Deployment Solution Architecture for Classification of Cancer Type with Gene Expression RNA-Seq Data

Zhanyang Zhu Zhanyang.Zhu@Gmail.com 12/06/2022 for UCSD Machine Learning Bootcamp Capstone Project

Design Requirements:

The tool will be web-based. There will be three steps:

1. Load the trained classification model (saved as pickle file). Follow the following link to see how the model is trained:
   https://github.com/ZhanyangZhuSD/UCSDMLCapstone/blob/main/GeneExpressionCancerRNA-Seq_FullData_XGBoost.ipynb
   - After loading the model, display the model accuracy and date when model is last update. New mode can be updated.
   - The feature (gene) list used in the model will be shown

2. Classification:
   In this step, user can upload their data file.
   - After loading the file, number of samples will be outputted.
   - Data will be transformed into proper format.
   - Classification result will be shown in table format.
   - User will have the ability to download the result into a csv file.

3. Distribution plot of predicted cancer types:
   This step will show the count distribution of the predicted cancer types.

Implementation Design:

StreamLit (https://streamlit.io/) will be used in the implementation.

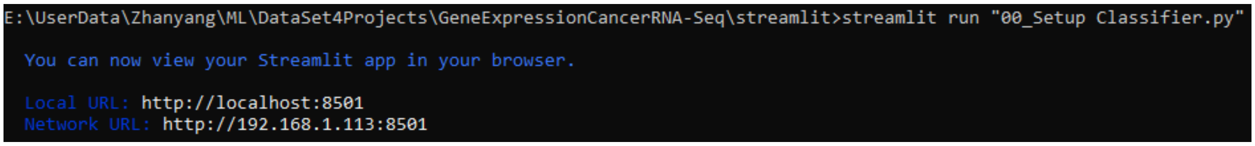
Streamlit is an open-source Python library that makes it easy to create and share beautiful, custom web apps for machine learning and data science.

Multiple pages will be created to match each steps described above. The distribution plot will be interactive by using *streamlit* plot function.

Deployment Result:

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

Step 0 - Start the *streamlit* server:

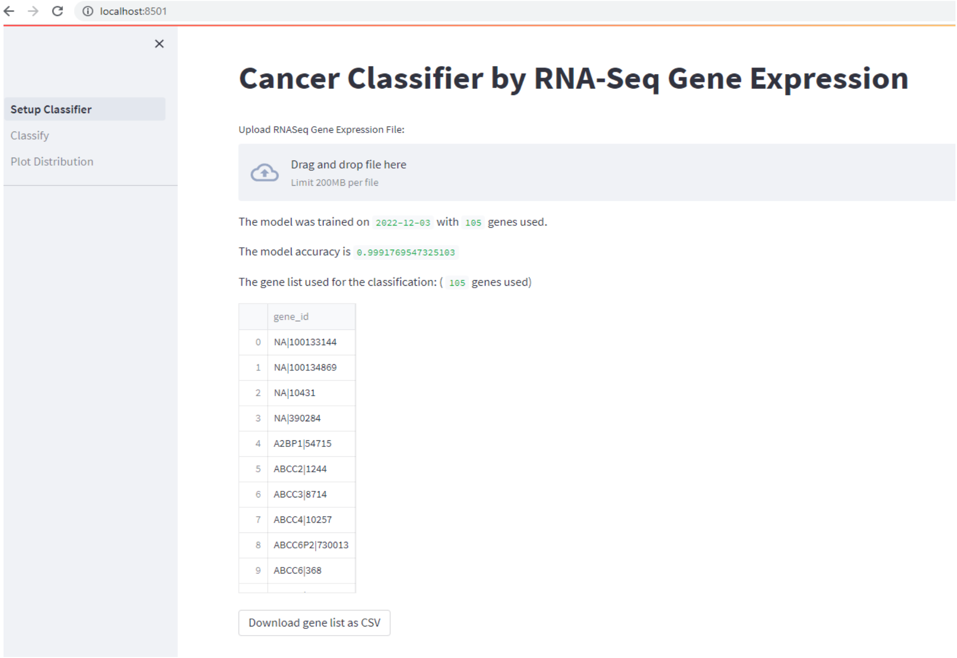


```
E:\UserData\Zhanyang\ML\DataSet4Projects\GeneExpressionCancerRNA-Seq\streamlit>streamlit run "00_Setup Classifier.py"

  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



Cancer Classifier by RNA-Seq Gene Expression

Upload RNASeq Gene Expression File:

Drag and drop file here
Limit 200MB per file

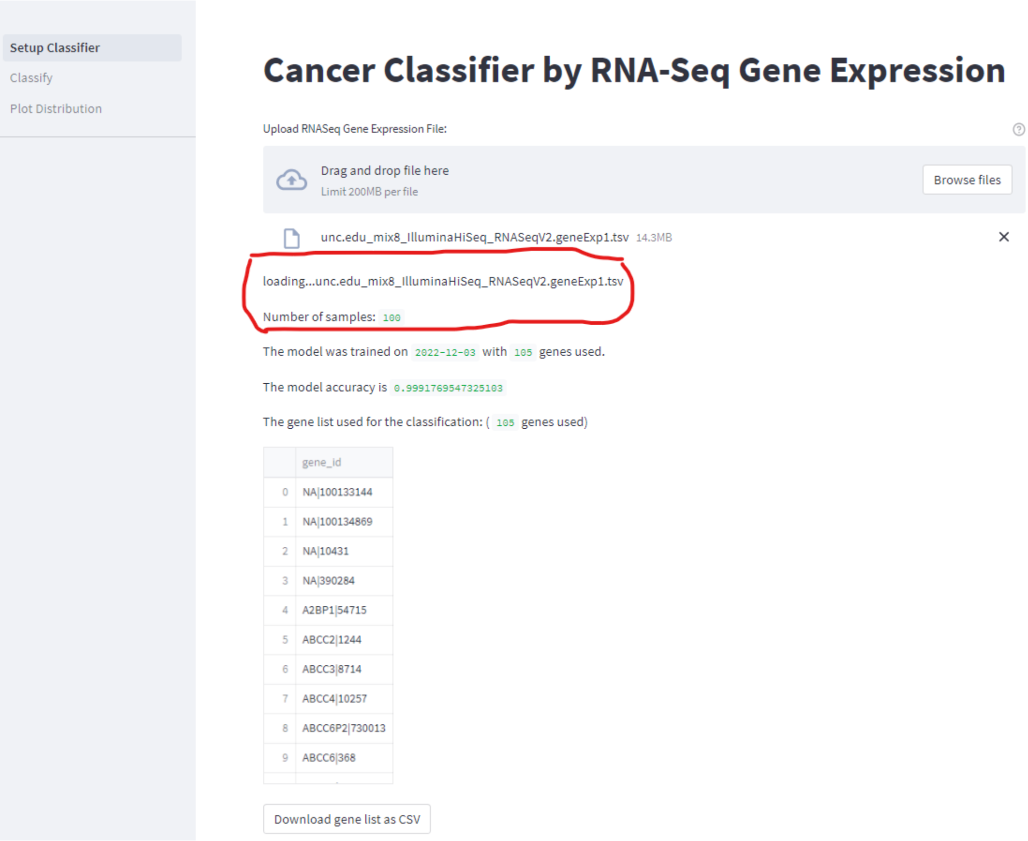
The model was trained on 2022-12-03 with 105 genes used.

The model accuracy is 0.9991769547325103

The gene list used for the classification: (105 genes used)

	gene_id
0	NA|100133144
1	NA|100134869
2	NA|10431
3	NA|390284
4	A2BP1|54715
5	ABCC2|1244
6	ABCC3|8714
7	ABCC4|10257
8	ABCC6P2|730013
9	ABCC6|368

Step 2 – Select input file and load:



Cancer Classifier by RNA-Seq Gene Expression

Upload RNASeq Gene Expression File:

unc.edu_mix8_IlluminaHiSeq_RNASeqV2.geneExp1.tsv 14.3MB

loading...unc.edu_mix8_IlluminaHiSeq_RNASeqV2.geneExp1.tsv

Number of samples: `100`

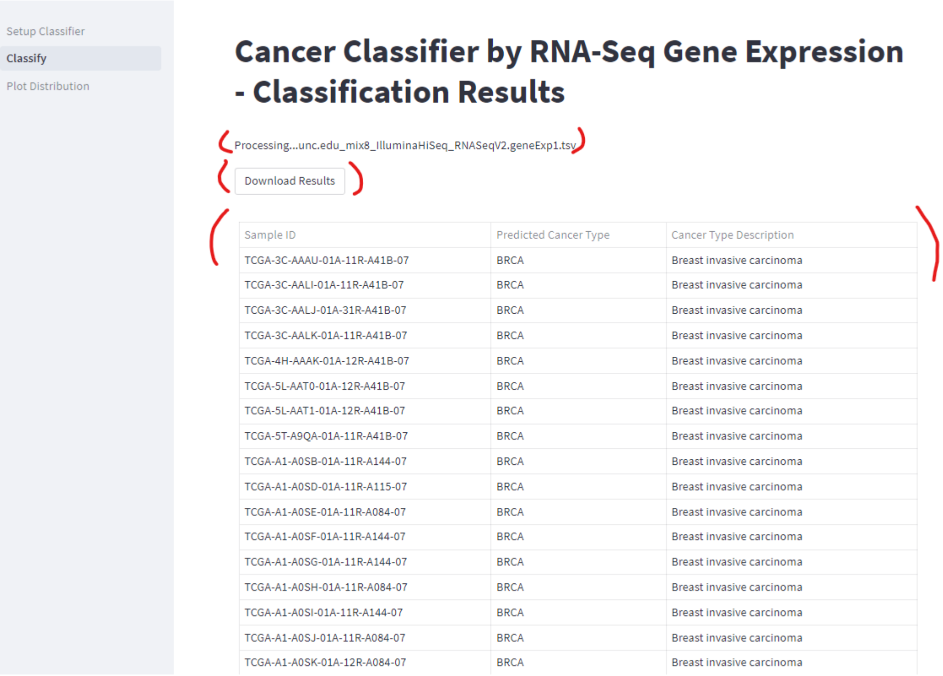
The model was trained on `2022-12-03` with `105` genes used.

The model accuracy is `0.9991769547325103`

The gene list used for the classification: (`105` genes used)

	gene_id
0	NA|100133144
1	NA|100134869
2	NA|10431
3	NA|390284
4	A2BP1|54715
5	ABCC2|1244
6	ABCC3|8714
7	ABCC4|10257
8	ABCC6P2|730013
9	ABCC6|368

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

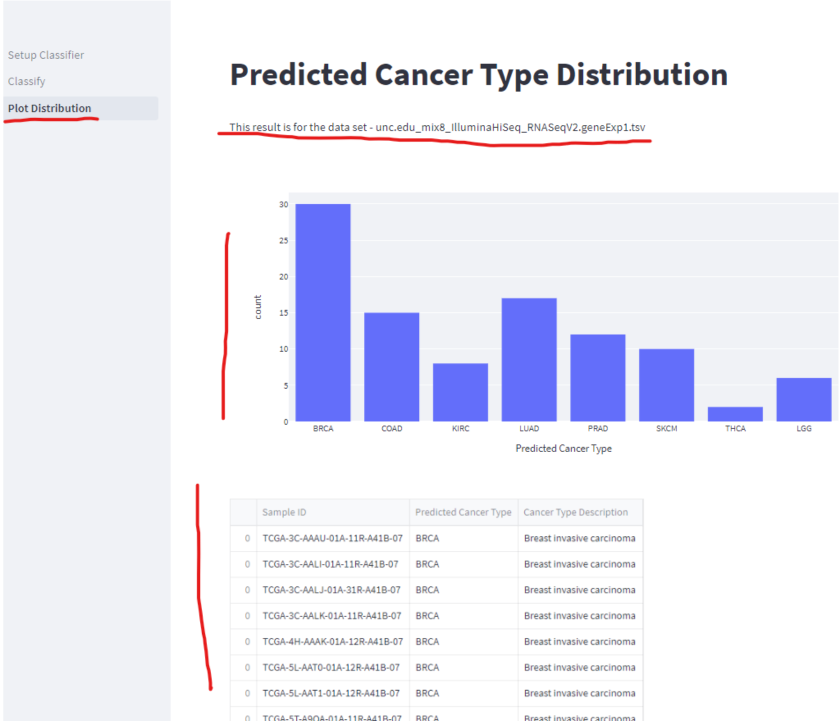


Cancer Classifier by RNA-Seq Gene Expression - Classification Results

Processing...unc.edu_mix8_IlluminaHiSeq_RNASeqV2.geneExp1.tsv

Sample ID	Predicted Cancer Type	Cancer Type Description
TCGA-3C-AAAU-01A-11R-A41B-07	BRCA	Breast invasive carcinoma
TCGA-3C-AALI-01A-11R-A41B-07	BRCA	Breast invasive carcinoma
TCGA-3C-AALJ-01A-31R-A41B-07	BRCA	Breast invasive carcinoma
TCGA-3C-AALK-01A-11R-A41B-07	BRCA	Breast invasive carcinoma
TCGA-4H-AAAK-01A-12R-A41B-07	BRCA	Breast invasive carcinoma
TCGA-5L-AAT0-01A-12R-A41B-07	BRCA	Breast invasive carcinoma
TCGA-5L-AAT1-01A-12R-A41B-07	BRCA	Breast invasive carcinoma
TCGA-5T-A9QA-01A-11R-A41B-07	BRCA	Breast invasive carcinoma
TCGA-A1-A0SB-01A-11R-A144-07	BRCA	Breast invasive carcinoma
TCGA-A1-A0SD-01A-11R-A115-07	BRCA	Breast invasive carcinoma
TCGA-A1-A0SE-01A-11R-A084-07	BRCA	Breast invasive carcinoma
TCGA-A1-A0SF-01A-11R-A144-07	BRCA	Breast invasive carcinoma
TCGA-A1-A0SG-01A-11R-A144-07	BRCA	Breast invasive carcinoma
TCGA-A1-A0SH-01A-11R-A084-07	BRCA	Breast invasive carcinoma
TCGA-A1-A0SI-01A-11R-A144-07	BRCA	Breast invasive carcinoma
TCGA-A1-A0SJ-01A-11R-A084-07	BRCA	Breast invasive carcinoma
TCGA-A1-A0SK-01A-12R-A084-07	BRCA	Breast invasive carcinoma

Step 4 – Plot Predicted Cancer Type Distribution:



Predicted Cancer Type Distribution

This result is for the data set - unc.edu_mix8_IlluminaHiSeq_RNASeqV2.geneExp1.tsv

	Sample ID	Predicted Cancer Type	Cancer Type Description
0	TCGA-3C-AAAU-01A-11R-A41B-07	BRCA	Breast invasive carcinoma
0	TCGA-3C-AALI-01A-11R-A41B-07	BRCA	Breast invasive carcinoma
0	TCGA-3C-AALJ-01A-31R-A41B-07	BRCA	Breast invasive carcinoma
0	TCGA-3C-AALK-01A-11R-A41B-07	BRCA	Breast invasive carcinoma
0	TCGA-4H-AAAK-01A-12R-A41B-07	BRCA	Breast invasive carcinoma
0	TCGA-5L-AAT0-01A-12R-A41B-07	BRCA	Breast invasive carcinoma
0	TCGA-5L-AAT1-01A-12R-A41B-07	BRCA	Breast invasive carcinoma
0	TCGA-5T-A9QA-01A-11R-A41B-07	BRCA	Breast invasive carcinoma

The graph is interactive.