Capstone Project Report:

Cancer Classification from Gene Expression RNA-Seq Data

Zhanyang Zhu, PhD

12/06/2022

Zhanyang.Zhu@Gmail.com

Goals:

1. Train multi-class classification models to determine a cancer type given gene expression data of a patient
2. Analyze the importance genes (features) that can distinguish the cancers
3. Perform feature dimension reduction to determine a minimal set of features (i.e., genes) that can be for testing

In this exercise, the cancer types are limited to BRCA, KIRC, COAD, LUAD and PRAD, SKCM, THCA, LGG. More cancer types can be included.

TCGA Study Abbreviations https://gdc.cancer.gov/resources-tcga-users/tcga-code-tables/tcga-study-abbreviations

Data Selection:

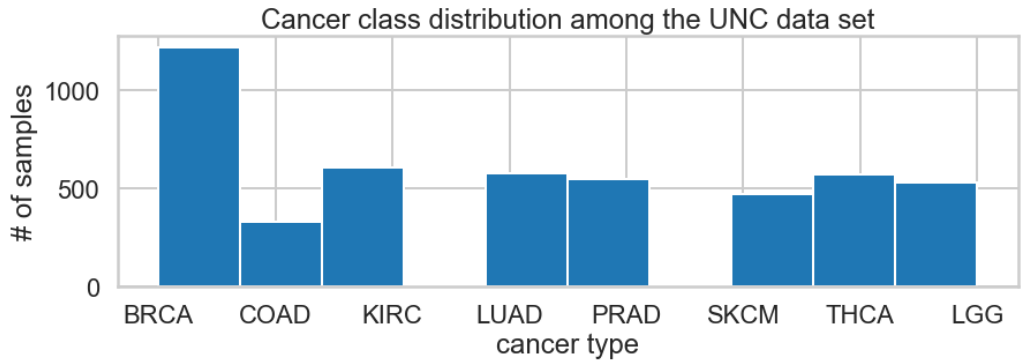
For this UCSD Data Bootcamp projects, I downloaded eight data sets from https://www.synapse.org/#!Synapse:syn2812961. All eight data sets are Illumina HiSeq RNASeq V2 data collected by unc.edu (data set name and cancer type)

1. unc.edu_BRCA_IlluminaHiSeq_RNASeqV2.geneExp.tsv - BRCA: Breast invasive carcinoma
2. unc.edu_COAD_IlluminaHiSeq_RNASeqV2.geneExp.tsv - COAD: Colon adenocarcinoma

3. unc.edu_KIRC_IlluminaHiSeq_RNASeqV2.geneExp.tsv - KIRC: Kidney renal clear cell carcinoma
4. unc.edu_LUAD_IlluminaHiSeq_RNASeqV2.geneExp.tsv - LUAD: Lung adenocarcinoma
5. unc.edu_PRAD_IlluminaHiSeq_RNASeqV2.geneExp.tsv - PRAD: Prostate adenocarcinoma
6. unc.edu_SKCM_IlluminaHiSeq_RNASeqV2.geneExp.tsv - SKCM: Skin Cutaneous Melanoma
7. unc.edu_THCA_IlluminaHiSeq_RNASeqV2.geneExp.tsv - THCA: Thyroid carcinoma
8. unc.edu_LGG_IlluminaHiSeq_RNASeqV2.geneExp.tsv - LGG: Brain Lower Grade Glioma

Samples (instances) are stored column-wise. Variables (attributes in rows) of each sample are RNA-Seq gene expression levels measured by illumina HiSeqV2 platform.

- 4859 samples with 20530 features



- For more data summary see https://github.com/ZhanyangZhuSD/UCSDMLCapstone/blob/main/GeneExpressionCancerRNA-Seq.ipynb

Classification Model Selection

For explorative analysis with 40% of the total data, the following classification models are trained, and the best model is selected.

Without hyperparameters optimization, the order of accuracy is

1. XGBoost (0.9979423868312757)
2. Support Vector Machine (0.9958847736625515)

3. Stochastic Gradient Descent (0.9938271604938271),
4. Random Forest (0.9917695473251029),
5. Naïve Bayes (0.9917695473251029),
6. Decision Tree (0.9835390946502057),
7. K-Nearest Neighbours (0.9506172839506173)

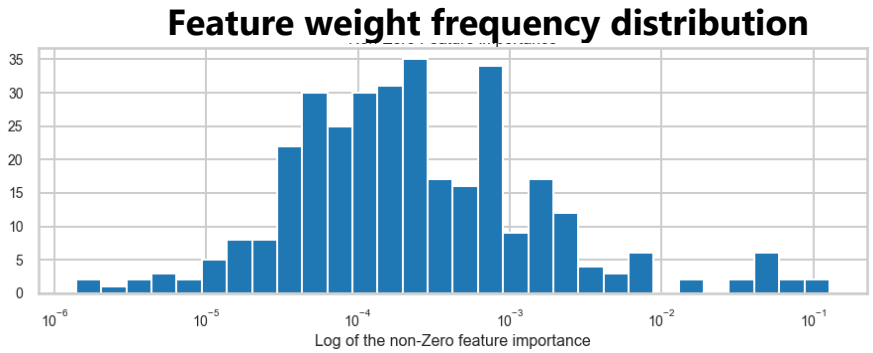
For details, see https://github.com/ZhanyangZhuSD/UCSDMLCapstone/blob/main/GeneExpressionCancerRNA-Seq.ipynb

Results

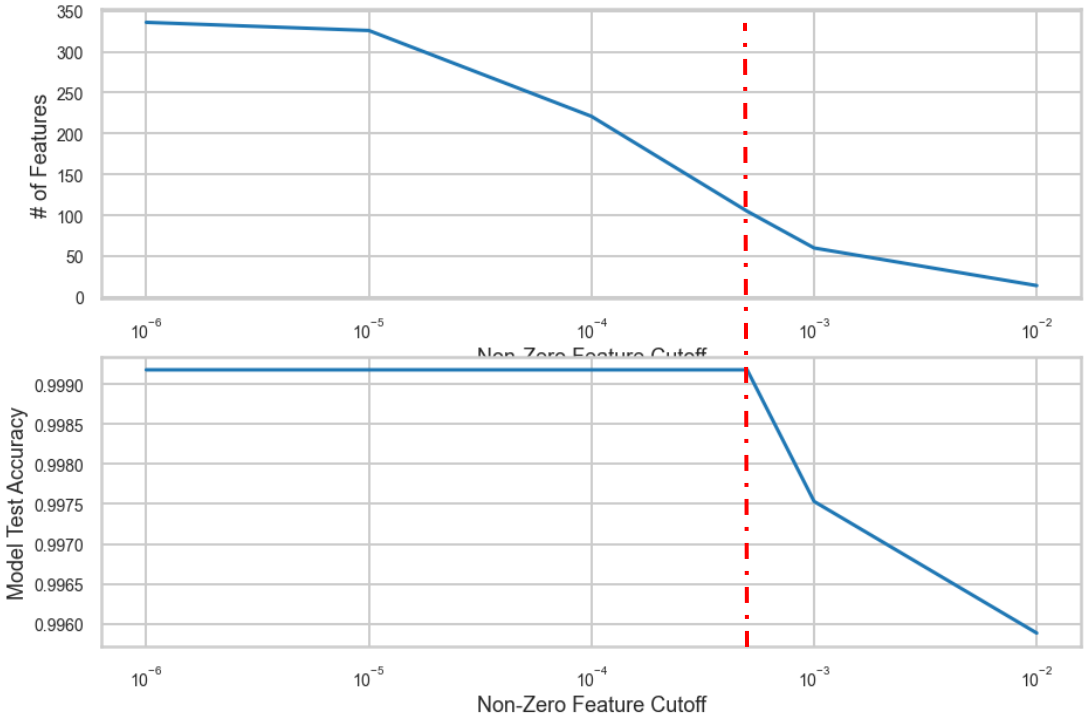
- Best classification model is XGBoost based on the explorative analysis. (https://github.com/ZhanyangZhuSD/UCSDMLCapstone/blob/main/GeneExpressionCancerRNA-Seq.ipynb)
- The accuracy is 0.99918 when XGBoost model is applied to the full data set with all the features are used

	precision	recall	f1-score	support
BRCA	1.00	1.00	1.00	326
COAD	1.00	1.00	1.00	61
KIRC	0.99	1.00	1.00	157
LGG	1.00	1.00	1.00	133
LUAD	1.00	1.00	1.00	143
PRAD	1.00	1.00	1.00	139
SKCM	1.00	0.99	1.00	106
THCA	1.00	1.00	1.00	150
accuracy			1.00	1215
macro avg	1.00	1.00	1.00	1215
weighted avg	1.00	1.00	1.00	1215

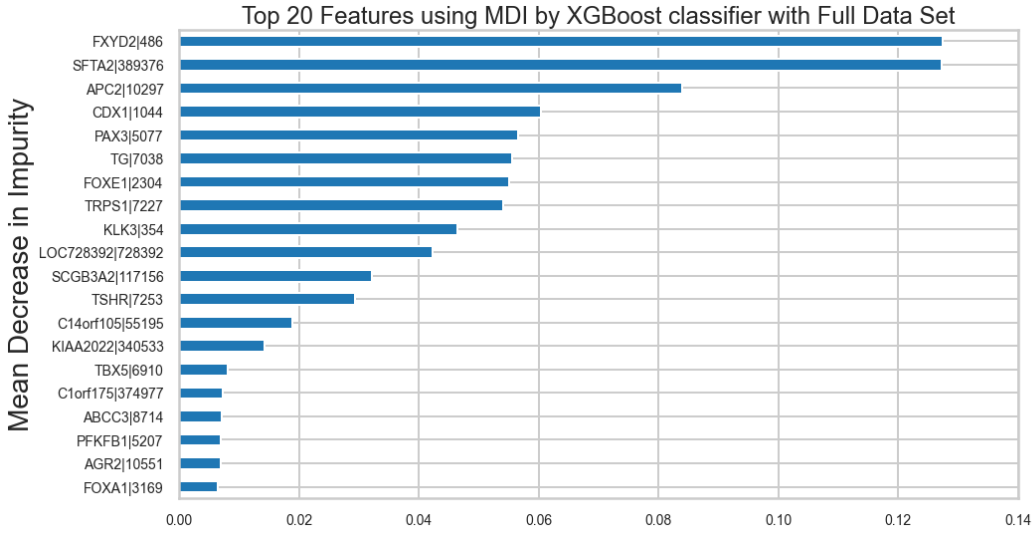
- With feature importance cutoff at 0.0005, the number of features (i.e., genes) is 105. The classification accuracy remains at 0.99918:



Changes of model accuracy and number of features as the function of feature weight cutoff:



- Given the number of genes reduced to 105 from 20530, the cancer test will be cheaper and easier to run. ¶



The following are functional annotations of 10 selected genes:

1. TRPS1 - Expression in Breast Carcinomas: Focusing on Metaplastic Breast Carcinomas..

2. TSHR - thyroid stimulating hormone receptor. Somatic mutations in the TSHR gene have been identified in thyroid tumors. These mutations are found only in the tumor cells.

3. KLK3 - used in the diagnosis and monitoring of prostate cancer. Elevated PSA levels are seen in some breast and gynecologic cancers.

4. PAX8 - Overall, PAX8 is expressed in primary and metastatic pancreatic well-differentiated neuroendocrine tumors, enabling reliable differentiation between pancreatic and ileal and pulmonary well-differentiated neuroendocrine tumors using immunostaining methods..

5. NAPSA - Diseases associated with NAPSA include Ovarian Clear Cell Adenofibroma and Adenocarcinoma.

6. SFTPB - This gene encodes the pulmonary-associated surfactant protein B (SPB), an amphipathic surfactant protein essential for lung function and homeostasis after birth.; Pro–Surfactant Protein B As a Biomarker for Lung Cancer Prediction

7. NCAN - Diseases associated with NCAN include Bipolar Disorder and Schizophrenia.

8. TG - The TG gene provides instructions for making a protein called thyroglobulin, one of the largest proteins in the body. This protein is found only in the thyroid gland, a butterfly-shaped tissue in the lower neck. Mutations within the Tg gene cause defective thyroid hormone synthesis, resulting in congenital hypothyroidism. Thyroid carcinoma may develop from dyshormonogenic goiters due to Tg mutation..

9. *LOC728392* - The genomic region with the most differentially methylated sites (LOC728392) does not have a defined function but does have predicted gene coding regions and an identified CpG island. Figure 4 plots 8 out of 27 CpG sites examined across 776 bps of a CpG island that all have higher methylation in high-risk tumors.

10. PAX3 - Rearrangements of genetic material involving the PAX3 gene are associated with a cancer of muscle tissue called alveolar rhabdomyosarcoma, which typically affects adolescents and young adults.

For detail, see https://github.com/ZhanyangZhuSD/UCSDMLCapstone/blob/main/GeneExpressionCancerRNA-Seq_FullData_XGBoost.ipynb

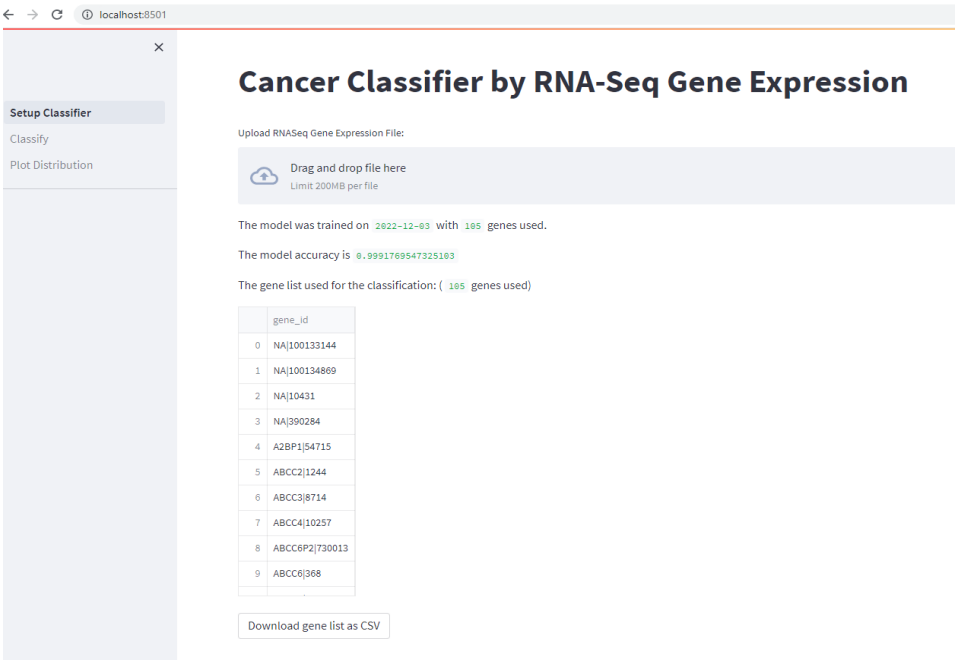
Deployment:

The deployment architect design is described here <link>. Here are a few screen captures of a classification run:

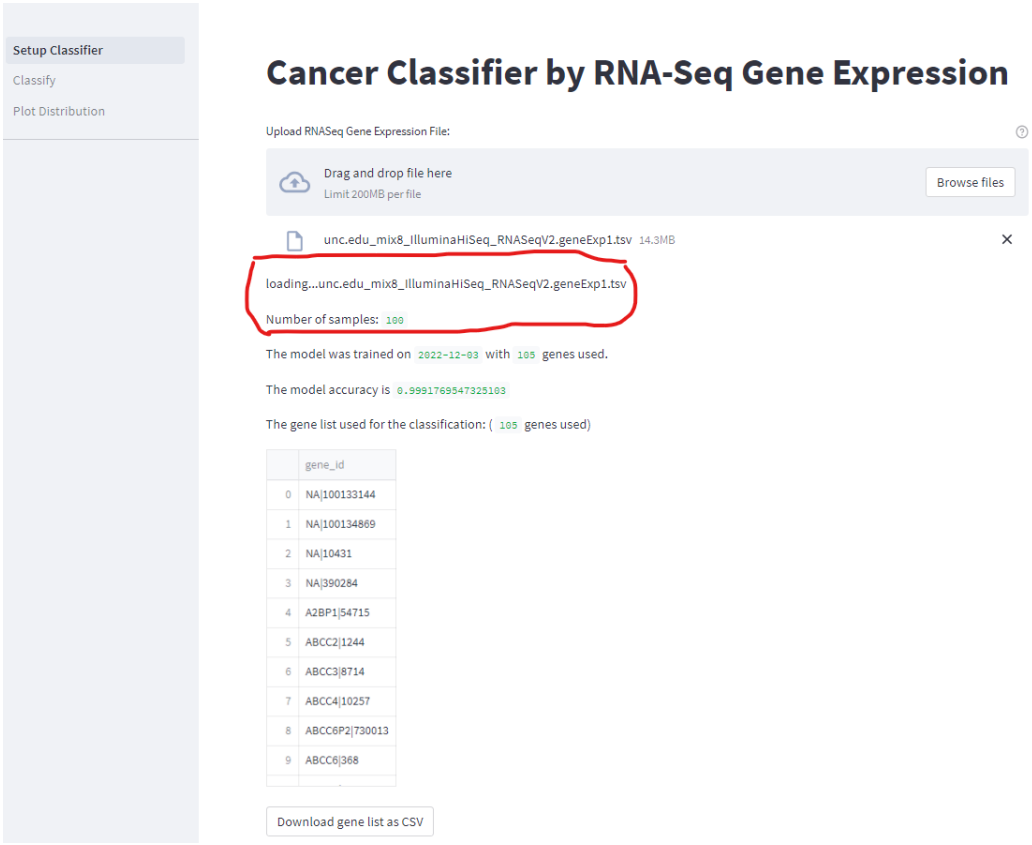
Step 0 - Start the streamlit server:

```
E:\UserData\Zhanyang\ML\DataSet4Projects\GeneExpressionCancerRNA-Seq\streamlit>streamlit run "00_Setup Classifier.py"

  You can now view your Streamlit app in your browser.

  Local URL: http://localhost:8501
  Network URL: http://192.168.1.113:8501
```

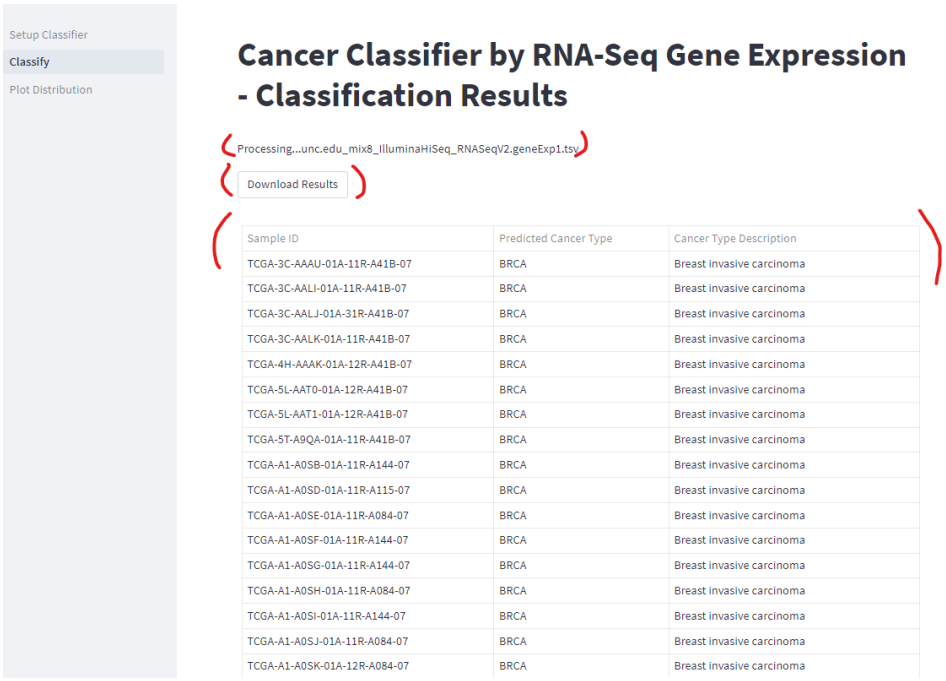
Step 1 - Setup Pre-trained Classifier



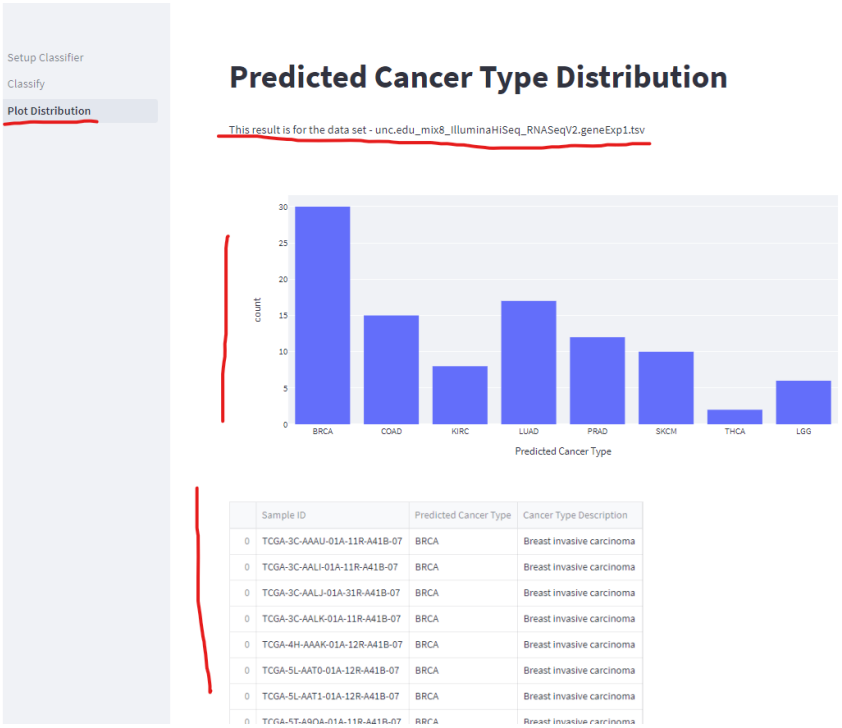
Step 2 – Select input file and load:



Step 3 – Classify – the result can be download to a csv file



Step 4 – Plot Predicted Cancer Type Distribution:



The graph is interactive.

Others Notes:

This project was inspired by a similar UCI data set - a random extraction of gene expressions of patients having different types of tumor: BRCA, KIRC, COAD, LUAD and PRAD

https://archive.ics.uci.edu/ml/datasets/gene+expression+cancer+RNA-Seq

Relevant Papers:

Weinstein, John N., et al. 'The cancer genome atlas pan-cancer analysis project.' Nature genetics 45.10 (2013): 1113-1120.