End-to-End Brain Morphometry

Varun Joshi
Data Science
Georgia State University
Atlanta Georgia United States
vjoshi6@student.gsu.edu

Brad Baker
Computer Science and Engineering
Georgia Institute of Technology
Atlanta Georgia United States
bbaker43@gsu.edu

Kevin Wang
Research
Tri-Institutional Center for Translational Research in Neuroimaging and Data Science
Chicago Illinois United States
kwang26@gsu.edu

Sergey Plis
Computer Science
Georgia State University
Atlanta Georgia United States
splis@gsu.edu

ABSTRACT

Brain morphometry provides a powerful way to better research and understand various neurological diseases. It revolves around extracting valuable morphological characteristics from a structural magnetic resonance imaging (MRI) scan of a human brain (such as cortical thickness, mean curvature of the brain, and related measures). Many of the popular approaches to calculating gray matter volumes from brain scans avoid manual measurements and utilize automated software for a more efficient and reliable extraction. The prevalent approach is to first segment the MRI volume into semantic regions followed by computing required metrics. Semantic segmentation needs to be performed precisely for each region, although most of the obtained fine structure is subsequently discarded. Another problem with this approach lies largely in the fact that it can take several hours to complete the calculations required to estimate the cortical gray matter volume for one brain scan. Clinical and research settings that use MRI data require fast and accurate calculations of these morphological characteristics, thus limiting brain morphometry's practical application. The ability of deep learning models to generate accurate predictions within seconds on complicated data like MRI scans remedies the issue of time that many current approaches face and bypasses statistical issues of the two-stage methods. We developed a fast, high performing supervised machine learning model that predicts cortical gray matter volumes. Our model was trained using 670 brain scans each with 68 cortical gray matter labels generated by a popular brain morphometry software called FreeSurfer. We show high quality predictions of cortical gray matter volumes that are generated within mere seconds, rendering deep learning a much more viable alternative to current methods of brain morphometry.

CCS CONCEPTS

• Deep Learning • Convolutional Neural Network • Machine Learning

KEYWORDS

Brain Morphometry, Neuroscience, Deep Learning, Magnetic Resonance Imaging

ACM Reference format:

Varun Joshi, Brad Baker, Kevin Wang and Sergey Plis. 2022. End-to-End Brain Morphometry. In *KDD Undergraduate Consortium. 6 pages.*

1 Introduction

MRI scans have been a boon to both research and clinical work, providing an accurate, reliable and non-invasive technique to understand and view the brains of patients and subjects in-vivo [1-3]. The advances in MRI's usefulness largely come due to a popular method of analyzing MRI scans: brain morphometry [4].

Brain morphometry focuses on measuring various characteristics of an MRI, such as gray matter density, cortical thickness, segmentation, and other morphological characteristics [4]. This method of analyzing MRI largely uses the T1-Weighted MRI scan, since these typically have higher contrast, enabling easier calculation and detection of many characteristics of the brain [2]. Current approaches to brain morphometry like Voxel-Based-Morphometry (VBM) [5] and Surface-Based-Morphometry (SBM) [6] have high accuracy when calculating these various measurements but suffer largely due to the slow calculation time. One of the most popular software for brain morphometry, FreeSurfer [7], requires 8-9 hours of calculation to get measurements for one MRI scan. The relatively high time commitment leads to problems in clinical settings where many patients need quick results, but also can slow down research on large datasets (for example, processing 250 MRI scans would take over 83 days on a single machine).

To address the issue of low speed while maintaining accuracy, some researchers have turned to machine learning and deep learning techniques due to their ability to quickly output hundreds

of predictions within minutes. The software FastSurfer [8]- FreeSurfer's faster version- utilizes deep learning to boost its prediction time. FastSurfer achieves a significant speed up, requiring only around 1 hour for each brain scan. While this approach boasts a high accuracy and speed, the time for evaluation is still rather limiting.

Another approach used random forest regression and was able to reduce prediction time for each brain scan to around 15 minutes, but still required the use of some preprocessing measures from FreeSurfer [9]. This approach focused on predicting the thickness and mean curvature of the cerebral cortex since both are valuable measures in detecting and monitoring neurodegenerative diseases [9].

Recently, Rebsamen et al. [10] proposed a deep-learning based approach that utilized 3D Convolutional Neural Networks (CNNs) to predict volumes of subcortical regions, mean thickness and mean curvature of cortical parcellations directly from the T1-weighted MRI with less FreeSurfer preprocessing than the previous approach. The approach was able to generate suitable performance for many individual regions but performed poorly overall. Another paper published by Cruz et al. [11] attempted to improve the results of the former paper. This paper focused on using a 3D ResNet-based neural network called HerstonNet and scores an overall improved performance on volume, cortical thickness and mean curvature prediction.

These approaches all have been able to contribute to the much faster calculations of morphological measures but lack accuracy and focus largely on cortical thickness and mean curvature. While cortical thickness and mean curvature provide valuable insight into many different conditions, cortical gray matter volume is arguably just as useful as a marker for various other conditions and for research purposes [12-15]. We propose a deep learning-based approach using 3D CNNs to predict gray matter volumes for various cortical regions of interest (ROI). Using the results from [11], we utilize a MeshNet architecture (inspired largely from previous work by those in our lab demonstrating this architecture's efficiency for MRI segmentation [16,17]) to form our predictions. We also use the model architecture from Rebsamen et al. [10] to act as a baseline that we can use to better understand and compare the MeshNet model's performance since there is little to no research focused on using deep learning to predict cortical gray matter volumes. Our task is a multi-class linear regression task, with the inputs being MRI scans and the outputs being gray matter volumes for regions of interest.

2 Methodology

2.1 Data and Preprocessing

The Human Connectome Project (HCP) [18] was the data source used for this project. The HCP is a large dataset that contains MRI scans and labels generated using FreeSurfer [7]. Out of this dataset, 670 healthy subject's brains were used, with 80% in the training data and 20% in the test data. Exclusively T1-Weighted images were used, and three main preprocessing steps were applied: normalizing the images to 1x1x1 mm thickness using FreeSurfer; zero-padding to create images with consistent dimensions of 256x256x256; and Min-Max normalization on the volumes.

2.2 Models

Two main model architectures were used: the model from Rebsamen et al. [10] that predicted cortical thickness and mean curvature and a MeshNet-based model [16,17]. The model architecture from Rebsamen et al. [10] is used as a baseline since the paper attempted to predict different morphological characteristics for the same regions of interest as our paper. The MeshNet model (pictured in Fig. 2) is the primary focus of this paper, with the model from Rebsamen et al. acting as a baseline model (pictured in Fig. 1) to better understand how well the MeshNet model performs. Since there are little to no existing papers on cortical gray matter prediction, utilizing two models gave us greater insight into overall evaluation and understanding of relative performance. Nearly the entirety of model creation, training, and testing utilized PyTorch [19] and its extension, Catalyst [20]. For all testing, two Tesla V100 PCle 16 GB GPUs were used.

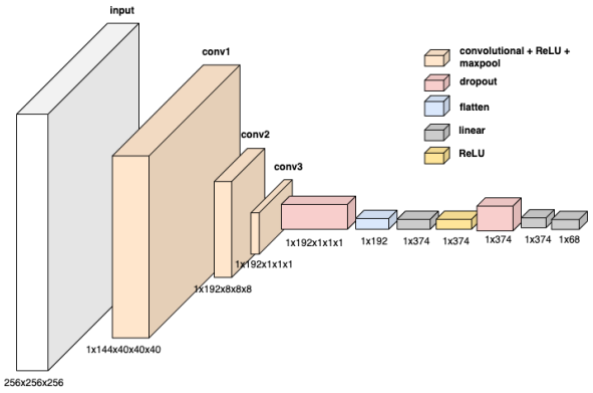

Figure 1: Baseline model architecture [10]

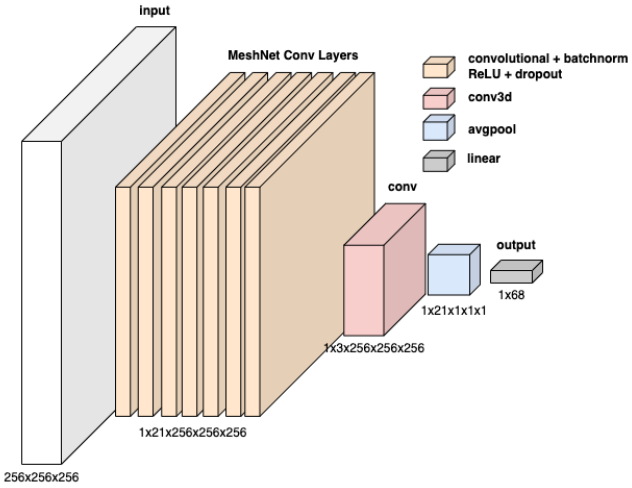

Figure 2: MeshNet based model architecture [16, 17]

2.3 Evaluation

There were three main evaluation metrics through which the models were evaluated: mean squared error (MSE), normalized loss, and saliency. MSE was used largely to monitor model performance during training and to pinpoint regions in the output that were difficult for our models to predict. While MSE was used for training, it was not the most important metric due to how it is calculated.

Gray matter prediction is done across brain regions with a wide range of volumes, with some being several thousands of voxels (1x1x1 mm cubes) and others being only a few hundred. Using MSE over-exaggerates the differences between large regions over small regions when attempting to gauge performance. To remedy this and make evaluation more consistent, a normalized loss was utilized (1). The normalized loss simply took the MSE for the actual and predicted values, but also divided the MSE by taking the MSE of the actual value and zero. This method removed the confusion caused by differences in scale between different regions. This metric was not useful in analyzing individual model performance, but rather, for comparing our baseline and MeshNet models and understanding which regions were truly difficult for the model to predict.

$$\frac{\frac{\sum_{i=0}^{n}(y-\hat{y})^2}{2n}}{\frac{\sum_{i=0}^{n}(y-0)^2}{2n}} \quad (1)$$

Saliency was the final piece in understanding model performance. Saliency allows one to understand which parts of the input influence the output of the model the most. Generation of saliency and saliency visualizations was done using the Captum library [21]. By peering into what part of the input is the most important for predictions, it is much easier to understand if the model's predictions will be able to generalize well to unseen data. For example, if the model is largely focusing on the skull to generate its predictions, this likely will translate poorly to a clinical setting even if the predictions the model makes are fairly accurate. Likewise, if a model focuses entirely on the wrong part of the brain to predict a given region, clinical and research application becomes difficult to justify. Saliency can also provide insight into what may be causing a model's performance to suffer by showing exactly how the model is generating a lackluster result.

To generate saliency images that were easy to interpret, some processing was applied on the visualizations. To calculate saliency, the Captum library creates an N-dimensional tensor that is the same size as the input. This tensor has a value for each pixel in the input that lets the user know which pixels were most important in predicting the output (a more positive value indicates higher importance, while more negative values indicate the opposite). While the saliency tensors for each output region were different, they initially looked very similar when visually plotted. To make more apparent visual differences, the mean saliency tensor (the sum of each region's tensor divided by the number of regions), was subtracted from each individual region's saliency tensor. This subtraction made it easier to see the saliency for one region, since the overall model's saliency for all the regions was removed. The saliency images presented in the paper were created by taking the saliency tensor generated by Captum, finding the coordinate with the highest value, and then slicing the 3D images at that coordinate to generate 2D representations of the saliency.

3 Results

This section will go over the performance of both models by analyzing the average absolute difference of predictions and targets for each region as well as looking at saliency visualizations for both models. Overall performance and saliency will be analyzed for both models.

The graphs in Figs. 3 and 4 were generated by taking each brain scan in the test data, taking the absolute difference between the predicted and actual value for each brain region, and then plotting box plots to show the distribution of this absolute difference for each region. These graphs show how incorrect the model was in voxels for each brain region and gives insight into how each region performed on average.

3.1 Performance Analysis

3.1.1 Baseline Model. On average, the baseline model took 6.2 seconds per scan to make its predictions over the 135 scans in the testing data: a substantial improvement over the several hours taken by software like FreeSurfer. The final, normalized loss obtained for this model was 0.03. To better understand this loss, we can look at the absolute differences between targets and predicted values as well as saliency analysis.

Using the baseline model's absolute difference plot (Fig. 3), we can see that the baseline model had some regions for which it performed quite well, but many regions were on average hundreds, if not thousands, of voxels off the true value. Some regions (like the superior frontal) were often off by tens of thousands of voxels. These results alone indicate that this model may not be suitable for a clinical or research setting.

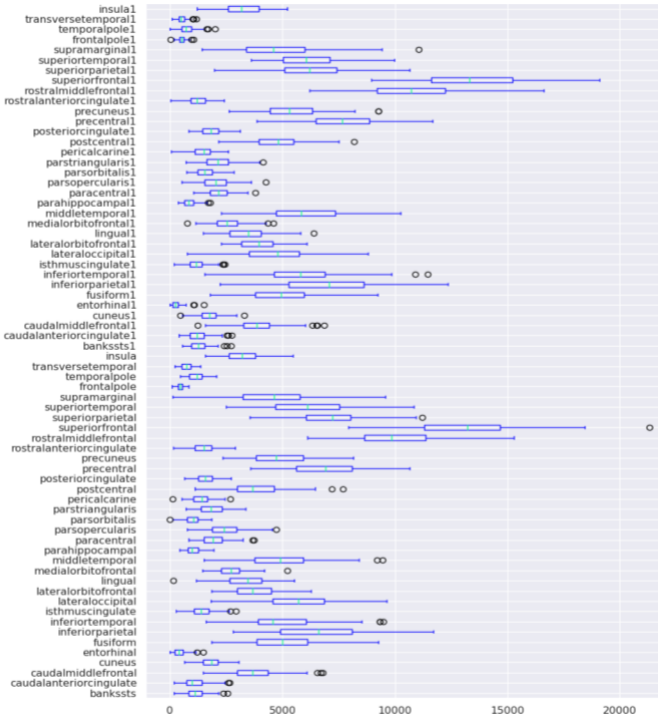

Figure 3: Absolute difference distribution for each region; x-axis is in voxels (baseline model)

3.1.2 MeshNet Model On average, this model took slightly longer than the baseline model to make predictions: 8.5 seconds on average over the 135 images used in training. The difference in prediction time is relatively small and is still a great improvement over the hours taken per scan by FreeSurfer. The model's overall normalized loss was 0.015; approximately half of the loss from our baseline model. The model seems to have much better performance purely from this number, but again, analyzing the distributions of absolute difference between labels and predicted values for each region provides much better insight overall.

In Fig. 4, we can see that the MeshNet model drastically outperforms the baseline model in every region. The MeshNet model is much closer to the actual target values on average, with most regions being within a thousand voxels of their target value on average. While the accuracy of this model seems to have improved when compared to the baseline model, saliency is still required to fully justify both models' usage in a clinical or research setting.

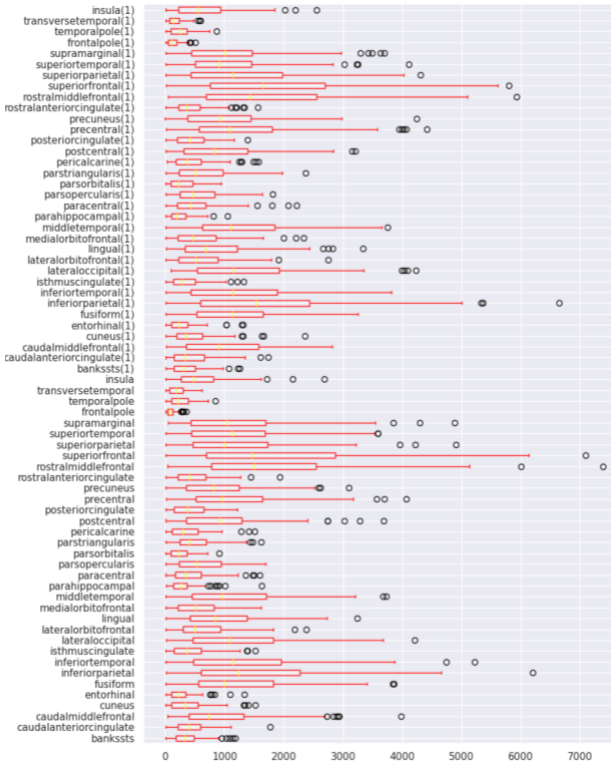

Figure 4: Absolute difference distribution for each region; x-axis is in voxels (MeshNet model)

3.2 Saliency Analysis

In this section, we will examine the saliency visualizations for the MeshNet and baseline model. We will compare the saliency visualizations of the best, average, and worst performing regions (in terms of normalized loss) for both models. Examining these visualizations will reveal why the MeshNet model is able to have better performance and where the baseline model fails.

3.2.1 Saliency for best regions. Fig. 5 plots saliency visualizations for the best performing region for the baseline model. In these images, the brighter the color of the saliency, the more important that part of the image was to predict the final output. This saliency visualization shows that the baseline model focuses on a part of the brain that is somewhat in the area of the region it is trying to predict, but the model also focuses heavily on the skull to make its predictions. Since the model seems to be focusing on irrelevant parts of the MRI to predict its best performing region, it helps to explain the lackluster performance from this model.

Immediately in the best performing region's saliency image (here, the darker the saliency, the more relevant that part of the image is to making predictions), we can see that the MeshNet model not only focuses on the most relevant parts of the brain for the region it is predicting, but also does not unnecessarily focus on the skull (Fig. 6). The left frontal pole is right at the tip of the brain, and that is exactly where the model is looking to make its predictions

on this region. The saliency visualizations clearly demonstrate why the MeshNet model outperforms the baseline model.

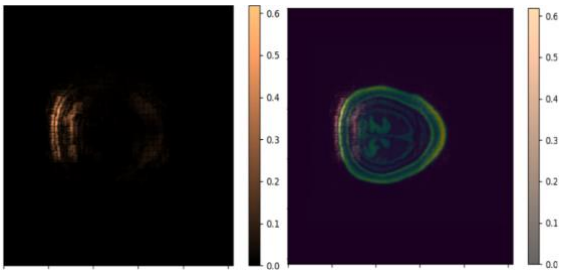

Figure 5: A Saliency visualization for Right Entorhinal Region. Both the pure saliency image (left) and the saliency image laid over a visualization of the brain scan (right) are shown (normalized loss: 0.001)

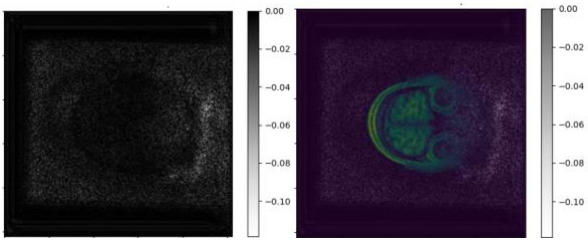

Figure 6: Saliency visualizations for Left Frontal Pole region. Both the pure saliency image (left) and the saliency image laid over a visualization of the brain scan (right) are shown (normalized loss: 0.000166)

3.2.2 Saliency for average regions. In Fig. 7, we have an average performing region for the baseline model. Here, we can see that the model is almost completely focusing on the skull to make its prediction and lightly focuses on some random areas of the brain unrelated to the region being predicted. Once more, the model focuses too much on the skull and irrelevant parts of the brain to make its predictions, explaining its poor performance and deeming it unreliable for a clinical or research setting.

Even in the average performing region's saliency for the MeshNet model (Fig. 8), the model still focuses most of its attention on the brain and on the most relevant parts of the brain to the region it is predicting. In contrast to the baseline model, the MeshNet model is still recognizing relevant regions of the brain to make predictions even in average performing regions.

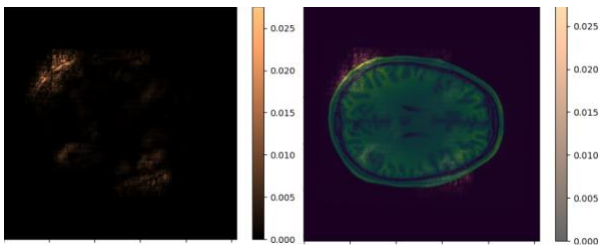

Figure 7: Saliency visualizations for Left Lingual Region. Both the pure saliency image (left) and the saliency image laid over a visualization of the brain scan (right) are shown (normalized loss: 0.1359)

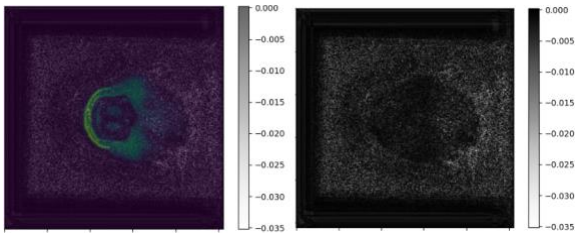

Figure 8: Saliency visualizations for Right Pars Triangularis region. Both the pure saliency image (left) and the saliency image laid over a visualization of the brain scan (right) are shown (normalized loss: 0.0076)

3.2.3 Saliency for worst regions. For the worst performing region for the baseline model in Fig. 9, we see that the model completely focuses on a small region in the skull. This model clearly lacks the sophistication to be used in a clinical or research setting both due to its lack of accuracy in predictions shown from the absolute difference plot, but also due to its poor method of forming predictions shown by the saliency.

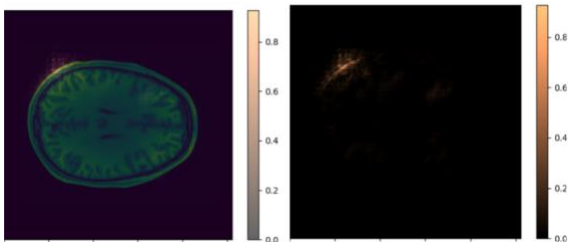

Figure 9: Saliency visualizations for Right Superior Frontal Region. Both the pure saliency image (left) and the saliency image laid over a visualization of the brain scan (right) are shown (normalized loss: 2.1049)

In the worst performing region (Fig. 10), the MeshNet model focuses on the skull to form its predictions, completely ignoring the brain and relevant regions. Even though the performance of the MeshNet model is superior to the baseline model, it still requires further training to be reliable for all regions. Regions such as the left superior frontal region indicate the MeshNet model still requires further training and tweaking before it can consistently predict all the regions of interest this paper focuses on.

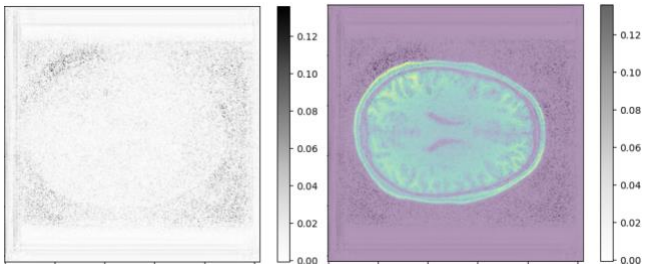

Figure 10: Saliency visualizations for Left Superior Frontal region. Both the pure saliency image (left) and the saliency image laid over a visualization of the brain scan (right) are shown (normalized loss: 0.069)

4 Discussion and Future Work

Our MeshNet-based architecture boasts impressive performance on many regions of the brain and demonstrates the potential for deep learning to be applied to brain morphometry. The MeshNet architecture also proves its utility in a brain morphometry application once more, as done so in Cruz et al. [11], and provides a starting point for researchers exploring other applications of brain morphometry such as segmentation.

The main limitations now are both the need for better, more consistent performance across regions as well as the need for saliency images showing the model utilizes relevant parts of the MRI scan to make its predictions. While we can estimate the volumes of some regions well, many regions still could have much better predictions. The focus moving forward will be to improve model predictions on underperforming regions.

To improve model predictions and saliency, there are several avenues of future work to explore. The main way to further the results of current work is focusing on further tweaking and expanding the MeshNet model. The model seems to have reached convergence, but could benefit from hyper-parameter optimization, data augmentation, and other measures which may improve performance. We are currently also exploring multi-task learning, which operates on the theory that when one model performs two similar but separate tasks, performance on both objectives increases. Our initial work in this area focuses on developing a multi-task model that predicts both gray matter volumes of cortical regions as well as gray matter and white matter segmentation.

5 Conclusion

Our work is among the first to demonstrate that deep learning models provide fast and reliable estimations of cortical gray matter volumes using T1-weighted MRI scans. The results from this paper provide hope for efficient methods of morphometry in both clinical and research settings. With further work, AI-based morphometry will only continue to become a more viable option.

ACKNOWLEDGMENTS

This work was in part supported by the NSF grant 2112455 and NIH grant RF1MH121885.

REFERENCES

[1] Gordon J.G. Asmundson, Amy Krain Roy, Erica Ferrara, and Rodolfo Keesey. 2022. 3.04 - fMRI and Other Neuroimaging Methods. In *Comprehensive clinical psychology*. Amsterdam: Elsevier, 62–82

[2] Arthur W. Toga, A.van der Kouwe, and B. Fischl. 2015. Anatomical MRI for Human Brain Morphometry. In *Brain mapping: An encyclopedic reference*. Amsterdam: Elsevier/Academic Press, 3–28.

[3] Arne May and Christian Gaser. 2006. Magnetic resonance-based morphometry: A window into structural plasticity of the brain. (August 2006). Retrieved May 20, 2022 from https://pubmed.ncbi.nlm.nih.gov/16914981/

[4] Edward R. Hirt, Joshua J. Clarkson, Lile Jia, Dylan D. Wagner, and Todd F. Heatherton. 2016. Chapter 14 - What Can Cognitive Neuroscience Tell Us About the Mechanism of Ego Depletion? In *Self-regulation and ego control*. London, England: Academic Press, 281–300.

[5] John Ashburner and Karl J. Friston. 2000. Voxel-based morphometry—the methods. *NeuroImage* 11, 6 (June 2000), 805–821. DOI:http://dx.doi.org/10.1006/nimg.2000.0582

[6] Anders M. Dale, Bruce Fischl, and Martin I. Sereno. 1999. Cortical surface-based analysis. *NeuroImage* 9, 2 (1999), 179–194. DOI:http://dx.doi.org/10.1006/nimg.1998.0395

[7] Bruce Fischl. 2012. Freesurfer. *NeuroImage* 62, 2 (August 2012), 774–781. DOI:http://dx.doi.org/10.1016/j.neuroimage.2012.01.021

[8] Leonie Henschel, Sailesh Conjeti, Santiago Estrada, Kersten Diers, Bruce Fischl, and Martin Reuter. 2020. FastSurfer - a fast and accurate deep learning based neuroimaging pipeline. *NeuroImage* 219 (2020), 117012. DOI:http://dx.doi.org/10.1016/j.neuroimage.2020.117012

[9] Yannick Suter, Christian Rummel, Roland Wiest, and Mauricio Reyes. 2018. Fast and uncertainty-aware cerebral cortex morphometry estimation using random forest regression. *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)* (2018). DOI:http://dx.doi.org/10.1109/isbi.2018.8363752

[10] Michael Rebsamen, Yannick Suter, Roland Wiest, Mauricio Reyes, and Christian Rummel. 2020. Brain morphometry estimation: From hours to seconds using Deep Learning. *Frontiers in Neurology* 11 (April 2020). DOI:http://dx.doi.org/10.3389/fneur.2020.00244

[11] Rodrigo Santa Cruz et al. 2021. Going deeper with brain morphometry using neural networks. 2021 *IEEE 18th International Symposium on Biomedical Imaging (ISBI)* (2021). DOI:http://dx.doi.org/10.1109/isbi48211.2021.9434039

[12] Stephen Ramanoël et al. 2018. Gray matter volume and cognitive performance during normal aging. A voxel-based morphometry study. *Frontiers in Aging Neuroscience* 10 (August 2018). DOI:http://dx.doi.org/10.3389/fnagi.2018.00235

[13] Katharina Wittfeld et al. 2020. Cardiorespiratory fitness and gray matter volume in the temporal, frontal, and cerebellar regions in the general population. *Mayo Clinic Proceedings* 95, 1 (2020), 44–56. DOI:http://dx.doi.org/10.1016/j.mayocp.2019.05.030

[14] Emma M. Coppen, Milou Jacobs, Annette A. van den Berg-Huysmans, Jeroen van der Grond, and Raymund A.C. Roos. 2018. Grey matter volume loss is associated with specific clinical motor signs in Huntington's disease. *Parkinsonism & Related Disorders* 46 (2018), 56–61. DOI:http://dx.doi.org/10.1016/j.parkreldis.2017.11.001

[15] Xu Jiang et al. 2021. Abnormal gray matter volume and functional connectivity in Parkinson's disease with Rapid Eye Movement Sleep Behavior disorder. *Parkinson's Disease* 2021 (2021), 1–11. DOI:http://dx.doi.org/10.1155/2021/8851027

[16] Alex Fedorov, Eswar Damarajua, Vince Calhoun, and Sergey Plis. 2017. An (almost) instant brain atlas segmentation for large-scale studies. *arXiv* (2017).

[17] Alex Fedorov, Jeremy Johnson, Eswar Damaraju, Alexei Ozerin, Vince Calhoun, and Sergey Plis. 2017. End-to-end learning of brain tissue segmentation from imperfect labeling. *2017 International Joint Conference on Neural Networks (IJCNN)* (2017). DOI:http://dx.doi.org/10.1109/ijcnn.2017.7966333

[18] David C. Van Essen, Stephen M. Smith, Deanna M. Barch, Timothy E.J. Behrens, Essa Yacoub, and Kamil Ugurbil. 2013. The Wu-minn human connectome project: An overview. NeuroImage 80 (2013), 62–79. DOI:http://dx.doi.org/10.1016/j.neuroimage.2013.05.041

[19] Adam Paszke et al. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. 33rd Conference on Neural Information Processing Systems (NeurIPS 2019) (2019).

[20] Sergey Kolesnikov. 2020.(2020). Retrieved May 23, 2022 from https://catalyst-dl.readthedocs.io/en/latest/

[21] Narine Kokhlikyan, Vivek Miglani, Miguel Martin, and Edward Wang. 2020. Captum: A unified and generic model interpretability library for PyTorch. (2020).