COMP 755 Project Abstract:
Classification Methods for Breast Cancer Subtype

Kevin O'Connor

For many years now, successful treatments for various types of cancers have eluded researchers. Breast cancer in particular continues to affect millions of women and men. An important characteristic in the study of breast cancer is the subtype of disease, identified as *Luminal A*, *Luminal B*, *Basal*, *HER2*, or *Normal*. For a single tumor, the subtype determines both how the disease may develop in the patient as well as the proper course of treatment. Thus, there is significant incentive for researchers to find distinctive characteristics of each subtype to more clearly identify them in future patients.

In the past decade, the cost of obtaining genetic information from an individual has decreased substantially. This has led to a boom in the amount of genetic data available in many areas of medicine. In 2005, a database called *The Cancer Genome Atlas* (TCGA) was established in order to organize this data and make it widely available to researchers around the world for analysis. At this point, the database contains genetic expression levels for over 20,000 genes from patients with tumors in each of the 5 breast cancer subtypes. It offers an excellent opportunity to look for patterns in the genetic data that distinguish the subtypes.

In my project, I propose to use the classification methods we are learning in the course to find distinguishing patterns in the gene expression data to predict breast cancer subtype. Specifically, after processing the data and preparing it for analysis, I will apply logistic regression, an SVM, and a deep neural network model. In doing so, I will discuss the potential advantages and disadvantages of each model and compare their performances. I expect to see the misclassification rate decrease as we move from logistic regression to SVM to deep neural network. If time permits and interesting sets of genes are found to be important predictors, I will also try to compare them to known gene pathways associated with some biological process to draw connections between cancer subtype and function elsewhere in the body.

Subtype	TCGA Sample Size
Luminal A	572
Luminal B	219
Basal	191
HER2	82
Normal	137