Diagnosing Brain Cancers with Gene Expression Data using a Novel Neural Network Method

Rianna Santra

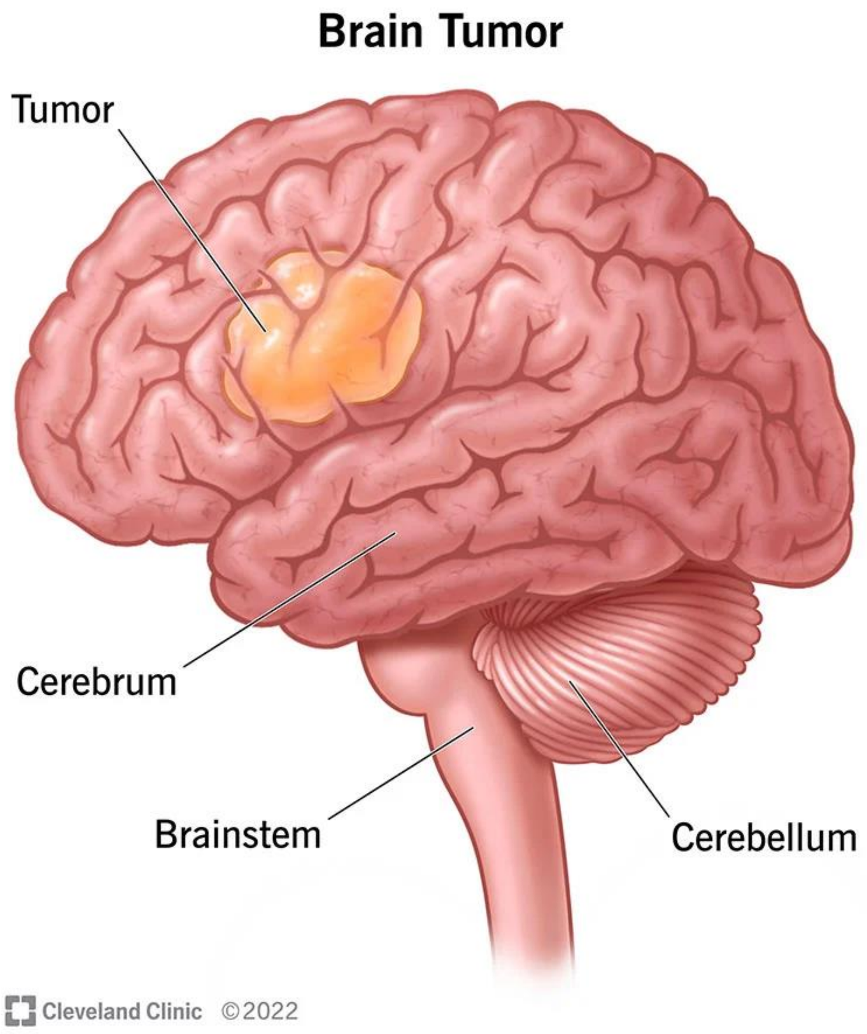
Mentor: Gil Alterovitz

MIT PRIMES Fall Conference

October 15-16, 2022

Context | Brain Cancer

- Brain cancer is one of the deadliest cancers in the US
- Treating brain cancer is also very difficult, due to the tumors being near sensitive areas of the brain and spinal cord
- How can we diagnose brain cancer and administer drugs to patients effectively?

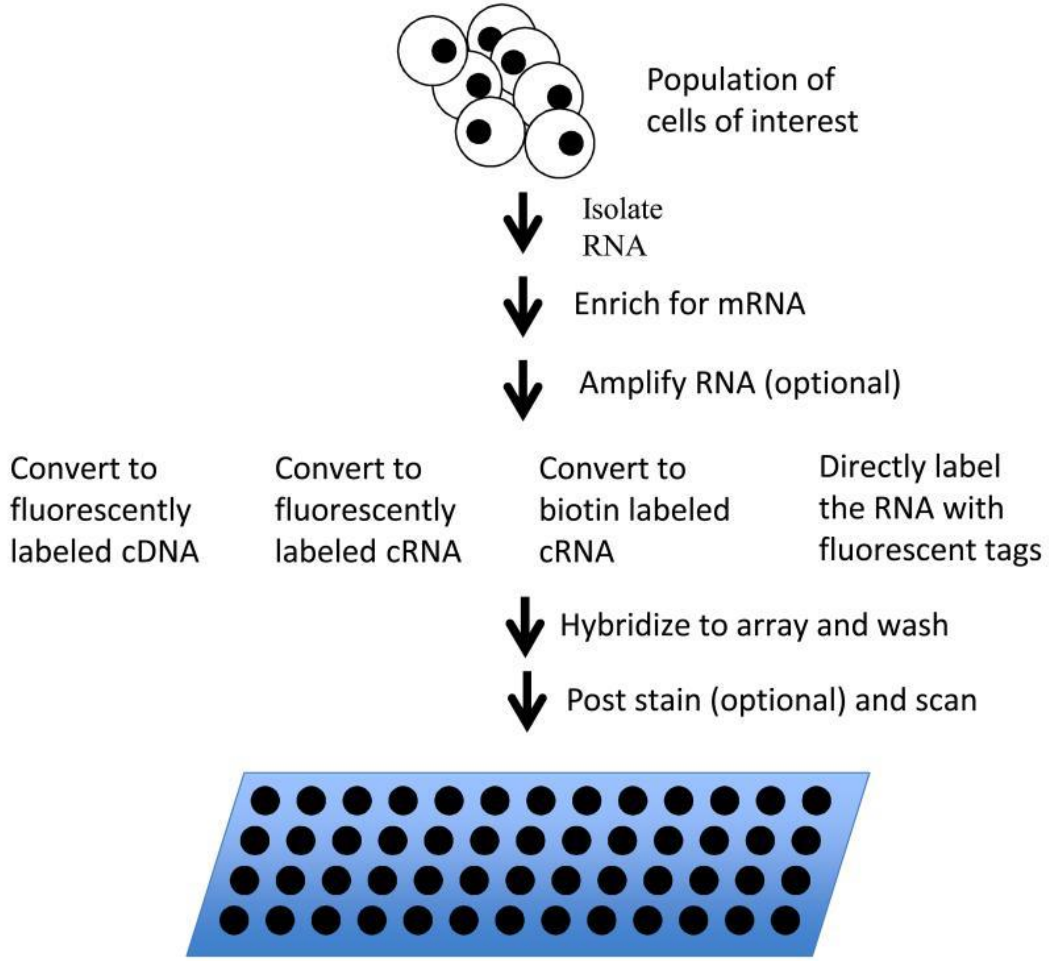


Brain Tumor

Source: Cleveland Clinic

Context | Microarray Data

- One of the latest ways to measure gene expression levels is through DNA microarrays
- These can often show a difference of gene expression levels between cancer cells and healthy cells



Population of cells of interest

Isolate RNA

Enrich for mRNA

Amplify RNA (optional)

Convert to fluorescently labeled cDNA

Convert to fluorescently labeled cRNA

Convert to biotin labeled cRNA

Directly label the RNA with fluorescent tags

Hybridize to array and wash

Post stain (optional) and scan

Process of how microarrays work

Image source: (Bumgarner, 2013)

Research Goals

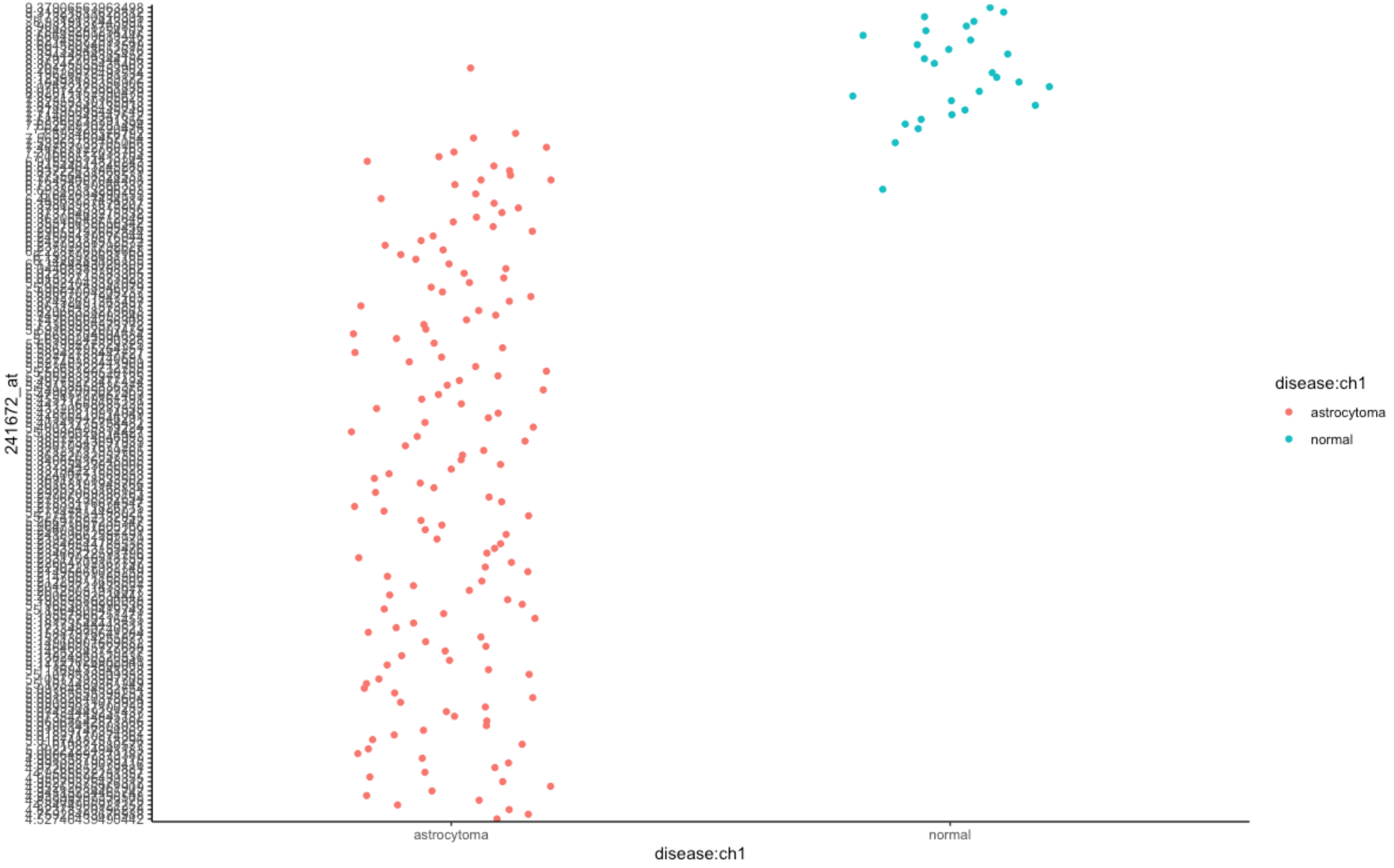
- Identify the specific subtype of brain cancer a patient has, to provide enough time for treatment
- Identify genes that have a high variance between cancer cells and healthy cells

Methodology | Data Collection

- First, microarray data was collected from the Repository for Molecular Brain Neoplasia Data (REMBRANDT)[1]
- This data was normalized using quantile normalization and RMA normalization
- The most expressed genes were selected to move on to the neural network phase

[1] (Gusev et al., 2018)

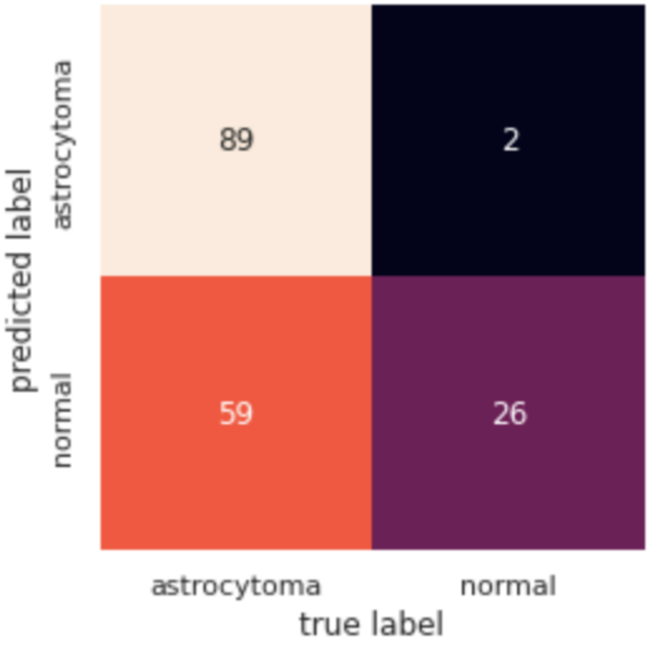
Methodology | Data Collection

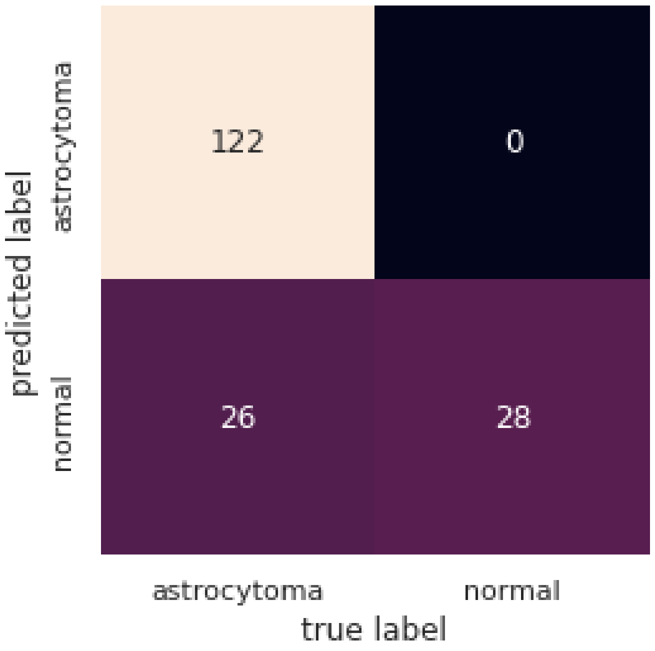


Scatter plot of gene expression between astrocytoma and normal cells for the gene SERTM1

Methodology | Neural Network

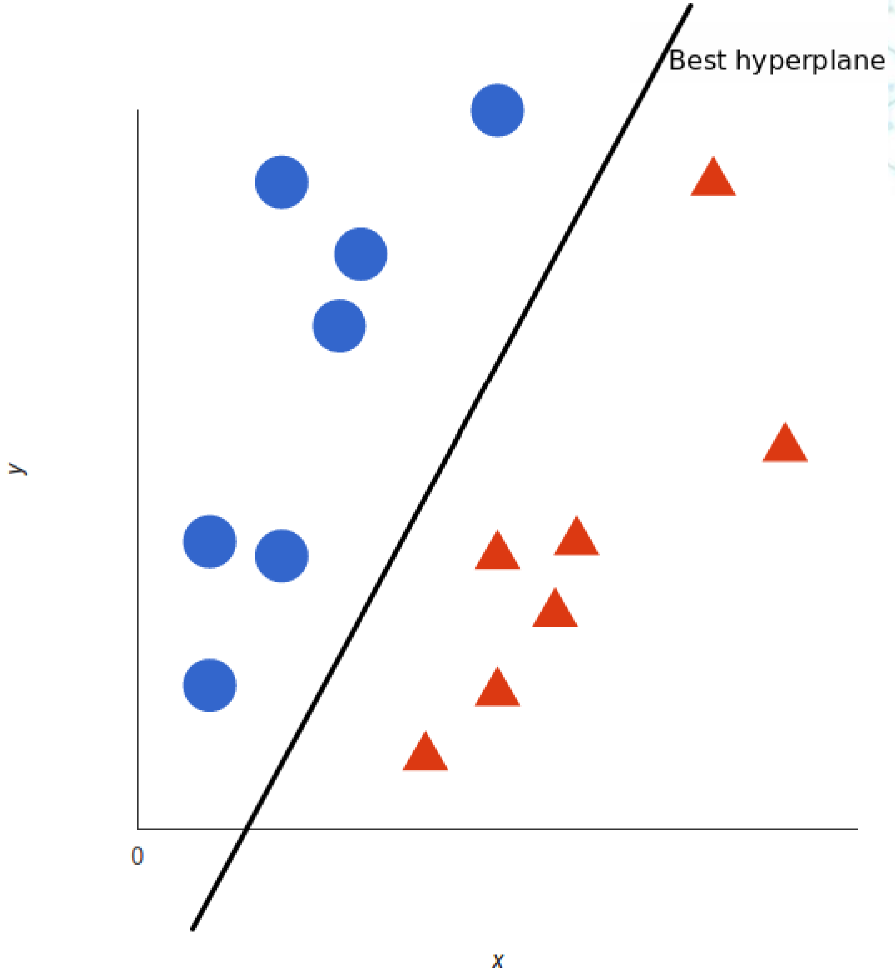
- The most expressed genes were selected from the dataset to be used in a support vector machine (SVM) neural network
- From these genes, an accuracy was formed





Context | Support Vector Machines (SVM)

- Support vector machines work best with binary classification (2 labels)
- They make a **hyperplane** between the two clusters
- It’s much easier to classify a new sample by seeing where it falls against the plane



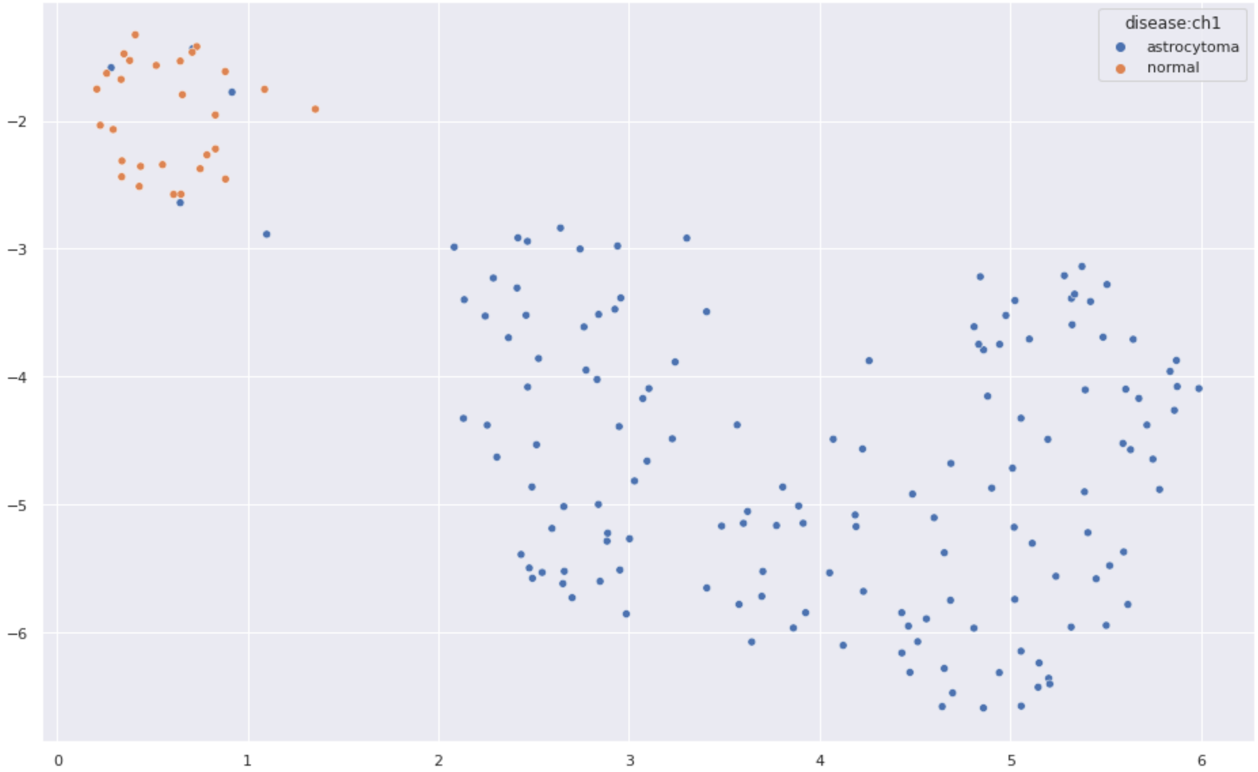
Support vector machines

Source: MonkeyLearn.com

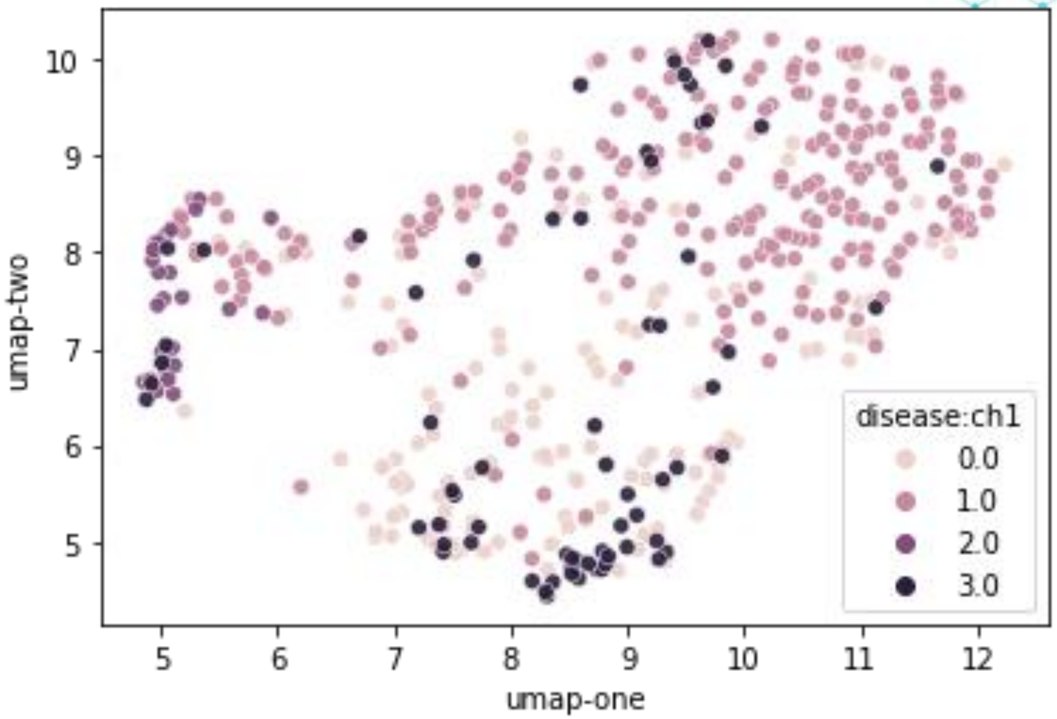
Results | Why Neural Network Method

- The data was filtered into three subsets: astrocytoma vs. normal, oligodendroglioma vs. normal, and glioblastoma vs. normal.
- These data are put into their own neural network using the One vs. All method.
- This accuracy is much higher rather than having a single neural network

Results | Why Neural Network Method



UMAP plot for astrocytoma vs. normal

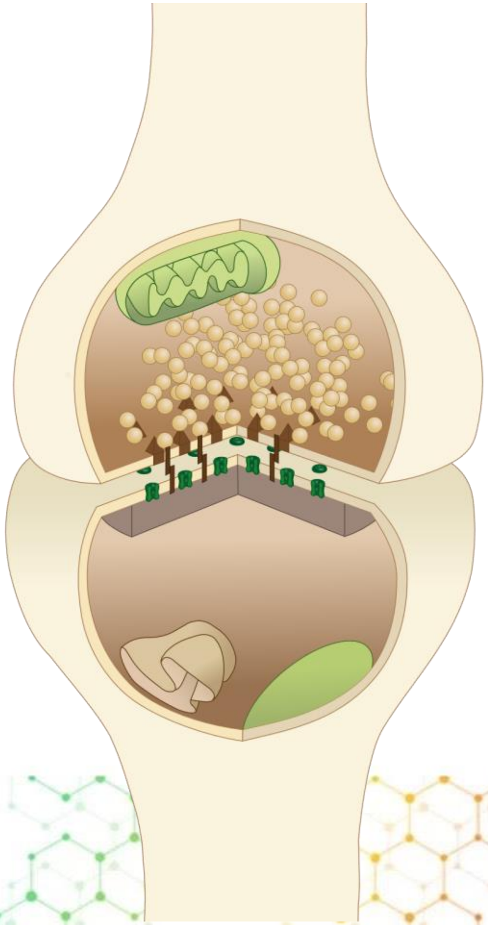


UMAP plot for all subtypes

Conclusions | Neural Networks

- The separate neural networks each had a much higher accuracy than the single multiclass neural network
- To classify a single sample, the sample is put into the 3 neural networks of the probability it is either normal or that specific subtype
- The highest probability will be used and returned

Conclusions | Biomarkers



- Biomarkers were also found from the genes that influenced the neural network the most
- These genes made proteins that made up the structural part of a presynaptic active site
- These proteins may not have been adequately made, leading to a deformation of neurons

Future Extensions

- This can be extended to all cancer subtypes
- Housekeeping genes (genes that do not have a high variance between subtypes) can also be identified for other experimental types

Acknowledgements

I want thank **Prof. Gil Alteroviz** for his mentorship throughout the year, **Ning Xie** for helping connect us and my school STEM teacher **Dr. Kevin Crowthers** for his support on this project. I want to thank **Dr. Slava Gerovitch, Prof. Srini Devadas,** and the **MIT PRIMES** program for providing me this opportunity.

Any Questions?