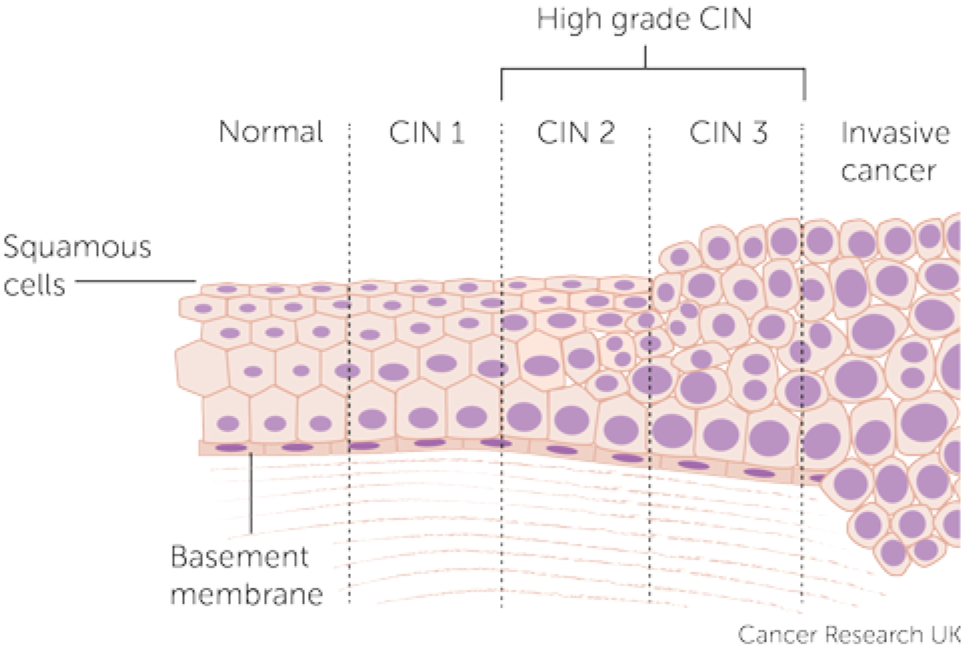
Data Mining Trends from Cervical Cytology using a 3-type HPV mRNA Cervical Intraepithelial Neoplasia Grade 2+ in Young Norwegian Women

Andrew Vuong, Henry Wang, Aaron Johnson

Problem Statement

The project aims to create a classification model to classify whether the person has cervical cancer or not based on a collection of data from a cohort study of cervical cytology with or without HPV mRNA biomarkers examined to detect cervical intraepithelial Neoplasia in young Norwegian women. There are two diagnostic tests to determine the presence of cervical cancer, through a cervical cytology (pap smear) or an HPV mRNA DNA test. The problem is important to determine whether HPV mRNA biomarkers are significant enough to influence positively the sensitivity of cervical cancer screening. Needless to say, early detection is key to high survival rate in cancer patients and having a higher sensitivity would result in less false negatives. The model could be trained to predict whether the patient has cervical cancer given a set of features such as HPV mRNA biomarkers (SEE_01, SEE_16, SEE_18, and SEE_45) and the types of diagnostic study exam.



https://www.cancerresearchuk.org/about-cancer/cervical-cancer/treatment-for-abnormal-cervical-cells/what-are-abnormal-cervical-cells

Data

The dataset collected 4366 records of young Norwegian women (under the age of 40) in a cohort study. The features included were four mRNA biomarker and type of diagnostic study exams. The dataset contains biopsy results which are considered the label. Half of the records were missing the DNA test results which later is filled through a novel technique, all had at least a pap smear.

Method

The features included were four mRNA biomarker, type of diagnostic study exam, and age. The labels were a graded biopsy result.

There are 7 total labels for the data, with 3 of them indicating the presence of the disease. We iterated over the dataset and converted instances of those 3 labels to ‘1’ and the rest to ‘0’ so that we could have a binary classification problem that simply predicts whether a patient has the disease or not.

We assumed several things: the grade of cervical cancer from the biopsy label may be binned as positive because given enough time and lack of diagnosis all lower grade cancer progress to a higher grade naturally and therefore was dependent on time of identifying the grade of cancer, the positive HPV mRNA values does not guarantee certainty of cancer (even the science is not absolute that presence of HPV mRNA means cervical cancer, this is where our data mining experiment becomes interesting), that we fully understand the relationship between pap smear cytology and mRNA test which the researchers aim to identify, and when considering results that may impact life or death of a patient we assume that we have high enough accuracy even though sometimes 99.9% is not enough when talking about someone’s life.

Around half of the data points had values missing in 3 of the features, so to fill them in, we looked at the points in a subset “B” that had those features filled in but also matched all the other features and assigned the missing values a random number, based on the distribution of the points in B had for that feature. To signal missing mRNA values, the data collectors used a value of -1 for the first feature of the mRNA SEE_01 and NaN for the rest of the mRNA’s. In order to have symmetry in data structure, this was the only case where we manually modified the dataset by removing all of the -1 so that if mRNA values were missing they would all be NaN.

The distribution was created using DataFrame groupby() around ['study_gr', 'study_gr1', 'Diag_cyt', 'Diag_cyt_rev'] as the keys which all data shared and then the mean of the respective keys. The mean on binary values will yield probability that the missing values are 1 (pos) or 0 (neg). This way when an encountered record missed mRNA values, the project used the keys to generate the distribution of mRNA values.

Following data cleaning and data prepping we applied several classifiers to form the following network architecture: Decision Tree, Naive Bayes, KNN, SVM, NN, Ensemble, AdaBoost. The architecture also produced a ROC Curve for the Naive Bayes and a confusion matrix as well as nested cross validations.

Challenges

One of the biggest challenges we encountered was simply the time it took to build certain models. For example, for our SVM model we wanted to have a pipeline that included dimensionality reduction as well as scaling, and pass that into a grid search in order to find the best parameters. However, it was taking an extremely long time to run, so we had to forego the pipeline and pass the scaled data directly into the SVC. Fortunately, we were still able to obtain a respectable accuracy.

Another challenge was understanding the domain we were working with, fortunately we do have some medical background to understand the data and research, so all in all that went well. The classification was straightforward given we binned the biopsy result between positive or not rather than a graded multi classification between the level of cancer presence. Binary classification actually made more sense than a multi classification problem because whether high grade or low grade cancer, it is utmost important to determine whether cancer is presence or not and to treat it immediately. The added complexity of multi classification fail to offer a sufficiently strong enough trade off because given enough time all low grade cancer naturally progresses to a higher grade cancer when left untreated, therefore is reliance on time of identification and therefore beg the question of whether cancer or not, a binary classification problem.

Results

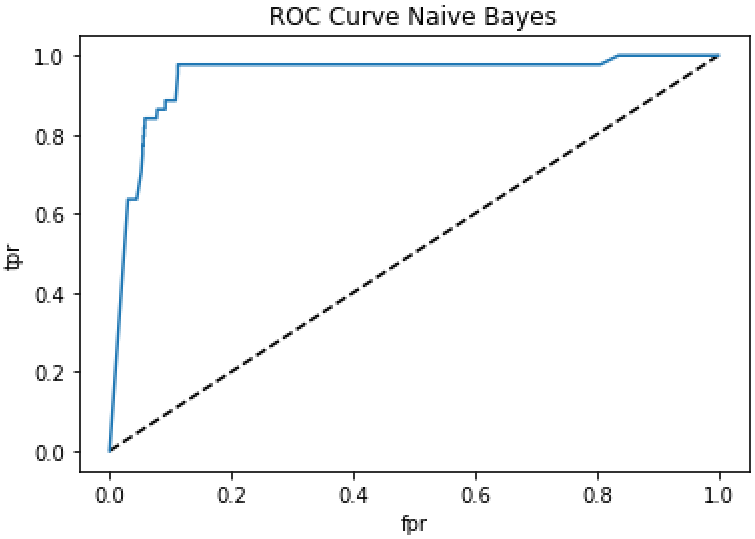
Our results were extremely good, as during our 5-fold cross validation we had an accuracy of over 95%. The diagnostic test used in the study was very accurate in detecting if the patient had cancer or not, with both high precision and recall.

5-fold cross validation results:

```
[0.96137339 0.95851216 0.95702006 0.95845272 0.95845272]
```

The ROC Curve has a great score of 0.9471 meaning the true negatives minimally overlaps the true positives meaning there are low amounts of false negatives and false positives. This is tremendous because the initial goal of the researchers and data collectors wished to maximize sensitivity which is more important in medicine than specificity. It is much more important to minimize false negatives because that would mean telling a cancer patient they don't have cancer and therefore allowing their cancer to progress to a further grade, ultimately exponentially reducing the rate of survival with the progression of the cancer to a further grade. Telling someone they have cancer while they actually don't have, most likely does not kill anyone, but is a massive failure and extremely cruel to make the mistake of clearing a patient who actually has cancer.

roc_auc_score: 0.9470564074479737



Our model has high accuracy based on the diagnostic tests, pap smear (cervical cytology), and mRNA features in identifying patients with cervical cancer.

Next Steps

The researchers and data collectors have our model to classify whether a patient has cancer or not given that we have input of diagnostic tests done, pap smear, and mRNA test. Pap smears are not favorable to women therefore a follow up experiment would be to compare results of mRNA results alone and see if that is sufficient to replace the current pap smear test. The common method as seen from the dataset is to perform a pap smear test with or without a mRNA test. This would save money and minimize discomfort in women who do not want to do a pap smear.

There could always be more data feature and data record to build an even stronger model from the current dataset, though not too much because that could add unnecessary dimensionality.

Healthcare is an extremely interesting and personal domain where we are starting to see an explosion of personalized data collection. Medicine's dream is to have personalized medicine, data mining is the key to this dream.