, and Orion Reblitz-Richardson. Captum: A unified and generic model interpretability library for pytorch, 2020. 3

[8] F. Márquez and M. A. Yassa. Neuroimaging biomarkers for alzheimer's disease. *Molecular Neurodegeneration*, 14(1):21, 2019. 3, 4

[9] Agustin Ibá nez et al. Dementia caregiving across latin america and the caribbean and brain health diplomacy. *The Lancet Healthy Longevity*, 2(4):222–231, April 2021. 1

[10] World Health Organization. Dementia, March 2025. 1

[11] P. Prado, V. Medel, R. Gonzalez-Gomez, et al. The brainlat project, a multimodal neuroimaging dataset of neurodegeneration from underrepresented backgrounds. *Scientific Data*, 10, 2023. 1

[12] Garima Pruthi, Frederick Liu, Satyen Kale, and Mukund Sundararajan. Estimating training data influence by tracing gradient descent. *Advances in Neural Information Processing Systems*, 33:19920–19930, 2020. 3, 5

[13] Kerstin Ritter, Matija Bosnjak, Sergio Ruiz-España, and et al. Interpretable brain disease classification and relevance-guided deep learning. *Scientific Reports*, 12:1–12, 2022. 2

[14] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. *arXiv preprint arXiv:1711.05225*, 2017. 1

[15] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 618–626, Venice, Italy, 2017. 3

[16] Bas H.M. van der Velden, Hugo J. Kuijf, Kenneth G.A. Gilhuijs, and Max A. Viergever. Explainable artificial intelligence (xai) in deep learning-based medical image analysis. *Medical Image Analysis*, 79:102470, 2022. 1

[17] M.E. Vlontzou, M. Athanasiou, K.V. Dalakleidi, et al. A comprehensive interpretable machine learning framework for mild cognitive impairment and alzheimer's disease diagnosis. *Scientific Reports*, 15:8410, 2025. 1

[18] Di Wang, Nicolas Honnorat, Peter T. Fox, Kerstin Ritter, Simon B. Eickhoff, Sudha Seshadri, and Mohamad Habes. Deep neural network heatmaps capture alzheimer's disease patterns reported in a large meta-analysis of neuroimaging studies. *arXiv preprint arXiv:2207.11352*, 2022. 2

[19] Ying Zhao, Sarah Sisco, Allison Reuben, and et al. Quantification of race/ethnicity representation in alzheimer's disease neuroimaging research in the usa: a systematic review. *Communications Medicine*, 3(1):1–11, 2023. 2