Detecting Alzheimer's Disease from T1-weighted MRI Coronal Brain Scans

Shakthi Visagan

Dr. Aaron Meyer

Bioengr 188: Machine learning & data-driven modeling in bioengineering

Final Project

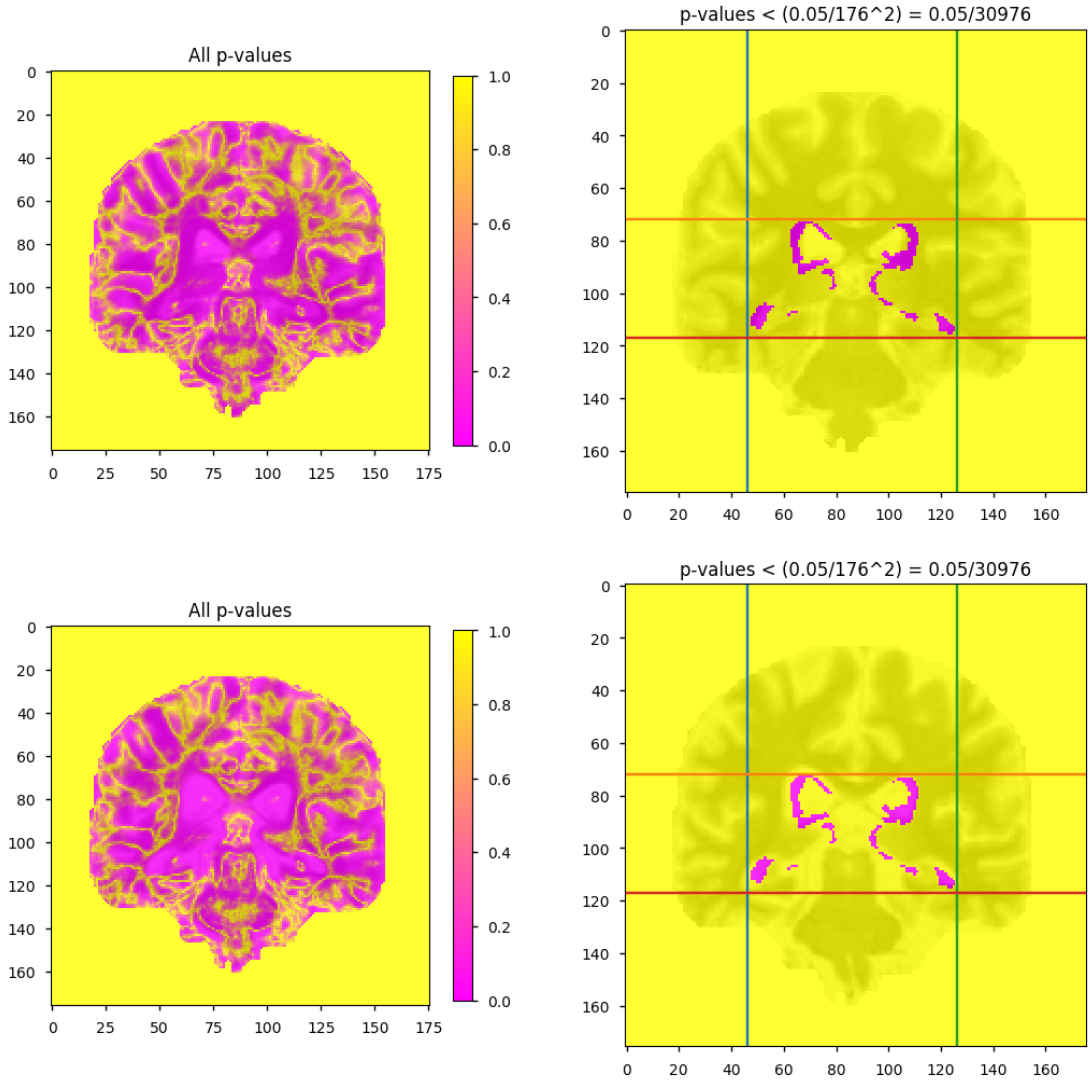
23 March 2018

I. Introduction

Some studies have shown that computer-aided-diagnosis using Support Vector Machines (SVMs) can outperform trained radiologists in identifying patients with Alzheimer's Disease (AD) and other illnesses (Klöppel et al., 2008). Other groups have tried using convolutional neural networks (CNNs) to diagnose and discriminate AD from healthy, no condition (NC), patients using MRI scans as an input, while others have used fMRI time-series data (Payan et al., 2015, Sarraf et al., 2016). A study that is similar to the analysis that follows in this project was done by Gunawardena et al. in 2016, and is presented in their paper titled "Applying convolutional neural networks for pre-detection of Alzheimer's disease from structural MRI data." In this study, researchers used SVMs and CNNs on segmented regions of interest after post-processing using edge-detection algorithms on assorted structural MRI scans. This study was able to achieve a sensitivity of 96% and a specificity of 98% on a dataset of 1615 images using their best performing workflow, which was utilizing segmented images with edge-detection post-processing. Although the analysis that follows is not a reimplementation of the aforementioned study, it draws upon many of the design decisions and hyperparameter choices Gunawardena et al. make.

Brookmeyer et al. forecasts that by 2050, the prevalence of AD will quadruple from 26.6 million patients, where approximately 43% of prevalent cases will need a high level of care (Brookmeyer et al., 2007). If the diagnosis of AD can be improved upon and treated at an earlier, more manageable stage, where therapeutics and preventative care are more useful, then the number of future AD patients will most likely decrease. The

following exploratory analysis will be concerned with diagnosing patients as AD or NC using CNNs. The data consisting of T1-weighted coronal brain MRI scans and the corresponding table of patient information come from the Open Access Series of Imaging Studies (OASIS): Cross-sectional MRI Data in Young, Middle Aged, Nondemented, and Demented Older Adults (Marcus et al., 2007). This analysis specifically chooses slices 90 through 99, as it has been shown in previous studies to show the most significant differences in terms of grey and white matter intensity between AD and NC patients (Zhang et al., 2015).



[1]**Figure 1**

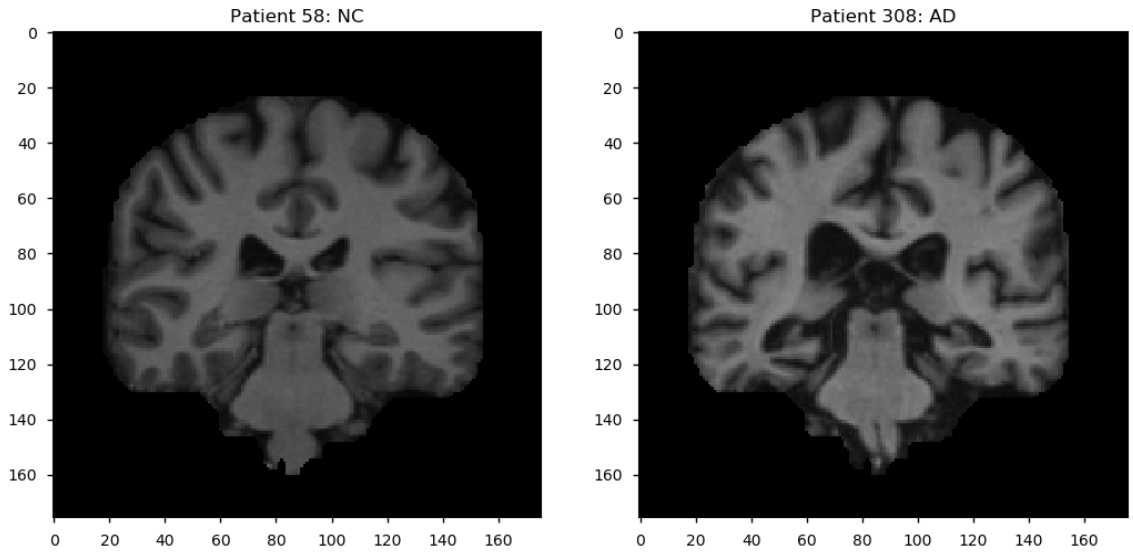
The data also contains Clinical Dementia Rating (CDR) and Mini-Mental State Examination (MMSE) scores for each patient, which are used to classify the patient as AD or NC. The question to be studied is whether or not data like patient information and the images of patient's brains are sufficient enough to provide a valuable diagnostic for AD when used as an input into a well trained machine learning algorithm. This also

[1]**Figure 1**: Results of conducting independent two sample T-tests for pixel intensity for each pixel location against the AD and NC group of coronal slice 90. Bonferroni correction leads to an alpha value of α=0.05/30976, highlighting the most significant areas as the hippocampi and the ventricles. The first row of images is overlaid on a NC healthy brain, while the second row is overlaid on a AD unhealthy brain.

raises the question of whether or not the complexity of the brain can be simplified into a relatively small set of interpretable variables, like the values of the pixel intensities in MRI scans.

II. Methods

After removing, NaN values and patients who did not have a CDR score, the number of patients reduced to 216 from 416. Each patient was then bifurcated into either an AD or NC class based on their CDR score. Due to the lack of a diverse range of scores, patients with a CDR score of 0 were labeled as healthy NC, while any other score, 0.5 through 2, were labeled as AD. Each patient had 10 coronal slices from slice in total making 2160 images. Patient 58 is a representative NC patient, since she is female, 46 years old and has a CDR score of 0. Patient 308 is a representative AD patient, since she is female, but is 78 years old, and has a CDR score of 2.

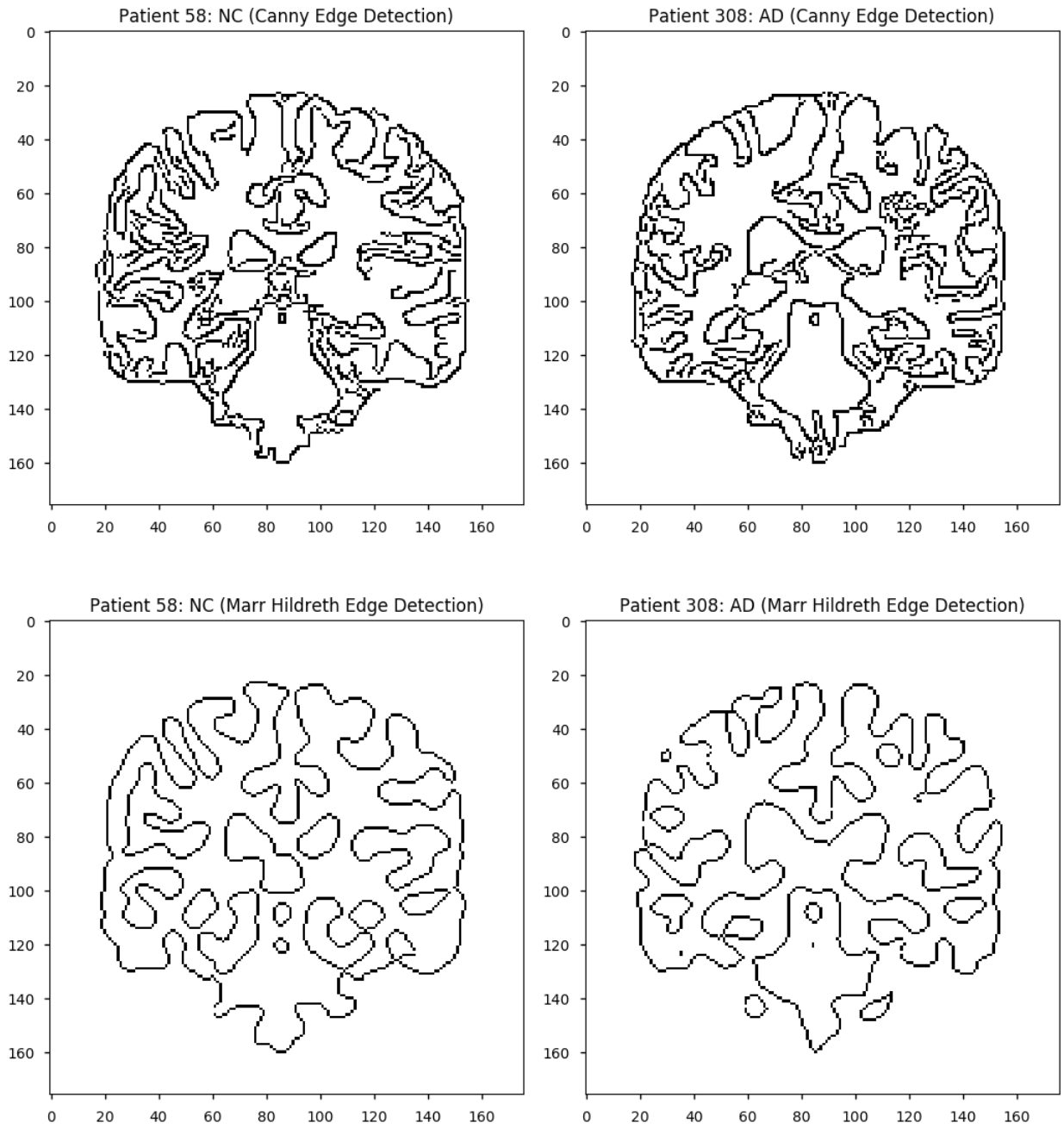


[2]**Figure 2**

[2]**Figure 2**: T1-weighted coronal 90th slices for two representative patients. Patient 58 is NC and Patient 308 is AD. The atrophy in Patient 308 is clearly visible and is in stark contrast to the size of the ventricles, hippocampi, and brain volume in Patient 58.

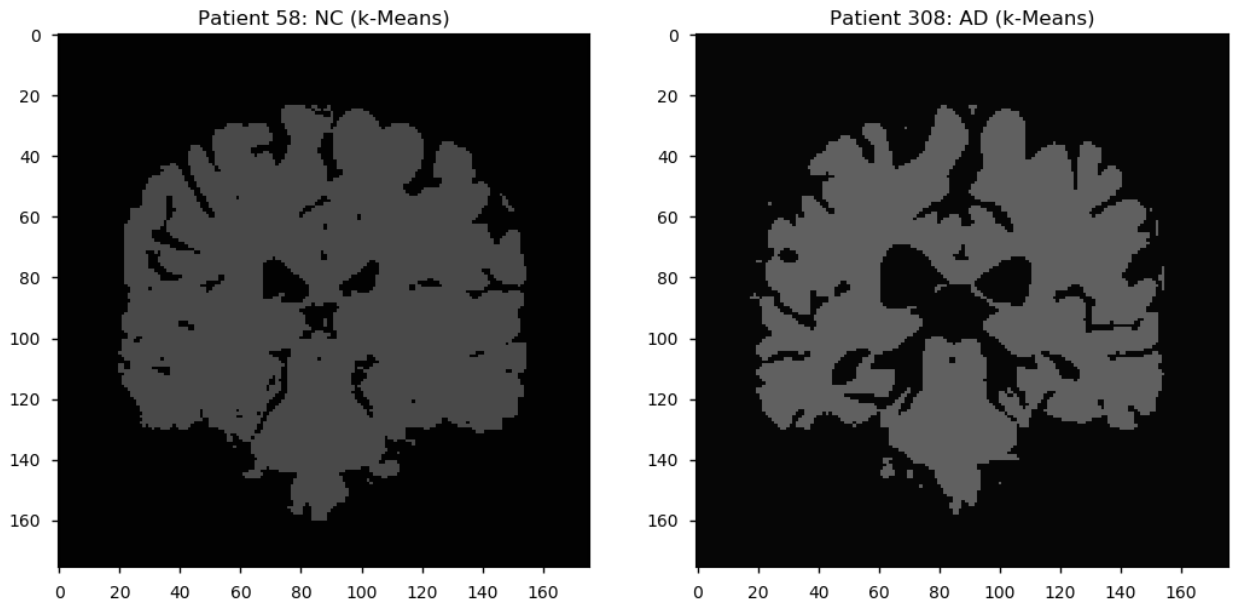
Using the python module OpenCV, the image was read in as a three-dimensional numpy array of size 176x176x3, where 3 corresponds to the 3 red, green, and blue (RGB) channels. This was then normalized by dividing each intensity by 255, the maximum intensity. This scaling allows `matplotlib.pyplot.imshow()` to display pixel values with float data types as long as they are in the range between 0 and 1.

What may be novel about this study, regarding AD diagnosis, is the following preprocessing step of running *k*-means on the pixel space of each image using `sklearn.cluster.KMeans()` from scikit-learn. Since each image's pixel intensities live on a three-dimensional space defined by their RGB channel values, this spread of points of color can be clustered into regions of dominating colors. In this analysis, only 2 clusters were chosen for simplicity and maximum noise reduction, to identify brain and not brain. Each pixel was then assigned to the color of the cluster it was in. It would be interesting in the future to look at more clusters, each cluster of color corresponding to a particular anatomical feature. For example, one could use 4 clusters to identify the background, white matter, gray matter, and cerebrospinal fluid (CSF). The advantages of *k*-means for segmentation processing over a method like gray-scaling the image, or gray-scaling the image and then performing edge-detection is that information of volume and anatomical similarity are retained. Furthermore, since certain colors should be abundant (background black) or scarce (gray and white matter) in AD brains than in NC brains, due to atrophy, the color clusters should differ. This information would likely be lost in edge detection processing since it outputs binary information: edge or not edge.



[3]**Figure 3**

[3]**Figure 3:** Two edge detection algorithms and their output when given binarized images of those from Figure 2. Canny Edge detection uses hysteresis thresholding, after applying a gaussian filter to blur the image and finding changes in the gradient of intensity. Marr-Hildreth does something similar but calculates the second derivative of the intensities. Marr-Hildreth has been shown to create false edges but also provides a less noisy image.



[4]**Figure 4**

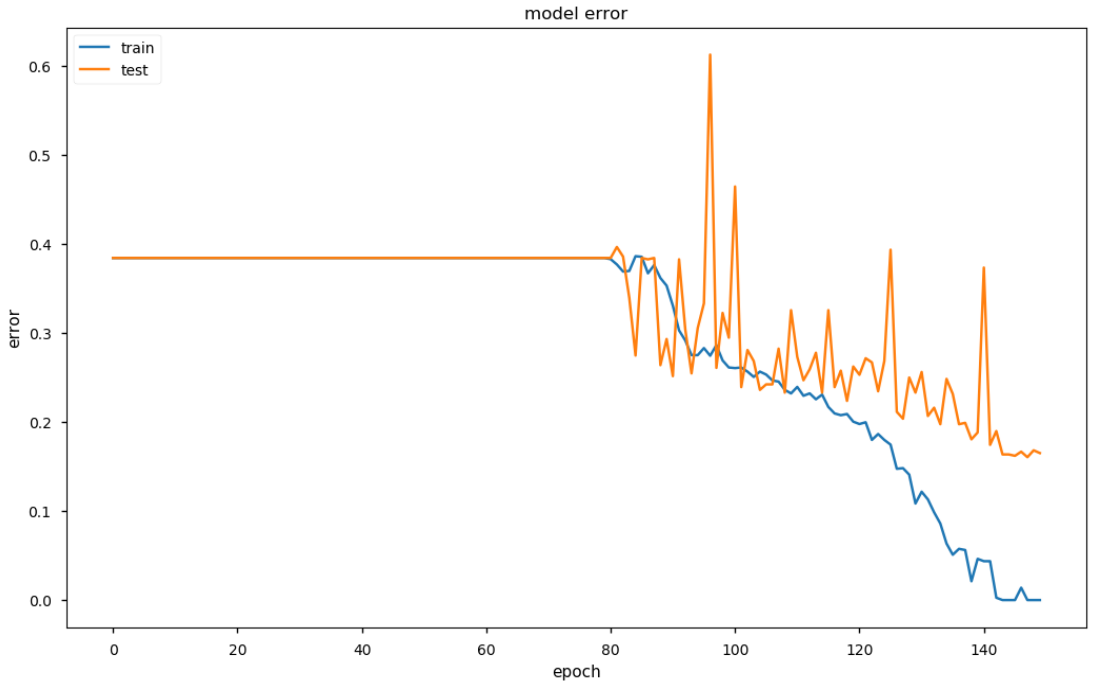
The 2160 images were then partitioned into a training set and a validation set, where 70% of the 2160 images, or 1512 images, were used to train the model in folds or batches, while the other 30%, or 648 images, were used to validate the model. 70% was chosen as it is not only a good partition between training and testing, but also allows the partitioning of each set of training and testing data to have similar prevalence of AD and NC cases. In both sets, 61.57% of the images, or 931 images in the training set and 399 images in the testing set were NC, while the remaining 38.43% of each set were AD. These 1512 images in batches of 32 were then fed into the convolutional neural network VGG16, which uses alternating blocks of convolutional and pooling layers for image recognition (Simonya et al., 2014). This model was created using the Keras `Sequential()` API which lets one add layers to their model sequentially. After 5

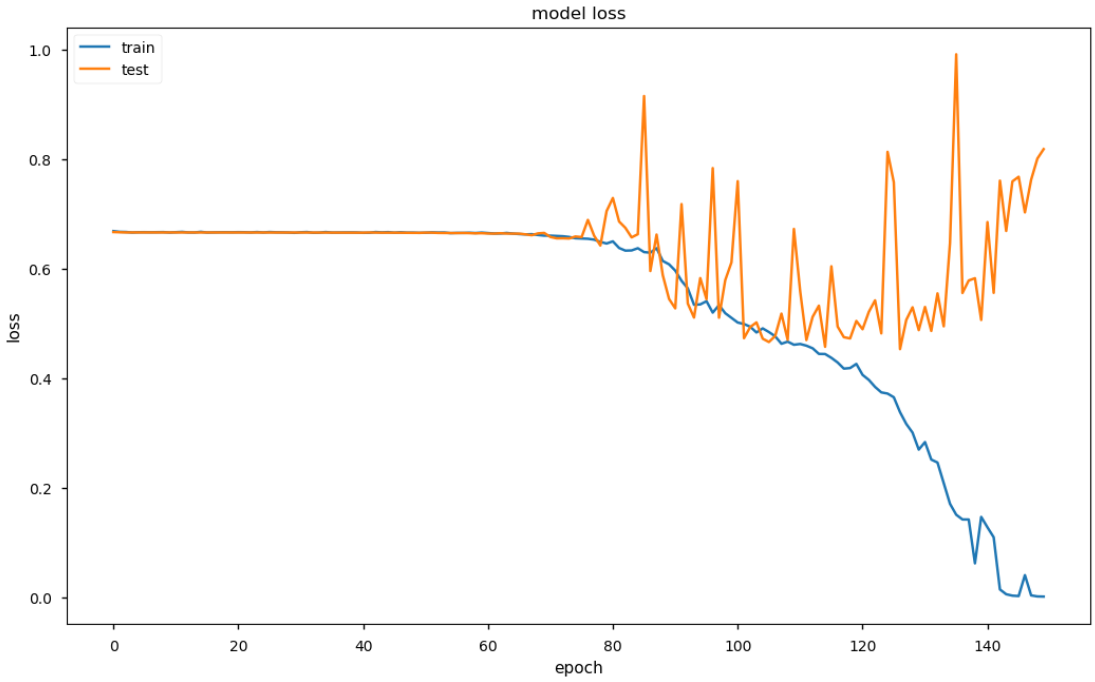
[4]**Figure 4:** The result of applying *k*-means on the images from Figure 2. Though it is difficult to see, Patient 308 has a slightly lighter shade of gray for her brain than the gray used in Patient 58. A difference like that should be one of the features the model learns.

blocks of repeating convolutions and poolings, the model flattens the output and feeds it into two fully-connected or dense layers each containing 1028 connecting units. Following each dense layer is a dropout mechanism, which prevents certain weights from overfitting to the training data by not training a random set of units each training batch, and is a form of regularization. Using a softmax activation function in the final layer, stochastic gradient descent as an optimizer, categorical cross entropy as a loss, and training over 150 epochs, the model trained on a Windows 10 Dell laptop with Intel© Core™ i7-6700HQ CPU @ 2.60GHz, 2601 Mhz, 4 Core, 8 Logical Processors, and had GPU boosted training utilizing the NVIDIA© GeForce GTX 960M enabled by the cuDNN library provided by NVIDIA©. In total, the network had 28,933,986 trainable parameters (weights for each connected unit). Any larger batch size, and any larger network architecture would throw `ResourceExhaustedError` errors and run out of memory when allocating the tensors used in the TensorFlow backend.

III. Results

Each epoch took about 115 seconds to train with the default learning rate on stochastic gradient descent at 0.01. Over a 150 epochs, training took approximately 290 minutes or almost 5 hours.

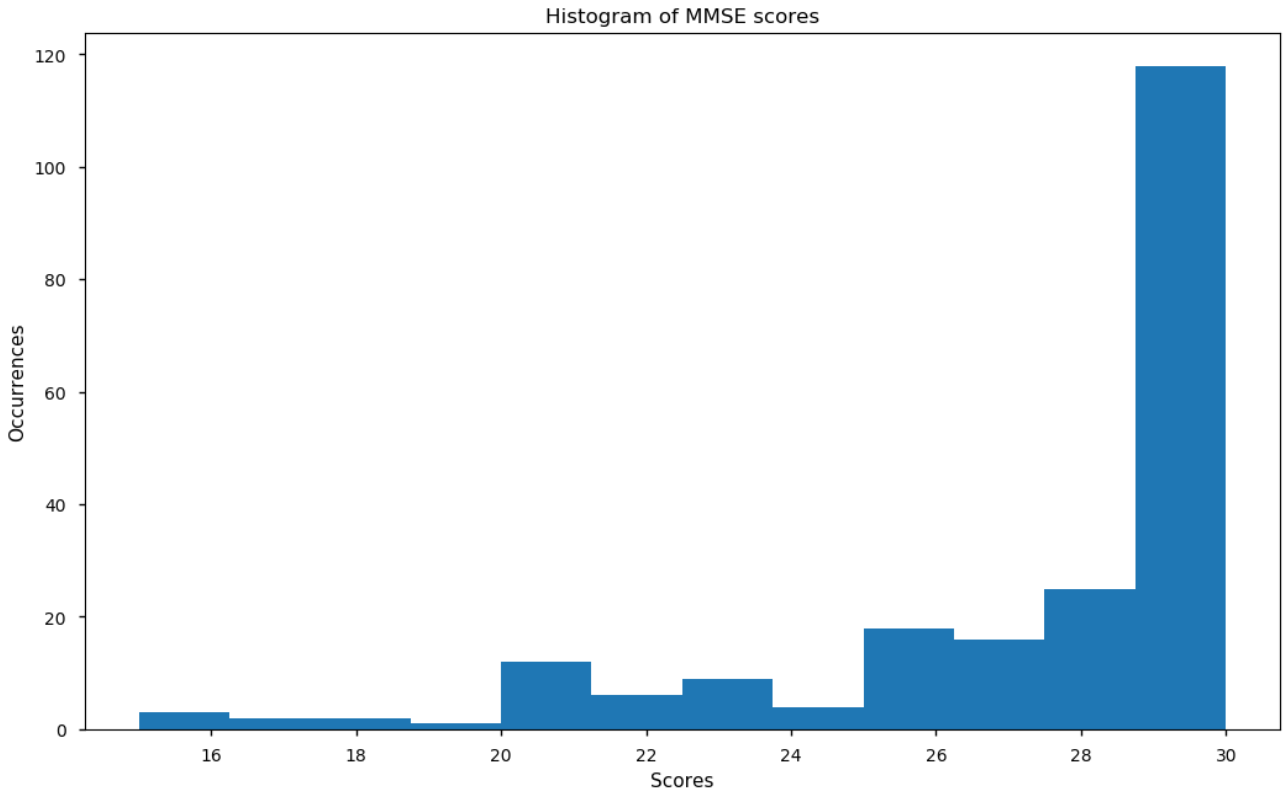




[5]**Figure 5**

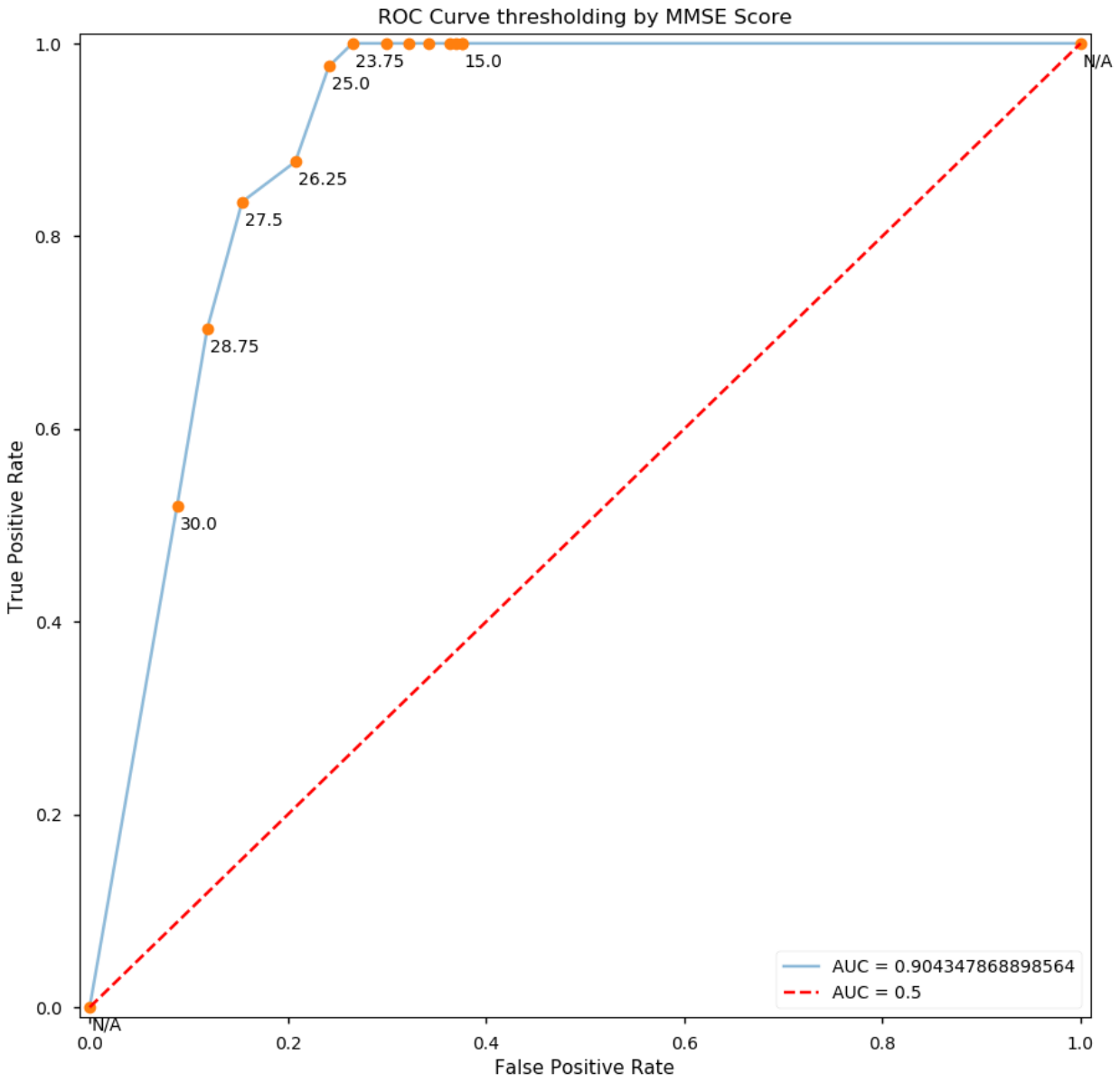
[5]**Figure 5:** The results of training the model over 150 epochs. The fitting error becomes 0 after 144 epochs while the prediction error continues to decrease past 150. Test loss increases after 150 epochs and indicates overfitting as early as epoch 110.

At each epoch, training and prediction accuracy are measured. For the first 81 epochs, the model stagnates by predicting that all brains are NC, but ultimately moves out of the local minima and begins increasing its accuracy. At epoch 144, training error reaches 0%, meaning it has correctly diagnosed all the brains and is evidence of overfitting on the training set. However, by epoch 148, the prediction error is the lowest at 16.05%. It is possible to say that the prediction error could decrease further, but the model was stopped prematurely because of it achieving 0% training error. To ultimately evaluate the power of the model, a Receiver Operating Characteristic Curve (ROC) was constructed by thresholding against a patient's MMSE scores.



[6]**Figure 6**

[6]**Figure 6:** The CDR scale is less continuous and more discrete than the MMSE scale. The MMSE scale also shows a (not as clear but apparent) separation between the two groups in terms of distributions.



[7]**Figure 7**

[7]**Figure 7:** The ROC curve generated by calculating the False Positive Rate (FPR) and True Positive Rate (TPR) for each threshold based on the possible scores on the MMSE. Values larger than 30 and less than 15 resulted in dividing by 0, so the points (0,0) and (1,1) had to be extrapolated. However, a TPR of one was already achieved by the sixth threshold, and hence the curve was bound. The ROC yields an AUC of 0.9. Higher MMSE scores indicate healthier patients. Scoring below a 25 indicates mild impairment and scoring less than 20 and 10 indicates moderate to severe cognitive impairment. Scoring less than a 24 in general is considered abnormal.

ROC analysis by thresholding against the MMSE scores and by using `numpy.trapz(y,x)` to calculate the area under the curve (AUC) yields an AUC of about 0.9 which is good, but is not good enough for clinical settings. Interestingly, several studies often use a MMSE score of 24 as a single cutoff for healthy and unhealthy cognitive state assessment and, the ROC curve indicates peak model accuracy at thresholds below 24, validating their claims (Folstein et al., 1975).

In conclusion, the model of preprocessing with *k*-means and training with VGG16 performed well (training accuracy = 100%, prediction accuracy = 83.95%, AUC = 0.904) and has a clear possibility of performing better, by increasing the number of clusters in the *k*-means step, combining edge-detection and segmentation steps, using better hardware and larger networks in terms of number of layers and number of epochs run and in general obtaining more data. In terms of engineering the biological system of the brain in the context of AD, this model does well in interpreting the brain as a set of pixels and their intensities but easily disregards the numerous other variables that assemble a diagnosis, especially considering that MRI scans are already a large abstraction from the vastly complex human brain.

IV. References

Mukherjee, Siddhartha. “A.I. Versus M.D.” The New Yorker, The New Yorker, 19 June 2017, www.newyorker.com/magazine/2017/04/03/ai-versus-md.

Klöppel, Stefan et al. “Accuracy of Dementia Diagnosis—a Direct Comparison between Radiologists and a Computerized Method.” Brain 131.11 (2008): 2969–2974. PMC. Web. 11 Mar. 2018.

Payan, Adrien et al., “Predicting Alzheimer's disease: a neuroimaging study with 3D convolutional neural networks.” Computer Vision and Pattern Recognition on arXiv. (2015).

Sarraf, Saman et al., “Classification of Alzheimer’s Disease Using fMRI Data and Deep Learning Convolutional Neural Networks.” Computer Vision and Pattern Recognition on arXiv. (2016).

K. A. N. N. P. Gunawardena, R. N. Rajapakse and N. D. Kodikara, “Applying convolutional neural networks for pre-detection of Alzheimer's disease from structural MRI data,” *2017 24th International Conference on Mechatronics and Machine Vision in Practice (M2VIP)*, Auckland, 2017, pp. 1-7. doi: 10.1109/M2VIP.2017.8211486

Ron Brookmeyer, Elizabeth Johnson, Kathryn Ziegler-Graham, H. Michael Arrighi, “Forecasting the global burden of Alzheimer’s disease,” Alzheimer's & Dementia, Volume 3, Issue 3, 2007, Pages 186-191, ISSN 1552-5260, https://doi.org/10.1016/j.jalz.2007.04.381.
(http://www.sciencedirect.com/science/article/pii/S155252600700475X)

Marcus, Daniel S., Tracy H. Wang, Jamie Parker, John G. Csernansky, John C. Morris, and Randy L. Buckner. 2007. “Open Access Series of Imaging Studies (OASIS): Cross-Sectional MRI Data in Young, Middle Aged, Nondemented, and Demented Older Adults.” Journal of Cognitive Neuroscience 19 (9) (September): 1498–1507. doi:10.1162/jocn.2007.19.9.1498

Zhang Yudong, Dong Zhengchao, Phillips Preetha, Wang Shuihua, Ji Genlin, Yang Jiquan, Yuan Ti-Fei. “Detection of subjects and brain regions related to Alzheimer's disease using 3D MRI scans based on eigenbrain and machine learning.” Frontiers in Computational Neuroscience 9. (2015):66. https://www.frontiersin.org/article/10.3389/fncom.2015.00066DOI: 10.3389/fncom.2015.00066. ISSN 1662-5188.

Simonyan, Karen, and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." arXiv preprint arXiv:1409.1556 (2014).

Folstein MF, Folstein SE, McHugh PR: “Mini-mental state: A practical method for grading the cognitive state of patients for the clinician.” J Psychiatr Res. 1975;12:189-198.