Julian Romero

Early Cancer detection using machine learning

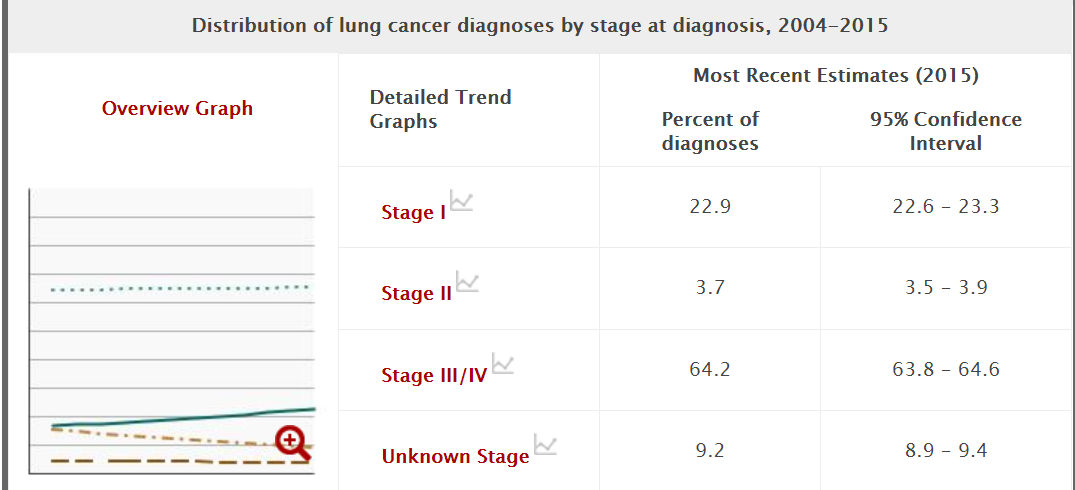
In medicine one of the most important steps a physician will take is diagnosis. For the most part, diagnosis is not very quick when it comes to detecting a certain disease until symptoms appear. Cancer being one of the most prevalent diseases of our time is often always diagnosed late. For example, lung cancer is diagnosed mostly at stage 3 and 4 with a percent diagnosis being at 65 % compared to all lung cancer stages. This can be seen in figure 1 in the appendix. This is according to the national institute of cancer that conducted a study in 2015. The ability to predict cancer before later stages occur is critical in saving patients from death.

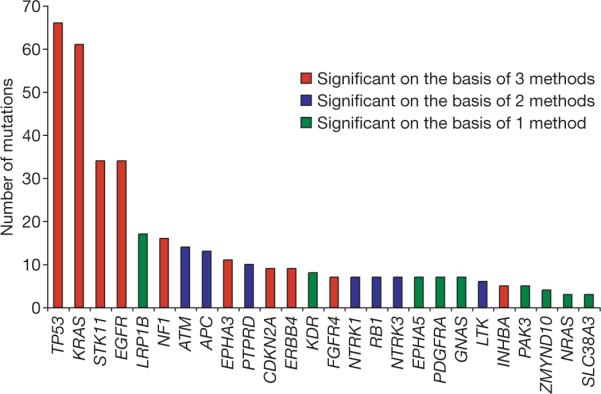
It is known that there are certain gene mutation signatures that are an indicator that cancer is present in certain cells. From this fact I will try to use machine learning to find patterns that are associated with these gene mutation signatures by analyzing different genetic sequences in a dataset. The dataset will consist of people who have a certain cancer and people who do not have a certain cancer. The specific cancer will most likely be lung cancer. From these gene mutations I will try to use a system to predict whether cancer is present in certain cells. Once the system is done training, I will try to test it to detect cancer in another data set like the previous data set. From here when I have validation that my system is acceptable. I will test it in patients who have stage 1 cancer. If successful, this has the potential to be used to diagnose many patients with cancer before stage 3 and 4 cancer.

I will implement this as follows. The inputs will be the gene sequences found from data bases of cancer patients. The databases used will be from <http://portals.broadinstitute.org/cgi-bin/cancer/datasets.cgi>. The outputs will be a Boolean value of whether someone has cancer or not. I will encode the system by trying to parse through different genetic sequences and using set mutagenic signatures as features. These signatures will be based by the signatures found in figure 2. This is the basis of the encoding. More practically I will train the system by finding mutagenic sequences of cells that can find a pattern from sequences that have tested positive for cancer.  I will also parse gene sequences that do not test positive for cancer. If an input is guessed correctly, we will associate a greater weight with the feature used. If guessed incorrectly we will associate less of a weight to that feature. Once trained efficiently we will test it against a new set of test values to see how accurate our trained system is. This is basically using the perceptron algorithm in machine learning.

I hope to get out of this a network that can detect cancer in the gene code of cells before patients reach stage 3 or 4 cancer. If the network is effective, it can help doctors to diagnose early stages of cancer to begin treatment sooner and catch the cancer before it can become more dangerous.

**Appendix**

**Figure 1.** Cancer stage diagnosis trends for lung cancer

**Figure 2**. Mutation signatures in lung cancer cells with the greatest number of mutations

Cancer stats:

[**https://progressreport.cancer.gov/diagnosis/stage**](https://progressreport.cancer.gov/diagnosis/stage)

Cancer staging summary

<https://www.cancer.gov/about-cancer/diagnosis-staging/staging>

Mutation sequence signature resource:

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2694412/>

<https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0007485>

Database resource: