Results:

* For the test data the best accuracy rate is: 85.7% with brain threshold to be 0.4, PCA parameter set to be 78, SVM using linear kernel, and 19 stratified folds.
* For the retest data the best accuracy rate is: 87.5% with PCA parameter set to be 90, SVM using linear kernel, and 16 stratified folds.

1. How to run your code? Any variable needs to be changed before we run your code?

For this assignment, I chose to apply svm with pca which give the better accuracy rate for test and retest dataset. Below are all the functions:

1. Brain region extraction: get\_masked(file,threshold)
2. Feature extraction: extract\_features(masked, n)
3. SVM with PCA : svm\_with\_pca(x\_reduced, labels, c, kernel, splits)
4. SVM without PCA: svm\_wth\_pca(masked, labels, c, kernel, splits)

(used to compare the results with results from svm+pca)

The first four lines are the file paths and loading of labels.

Following all the functions are the lines that calls all the functions, just run them line by line and it will give the best accuracy rate for test and retest data set.

1. Briefly discuss the limitations. How can you improve it?

The first limitation is that the sample size is relatively small, especially when we do the higher number of folds to do cross validation, the sample is divided into even smaller groups. With a larger data set, the result could be more persuasive, and the method can be tested with stronger support.

Without using the dimension reduction, SVM would take a longer time. Therefore, it may be very costly when sample size grows very large. To improve the efficiency, we could try different methods of classification.

For my homework, I did not realign the brain images, which could possibly improve the accuracy rate.

1. Describe the results of the experiment in your own words. Compare the results of two approaches (SVM vs PCA+SVM) and briefly discuss why one works better/worse than the other?

From my results, PCA+SVM works better than SVM using the same parameters.

PCA is not only reducing the dimension but also maintaining the relationship between data points. When each group’s essential information is kept, the main features are selected which facilitates svm can do a better job in classifying them. When we only use svm, the original data points are high dimensional, which could add many noises and make the classification difficult and inaccurate.