Thesis Project Portfolio

Machine Learning with p53 Gene Somatic Mutations

(Technical Report)

Invisibility of Groups in Relation to Medicine and Toxin Dumping

(STS Research Paper)

An Undergraduate Thesis

Presented to the Faculty of the School of Engineering and Applied Science
University of Virginia • Charlottesville, Virginia

In Fulfillment of the Requirements for the Degree
Bachelor of Science, School of Engineering

Rupal Saini

Spring, 2022

Department of Computer Science

Table of Contents

Sociotechnical Synthesis

Cancer is caused by mutated DNA in cells. The mutations make it so that afflicted cells grow uncontrollably, repeatedly dividing and metastasizing. But why does this DNA mutate? Is there some attribute of the cells themselves which causes the division? Does the environment play a role in determining whether someone will be diagnosed with cancer? The following technical and STS theses address these questions by discussing what variables make someone more susceptible to cancer.

The technical thesis set out to determine which features of a cell are the most important when determining whether a p53 cell is malignant or benign. The method employed to do this is clustering in machine learning, grouping unlabeled samples by measuring similarity. After finding a dataset of p53 cells and formatting it to be applicable to a clustering and random forest algorithm, feature selection was performed. This showed that the top five attributes of a cell when determining which cluster (cancerous or non-cancerous) a cell fell into are mutation_ID, morphology, sub-topography, topography, and topo_code. These features make it seem as if the physicality of the cell is more important than aspects such as codon, location, type of cell, and more when looking at classification.

Further, the features were used to cluster the data. The classifications then allowed for the use of a random forest algorithm to determine how accurate a predictive method would be in determining classification. The accuracy was very high leading to the assumption that machine learning with these specific attributes (the top 5) would be extremely precise when attempting to diagnose a patient. To further verify this, random forest was employed with different datasets to confirm that the calculated accuracy was not a coincidence.

The STS thesis explores how the diagnosis of cancer could be environmental and how certain groups are disproportionately diagnosed and treated. There is a connection between the cause and treatment of cancer and the overlooking of certain groups in society. The main idea of this paper is that perhaps minorities are more likely to be afflicted with cancer. This is because of toxin dumping in low-income areas. Toxin dumping is extremely dangerous and leads to water contamination which mostly affects minorities. When these minorities seek medical help, they are confronted by racially tinged treatment plans through the use of race correction in diagnoses. Again, this mostly affects minorities. This paper explains how race and socioeconomic status may be important when looking at whether a person will be diagnosed with cancer. Society is keeping groups of people in conditions where cancer is a very common illness without proper treatment.

In conclusion, it is seen that physical characteristics of cells and environmental/social issues are causes of cancer. All of these should be considered when determining if someone will be diagnosed with the illness. The main question was thoroughly answered. I am grateful to Dr. Sean Ferguson and Dr. Nada Basit for guiding me through topics that I had little to no knowledge on.