# CS221 Fall 2016 Project Progress Report

SUNet IDs: udai, nng1, bcui19
Names: Udai Baisiwala, Natalie Ng, Brandon Cui

## Project Scope

In our project, we want to consider the use of artificial intelligence algorithms to investigate gene express expression on data. Many medical papers include small dataset ($< 100$ samples) of gene expression data that they use to postulate a certain relationship between combinations of gene expressions and specific medical conditions (frequently specific cancers). We want to run several generic algorithms on multiple of these datasets to see whether we get the same classifications as the paper authors and evaluate the efficacy and drawbacks of those various algorithms. With this, we hope to get a better understanding of what techniques are optimal for gene expression data and potentially design an artificial intelligence algorithm that will be a good classifier for all gene expression classification problems.

## Input-Output Behavior

For each artificial intelligence technique that we use, our input will be a curated dataset that connect the expression levels of 20,000 genes to whether or not that sample is cancerous or not. We will collect datasets from the Gene Expression Omnibus, which is a functional genomics data repository of high throughput gene expression data and hybridization arrays, chips, and microarrays. We will split datasets into training and validation sets, and only perform training on the training set. The output will be whether a particular sample is classified as cancerous or not. An example datapoint with only 4 genes for a patient is as follows:

| Gene Name | Expression Level |
|---|---|
| $A23P100011$ | $-1.584$ |
| $A23P100022$ | $-0.695$ |
| $A23P100056$ | $-0.661$ |
| $A23P100074$ | $-0.641$ |

where there is mapping between the gene name and the level of gene expression.

## Evaluation Metric for Success

We want to have a good understanding of why certain algorithms perform or do not perform well in classifying the presence of cancer. If we can give credible explanations for why algorithms do or do not perform well on various gene expression data sets, we would consider this project a success. As a reach goal, we would like to design an algorithm that would perform well on all gene expression datasets.

## Concrete Sample Inputs and Outputs

We performed a proof of concept on one dataset to gain a baseline value. From this data repository we have selected an RNA dataset (Acession: GSE16449) on renal carcinoma (kidney cancer), where we will get features based upon gene expression that will be used for our input. The output will be a classification of whether or not the gene expression is cancerous. We implemented a naive logistic regression algorithm that naively runs through the genetic data and for each gene considers the gene expression as the feature weight. With the false positive rate being 85.71% and a false negative rate of 36.36% on an external test set. This shows that the dataset believes that everything is cancer.

## Baseline and Oracle

Our baseline is our naive logistic regression method that we have implemented. Results from this are described in more detail above.

Our Oracle is the false positive and false negative rates of current cancer screens in the US. Most of these screens range from 5% to 20% false positive rates, and 5% to 10% false negative rates. The best one that we found for kidney cancer is a renal biopsy with a 4.4% false positive rate and a 1.2% false negative rate. We keep this as our oracle.

## Possible Challenges

Since our datasets are relatively small, in order to have enough samples for a test and validation set, we might need to experiment with resampling techniques. If this does not work, we probably could still have enough statistical power with our datasets.

Another possible challenge is the scalability of our project. Each one of our datasets comes with 20,000 features, so we need to be intelligent with which features we use in our artificial intelligence techniques. Or, we may not have enough computaitonal power.

Finally, we forsee trying to gain insights on why particular methods work better on certain datasets. This will require us to gain quite a bit of intuition on how the dataset is formed and the techniques used to gather this data. If we run on upwards of 10 datasets, it would take quite a bit of time and data exploration to gain this intuition.

## Related Work

JA Cruz, et. al. "Applications of Machine Learning in Cancer Prediction and Prognosis" This is a review paper that shows that "old" machine learning techniques can be used to improve prediction by 10-15% percent. (2006)

K. Korou et al. "Machine learning applications in cancer prognosis and prediction." This article shows that ML techniques have been used, but additional validation of these

techniques is needed. This shows that there is a need for a project like ours to gain intuition about which techniques work for certain gene expression datasets. (2015)