Automated quality control of Freesurfer cortical parcellation

Howard Chiu

Abstract

Magnetic resonance imaging (MRI) is one of many imaging methods frequently used to study the brain across the lifespan, and segmentation and parcellation of these MR images are commonly processed by automated pipelines. However, these algorithms are not perfect, and quality control is performed manually by visual inspection of brain volumes, typically by multiple human raters, or by quantitative analyses of derivatives. Computer vision models are increasingly being used in computer-assisted diagnosis, and could conceivably be used upstream during image processing for quality control. This paper demonstrates that out-of-the-box classifiers can perform significantly better than chance at perceiving small differences that separate good and bad Freesurfer segmentations.

Background

Magnetic resonance imaging (MRI) is one of many imaging methods frequently used to study the brain across the lifespan. Raw MR images must be segmented to identify the various types of cortical tissue. This includes gray and white matter, as well as subcortical structures that are differentially implicated in brain function. Currently, this segmentation and parcellation is done by computing where the boundaries of various surfaces lie.

Incorrect edge detection leads to changes in variables such as volume, cortical thickness, and cortical surface area, and subsequent statistical analyses are dependent on these derivatives.

Currently, quality control is performed manually by visual inspection of every brain volume (Raamana et al., 2022), typically by multiple human raters to ensure inter-rater reliability if possible. There are current efforts for quality control using summary statistics from these brain volumes such as cortical thickness or surface area (e.g. Qoala-T: Klapwijk et al., 2019), but they are insufficient to guarantee good quality scans.

Computer vision models are increasingly being used in computer-assisted diagnosis (e.g. MedSAM for tumour segmentation: Ma et al., 2024; GoogLeNet for cancer classification: Deepak & Ameer, 2019), and could conceivably be used upstream during image processing for quality control. Having a computer vision model perform this task as a first pass could

reduce the time required to perform quality control on large datasets for neuroscience researchers.

There is evidence that computer vision models benefit from pre-training on large datasets, especially when domain-specific datasets are small. Convolutional neural networks (CNNs) have been previously used to detect brain tumours as part of computer-assisted diagnosis (Krishnapriya & Karuna, 2023).

Dataset

The dataset comprised a mix of anatomical magnetic resonance images (MRIs) of real human subjects, as well as synthetic MRIs from Hugging Face (https://huggingface.co/datasets/SinKove/synthetic_brain_mri), to increase the diversity of the training input received by the model. The synthetic T1-weighted images were "sampled from generative models trained on data originally from the UK Biobank dataset".

Cortical reconstruction and volumetric segmentation of the real and synthetic MRIs was performed with the Freesurfer image analysis suite, which is documented and freely available for download online (http://surfer.nmr.mgh.harvard.edu/). The technical details of these procedures are described in prior publications (Dale et al., 1999; Dale and Sereno, 1993; Fischl and Dale, 2000; Fischl et al., 2001; Fischl et al., 2002; Fischl et al., 2004a; Fischl et al., 1999a; Fischl et al., 1999b; Fischl et al., 2004b; Han et al., 2006; Jovicich et al., 2006; Segonne et al., 2004, Reuter et al. 2010, Reuter et al. 2012).

The 3-dimensional MRIs were sliced along axial, sagittal, and coronal axes in Freeview (graphical user interface bundled with Freesurfer, see Figure 1) to provide an array of 2-dimensional images before these were labelled as 'good' or 'bad' segmentations (Figure 2). The output from Freeview was 372 by 311 pixels, but these were resized to 256 by 256 pixels and center cropped at 224 x 224 pixels for the two models used in this study. The data in the training set was also augmented with random horizontal flips using torchvision.transforms.

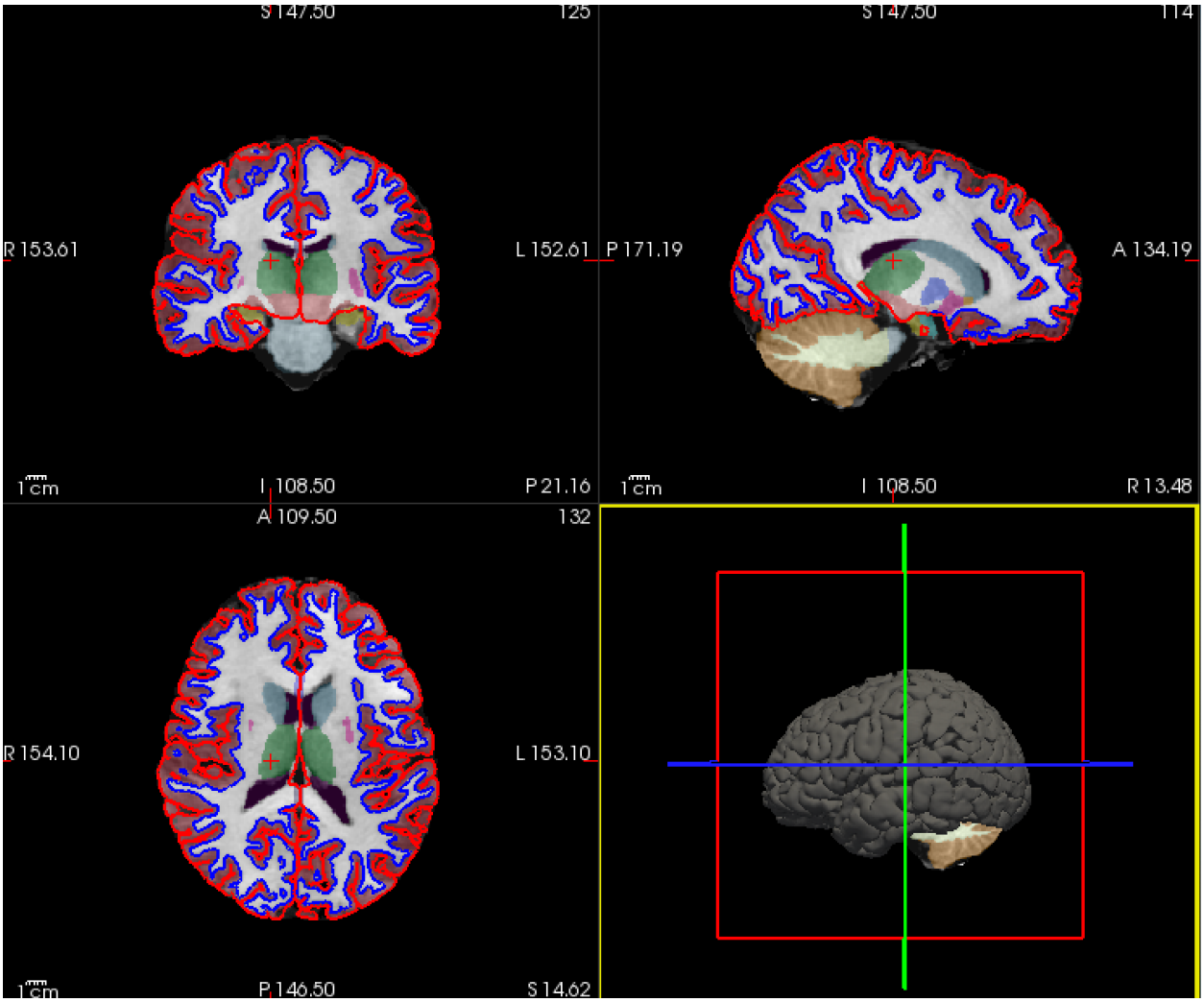


Figure 1: Example of 3 different views of the brain – coronal (top left), sagittal (top right), axial (bottom left), and the 3-dimensional volume with the location of the 3 planes denoted in the three primary colors (red marking location of sagittal slice, green marking location of coronal slice, and blue marking location of axial slice). The represented here included the surfaces – the pial (connective tissue layer that closely follows the surface of the brain and the spinal cord) and white matter masks for both hemispheres, as well as the volumes – the T1-weighted image, the whole brain mask, and the segmentation mask.

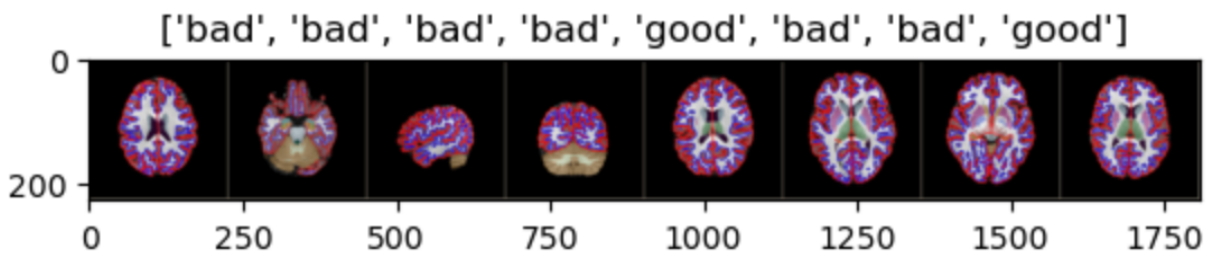


Figure 2: Example of various 2-dimensional slices of brains and the labels assigned to each of these exemplars.

Common errors that qualified as 'bad' segmentations included the pial borders extending into the skull or meninges, as well as cutting through a gyrus, as demonstrated in the figure below (Figure 2 in Pulli et al., 2002).

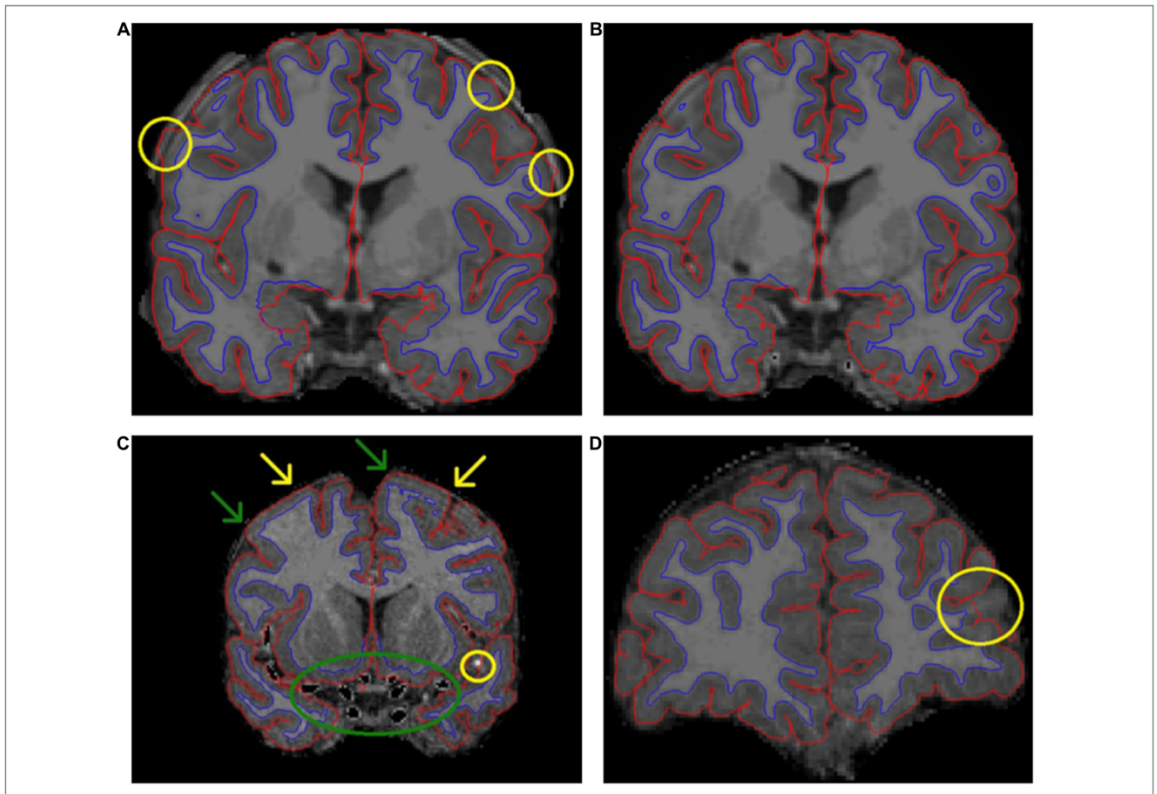


FIGURE 2 | A presentation of some common errors and fixes related to the pial border and non-brain tissues. **(A)** Demonstrates how skull fragments can cause errors in pial border (yellow circles). **(B)** Presents the same subject with skull fragments removed. In panel **(C)**, arteries were removed (green circle). We removed voxels with an intensity between 130 and 190, and therefore some parts of arteries were not removed (yellow circle). **(C)** Also demonstrates the challenges with artifact, meninges, and the pial border. In some areas, the pial border may extend into the meninges (yellow arrows). Meanwhile, at the other end of the same gyrus, the border may seem correct (green arrows). It is difficult to fix these errors manually. Additionally, the visible motion artifact adds further challenges to manual edits of the pial border. In panel **(D)**, the pial border cuts through a gyrus.

The composition of the dataset is detailed in Table 1 below, which contains a total of 684 images. This is more than twice that of the 305 images used in Krishnapriya & Karuna (2023).

	Good				Bad			
		Axial	Coronal	Sagittal		Axial	Coronal	Sagittal
Training	Human	0	0	17	Human	88	17	22
	Synthetic	46	55	51	Synthetic	26	9	20
Validation	Human	0	0	17	Human	87	16	22
	Synthetic	45	54	40	Synthetic	25	8	19

Table 1: summary of exemplars provided in each view grouped by their labels, as well as the source of the MRIs. Despite the author's best effort to balance across the cells, labelling sufficient data proved to be extremely time consuming. Bad exemplars were also oversampled relative to good exemplars as we typically do not expect Freesurfer segmentations to be frequently erroneous.

Methods

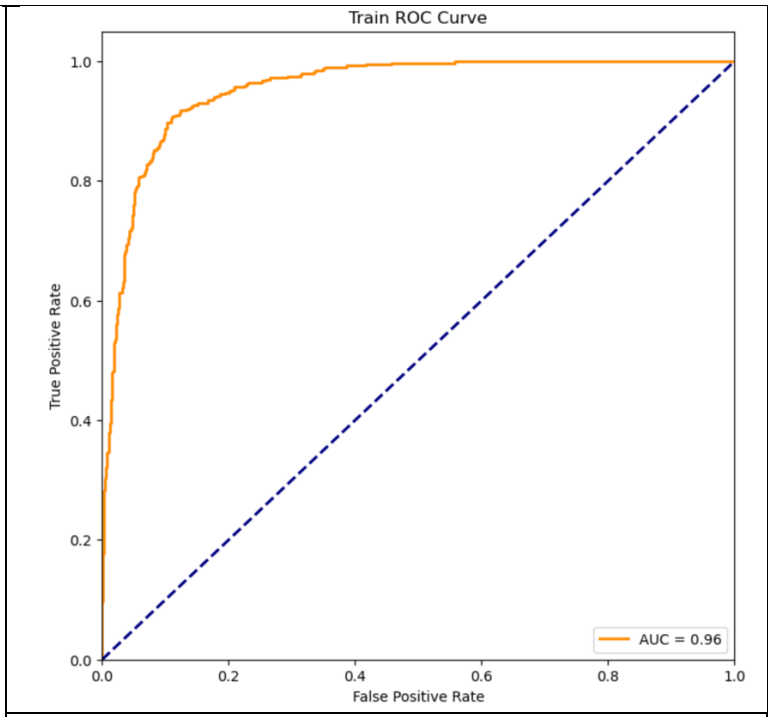
Two out-of-the-box classifiers pretrained on ImageNet images were used in this task – ResNet 19 (He et al., 2015) and VGG19 with batch normalization (VGG19_BN; Simonyan & Zisserman, 2014). The models were implemented on a cloud computing server with 32 GB of RAM using a Jupyter Notebook within a Miniconda environment. The environment.yml for the environment is provided in the GitHub repository associated with this paper.

The hyperparameters for both models were kept constant to allow for comparison. This included a learning rate of 0.001 which decayed by a factor of 0.1 every 7 epochs, a batch size of 6, and an adaptive moment estimation (Adam) optimizer. The models were trained for 25 epochs as in Krishnapriya & Karuna (2023).

Results

The first runs with a smaller and extremely imbalanced dataset resulted in significant overfitting. Increasing the number of epochs in that case did not improve model accuracy. The data supports this hypothesis, with a high area under the curve (AUC) but low accuracy, and the model may have just learnt to guess that category that is more prevalent.

```
Epoch 0/7
----------
train Loss: 0.4822 Acc: 0.7747
val Loss: 0.3922 Acc: 0.8946

Epoch 1/7
----------
train Loss: 0.1830 Acc: 0.9341
val Loss: 1.0484 Acc: 0.6757

Epoch 2/7
----------
train Loss: 0.1613 Acc: 0.9451
val Loss: 1.2381 Acc: 0.5730

Epoch 3/7
----------
train Loss: 0.2253 Acc: 0.8901
val Loss: 1.1645 Acc: 0.6027

Epoch 4/7
----------
train Loss: 0.2633 Acc: 0.9121
val Loss: 1.0704 Acc: 0.6676

Epoch 5/7
----------
train Loss: 0.1918 Acc: 0.9121
val Loss: 1.1734 Acc: 0.6081

Epoch 6/7
----------
train Loss: 0.2044 Acc: 0.9451
val Loss: 1.1944 Acc: 0.5838

Epoch 7/7
----------
train Loss: 0.2129 Acc: 0.9121
val Loss: 1.0500 Acc: 0.6595

Training complete in 5m 33s
```



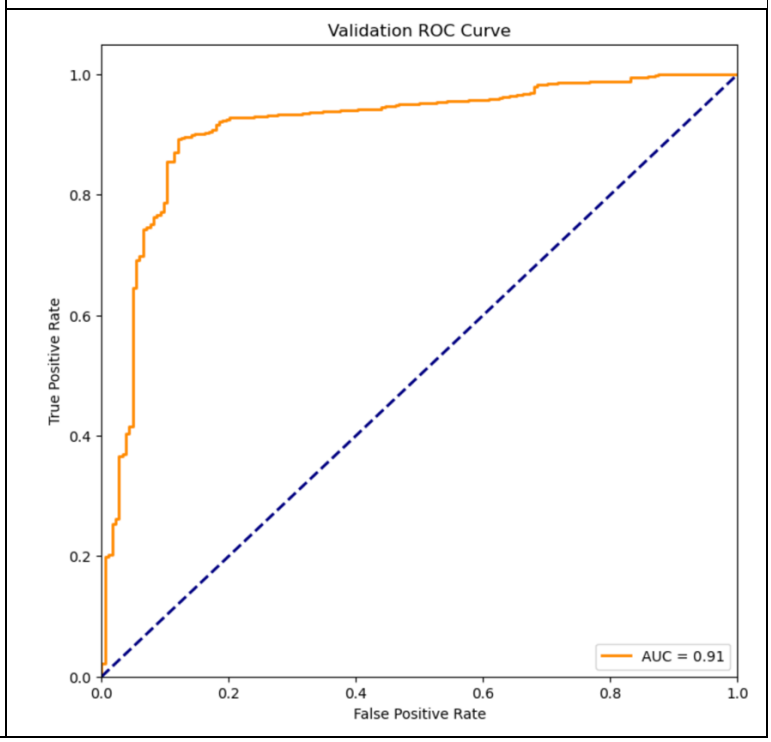
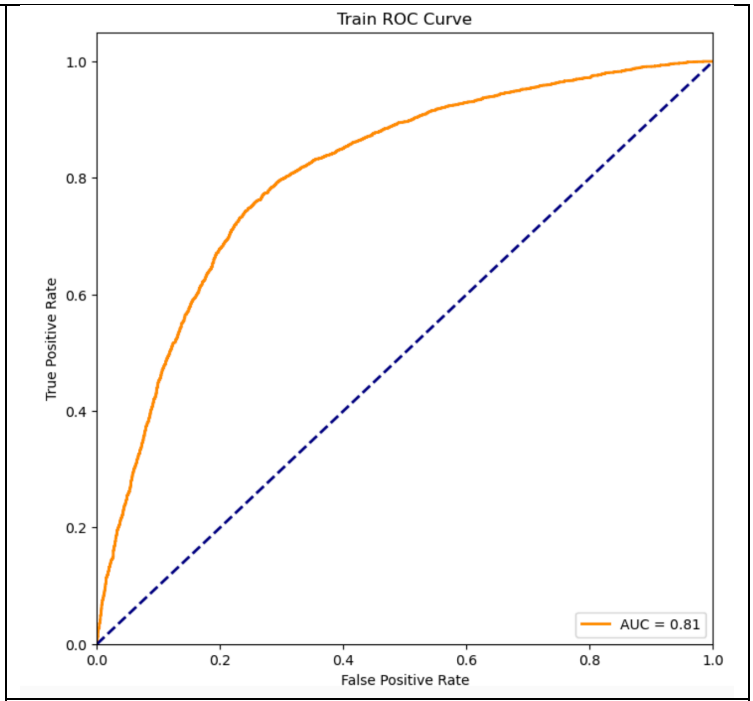


Figure 3: Performance of ResNet19 with highly imbalanced dataset on first run, and with 8 learning epochs

After augmenting the dataset with more exemplars, overfitting was reduced and accuracy improved at the cost of reduced AUC for the ResNet19 model (see Figures 3 and 4). Compared to an expert rater of 95% accuracy, this model has some way to go, but is still significantly better than chance at 50% accuracy.

```
Epoch 0/24
----------
train Loss: 0.7603 Acc: 0.7057
val Loss: 0.6051 Acc: 0.7898

Epoch 1/24
----------
train Loss: 0.6551 Acc: 0.7429
val Loss: 0.5859 Acc: 0.7808

Epoch 2/24
----------
train Loss: 0.6426 Acc: 0.7429
val Loss: 0.5679 Acc: 0.7898

Epoch 22/24
----------
train Loss: 0.6749 Acc: 0.7371
val Loss: 0.5932 Acc: 0.7868

Epoch 23/24
----------
train Loss: 0.6702 Acc: 0.7200
val Loss: 0.6022 Acc: 0.7808

Epoch 24/24
----------
train Loss: 0.6912 Acc: 0.7514
val Loss: 0.5861 Acc: 0.7838

Training complete in 29m 12s
```



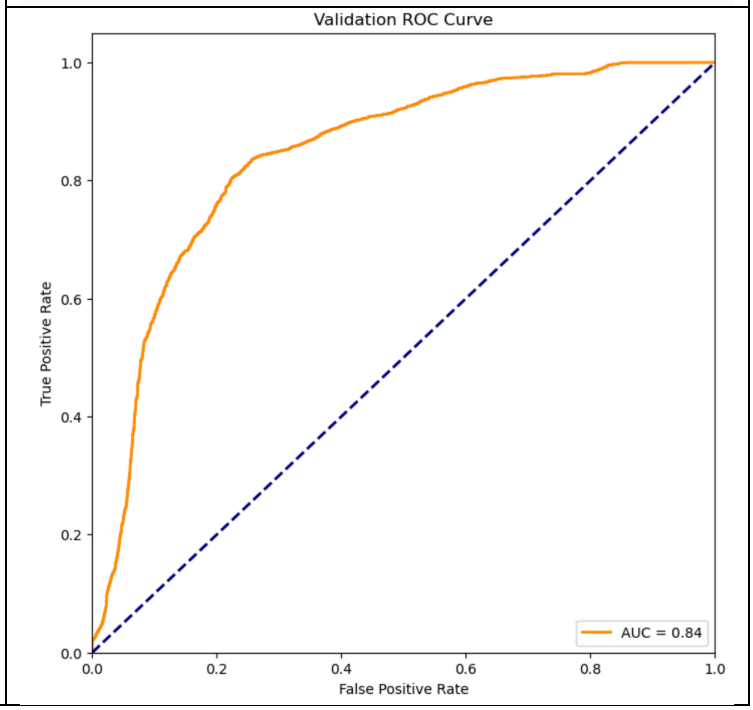
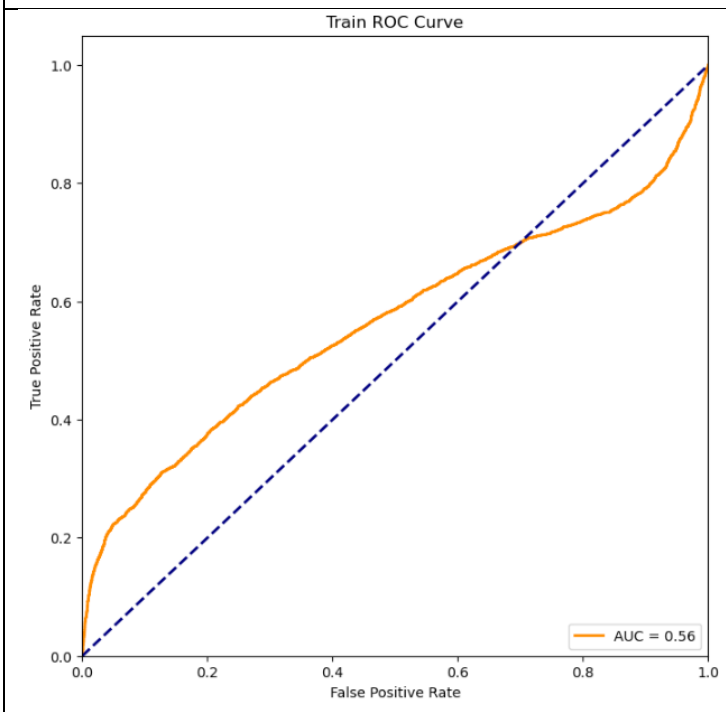


Figure 4: Performance of ResNet19 with augmented dataset and 25 training epochs.

```
Epoch 0/24
----------
train Loss: 1.2044 Acc: 0.5629
val Loss: 0.7742 Acc: 0.7027

Epoch 1/24
----------
train Loss: 1.0267 Acc: 0.5829
val Loss: 0.8017 Acc: 0.6787

Epoch 2/24
----------
train Loss: 1.0320 Acc: 0.6057
val Loss: 0.8473 Acc: 0.6697

Epoch 3/24
----------
train Loss: 1.0721 Acc: 0.6143
val Loss: 0.8415 Acc: 0.6697

Epoch 22/24
----------
train Loss: 1.0779 Acc: 0.6000
val Loss: 0.7974 Acc: 0.6667

Epoch 23/24
----------
train Loss: 1.1258 Acc: 0.5771
val Loss: 0.8096 Acc: 0.6697

Epoch 24/24
----------
train Loss: 1.1081 Acc: 0.6143
val Loss: 0.9073 Acc: 0.6486

Training complete in 294m 11s
```



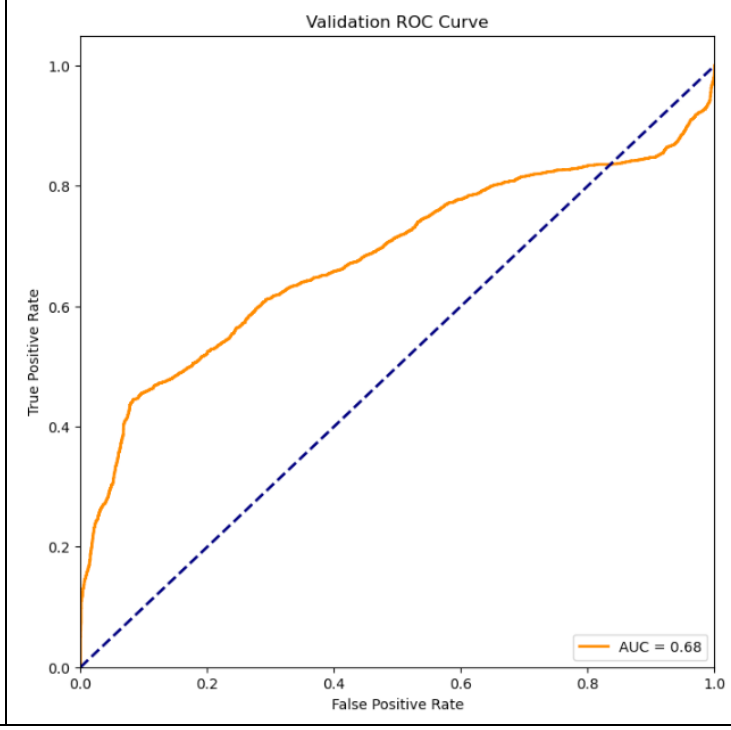


Figure 5: Performance of VGG19_BN with augmented dataset and 25 training epochs.

Surprisingly, the VGG19_BN model did not perform better than the ResNet19 model even though the training duration was 10 times as long. This also contradicted the previous findings of Krishnapriya & Karuna (2023). However, the hyperparameters used in this paper are far from optimal, as can be seen from the loss not decreasing over the epochs. Many other pre-processing decisions for the training and validation dataset could also affect the results (see Reflections section below).

Conclusion

This project demonstrates the potential of out-of-the-box computer vision models as an assistant to aid quality control of MR images. Beyond a classification ask, computer vision models could perhaps also be used to approach this as a segmentation task, where masks can tell us not just whether a segmentation is good or bad, but also highlight locations where it thinks the error is. This will also help to make the model slightly more transparent in the sense that we can visualize what it has learnt, especially for complex classification problems like this where the difference between a good or bad segmentation may only consist of differences in a few voxels.

Reflections

It was difficult for me to curate the dataset as there were many considerations that I was thinking about. The first was where to draw the threshold between a “good” and a “bad” segmentation. I observed myself to be perhaps more stringent than necessary, in which case I ended up with many slices that I labelled as being “bad” segmentations. However, other people may have different thresholds for acceptable segmentations. More importantly, human raters looking at a 3-dimensional volume can use adjacent slices as context, and errors in a single 2-dimensional slice may be “tolerated” in real life if the error does not appear on adjacent slices. Human raters use axial and coronal views more frequently compared to sagittal views, but I’ve attempted to give the model sufficient exemplars across all 3 views. Lastly, and maybe most importantly, the quality of a scan depends on the aim of the experiment. If the defects are in a region that is unrelated to the hypothesis, then the scan could still be rated as of acceptable quality.

Second was about parameters to be used in generating the 2-dimensional slices. I did not manage to establish the optimal image contrast for the computer vision models (in a process known as windowing; see Figure 6 below) or establish if my decision to color in the segmentation affects model performance (as opposed to leaving it in monochrome; see Figure 7). Instead of providing the models with screenshots, another possibility would be to feed the voxel signal intensities directly as a numpy array (since images are converted to arrays anyway), where perhaps 70 could be gray matter and 110 white matter and a decision boundary established somewhere between the two.

Third, Freesurfer errors are also not equally distributed throughout the brain – errors are more common in the temporal lobe, and less common in the anterior parts of the brain. Again, I decided to be agnostic about the priors and provide exemplars with a higher proportion of images taken from the medial segments of the brain.

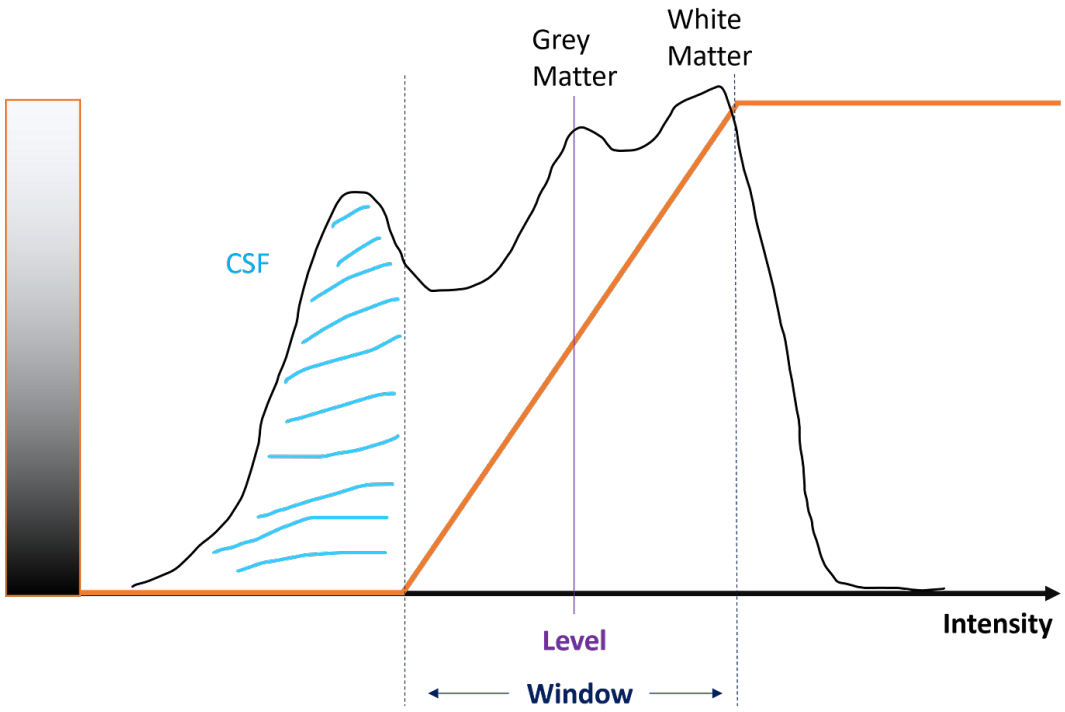


Figure 6: Windowing allows us to set a level and a window width for those gray values that we want visualize. Values outside this range are either black (left side of the window) or white (right side of the window). Only values inside the defined range are placed into the grayscale. (Image and caption taken from https://medium.com/@susanne.schmid/visualization-of-medical-images-adjusting-contrast-through-windowing-c2dd9abb1d5)

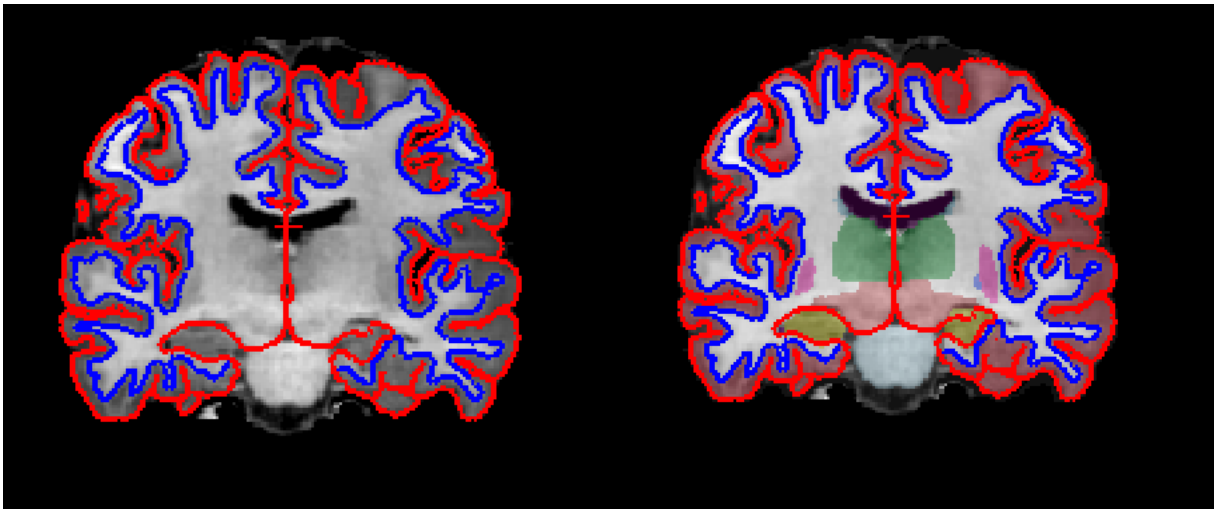


Figure 7: illustration of leaving the segmentations in monochrome (left) or filling it in with an arbitrary color map (right)

Open Science and Transparency

The Jupyter notebook containing the code, as well as the synthetic images for training and validation sets, and the environment.yml file for setting up the miniconda environment, has been pushed to GitHub at https://github.com/chiuhoward/cs432.

References

Dale, A.M., Fischl, B., Sereno, M.I., 1999. Cortical surface-based analysis. I. Segmentation and surface reconstruction. Neuroimage 9, 179-194.

Dale, A.M., Sereno, M.I., 1993. Improved localization of cortical activity by combining EEG and MEG with MRI cortical surface reconstruction: a linear approach. J Cogn Neurosci 5, 162-176.

Deepak, S., & Ameer, P. M. (2019). Brain tumor classification using deep CNN features via transfer learning. Computers in Biology and Medicine, 111, 103345. https://doi.org/10.1016/j.compbiomed.2019.103345

Fischl, B., Dale, A.M., 2000. Measuring the thickness of the human cerebral cortex from magnetic resonance images. Proc Natl Acad Sci U S A 97, 11050-11055.

Fischl, B., Liu, A., Dale, A.M., 2001. Automated manifold surgery: constructing geometrically accurate and topologically correct models of the human cerebral cortex. IEEE Trans Med Imaging 20, 70-80.

Fischl, B., Salat, D.H., Busa, E., Albert, M., Dieterich, M., Haselgrove, C., van der Kouwe, A., Killiany, R., Kennedy, D., Klaveness, S., Montillo, A., Makris, N., Rosen, B., Dale, A.M., 2002. Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain. Neuron 33, 341-355.

Fischl, B., Salat, D.H., van der Kouwe, A.J., Makris, N., Segonne, F., Quinn, B.T., Dale, A.M., 2004a. Sequence-independent segmentation of magnetic resonance images. Neuroimage 23 Suppl 1, S69-84.

Fischl, B., Sereno, M.I., Dale, A.M., 1999a. Cortical surface-based analysis. II: Inflation, flattening, and a surface-based coordinate system. Neuroimage 9, 195-207.

Fischl, B., Sereno, M.I., Tootell, R.B., Dale, A.M., 1999b. High-resolution intersubject averaging and a coordinate system for the cortical surface. Hum Brain Mapp 8, 272-284.

Fischl, B., van der Kouwe, A., Destrieux, C., Halgren, E., Segonne, F., Salat, D.H., Busa, E., Seidman, L.J., Goldstein, J., Kennedy, D., Caviness, V., Makris, N., Rosen, B., Dale, A.M., 2004b. Automatically parcellating the human cerebral cortex. Cereb Cortex 14, 11-22.

Han, X., Jovicich, J., Salat, D., van der Kouwe, A., Quinn, B., Czanner, S., Busa, E., Pacheco, J., Albert, M., Killiany, R., Maguire, P., Rosas, D., Makris, N., Dale, A., Dickerson, B., Fischl, B., 2006. Reliability of MRI-derived measurements of human cerebral cortical thickness: the effects of field strength, scanner upgrade and manufacturer. Neuroimage 32, 180-194.

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Deep Residual Learning for Image Recognition. https://doi.org/10.48550/ARXIV.1512.03385

Jovicich, J., Czanner, S., Greve, D., Haley, E., van der Kouwe, A., Gollub, R., Kennedy, D., Schmitt, F., Brown, G., Macfall, J., Fischl, B., Dale, A., 2006. Reliability in multi-site structural MRI studies: effects of gradient non-linearity correction on phantom and human data. Neuroimage 30, 436-443.

Klapwijk, E. T., Van De Kamp, F., Van Der Meulen, M., Peters, S., & Wierenga, L. M. (2019). Qoala-T: A supervised-learning tool for quality control of FreeSurfer segmented MRI data. NeuroImage, 189, 116–129. https://doi.org/10.1016/j.neuroimage.2019.01.014

Krishnapriya, S., & Karuna, Y. (2023). Pre-trained deep learning models for brain MRI image classification. Frontiers in Human Neuroscience, 17, 1150120. https://doi.org/10.3389/fnhum.2023.1150120

Ma, J., He, Y., Li, F., Han, L., You, C., & Wang, B. (2024). Segment anything in medical images. Nature Communications, 15(1), 654. https://doi.org/10.1038/s41467-024-44824-z

Pulli, E. P., Silver, E., Kumpulainen, V., Copeland, A., Merisaari, H., Saunavaara, J., Parkkola, R., Lähdesmäki, T., Saukko, E., Nolvi, S., Kataja, E.-L., Korja, R., Karlsson, L., Karlsson, H., & Tuulari, J. J. (2022). Feasibility of FreeSurfer Processing for T1-Weighted Brain Images of 5-Year-Olds: Semiautomated Protocol of FinnBrain Neuroimaging Lab. Frontiers in Neuroscience, 16, 874062. https://doi.org/10.3389/fnins.2022.874062

Raamana, P. R., Theyers, A., Selliah, T., Bhati, P., Arnott, S. R., Hassel, S., Nanayakkara, N. D., M. Scott, C. J., Harris, J., Zamyadi, M., Lam, R. W., Milev, R., Müller, D. J., Rotzinger, S., Frey, B. N., Kennedy, S. H., Black, S. E., Lang, A., Masellis, M., … C. Strother, S. (2022). Visual QC Protocol for FreeSurfer Cortical Parcellations from Anatomical MRI. Aperture Neuro, 76. https://doi.org/10.52294/1cdce19c-e6db-4684-97cb-ae709da06a3f

Reuter, M., Schmansky, N.J., Rosas, H.D., Fischl, B. 2012. Within-Subject Template Estimation for Unbiased Longitudinal Image Analysis. Neuroimage 61 (4), 1402-1418. http://reuter.mit.edu/papers/reuter-long12.pdf

Reuter, M., Rosas, H.D., Fischl, B., 2010. Highly Accurate Inverse Consistent Registration: A Robust Approach. Neuroimage 53 (4), 1181–1196. http://reuter.mit.edu/papers/reuter-robreg10.pdf

Segonne, F., Dale, A.M., Busa, E., Glessner, M., Salat, D., Hahn, H.K., Fischl, B., 2004. A hybrid approach to the skull stripping problem in MRI. Neuroimage 22, 1060-1075.

Simonyan, K., & Zisserman, A. (2014). Very Deep Convolutional Networks for Large-Scale Image Recognition. https://doi.org/10.48550/ARXIV.1409.1556