Using Machine Learning to Detect and Categorize the Presence of Cancer

New Mexico

Supercomputing Challenge

Final Report

April 11, 2021

Team 5

Taos High School

Team Members:

Haven Hennelly

Max Meadowcroft

Sawyer Solfest

Teacher:

Tracy Galligan

Executive Summary

Mostavi et al. describe the precise prediction of cancer types as “vital for cancer diagnosis and therapy” (Mostavi et al. 1). Mostavi et al. go on to site Siegel RL, Miller KD and Jemal A who state that “cancer is the second leading cause of death worldwide, an average of one in six deaths is due to cancer” (Siegel RL, Miller KD and Jemal A in Mostavi et al. 2). Standardized methods of cancer detection such as “lab tests, imaging tests (scans) [,] other tests or procedures” (National Cancer Institute), or a “biopsy” (National Cancer Institute) are inefficient. Genetic testing through computerized models has the potential to increase the accuracy and limit the duration of cancer detection. A machine learning model could also identify markers for hereditary subtypes of cancer. Our program uses supervised machine learning to intake data sourced from studies on certain markers and cancer. It is then trained on these markers to determine which are more likely to be correlated with cancer. The program would require several markers for certain types of cancer.

Problem Statement

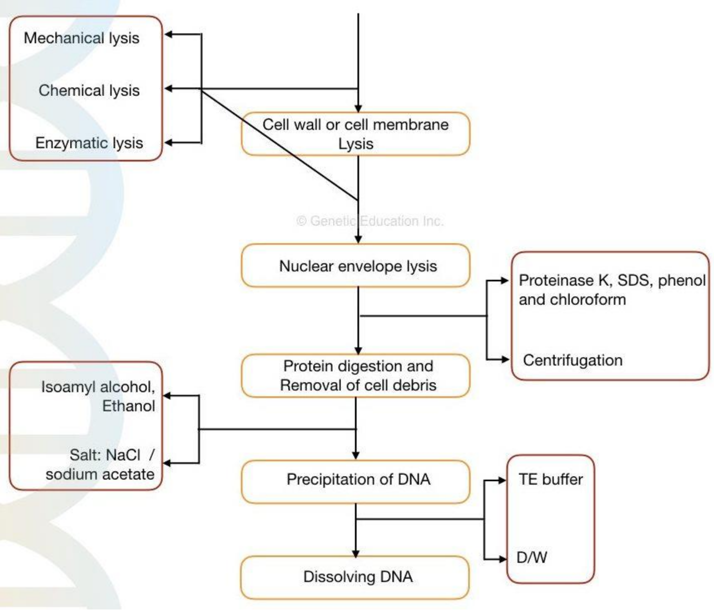
According to Sung et al., Cancer is the “first or second leading cause of death” (Sung et al. 1) for people under “the age of 70” (Sung et al. 1) for “112 out of 183 countries” (Sung et al. 1). 1 in every 3 men and 1 in every 2 females are expected to be diagnosed with cancer at some point in their life. Current cancer detection is generally inaccurate and requires significant clinical labor.

Methods:

Artificial intelligence is a growing field within Computer Science that according to IBM Cloud Education “enables computers and machines to mimic the perception, learning, problem-solving, and decision-making capabilities of the human mind” (IBM Cloud Education). According to Bell, “machine learning is a branch of artificial intelligence” (Bell 3) that learns from previous results to improve itself. The branch of machine learning we used is supervised learning which “refers to working with a set of labeled training data” (Bell 3) to train the model to classify something. We used portions of the human genome as our training data and used machine learning through TensorFlow to train a model to classify between people with cancer, people with a threat of cancer, and people without cancer.

The DNA samples are collected through fine needle aspiration, the same technique used in Biopsys. Extracting samples relies on the use of “ultrasound guidance to ensure accurate placement of the needle” (American Thyroid Association). Samples must originate from the affected cells to detect the presence of DNA markers associated with cancer. According to Mostavi et al. many methods of digital cancer detection “ignore the existence of tissue of origin within each cancer type. Without removing the influence of normal tissues during cancer classification, the implementation of a data interpretation scheme will unlikely to differentiate tissue-specific genes or cancer-type-specific genes” (Mostavi et al. 2). Mostavi et al. Conclude, “it is impossible to perform functional analysis or select biomarkers for cancer detection from such models” (Mostavi et al 2)

DNA can only be studied after an extensive chemical and physical process. To analyze the DNA, one must “break cell wall/ cell membrane and nuclear envelope as well” (Gene Education). The outer cell membrane can be dissolved using chemical disruption, this is accomplished by exposing the sample to ”phenol-chloroform” (Gene Education). Penetrating the nuclear membrane is completed during exposure to the phenol-chloroform. Isoamyl alcohol is used for protein digestion. The process also requires a ”lysis buffer” (Gene Education). The lysis buffer is used to maintain the integrity of the sample. The information is then processed via a sequencer.



Model Validation

Validating this model in the real world is an impossible prospect for our team, as data on individual patients suffering from cancer is both extremely confidential and not available in an aggregated form to high school students. This being the case, this team must rely on the information we gathered being accurate and use the standard technique of using different sets of data to first train the model and then test it to eliminate biases in the model.

Results

At the time of writing, our model has not produced any usable results. Due to a number of setbacks, including the fact that this subject matter is well beyond our level of education and available resources for raw data, and a simple lack of team cohesiveness and focus, this project has been confused and changed direction several times. At this moment, we are still assembling both the model and training data for the model. We expect to have at least a trained model, and be able to share our final 1results and conclusions at the time of the Expo.

Conclusions

Training data compiled from academic studies will allow us to associate data with the presence of cancer.

Works Cited

Bell, J. (2020). *Machine learning: Hands-on for developers and technical professionals* (Second ed.). INpolis, IN: Wiley.

Bini SA, Artificial Intelligence, Machine Learning, Deep Learning, and Cognitive Computing:

what do these terms mean and how will they impact health care?, The Journal Of Arthroplasty (2018), doi: 10.1016/j.arth.2018.02.067.

"Different Types of DNA Extraction Methods." *Genetic Education*, 10 Jan. 2020, geneticeducation.co.in/different-types-of-dna-extraction-methods/

"Fine Needle Aspiration Biopsy of Thyroid Nodules." *American Thyroid Association*, www.thyroid.org/fna-thyroid-nodules/.

"How Cancer Is Diagnosed." *National Cancer Institute*, www.cancer.gov/about-cancer/diagnosis-staging/diagnosis.

IBM Cloud Education. "What Is Artificial Intelligence (AI)?" *IBM*, www.ibm.com/cloud/learn/what-is-artificial-intelligence.

Mostavi, Milad, et al. "Convolutional Neural Network Models for Cancer TYPE Prediction Based on Gene Expression." *BMC Medical Genomics*, vol. 13, no. S5, 2020, doi:10.1186/s12920-020-0677-2.

Sung, Hyuna, et al. "Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries." *CA: A Cancer Journal for Clinicians*, 2021, doi:10.3322/caac.21660.