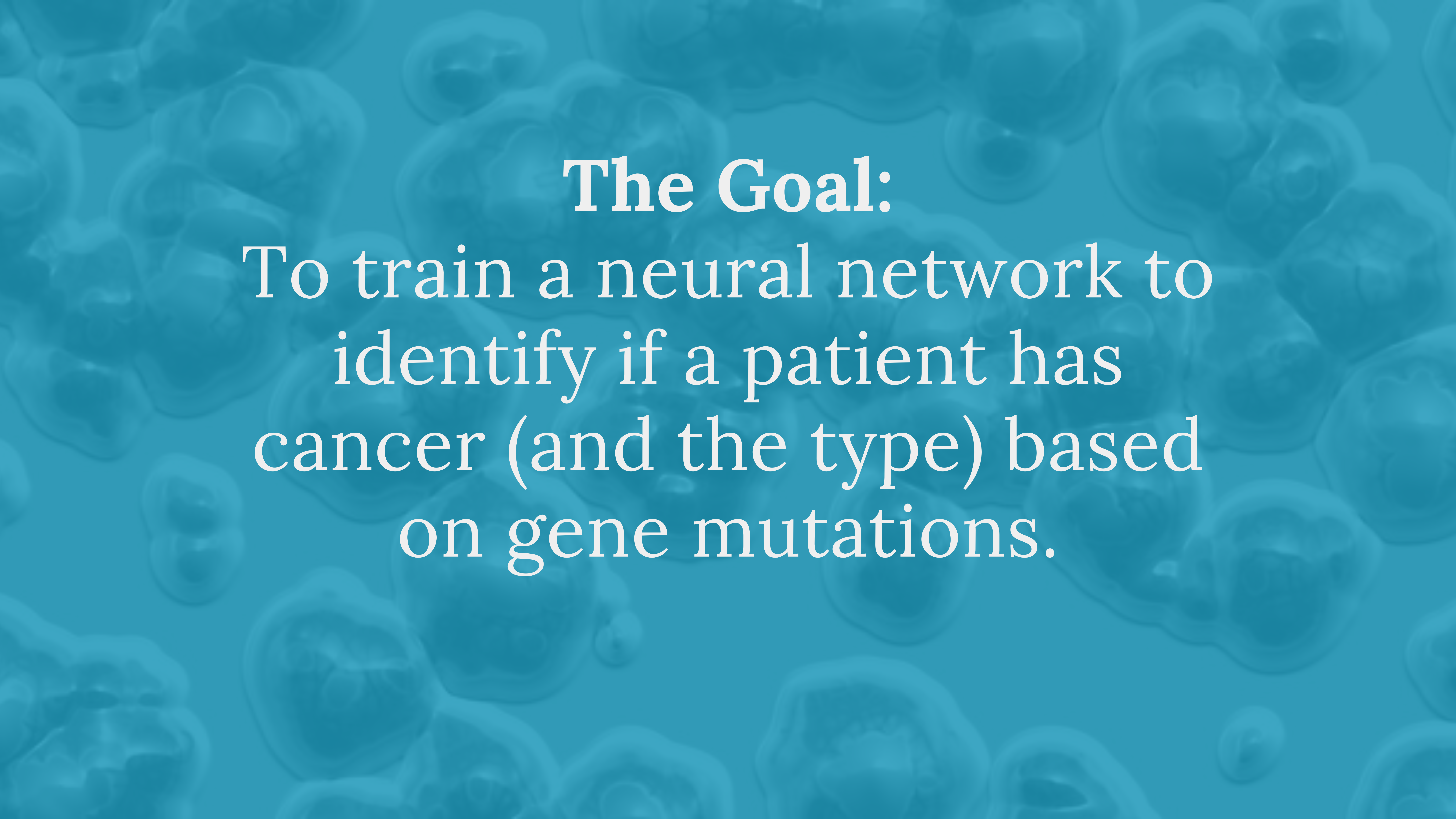
BIOINFORMATICS FINAL PROJECT

Identifying Cancer using Deep Learning

Nima, Kaitlyn, Katrina, Troy

Background

- Cancer is the 2nd leading cause of death in US
- Research shows relationship between mutations in the genome and cancer development
- A lot still unknown due to levels of gene expression, copies of genes across the genome, what leads to variety of mutations, etc.
- Being able to identify cancer through genome deep learning (GDL) will assist with early detection.
- “Identification of 12 Cancer Types through Genome Deep Learning” by Sun et al. in 2019 explores using individual and ensemble DNNs for cancer detection.



The Goal:

To train a neural network to identify if a patient has cancer (and the type) based on gene mutations.

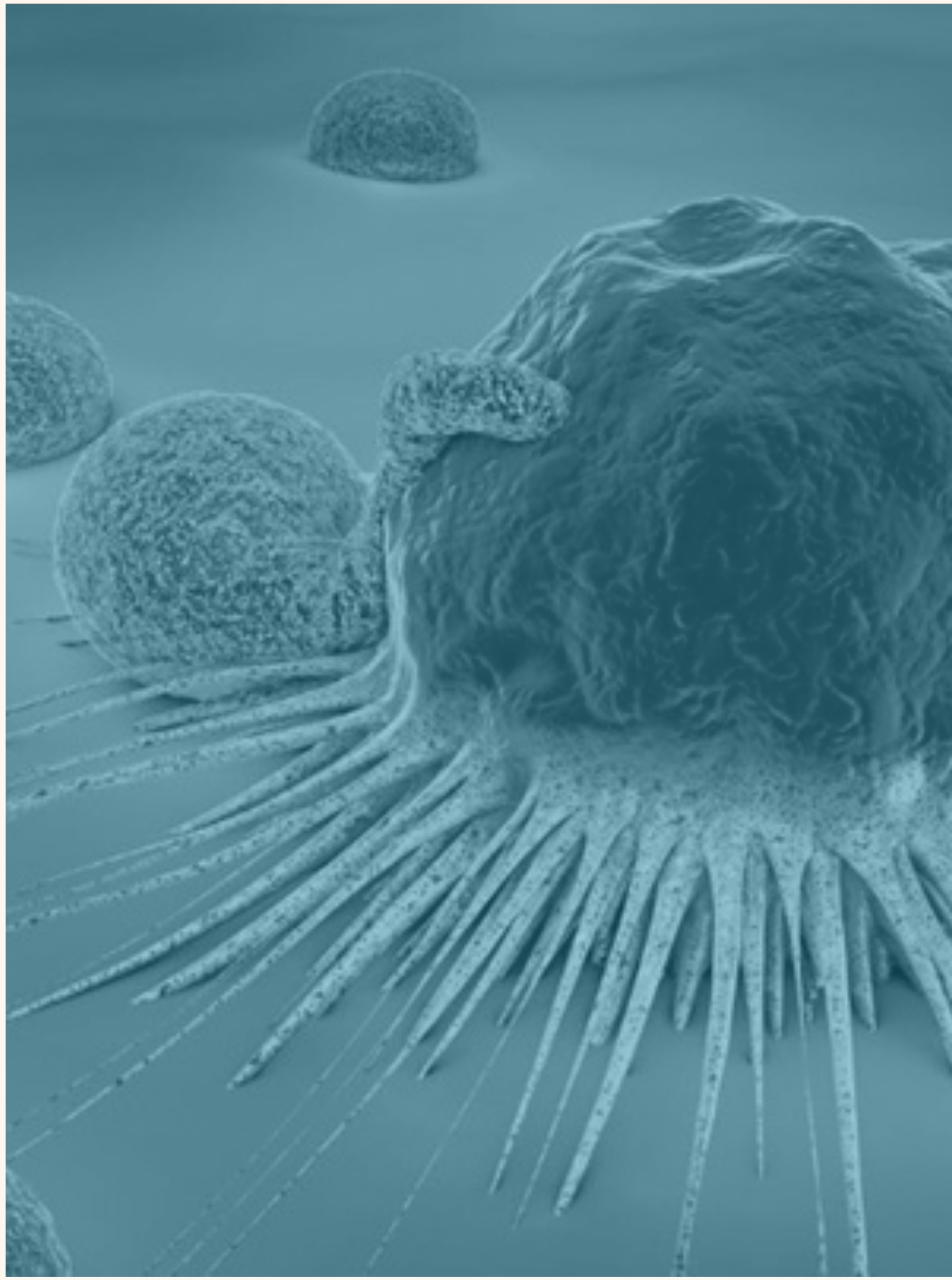
The Data

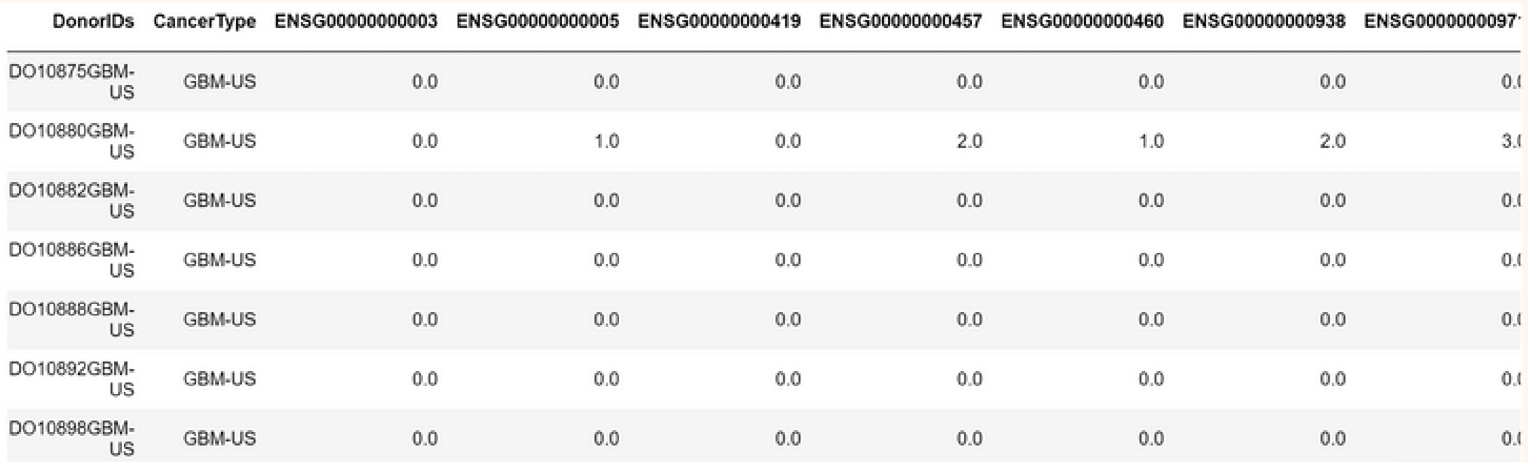
CANCER DATA

The data for about 6,000 donors with 12 different cancer types was collected from the International Cancer Genome Consortium. Data included cancer type, mutations, genes affected, etc.

HEALTHY DATA

Data for about 2,500 "healthy" donors was collected from the 1,000 genomes site.





DonorIDs	CancerType	ENSG00000000003	ENSG00000000005	ENSG00000000419	ENSG00000000457	ENSG00000000460	ENSG00000000938	ENSG0000000097
DO10875GBM-US	GBM-US	0.0	0.0	0.0	0.0	0.0	0.0	0.
DO10880GBM-US	GBM-US	0.0	1.0	0.0	2.0	1.0	2.0	3.
DO10882GBM-US	GBM-US	0.0	0.0	0.0	0.0	0.0	0.0	0.
DO10886GBM-US	GBM-US	0.0	0.0	0.0	0.0	0.0	0.0	0.
DO10888GBM-US	GBM-US	0.0	0.0	0.0	0.0	0.0	0.0	0.
DO10892GBM-US	GBM-US	0.0	0.0	0.0	0.0	0.0	0.0	0.
DO10898GBM-US	GBM-US	0.0	0.0	0.0	0.0	0.0	0.0	0.

The Process

Data Encoding

- Data filtered down to the mutations and genes affected per donor.
- Encoding rows include cancer type and count of mutations per gene for each donor.

  (paper only did binary encodings)
- This results in about 30K features.

  (paper only used top 10K for each cancer)

Initial Model Fitting

1. The UMAP algorithm was first used to reduce the dimensionality of the inputs from 30k down to 600.
2. A dense (fully-connected) neural network was then trained to perform multi-class classification on the UMAP transformed data
3. Model was only trained and tested on cancerous examples (we had yet to fetch healthy data).

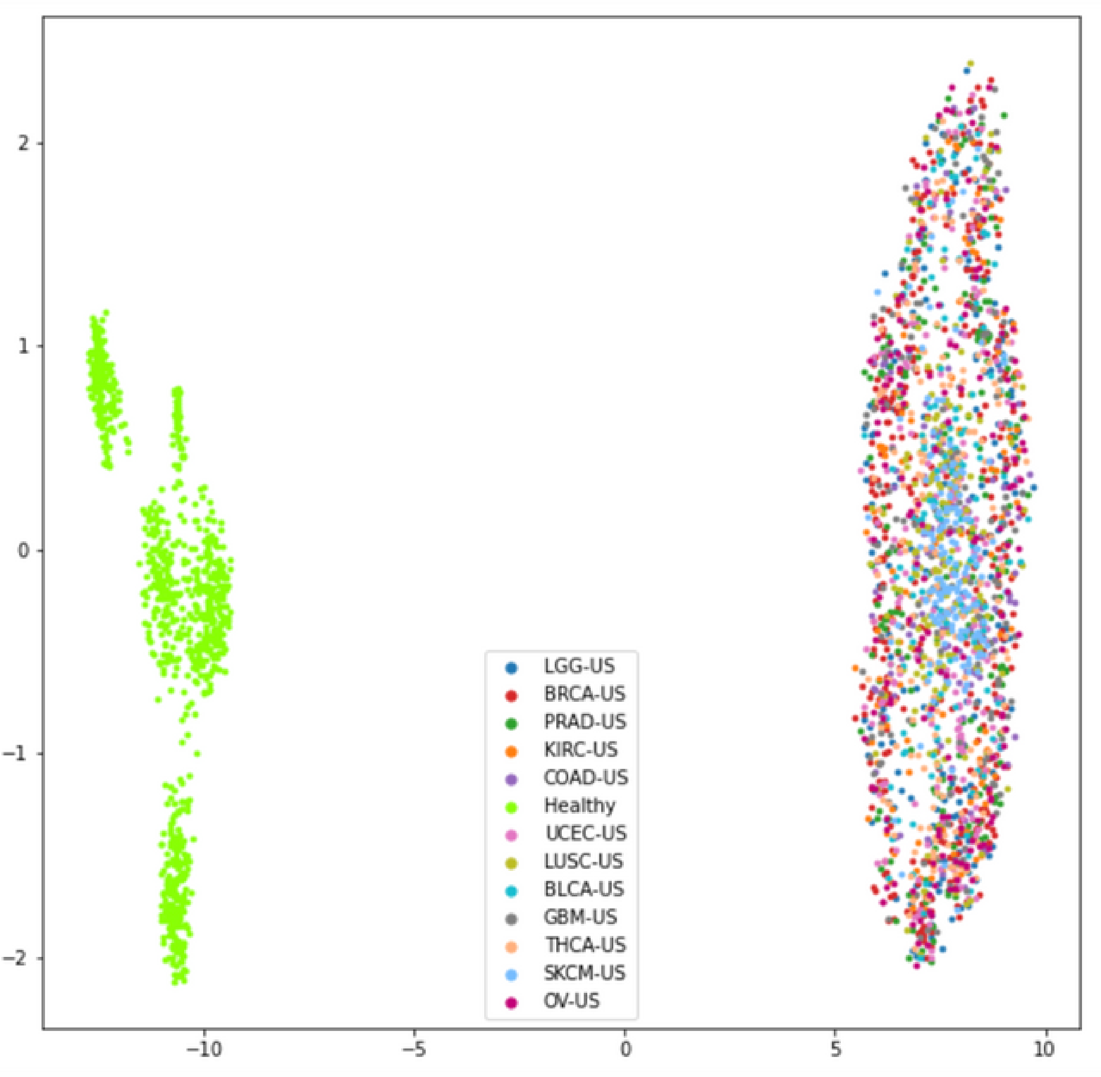
Training Method 2

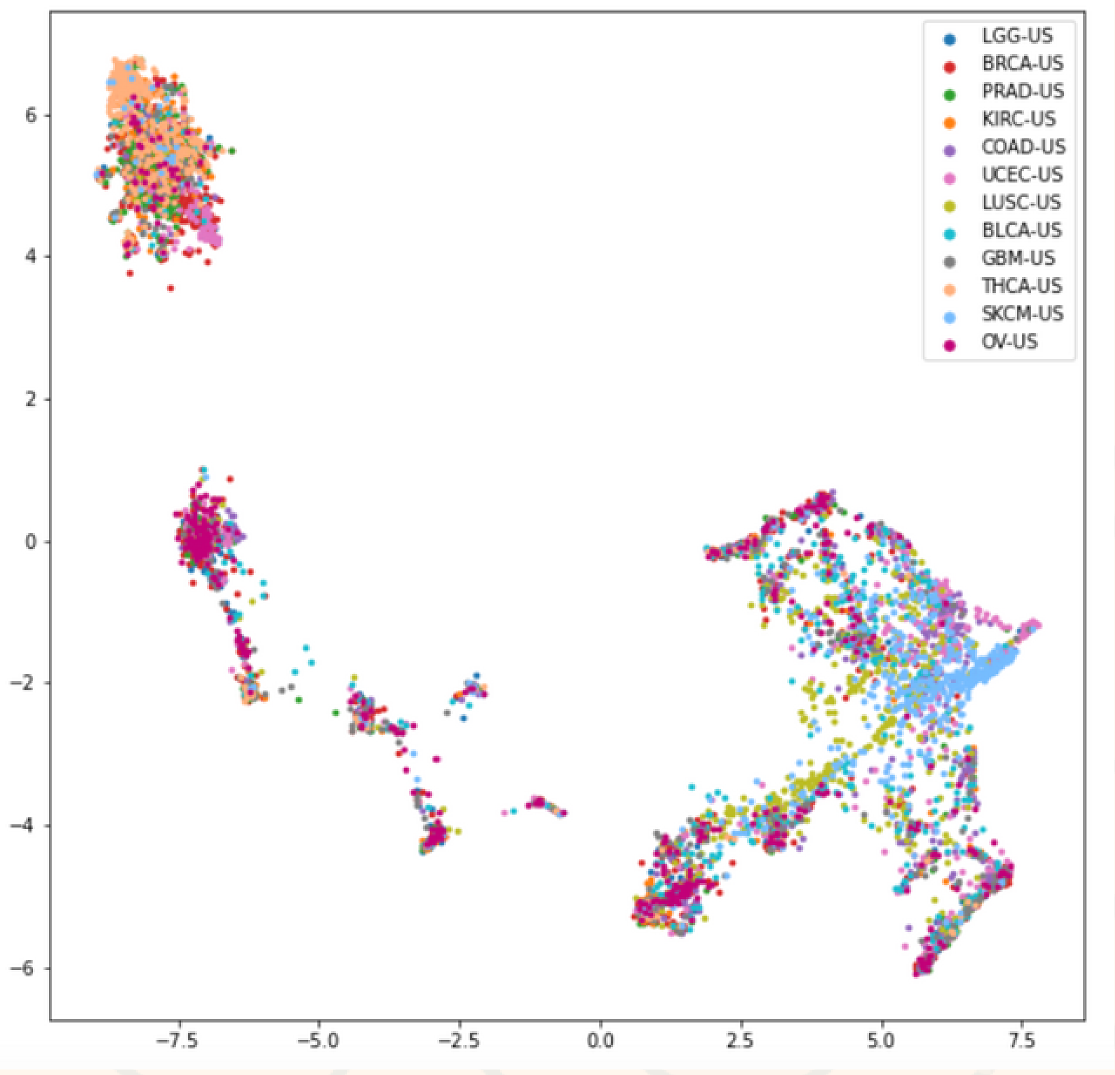
1. Attempting to train end-to-end, unsupervised training was first used to learn a latent representation of the raw inputs.
2. The encoder is then added as the front of a classifier and trained end-to-end.
3. This was done on both cancerous and healthy data.

Results

METHOD 1

- After training for 100 epochs, the model achieved a validation accuracy of 87%
- Accuracy on hold-out/test set of 85%
- Performed well, but was limited
  - Required initial transformation of data with UMAP
  - Was not trained on healthy examples

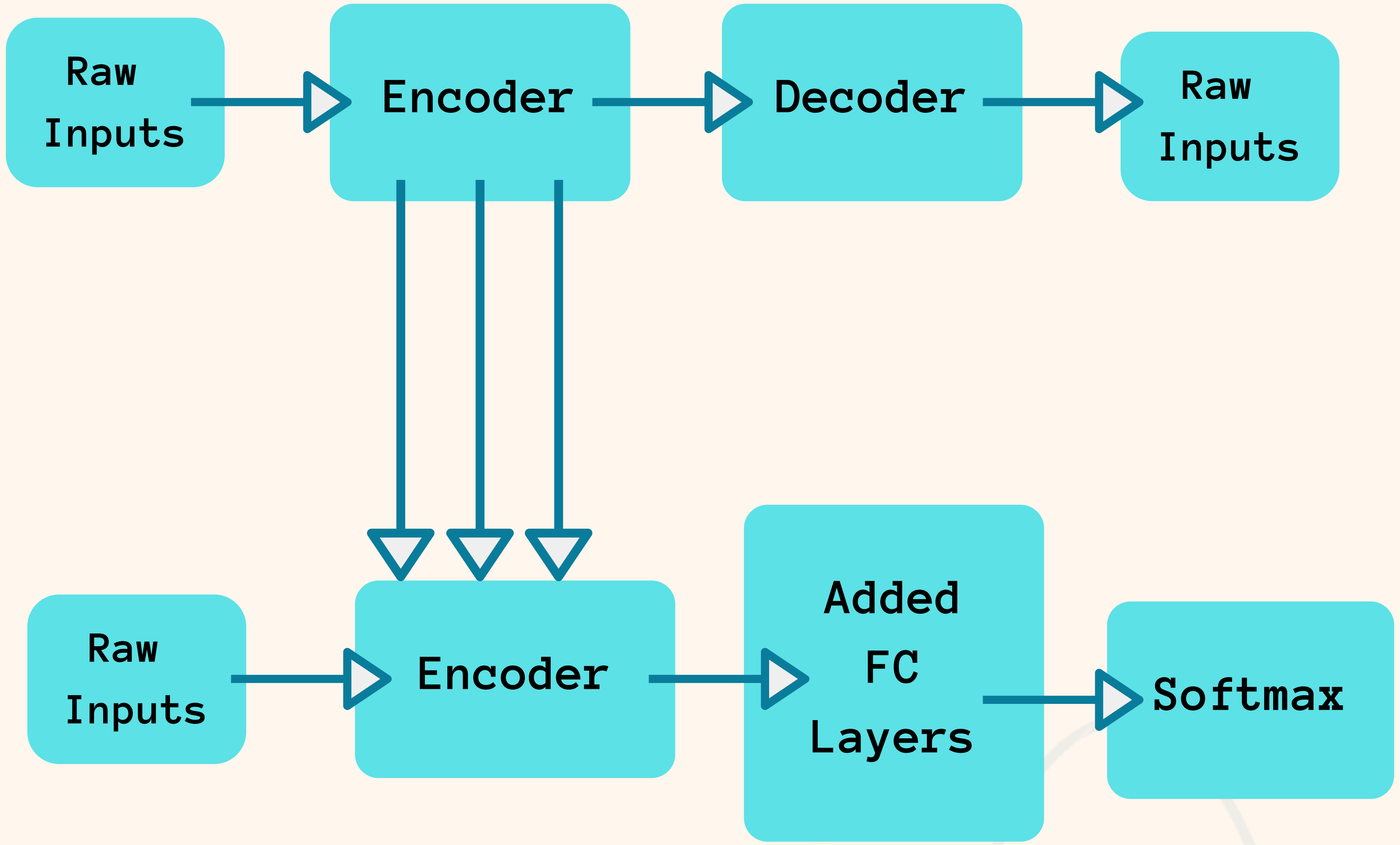




Results

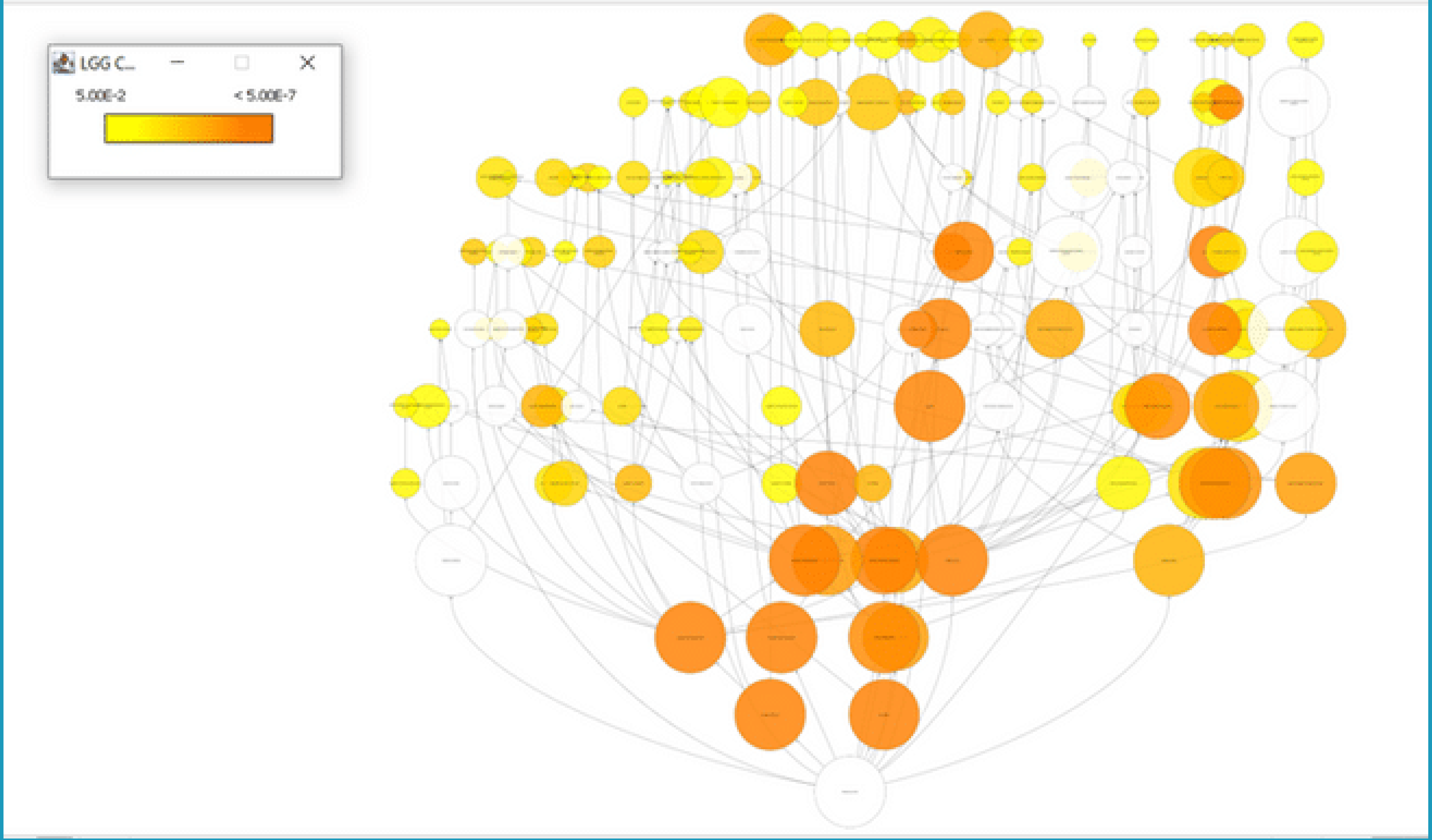
METHOD 2

- The autoencoder was trained for 50 epochs, the model achieved a validation loss (MSE) of 0.01.
- The classifier was trained for 100 epochs.
  - Validation and Test AUC of 0.98
  - Validation and Test Accuracy of 71%

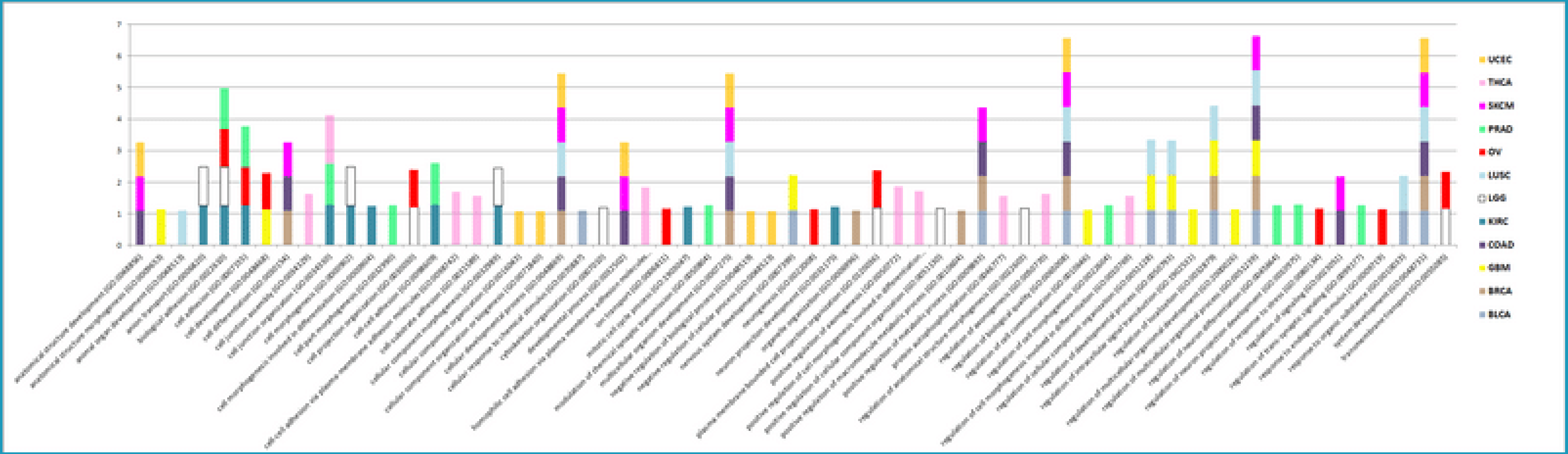


VISUALIZATIONS

Low grade gliomas
255 pathways represented



VISUALIZATIONS



Widest range over cancer types include broad groups like regulation of biological quality, regulation of multicellular organism processes, and system development.

Cancers with the most isolated pathways include thyroid carcinoma (THCA- light pink) and prostate adenocarcinoma (PRAD- light green)

Next Steps

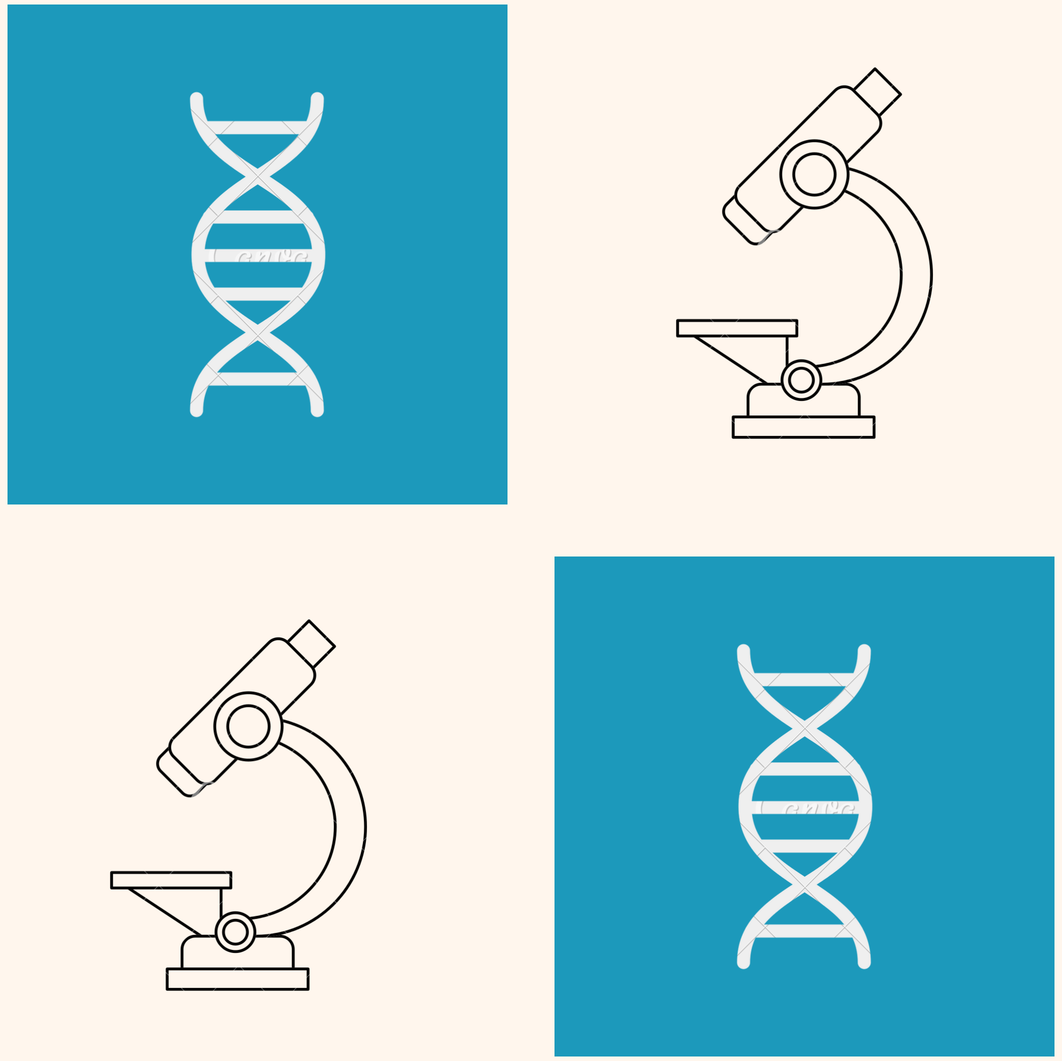
-Incorporate Tissue Specific Data

-Interpretability

-Binary and Multi-class Ensemble Approach

Tools Used

- Jupyter Notebooks
- pyensembl
- scikit-allel
- Cytoscape
    - Reactome
    - BiNGO
- Gene Ontology; Panther Classification
- Ensembl
- Keras
- SHAP
- Javascript
- Puppeteer.js



Thanks for listening! Any questions???

The Team

UNIVERSITY OF HAWAI'I AT MĀNOA

Nima Azbijari

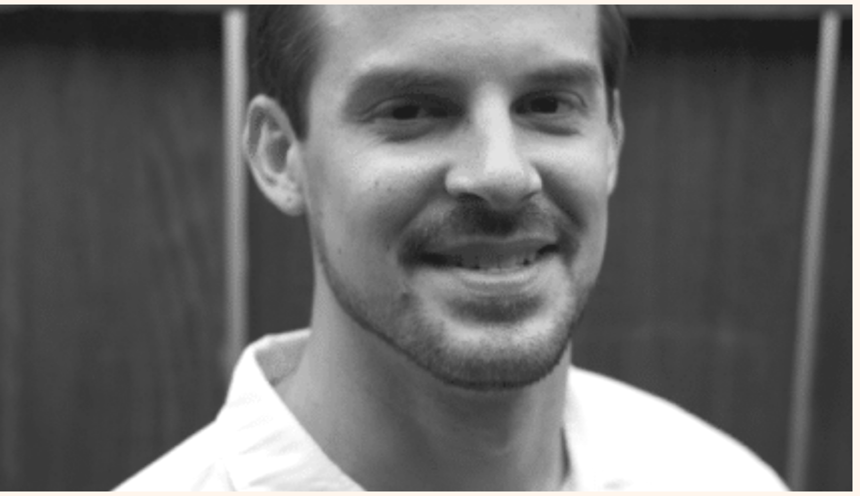
CS: Interpretability in Machine Learning

Kaitlyn Jacobs

MBIO: Data collection, interpretation, visualization

Katrina Turner

CS: Data Cleaning & Encoding



Billy Troy Wooton

CS: Machine Learning & Data Fetching

Resources

Papers

Sun, Yingshuai, et al. “Identification of 12 Cancer Types through Genome Deep Learning.” Scientific Reports, vol. 9, 2019, doi:10.1101/528216.

Angermueller, Christof, et al. “Deep Learning for Computational Biology.” Molecular Systems Biology, vol. 12, no. 7, 2016, p. 878., doi:10.15252/msb.20156651.

Yuan, Yuchen, et al. “DeepGene: an Advanced Cancer Type Classifier Based on Deep Learning and Somatic Point Mutations.” BMC Bioinformatics, vol. 17, no. S17, 2016,

Auton, A., Abecasis, G., Altshuler, D. et al. A global reference for human genetic variation. Nature 526, 68–74 (2015). https://doi.org/10.1038/nature15393

Websites

“CDC Cancer Data and Statistics.” https://www.cdc.gov/cancer/dcpc/data/index.htm

“Clinical Interpretation of Variants in Cancer.” CIViC, civicdb.org/home.Cosmic.

“COSMIC - Catalogue of Somatic Mutations in Cancer.” COSMIC | Catalogue of Somatic Mutations in Cancer, 5 Sept. 2019, cancer.sanger.ac.uk/cosmic.

“ICGC.” ICGC, https://dcc.icgc.org/.

“Variants.” DoCM, www.docm.info/.