ELEC 409 - Bioinformatic Analytics Assignment – due December 3, 2024, 5pm.

NB: **Supplementary Information** document (MS Word) obtained from Broad Institute site

https://portals.broadinstitute.org/cgi-bin/cancer/publications/view/52

has further information, including identification of Alive/Dead (page 24, entitled Dataset C – MD outcome). Note that the 1st 21 patients died, i.e. Brain_MD_1, Brain_MD_2, …, Brain_MD_21 are all non-responders. The remaining patients are responders.

This assignment is based on two research papers discussed in class: Golub et al, "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring", Science 1999; and Pomeroy et al, "Prediction of central nervous system embryonal tumour outcome based on gene expression", Nature 2002. You are to conduct a re-analysis of the gene expression data in Pomeroy et al, where a single data set representing 60 patients, and leave-one-out testing, had been used to predict medulloblastoma clinical outcome. See Dataset (opens with Excel), obtained from

https://portals.broadinstitute.org/cgi-bin/cancer/datasets.cgi

Instead of using a single data set, you are to rigorously separate the data into a training set representing about 30 patients, and a test set representing the remaining patients. Take about half of the patients for the training set and the other half for the test set, e.g. randomly choose the patients for the two sets. Ensure that you have about half the patients who survived, and about half who died, in each set. There must be no overlap between training and test sets. On the training set, use leave-one-out testing to discover optimal parameters for either weighted voting or k-nearest neighbor classifiers, and select the best weighted voting classifier or the best k-nearest neighbor classifier. Optionally, you may also train a Deep Learning or other machine learning algorithm on the training set. Then use the test set to evaluate the performance of the selected weighted voting classifier or k-nearest neighbor classifier, and compare performance with that of the machine learning algorithm if you have opted to develop one as another classifier. Use Fisher's exact test and Matthews' correlation coefficient Phi

http://vassarstats.net/tab2x2.html

to evaluate the performance of the selected classifier and of the optional machine learning algorithm. Your mark will not depend on your developing a machine learning classifier. Report your study in a mini-paper that you email to me as a pdf file by the due date. Include a copy of the software programs you used. Your report should have enough detail to enable replication of your results. Each student must submit their assignment wholly due to their own work.