Report on Interpolation of MRI Data for use in ADHD Classification in Children

Mattheus Bezerra

The Fall 2019 semester began with a fast reintroduction to Machine Learning Concepts with emphasis on its applications to MRI features of children. The semester's layout was planned for the research group headed by Martin Styner and Mahmoud Mostapha and each of the undergraduates began working on their respective projects. The project on interpolation of data points for the infant MRI data resulted in a plan consisting of the following: Firstly the number of relevant dimensions would be determined for the sake of finding the proper encoding space of the data. Next, this encoding dimensioned would be used in an autoencoder to use unsupervised learning in reconstruction of the data. This would be useful since an encoded representation can then be used for a more condensed representation of all the information. An autoencoder would then be generated using encoded representations of the data to predict future data points. This in turn allows for completion of the existing data and supplementation of the project to classify risk for ADHD in development.

Early into the semester, principal component analysis (PCA) was run on the data for both surface area and cortical thickness of the brain scans for the purpose of finding what a reasonable reduced dimensionality would be for the MRI data that still encapsulates the subject's information. The Gilmore data set provided had exactly 150 features ranging from the cortical thickness and surface area of the Fronto-marginal gyrus to divisions of the temporal sulcus. Focusing specifically on singleton children, the results are indicated in the figures below.

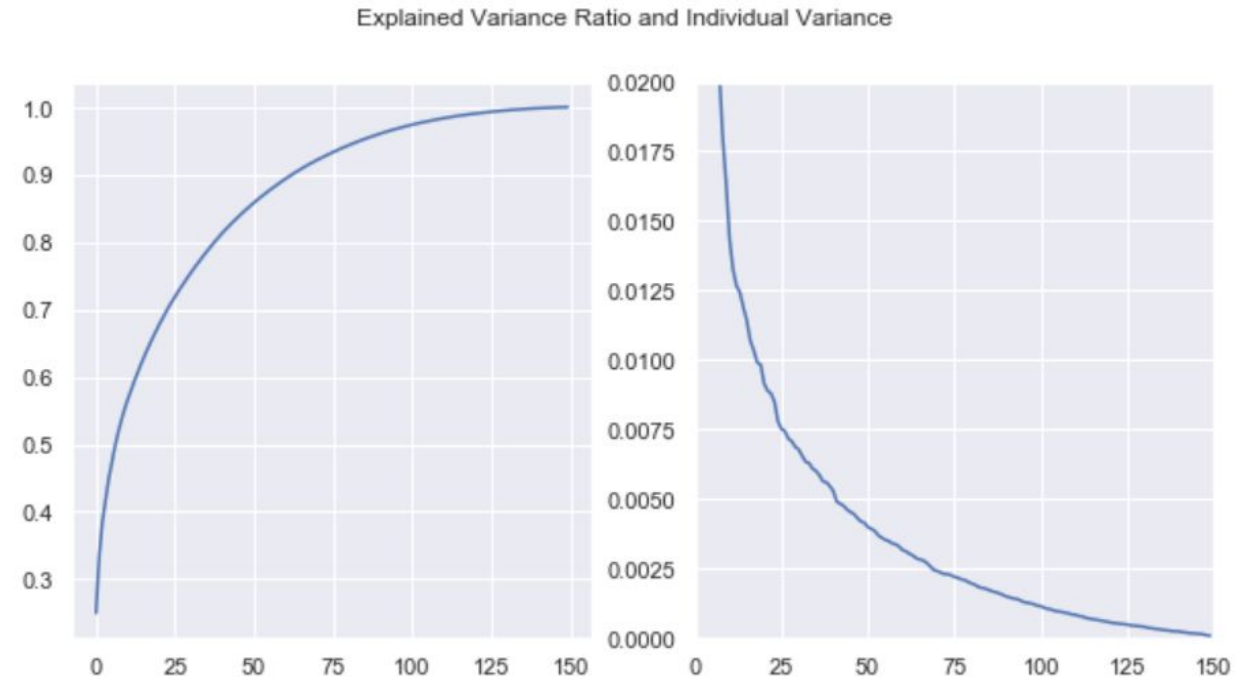


Figure 1a (PCA on cortical thickness at 1 year)

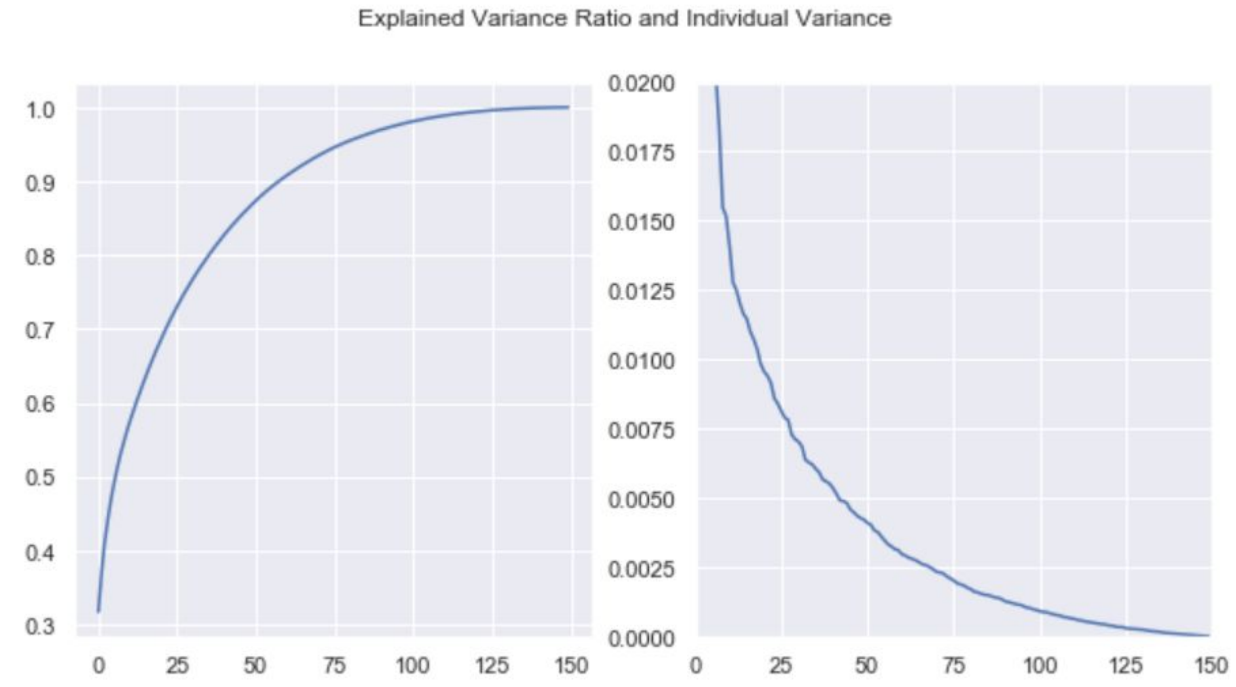


Figure 1b (PCA on cortical thickness at 2 years)

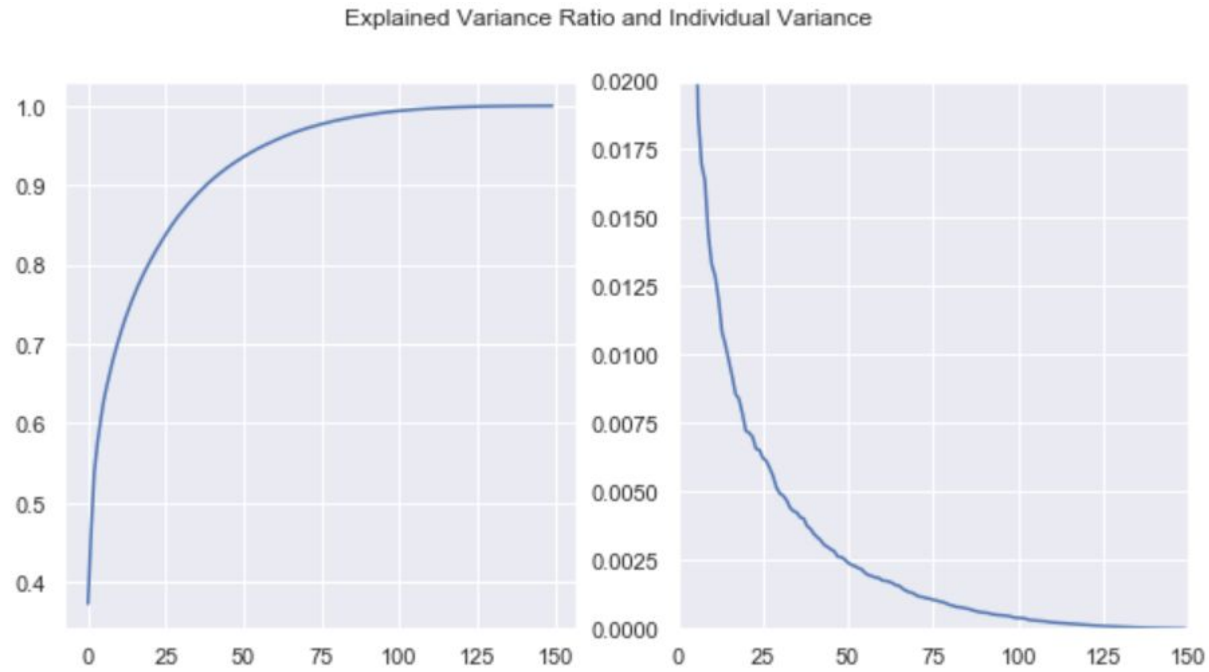


Figure 1c (PCA on cortical thickness at 4 years)

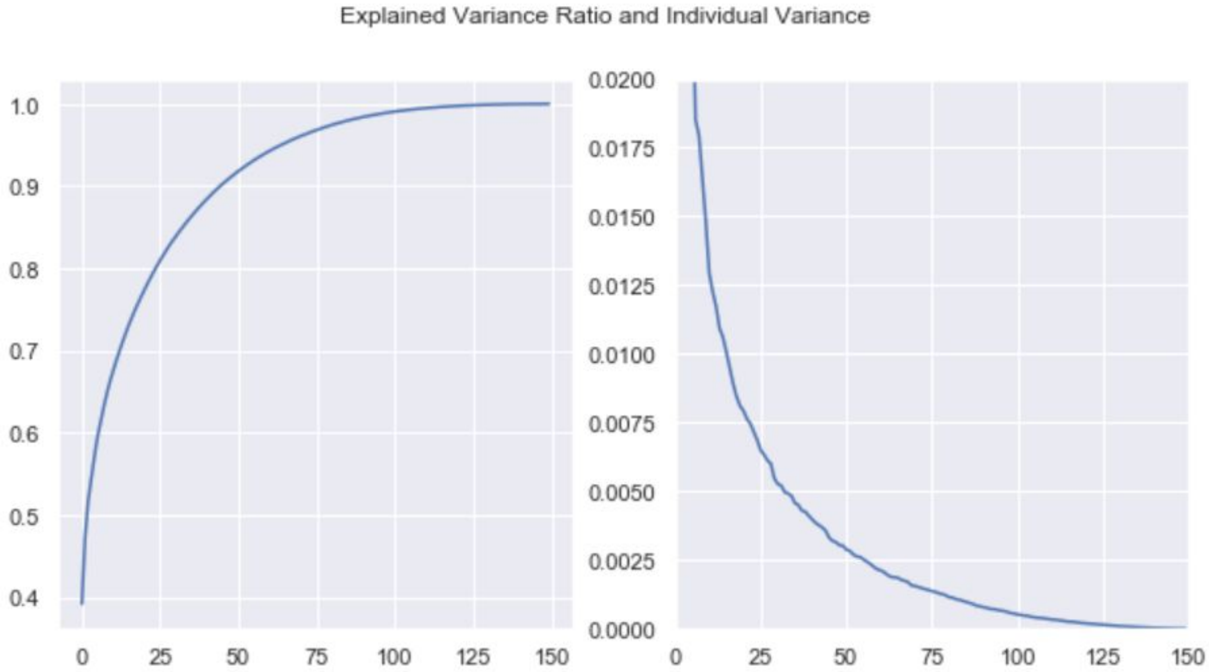


Figure 1d (PCA on cortical thickness at 6 years)

By locating where the greatest amount of variance could occur within the least amount of components the graphs indicate that for 1,2,4 and 6 years respectively:

Sum of PCs explained variance ratio at 25 PCs 0.719320215351573

Sum of PCs explained variance ratio at 50 PCs 0.8587291318117513

For 1 year.

Sum of PCs explained variance ratio at 25 PCs 0.7308269132341118

Sum of PCs explained variance ratio at 50 PCs 0.8739717278334065

For 2 years.

Sum of PCs explained variance ratio at 25 PCs 0.8384863911946363

Sum of PCs explained variance ratio at 50 PCs 0.9360940278535277

For 4 years.

Sum of PCs explained variance ratio at 25 PCs 0.8100507059168914

Sum of PCs explained variance ratio at 50 PCs 0.9185160553627404

For 6 years.

Given the individual variances it was decided that 13 dimensions make up for majority of individual variances in cortical thickness. The same procedure was taken for the surface area data also segmented into 150 features of the same regions of the brain. Its figures are below.

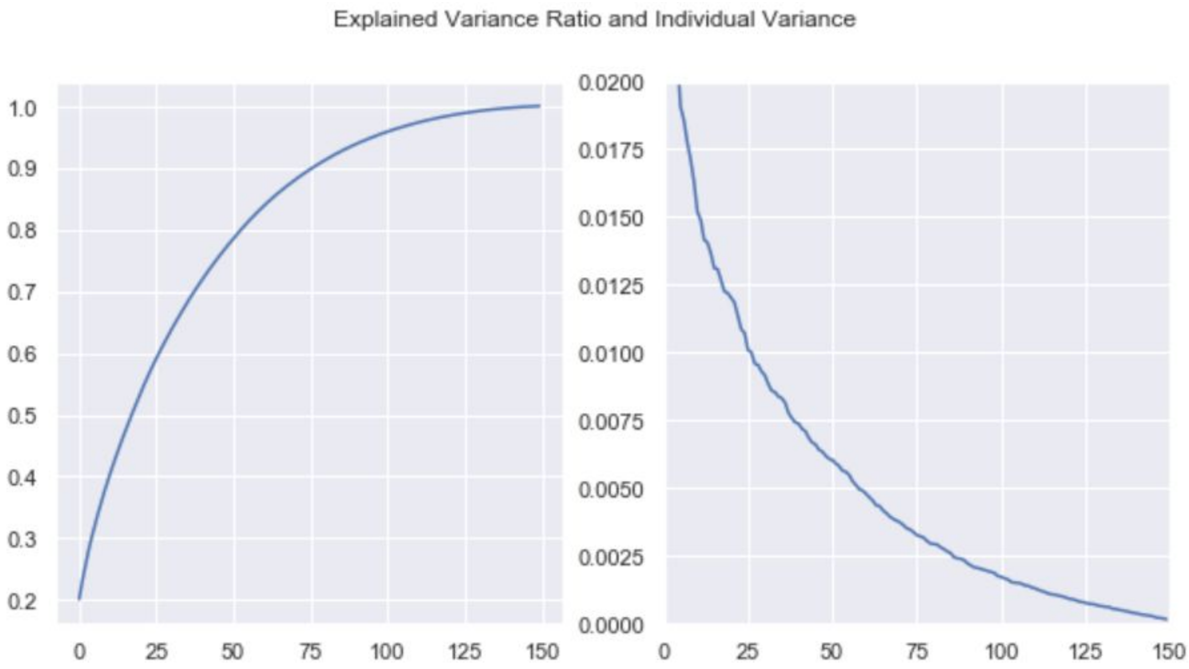


Figure 2a (PCA on surface area at 1 year)

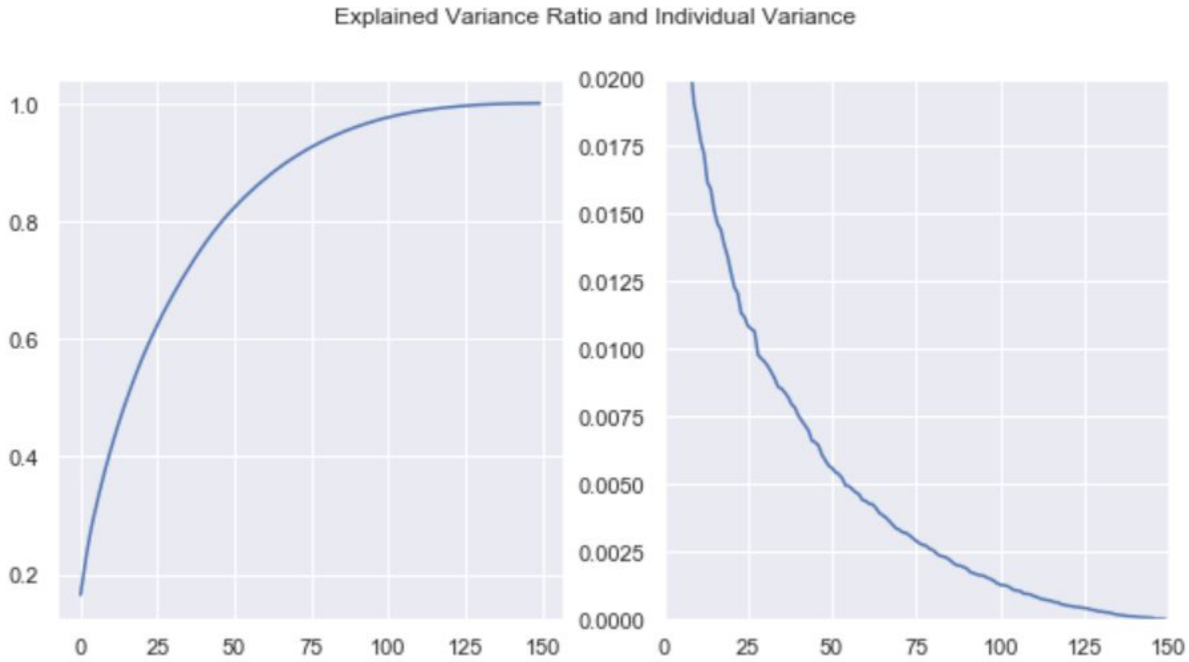


Figure 2b (PCA on surface area at 2 years)

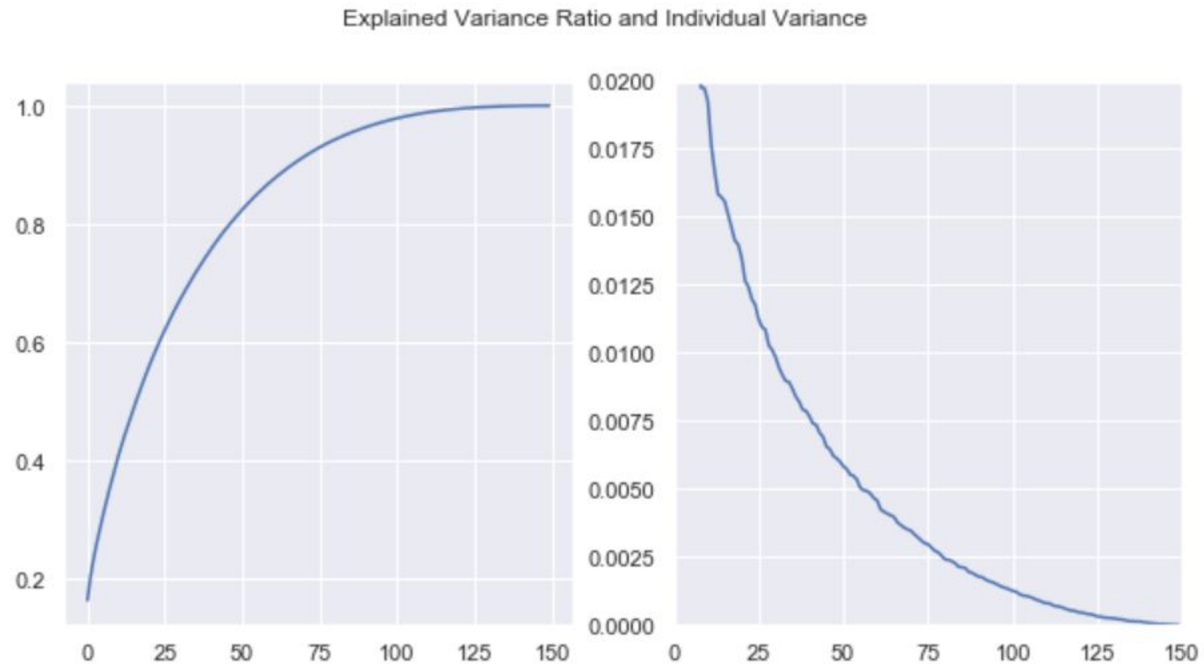


Figure 2c (PCA on surface area at 4 years)

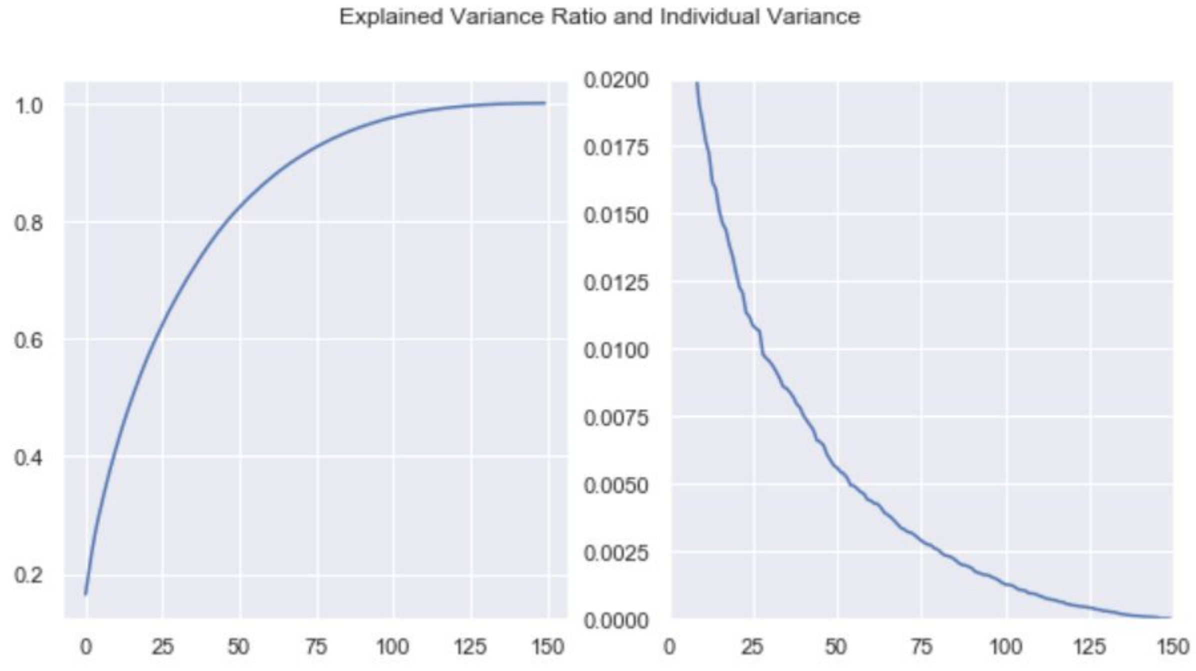


Figure 2d (PCA on surface area at 6 years)

By locating where the greatest amount of variance could occur within the least amount of components the graphs indicate that for 1,2,4 and 6 years respectively:

```
Sum of PCs explained variance ratio at 25 PCs 0.5916956818673522
Sum of PCs explained variance ratio at 50 PCs 0.7859582747751767
```

For 1 year.

```
Sum of PCs explained variance ratio at 25 PCs 0.6088394659542372
Sum of PCs explained variance ratio at 50 PCs 0.808407924285891
```

For 2 years.

```
Sum of PCs explained variance ratio at 25 PCs 0.6212662095691662
Sum of PCs explained variance ratio at 50 PCs 0.8245992088225031
```

For 4 years.

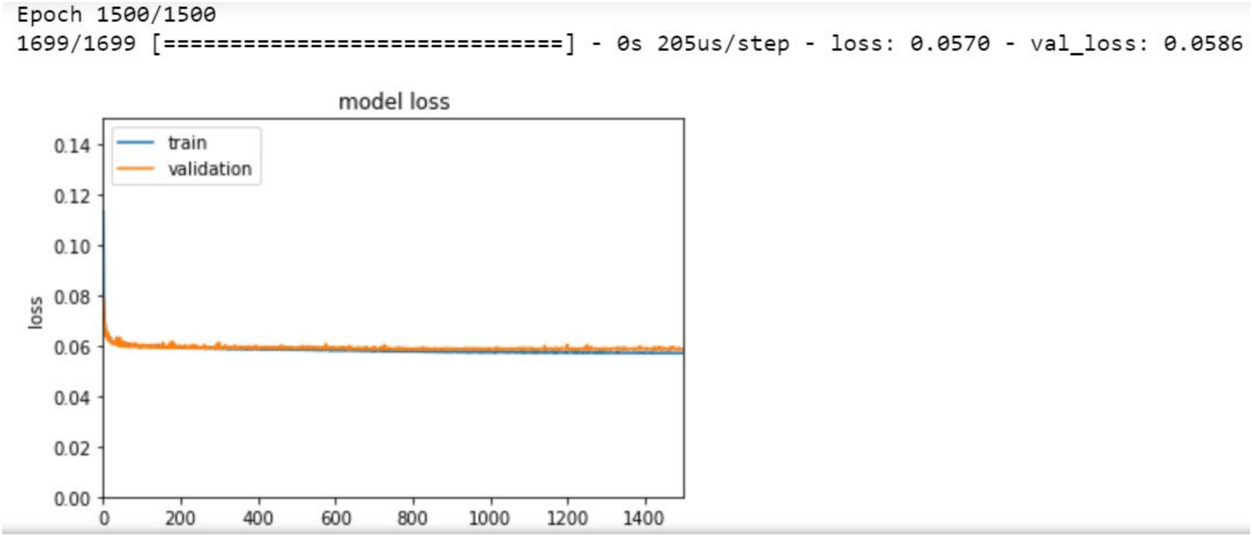
```
Sum of PCs explained variance ratio at 25 PCs 0.6241385394946164
Sum of PCs explained variance ratio at 50 PCs 0.823432744795252
```

For 6 years.

Given the individual variances it was decided that 25 dimensions make up for majority of individual variances in surface area.

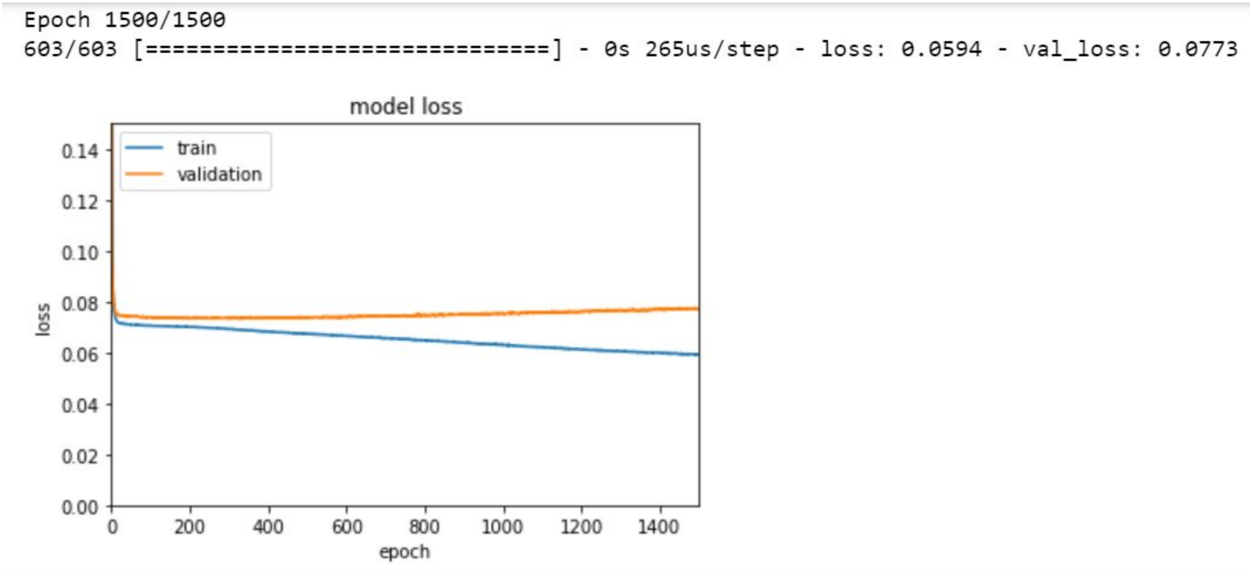
Following the determination of an encoded representation, the next step was to build an autoencoder for reconstruction of the data. This in turn determines a valid encoded representation for each data set and allows researchers to use the encoded representation as a starting point for prediction of future years. Much of the semester was spent optimizing this model and determining what gives the most accurate representation and least loss in reconstruction ensuring a strong model. The initial models began with high rates of loss, almost at 50%, indicating reconstruction was very inefficient. Loss was reduced by changing the model to use the Adam optimizer, along with choosing a different scaling of the data (standard scaler z-scoring to minmaxing). Lastly, it was found that a greater variance in subjects would reduce loss greatly. An adjacent data set known as the

IBIS data set was introduced as pretraining data for construction in the autoencoder to later be removed in testing. The final models ended up with 2 hidden layers, with the starting input layer consisting of 148 features (removal of the left and right medial walls as irrelevant information), followed by their 2 hidden layers to their respective encoding spaces. For cortical thickness this was 148 to 105 to 38 to 13. For surface area this was 148 to 107 to 66 to 25. Hidden layers were optimized by running multiple configurations and using models that had lower losses. These are roughly linear for 4 total layers in division between 148 and 13 or 148 and 25 respectively. Reconstruction ended in roughly a 6% loss.



```
Epoch 1500/1500
1699/1699 [==============================] - 0s 205us/step - loss: 0.0570 - val_loss: 0.0586
```

Figure 2a (Reconstruction loss for cortical thickness)



```
Epoch 1500/1500
603/603 [==============================] - 0s 265us/step - loss: 0.0594 - val_loss: 0.0773
```

Figure 2b (Reconstruction loss for surface area)

The models indicate overtraining which may need to be revisited before prediction.

Future steps will be to now construct a decoder that predicts future measurements from encoded representations of past ones. This will be done by using the reconstruction autoencoder and saving the encoded representation of the past measurements. These will then be decoded and validated against the future samples for prediction. In addition to this, work will be done on the long leaf cluster for efficiency and speed in training requiring a learning of slurm and the long leaf cluster. Predicted measurements can then be used to fill holes in current data as well as provide new, meaningful samples as added variance to classifying risk groups for ADHD.

References

[Goo] Google. Machine learning crash course. https://developers.google.com/machine-learning/crash-course/. Accessed on 2019-09-22.

[Say18] Sayantini. edureka: Autoencoders tutorial : A beginner's guide to autoencoders. https://www.edureka.co/blog/autoencoders-tutorial/, Oct 2018. Accessed on 2019-09-13.

[LAO] YangLi, ShoaibAkbar , Junier B. Oliva. Flow Models for Arbitrary Conditional Likelihoods. arXiv:1909.06319v1 [cs.LG] 13 Sep 2019. Accessed on 2019-09-27.

[LLZ+3] Mingxia Liu, Chunfeng Lian, Tao Zhou, Synthesizing Missing PET from MRI with Cycle-consistent Generative Adversarial Networks for Alzheimer's Disease Diagnosis: 21st International Conference, Granada, Spain, Sep 2018. Accessed on 2019-09-20.