Visualizing CNNs for Interpretable Alzheimer's Diagnosis Through Neuroimaging

Zi Ying Fan
Stanford University
zyfan@stanford.edu

Elissa Li
Stanford University
elissali@stanford.edu

Darian Martos
Stanford University
dtmartos@stanford.edu

Abstract

Research in the deep learning space as applied to disease diagnosis has tended to focus on classification accuracy. However, uptake in the medical community has been met with reticence due to the black-box nature of neural networks. Visualization and interpretation of deep learning models like convolutional neural networks (CNN) is crucial for increasing the credibility of such models for use in the healthcare industry. For Alzheimer's Disease (AD) in particular, it is important to not only visualize the saliency of the model, but also understand its performance in the presence of age, a tightly coupled confounding feature. In this study, we explore a confounder-aware visualization technique to elucidate the relationships and features that CNNs pick up on when diagnosing AD from neuroimages in the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset. We perform this technique on top of four baseline visualization methods from literature that are gradient-based and occlusion-based, for a total of eight visualizations. Our results highlight relevant areas in the brain scan, and we compare the methods qualitatively and quantitatively by referencing empirically relevant areas from neuroscience research. Our results demonstrate that the confounder-aware technique is able to remove age-related influences from the saliency visualization performed on a CNN model trained to classify AD/NC 3D MRI images. Namely, the confounder-free saliency maps do not contain the sprawling pink hue that could be indicative of age-related diffuse cortical atrophy, focus more strongly on the medial temporal gyrus out of the different sections of the temporal lobe, and produces a more symmetric visualization that is consistent with empirical evidence from AD patients. Overall, our results show promise for our visualization methods in facilitating CNN interpretation and in particular, towards a confounder-free understanding of Alzheimer's and the brain, in hopes of aiding medical professionals with diagnosis of this disease.

1. Introduction

Alzheimer's Disease (AD) is a chronic neurodegenerative disease which starts in the temporal lobe and is the cause of approximately 70% of cases of dementia in seniors [10]. Clinical diagnosis by medical imaging can be challenging due to the fact that the normal aging process induces visually-similar patterns of neurological atrophy compared to AD. Machine learning models present a natural solution to this problem by virtue of their ability to recognize brain tissue alterations that humans may have a harder time discerning. Deep learning approaches in particular, such as convolutional neural network (CNN) or recurrent neural network models, have the added benefit of being able to use neuroimaging data without relying on hand-crafted features and have proven to perform with good accuracy [6] [1]. However, are they good enough to avoid learning features that confound the effects of aging with the presence of AD?

Research in the deep learning space as applied to AD diagnosis has tended to focus on classification accuracy, particularly when pushing neural networks as a diagnostic tool. However, due to the black-box nature of deep learning models, uptake in the medical community has been met with some reticence. Medical professionals are interested not only in accuracy but also in the model's decision process, especially given the possibility of confounding image features. For example, AD may co-occur with Frontal Lobe Dementia (FLD), which shares similar symptoms and patient profiles [14]. More generally, a concern with AD classification is that the models predict confounding features like age instead of the disease itself, since AD is heavily skewed towards the elderly population. The opacity of deep learning models makes absolving these concerns difficult and also presents a challenge to medical professionals seeking to verify diagnoses. Clear interpretation of classification processes would increase the credibility of such models for adoption in the healthcare industry.

Towards that end, there is existing literature [9] that explores visualization techniques to elucidate the patterns that CNNs pick up on when diagnosing AD from neuroimages. These visualization methods (basic sensitivity analysis, guided backpropagation, occlusion, and brain area occlusion) use heatmaps to highlight regions in the brain scan that are particularly relevant for a trained CNN model's diagnosis of AD versus normal control (NC). However, these studies do not tell the story of whether age, in particular, exists as a confounding feature for these models.

Our study fills in the gap in literature by training a CNN to classify 3D MRI images from the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset, then applying a confounder-aware technique to this domain [5]. The confounder-aware technique is used in conjunction with basic backpropagation method for visualizing saliency and has previously shown to be insightful when applied against two other neuroimaging datasets on sexual dimorphism and Human Immunodeficiency Virus (HIV) [15]. In addition to applying this technique to a novel dataset, we hybridize the technique with the three other visualization methods used in [9]. By comparing the saliency maps produced by using the confounder-aware technique on top of each of the four baseline visualization methods, we are able to observe the effect of confounding features on the model's prediction. We find that both our baseline and confounder-aware models successfully illuminate areas of the brain that have been empirically shown to be relevant for AD, in particular the temporal lobe and insula. However, the confounder-aware method is able to produce more relevant saliency maps and removes three visual patterns we believe are associated more strongly with aging. Overall, our results show promise for our visualization methods in facilitating CNN interpretation and in particular, towards a confounder-free understanding of Alzheimer's and the brain, in hopes of aiding medical professionals with diagnosis of this disease.

2. Related Work

We innovate on the techniques proposed in [15] and [9] and draw upon various empirical studies that examine areas of the brain with respect to AD and aging, in order to validate our results from a medical perspective.

[15] tackles the challenge of algorithmic bias in CNN models when confounders (factors that correlate with both the input and label) are present due to the nature of the dataset or study by proposing a confounder-aware technique that is used on top of backpropagation-based visualization and can be adapted to various CNN architectures. More specifically, with this approach, a general linear model (GLM) is applied to each feature from the last convolutional layer to identify the variance in feature f_i explained by the prediction scores $s \in R^N$ versus a confounding variable $z \in R^N$, where N is the number of examples.

$$f_i = c + \beta_1 s + \beta_2 z$$

Equation 1: representation of the GLM, where c is a constant and β_1 and β_2 are coefficients to the variables of interest.

A feature is defined as "confounded" if $p < 0.05$ for the null hypothesis that the amount of variance explained by z is zero. A partial backpropagation method is then used for visualizing a confounder-free saliency map, where gradients only flow through un-confounded features.

The authors apply this visualization method to two 3D MRI datasets: one for diagnosing HIV (where age is a confounding variable) and one for analyzing sexual dimorphism in adolescents (where sex is a confounding variable). Their classification models achieve 73% and 89.3% accuracy respectively, and they find that the confounder-aware technique is able to remove brain areas that are visualized by vanilla backpropagation but empirically known to be attributed to the respective confounding factors.

Applying the confounder-aware technique to our dataset of interest is not only convenient, since the study in [15] was also performed on 3D MRI images, but also evaluative of the method itself, since the confounding effect of age on AD is believed to be much stronger than the confounders examined with the HIV and sexual dimorphism datasets. Furthermore, we are interested in expanding upon the work of [15] to examine whether the confounder-aware technique can be combined with other visualization methods.

To that end, we draw from [9], which examines four different methods of visualization on the ADNI dataset: sensitivity analysis (backpropagation), guided backpropagation, occlusion, and area occlusion. In this study, the authors train a four layer CNN to achieve a binary classification accuracy of 84% on the ADNI dataset and examine the results of all four visualization methods on this trained model, qualitatively via the saliency maps plotted over brain images and quantitatively via identifying top areas of the brain represented in those saliencies.

By running all four methods on our trained model and replicating the results of [9], we thus have a set of baseline visualizations and quantitative representations of the most relevant brain areas. We apply the confounder-aware technique on top of all four methods, producing four hybrid methods (three of which are novel contributions of this project), for a total of eight sets of results examined in our study.

3. Data

Our dataset comes from the ADNI database and consists of 1334 1.5T 3D MRI images collected across patients. Images of patients with AD are given positive labels, while NC patients serve as negative examples. Each example also comes with metadata, such as the date the image was taken and the age of the patient. Preprocessing steps were performed by Zhao et al. and included "denoising, bias field correction, skull striping, affine registration to the SRI24 template (which accounts for differences in head size), and re-scaling each image to a 64x64x64 volume" [15]. Finally, we apply data augmentation via rotations and shifts such that the number of positive and negative examples is balanced. The data augmentation has the additional benefit of increasing our sample size, since 3D datasets in the neuroimaging field are typically small by comparison to typical deep learning studies.

We split the data into train, validation, and test sets on the patient-level, in order to avoid overlap between images shown to the network at train and evaluation time. A summary of the train, validation, and test datasets used is provided in the figure below.

Training Set		
	NC	AD
original	538	276
augmented	486	748
total	1024	1024

Validation Set		
	NC	AD
original	180	89
augmented	76	167
total	256	256

Test Set		
	NC	AD
original	162	89
augmented	94	167
total	256	256

Table 1: Summary of augmentation and split of project dataset.

4. Methods

4.1. Baseline Classification

The visualization techniques that we are interested in require a trained CNN model as input. We train a CNN model to perform baseline binary classification. The architecture consists of four convolutional layers, each with relu activation, batch normalization, and maxpool of size 2. We also use dropout with probability of removing a neuron $p = 0.1$ on the middle two layers. Each layer uses the same convolutional filter size (3x3x3), stride=1, and padding=1 but differs in the number of filters (16, 32, 64, 16.) Finally, we use three fully connected layers with a sigmoid activation to provide a single classification output. Our network uses an Adam optimizer.

4.2. Baseline Visualization Methods

Backpropagation and occlusion-based methods are two common classes of visualization techniques. We adapt the visualization methods provided by [9], which include sensitivity analysis, guided backpropagation, occlusion, and brain area occlusion.

Sensitivity analysis (1) calculates the gradient of the model's output probability with respect to the input brain image, with the gradient value for each pixel representing the amount the classification probability changes when the pixel value changes. Guided backpropagation (2) sets negative gradients to 0 at ReLU layers in the backward pass, whereas (1) takes the absolute value of gradients. Occlusion (3) slides a black patch across the image and recalculates the output probability for each "occluded" image, then compares the probability of the resulting output class of the occluded image to the original probability of the raw image. Image area relevance is indicated by the "occluded" probability decreasing compared to the original raw probability. In our experiments, we use 8x8 occlusion patches with a stride of 8. Finally, brain area occlusion (4) occludes an entire area of the brain using the Automated Anatomical Labeling atlas [11].

4.3. Confounder-Aware Technique

For brevity, we refer to the convolutional layers of our model as the "feature extractor" portion and the fully connected layers that produce the prediction output as the "classifier" portion. We implement the confounder-aware technique by flattening the 4x4x4 output activations of the feature extractor. Since this final convolutional layer has 16 filters, this results in a total of 1024 features. For each feature f_i, we fit a GLM of the form in Eq. 1, where z is an array of ages corresponding to each example. We use the same criteria as [15] to identify confounders - namely, if $p < 0.05$ for a variable, we deem it a confounder. We construct a mask of size 1024 indicating which features should be removed, due to being confounded.

In order to perform partial backpropagation, we use the "refactorization trick" provided by [15], where we implement a mask layer that either allows gradient flow or disallows it, based on the confounding features identified above. We then insert this mask layer in between the feature extractor and classifier portions of our model, producing a modified version of the original model. This confounder-aware model is then used with the four methods from the section above to produce saliency maps where the effects of the confounded features are removed.

5. Results

5.1. Baseline Classification

We train our model with a binary cross-entropy with logits loss function and L2 regularization, and we use a scheduler that reduces the learning rate on plateaus based on the validation accuracy from each epoch.

Hyperparameter search was performed to improve upon the classification accuracy. Through examining the training and validation loss and accuracy curves, we determined that the most influential hyperparameters affecting model performance and convergence were learning rate and the number of training epochs. We used a grid search method to discover suitable ranges of values for the learning rate.

Our final model uses an initial learning rate of 0.001 and achieves 84% test classification accuracy over 40 epochs of training. This is consistent with comparable architectures in literature [5] and is sufficiently high for the purposes of our project, since we are interested in visualization and not optimizing classification accuracy.

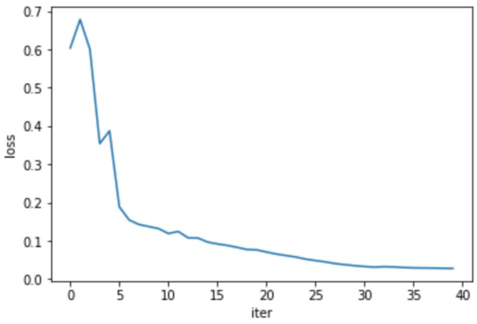


Figure 1: Training loss vs epoch for baseline classifier. Note that for each epoch, we only sample the loss of the last iteration.

5.2. Baseline Visualization

We apply the sensitivity analysis, guided backpropagation, occlusion, and brain area occlusion methods from [9] to our trained model. Red areas of the heatmaps produced by these methods indicate parts of the image that were especially relevant for the model's classification decision. We perform each method over our test set and average the results within the NC and AD classes, in order to obtain a representative saliency map. We plot these saliency maps over the mean image of all the test examples. Note that for both the baseline methods and confounder-aware technique, we use the unaugmented test set, since age information is unavailable for synthetic examples. The results produced by our network are shown in the figure below.

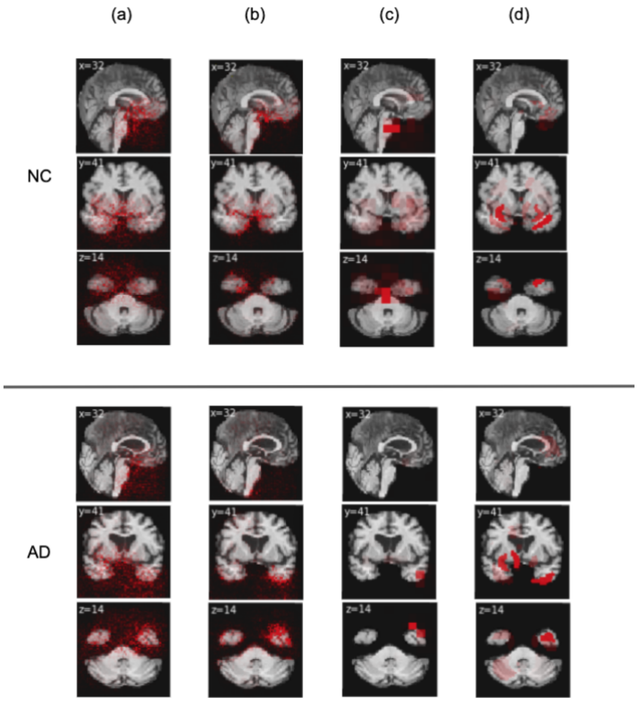


Figure 2: Sample baseline visualization results of different slices along the x, y, and z axes, averaged across all unaugmented test images per NC and AD conditions. (a) Sensitivity analysis (b) Guided backpropagation (c) Occlusion (d) Brain area occlusion.

In line with [9], we find that the relevance maps for AD and NC samples tend to illuminate similar areas of the brain. Intuitively, this is because regardless of the ultimate diagnosis, the model should focus on the same regions when determining the presence or absence of AD. We observe more variation between AD and NC samples for occlusion-based methods, but this instability is also in line with [9] and can be explained by the network's susceptibility to confusing the occlusion patch with brain atrophy and thereby increasing the AD-classification likelihood in those areas. In general, we find that the relevance maps focus on the temporal lobe, as expected. Unlike the results of [9], our model also identifies the insula as an area of interest, and it has also been shown that this region of the brain is affected since the early stages of AD [3]. The most relevant brain areas as identified by the visualization techniques on our model are shown in the table below.

	Sensitivity analysis	Guided backpropagation	Occlusion	Brain area occlusion
AD	TemporalInf (6.3%) FrontalInfOrb (5.1%) TemporalMid (4.1%) Insula (4.1%)	TemporalInf (8.3%) FrontalInfOrb (7.4%) TemporalPoleMid (7.2%) TemporalPoleSup (5.7%)	TemporalInf (7.3%) TemporalMid (6.8%) FrontalInfOrb (5.0%) Fusiform (4.6%)	Insula (12.4%) TemporalPoleSup (11.2%) TemporalMid (8.9%) TemporalSup (4.4%)
NC	TemporalInf (9.8%) FrontalInfOrb (8.6%) Insula (8.1%) TemporalMid (7.4%)	TemporalInf (12.3%) FrontalInfOrb (9.6%) TemporalPoleMid (7.9%) TemporalMid (7.8%)	TemporalMid (19.3%) TemporalSup (13.0%) TemporalInf (6.8%) Lingual (6.1%)	Insula (20.9%) TemporalSup (19.1%) TemporalPoleSup (17.7%) TemporalMid (17.6%)

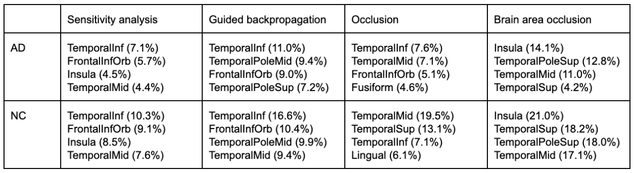
Table 2: Most relevant brain areas per baseline visual-

ization method, for average test image per AD and NC example.

5.3. Confounder-Aware Technique

The result of the GLM test masked out 426 confounding features in our model out of 1024. Since this is more than 5% of all features (the threshold for having $p < 0.05$ simply by chance), it is meaningful to apply the confounder-aware technique on top of the baseline visualization methods.

We find that the relevance maps for the confounder-aware visualizations replicate those of our baseline visualizations, with the top four areas staying unchanged for each of the visualization methods. However, we notice that overall percentages (measures of relevance) increase across the board for the confounder-aware model, indicating that the confounder-aware model places more focused attention on the important areas compared to the baseline model. TemporalSup in the brain area occlusion method is the sole exception, with its relevance dropping marginally, but the overall effect is negligible and the ordering among the top four relevant areas remains unchanged. The most relevant brain areas as identified by the visualization techniques on our confounder-aware model are shown in the table below.



	Sensitivity analysis	Guided backpropagation	Occlusion	Brain area occlusion
AD	TemporalInf (7.1%) FrontalInfOrb (5.7%) Insula (4.5%) TemporalMid (4.4%)	TemporalInf (11.0%) TemporalPoleMid (9.4%) FrontalInfOrb (9.0%) TemporalPoleSup (7.2%)	TemporalInf (7.6%) TemporalMid (7.1%) FrontalInfOrb (5.1%) Fusiform (4.6%)	Insula (14.1%) TemporalPoleSup (12.8%) TemporalMid (11.0%) TemporalSup (4.2%)
NC	TemporalInf (10.3%) FrontalInfOrb (9.1%) Insula (8.5%) TemporalMid (7.6%)	TemporalInf (16.6%) FrontalInfOrb (10.4%) TemporalPoleMid (9.9%) TemporalMid (9.4%)	TemporalMid (19.5%) TemporalSup (13.1%) TemporalInf (7.1%) Lingual (6.1%)	Insula (21.0%) TemporalSup (18.2%) TemporalPoleSup (18.0%) TemporalMid (17.1%)

Table 3: Most relevant brain areas per confounder-aware visualization method, for average test image per AD and NC example.

Correspondingly, side by side comparison of the saliency maps shows that across all four methods, the most important areas used by the classification model, such as the temporal lobe, are highlighted by both the confounder-aware as well as the baseline models. However, the confounder-aware methods produce less "noisy" visualizations that remove the diffuse pink hue seen in baseline results. Sample comparisons are shown below.

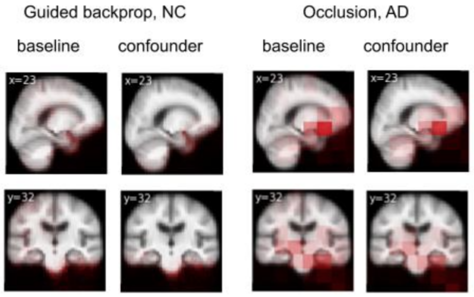


Figure 3: Side-by-side comparison of select slices in the baseline visualization vs. the confounder-aware visualization. The baseline visualization exhibits more diffuse coloring than the confounder visualization.

Baseline models may show a pinkish overall hue due to learning features corresponding to diffuse cortical atrophy, the overall shrinkage of the brain associated with aging [7]. In particular, in slices along the y-axis, baseline models show a noticeable pink hue around the frontal lobe. Empirical research has shown that the prefrontal cortex is one of the regions that is most severely affected by aging [8]. On the other hand, atrophic changes in the temporal lobe are frequently found in early and late stages of AD "but are less commonly related to normal aging" [7], so it is promising that our confounder-aware models produce visualizations that are more precisely focused around the temporal area.

Within the temporal lobe, some additional observations identify meaningful differences between the baseline and confounder-aware methods. For example, in the comparison of lateral views of brain area occlusion in Fig. 4 below, we see that both models are highly focused on the temporal_pole_sup area. However, whereas the baseline visualization seems to place similar importance upon the superior, medial, and inferior parts of the temporal lobe, the confounder-aware visualization is much darker in the medial temporal lobe – in particular, removing much of the importance from the superior temporal lobe. This is consistent with clinical evidence that atrophy of the medial temporal cortex is a hallmark sign of AD and strongly associated with its disease progression [4] [12].

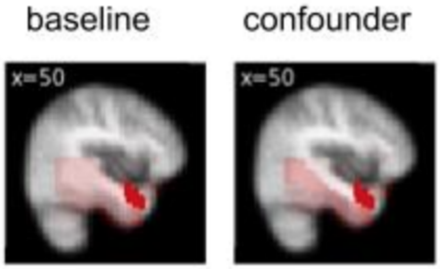


Figure 4: Comparison of lateral slice at x=50 for baseline and confounder models with the brain area occlusion visualization method on AD examples.

Additionally, we discover through the image slices along the y-axis that the baseline visualizations tend to produce asymmetric results, with more highlighting of the left side of the brain. However, the confounder-aware approach typically removes this emphasis and produces more symmetrical saliency maps, such as in Fig. 5 below. While there is evidence of asymmetrical atrophy in the natural aging progression and other aging-related diseases such as semantic dementia (where there is specifically more left-sided damage), neural decline in AD is generally observed to be symmetrical, so the results seen here are also supportive of the confounder-aware model producing more AD-specific saliency maps [2].

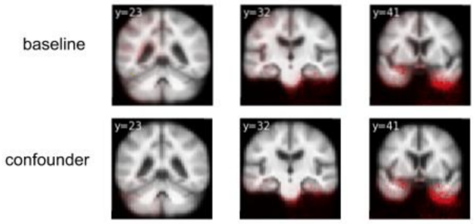


Figure 5: Comparison of baseline and confounder models with the guided backprop method on NC examples.

Finally, we note that the presence of AD as seen through MRI images relies mostly on symptoms of neural atrophy. However, the rate of loss in total brain volume accelerates noticeably after age 50-55, so the older the patient, the harder it becomes to disentangle the two causes [13]. Considering that our dataset has a relatively high average patient age (mean = 75.14, std dev = 6.46), we believe the confounder-aware technique was able to produce noteworthy results in this study.

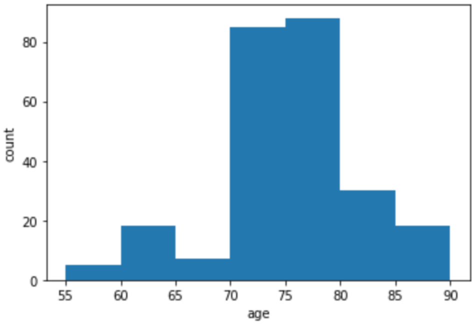


Figure 6: Histogram of patient age distribution corresponding to test images used.

6. Conclusion

Despite the high mean age of our test set, in this study, we achieve results that indicate the confounder-aware technique is able to remove age-related influences from the saliency visualization performed on a CNN model trained to classify AD/NC 3D MRI images. Although the overall patterns produced by baseline and confounder-aware visualizations appear similar, there are important distinguishing factors: namely, the confounder-free saliency maps do not contain the sprawling pink hue that could be indicative of age-related diffuse cortical atrophy, focus more strongly on the medial temporal gyrus out of the different sections of the temporal lobe, and produces a more symmetric visualization that is consistent with empirical evidence from AD patients.

Future work could delve into more detail to identify the specific cause of the asymmetrical saliency produced by the baseline methods, since there could be a variety of age-related natural and pathological explanations for such. This could give better insight into the specific features learned by the model and possibly illuminate the presence of latent disease patterns present in the database. With a more refined understanding of the behavior of the confounder-aware technique, future work could also apply it to classification tasks that are expected to exhibit more subtle differences, such as that of AD versus mild cognitive impairment, or AD versus FTD.

7. Contribution and Acknowledgements

Our code can be found in this repository: https://github.com/kathyfan/cs231n-adni Starter code from [15] and [9] are found at the below links: https://github.com/jrieke/cnn-interpretability and https://github.com/QingyuZhao/Confounder-Aware-CNN-Visualization

Kathy set up the GCP project and adapted the code for the baseline classification model. She wrote the code for data-related functions, the GLM, and running the baseline and confounder-aware models over the test set. She pair-programmed with Elissa to debug throughout the coding process. Elissa adapted the code for the baseline visualization methods from J. Rieke's repository. She wrote the code for incorporating the confounder mask layer into the classification model and was in charge of running all the code to produce the results. She pair-programmed with Kathy to debug throughout the coding process. Darian performed hyperparameter search for the baseline classification model and provided formatting and citations for both the milestone and final reports. All members contributed to writing both project reports.

We would like to express our deepest gratitude to Qingyu Zhao for his guidance, discussions, and draft reviews throughout this project, as well as for supplying us with the dataset and starter code for the confounder-aware technique.

8. Libraries

Python libraries used for the project included:

- Scikit-learn
- PyTorch
- Statsmodels
- Matplotlib
- Numpy
- Pandas

References

[1] S. Basaia. Automated classification of alzheimer's disease and mild cognitive impairment using a single mri and deep neural networks, 2018.

[2] D. Chan. Patterns of temporal lobe atrophy in semantic dementia and alzheimer's disease, 2001.

[3] A.L. Foundas. Atrophy of the hippocampus, parietal cortex, and insula in alzheimer's disease: A volumetric magnetic resonance imaging study, 1997.

[4] G. Frisoni. The clinical use of structural mri in alzheimer disease, 2010.

[5] Alzheimers Disease Neuroimaging Initiative. Data archive. Main database used for the modeling and training. http://adni.loni.usc.edu/data-samples/access-data/.

[6] W. Lin. Convolutional neural networks-based mri image analysis for the alzheimer's disease prediction from mild cognitive impairment, 2018.

[7] M. Park. Structural mr imaging in the diagnosis of alzheimer's disease and other neurodegenerative dementia: Current imaging approach and future perspectives, 2016.

[8] R. Peters. Ageing and the brain, 2006.

[9] J. Rieke. Visualizing convolutional networks for mri-based diagnosis of alzheimer's disease, 2018. Provided inspiration and a guide for much of our project and modeling. 1808.02874.pdf.

[10] R. Tarawneh. The clinical problem of symptomatic alzheimer disease and mild cognitive impairment, 2012.

[11] Various. Automated anatomical labeling atlas, 2019. https://www.gin.cnrs.fr/en/tools/aal/.

[12] S. Verfaillie. Thinner temporal and parietal cortex is related to incident clinical progression to dementia in patients with subjective cognitive decline, 2016.

[13] E. Vinke. Trajectories of imaging markers in brain aging: the rotterdam study, 2018.

[14] Y. Zhang. Joint assessment of structural, perfusion, and diffusion mri in alzheimer's disease and frontotemporal dementia, 2011.

[15] Q. Zhao. Confounder-aware visualization of convnets, 2019. Main resource for confounder-aware visualizations and related functions. 12727.pdf.