Capstone Project

Discovering Potential Biomarkers for Uterine and Cervical Cancers

Genevieve Ferguson - MS Data Science Student

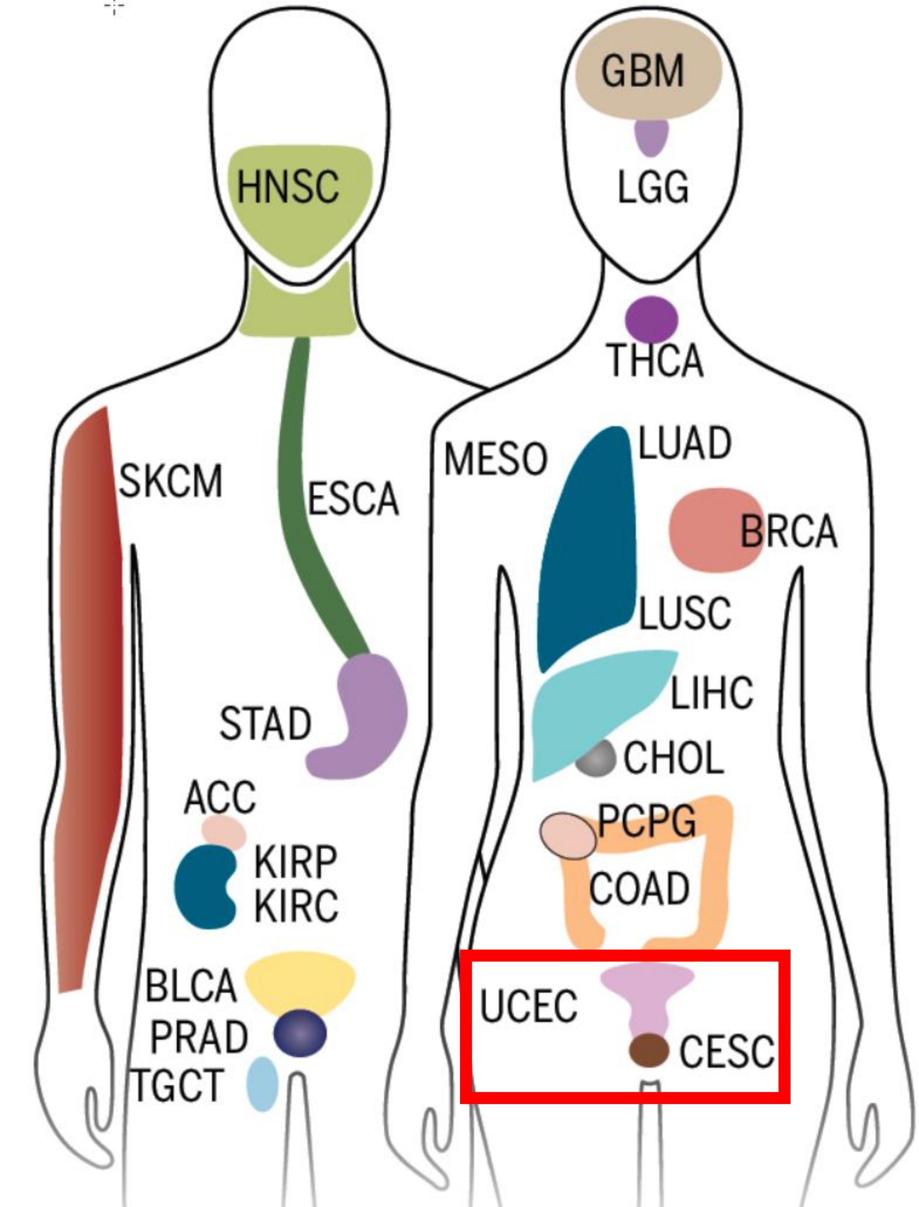
Mentor: Dr. Ananda Mondal

Guided by: Dr. Giri Narasimhan

Introduction

- Biomarker – biological measure that indicates a condition or disease
  - Can be used to identify risk/presence of cancer
- Potential biomarkers - datasets
  - Copy number alterations
  - Gene expression
  - miRNA
- Desired result:
  - Potential biomarkers for cancer (2 types)
    - UCEC/ Uterine Corpus Endometrial Carcinoma
    - CESC/Cervical Squamous Cell Carcinoma and Endocervical Adenocarcinoma

XenaBrowser



hhmi.org

Genome.gov

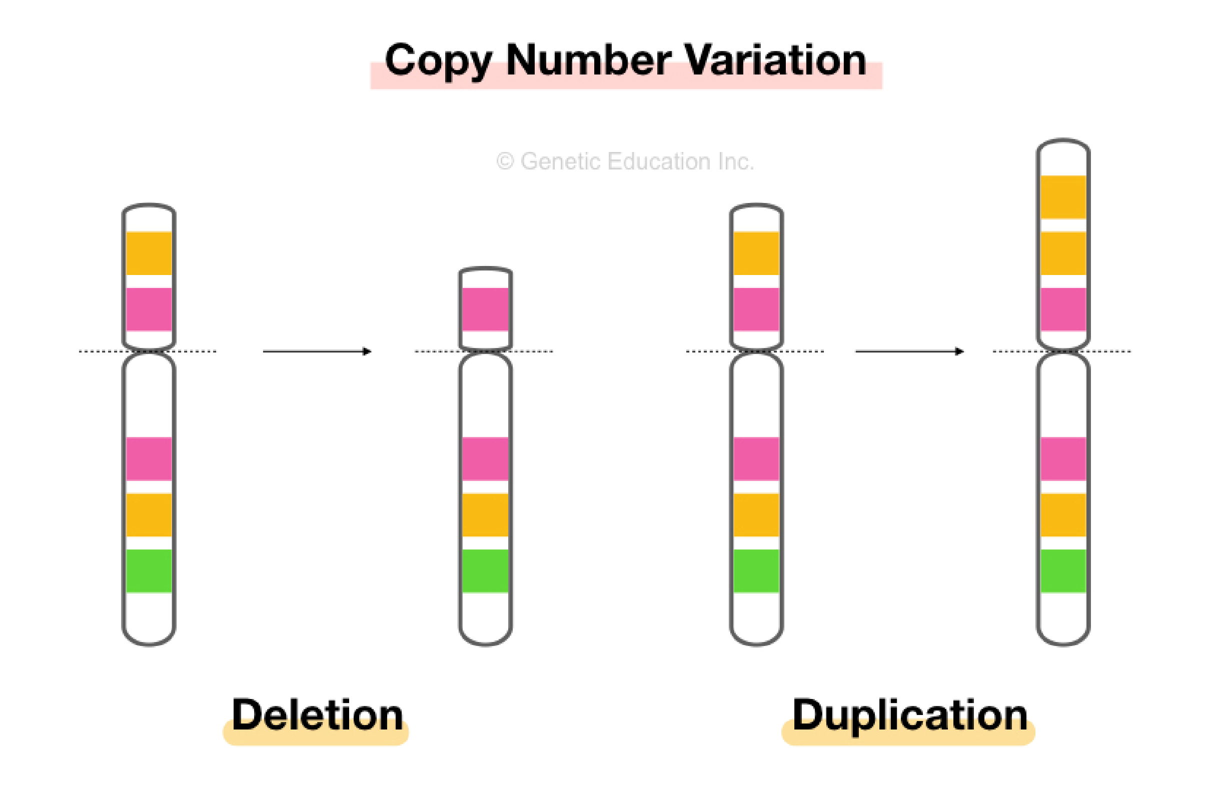
Background

- Each cell – 23 pairs chromosomes
- Chromosomes > DNA > genes
- DNA transcription to RNA
- RNA translation to proteins
- Changes to gene = changes to protein (when, where, how much)
- Changes to protein = changes to the body

Geneticeducation.co

Background – CNAs

- Copy number alterations (CNAs) - changes to chromosome structure
- Gain or lose number of genes
  - Certain CNAs associated w disease



Gene expression

- Central Dogma
- DNA -> RNA -> proteins
  - Process called gene expression
- Measuring protein concentration = hard
- Measure mRNA counts instead

Sources: Galaxyproject.org genome.gov

miRNA expression

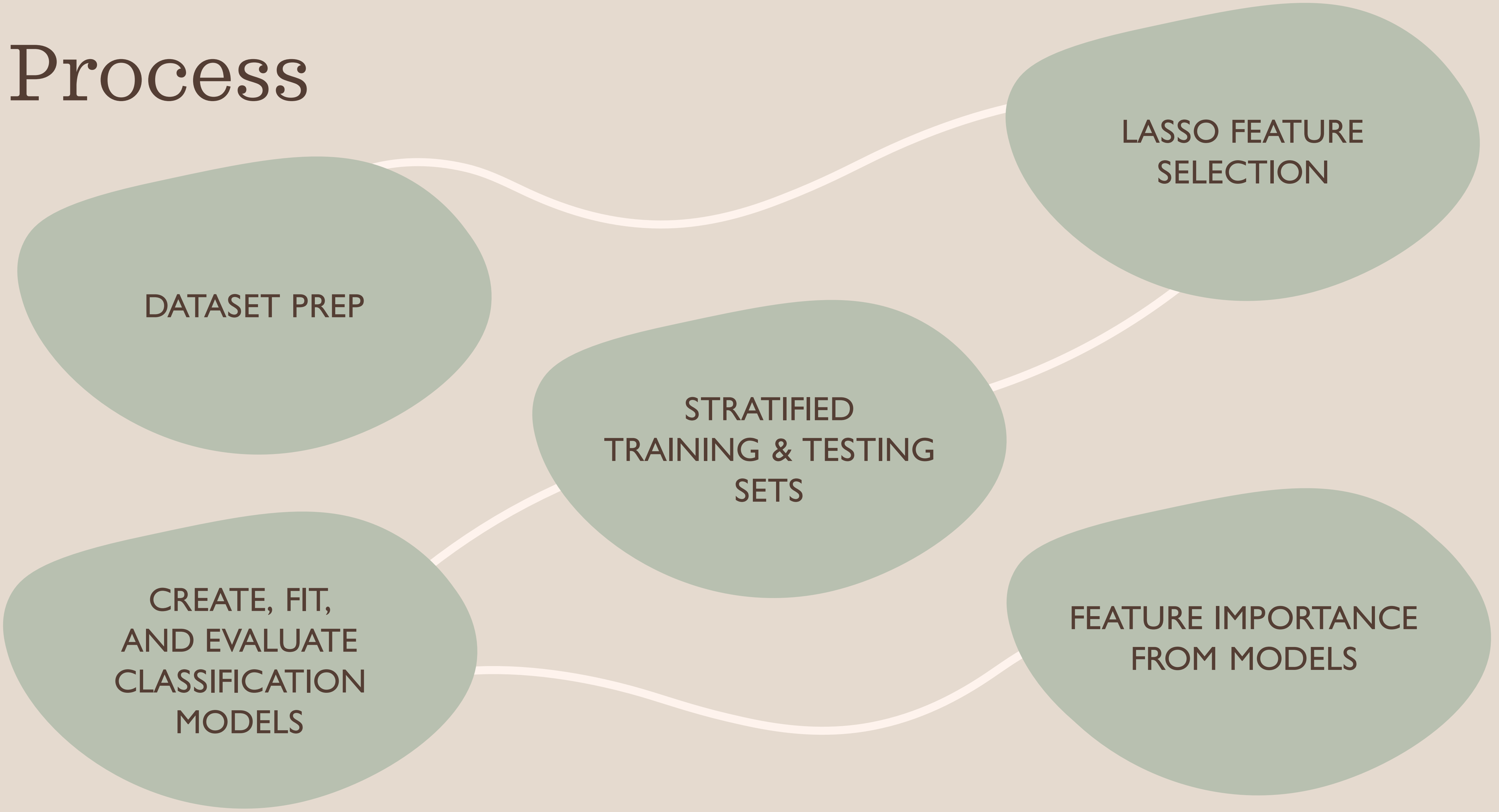
- miRNA – non-coding
- Bind to mRNA so cannot make proteins
- Can block oncogenes
  - stop cancer
- Can block tumor suppressors
  - promote cancer

Sources: PubMed. Computational Biology

Use Cases

- Springboard for further research
- Replicate for other cancers
- Early detection systems
  - Before symptoms start
  - Frequent screenings for high-risk individuals

Process



DATASET PREP

LASSO FEATURE SELECTION

STRATIFIED TRAINING & TESTING SETS

CREATE, FIT, AND EVALUATE CLASSIFICATION MODELS

FEATURE IMPORTANCE FROM MODELS

Goals & Objectives

Goals

BUILD ACCURATE CLASSIFICATION MODELS

FIND BIOMARKERS FOR BOTH CANCERS

DETERMINE DIRECTIONALITY OF BIOMARKER

Objectives

EVALUATE ALL CLASSIFICATION MODELS FOR ACCURACY

USE LASSO AND K-FOLD CV + TRAINING AND TEST ACCURACY

FIND AVERAGE SHAP VALUES FOR EACH FEATURE ACROSS ALL CLASSIFIERS

USE SHAP BEESWARM PLOTS TO IDENTIFY DIRECTION OF INFLUENCE

Motivation

2016-2020

- 11542 NEW CASES OF CERVICAL CANCER
  - **4272 women died**
- 54744 NEW CASES OF UTERINE CANCER
  - **11995 women died**

Source: CDC

Challenges

SLIGHTLY IMBALANCED DATA

FINDING SAMPLES ACROSS ALL DATASETS

LOW NUMBER OF SAMPLES

COMPUTATIONAL COST

Datasets from TCGA

miRNA

887 rows
1881 columns

signal intensity
miRNA

Gistic

845 rows
19729 columns

-1,0,1:
Evidence of copy del, no change, evidence of duplication

Protein-coding genes

Htseq

892 rows
60483 columns

mRNA mapping to Protein-coding, non-coding, & pseudo genes

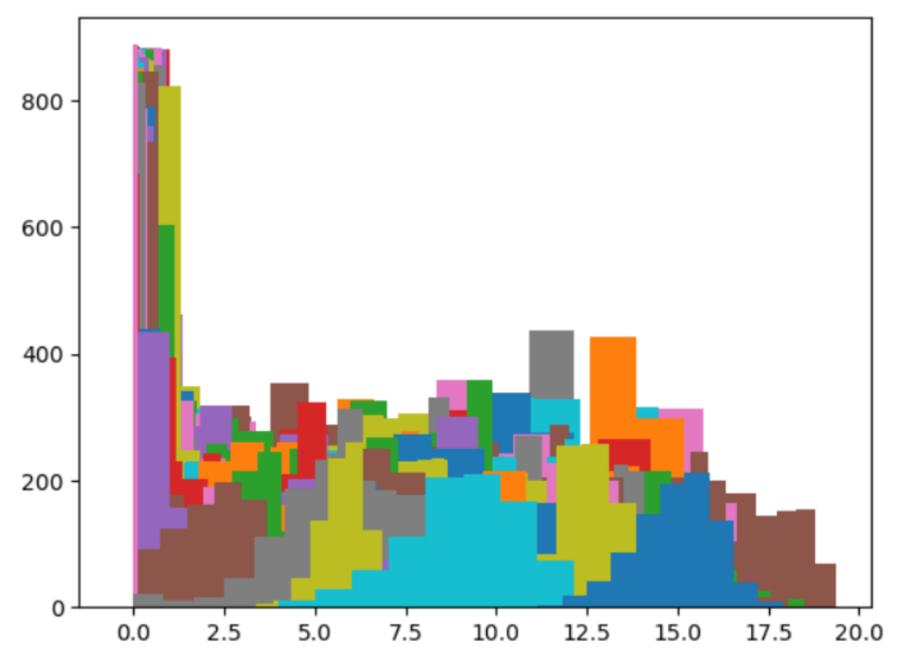
Combined

824 rows
(overlap)
82093 columns
(concatenated)

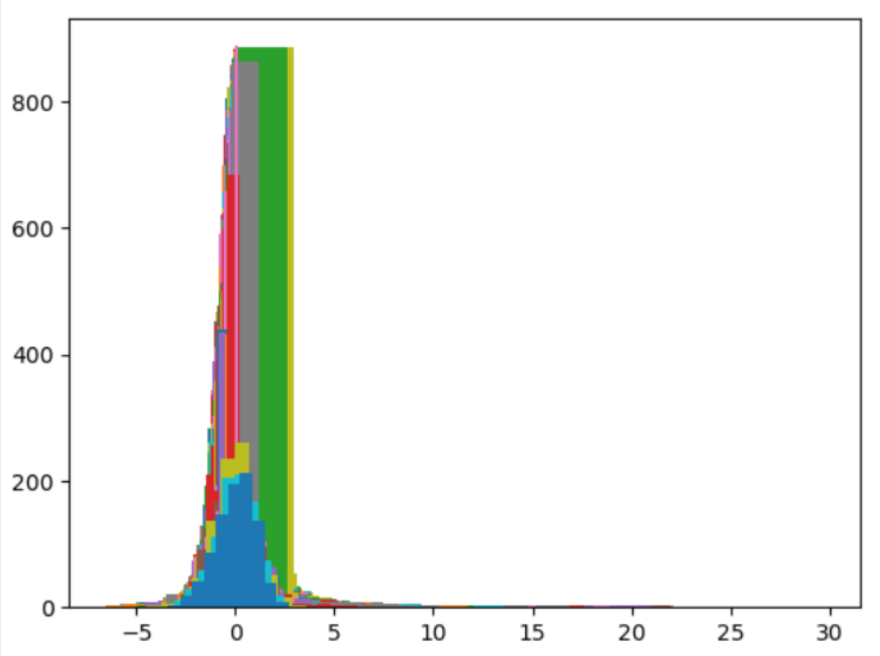
Participants in all datasets

miRNA dataset - Standard Scaling

Before



After



$$x_{standardized} = \frac{x - x_{mean}}{x_{standard\ deviation}}$$

LASSO (Least Absolute Shrinkage and Selection Operator)

FEATURE SELECTION – REDUCES INSIGNIFICANT OR REDUNDANT FEATURES TO ZERO

+OVERFITTING

10-FOLD CROSS VALIDATION ALSO FOR OVERFITTING

Datasets After LASSO

miRNA	Gistic	Htseq	Combined Dataset
1881 columns	19729 columns	60483 columns	82093 columns
Reduced to 131	Reduced to 77	Reduced to 118	Reduced to 152

Classification algorithms

- LogisticRegression
- RandomForestClassifier
- LinearDiscriminantAnalysis
- DecisionTreeClassifier
- AdaBoostClassifier
- ExtraTreesClassifier
- KNeighborsClassifier
- GaussianNB
- CalibratedClassifier
- SVM
- Perceptron
- PassiveAggressiveClassifier
- SGDClassifier
- XGBClassifier
- BaggingClassifier
- MLPClassifier
- RidgeClassifierCV
- ANN

Total = 18

ANN

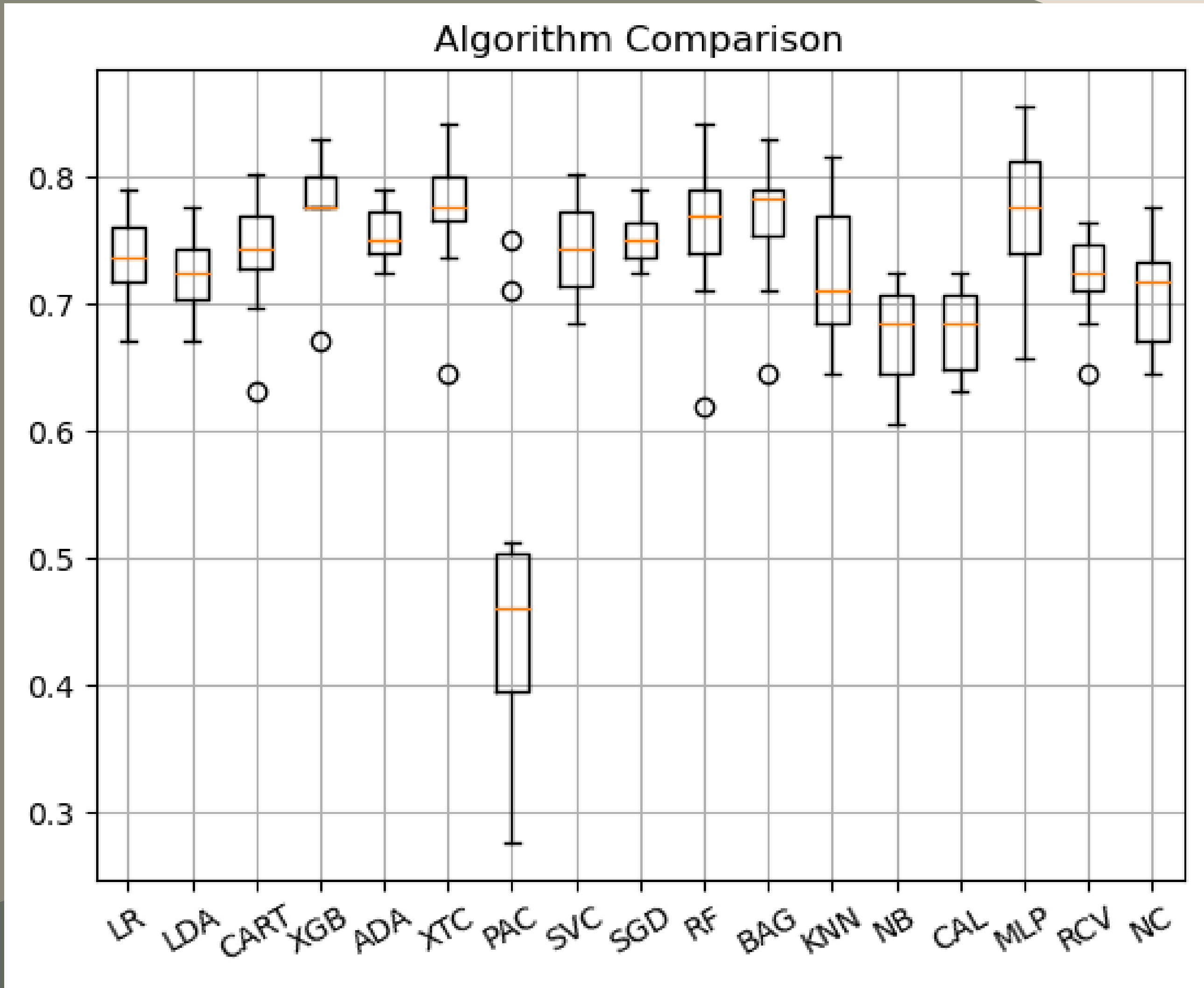
- Neural network for binary classification
- Layer 1: Input
  - Neurons: (# of features)
  - Activation Function: ReLU
  - Determined by the number of features in the input data
- Layer 2: Hidden
  - Neurons: (# of features)
  - Activation Function: ReLU
  - Further refines learned features from the previous layer
- Output Layer:
  - Neurons: 1
  - Activation Function: Sigmoid
  - Classifies cancer type into 0 and 1

Results

CNAs- 'Gistic' Dataset

CESC = '1' | count = 548
UCEC = '0' | count = 297

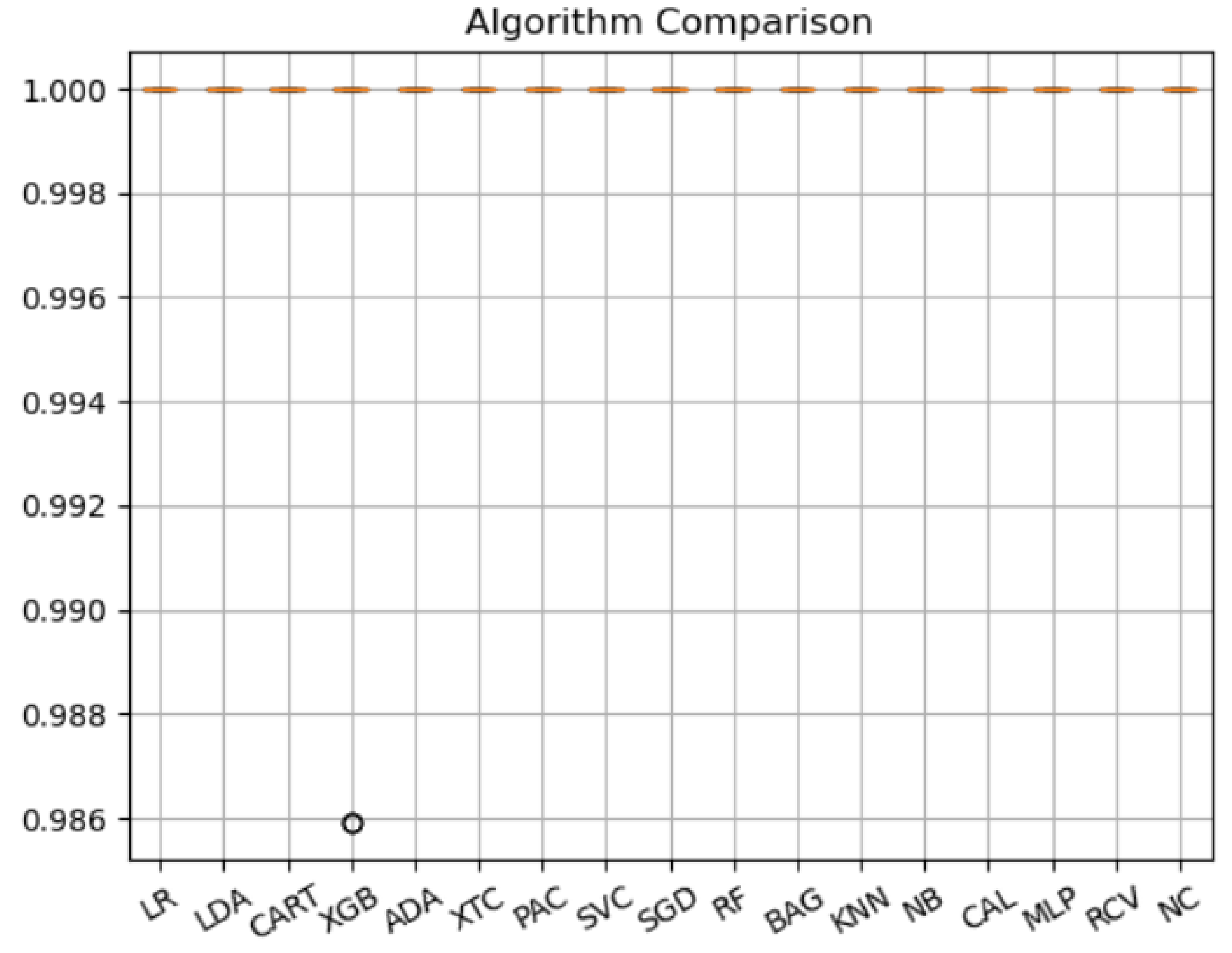
PAC removed



Results

Gene Expression – ‘Htseq’

CESC = ‘1’ | count = 583
UCEC = ‘0’ | count = 309

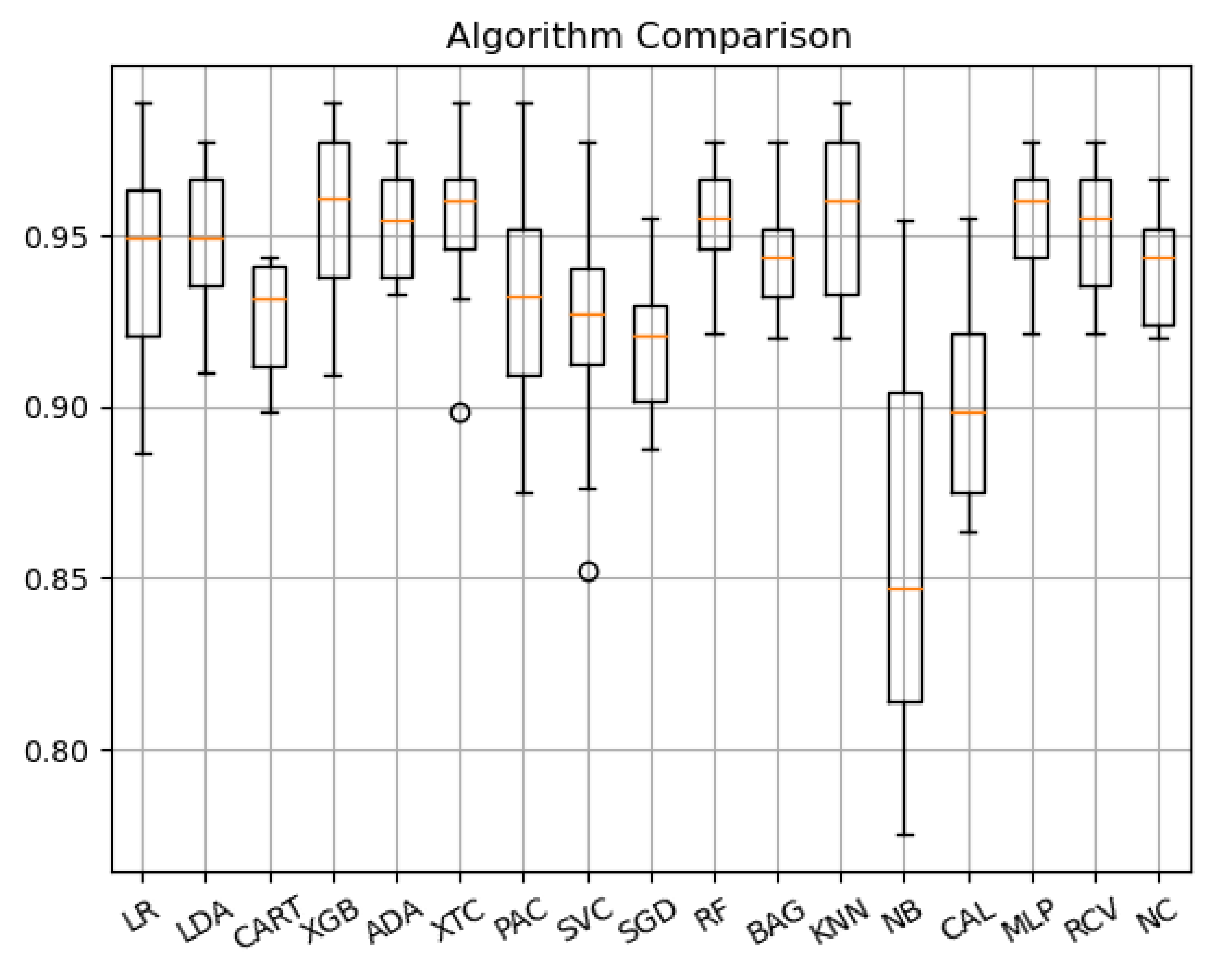


Results

microRNA – ‘mirna’

CESC = ‘1’ | count = 575
UCEC = ‘0’ | count = 312

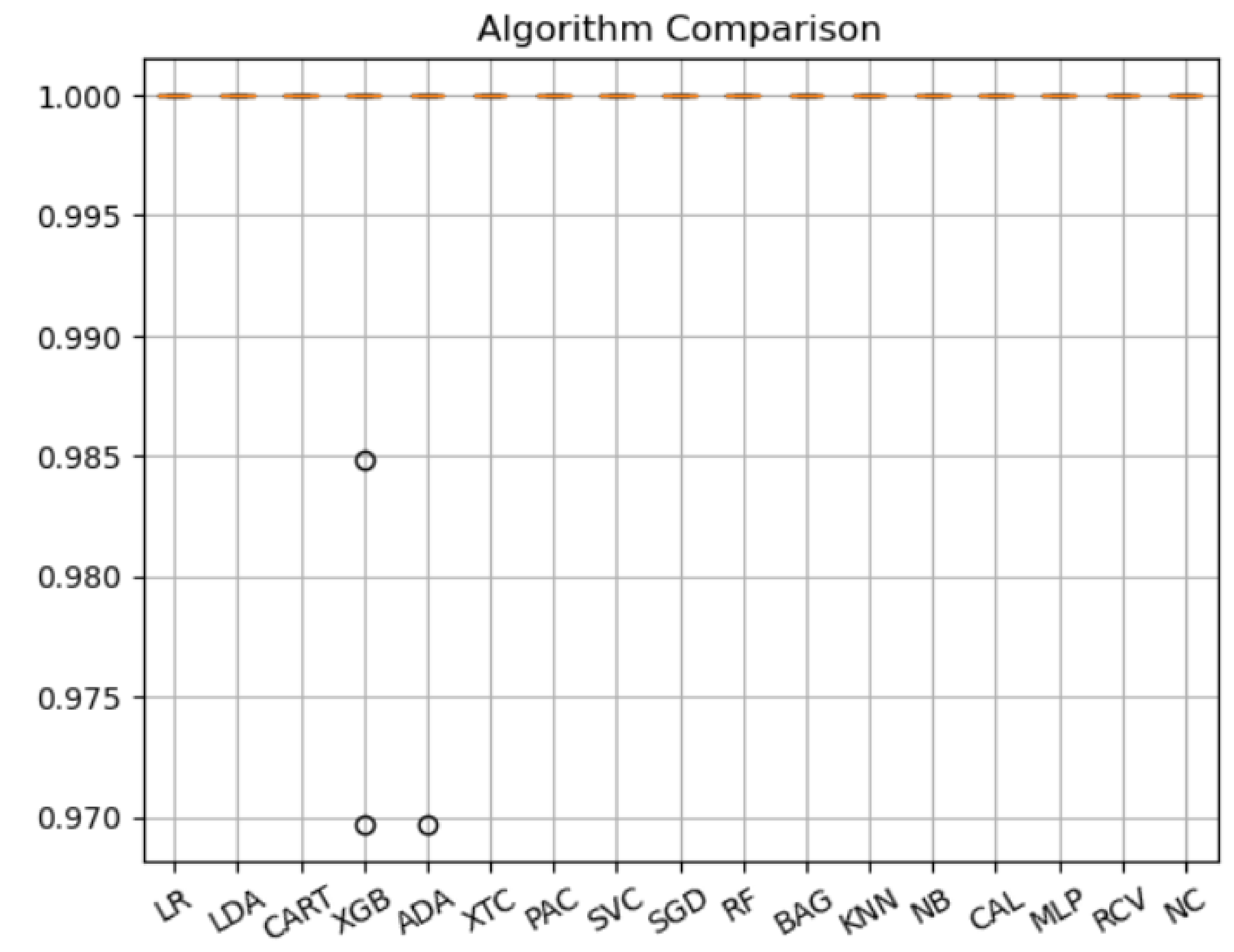
NB removed



Results

Combined Dataset

CESC = ‘1’ | count = 530
UCEC = ‘0’ | count = 294
Total = 824



Accuracies

Gistic Dataset	
Algorithm	Acc
LR	0.728
LDA	0.717
CART	0.724
XGB	0.767
ADA	0.741
XTC	0.755
SVC	0.728
SGD	0.722
RF	0.759
BAG	0.753
KNN	0.724
NB	0.671
CAL	0.667
MLP	0.763
RCV	0.714
NC	0.704
ANN	0.711

HtSEQ Dataset	
Algorithm	Acc
LR	1.0
LDA	1.0
CART	1.0
XGB	0.999
ADA	1.0
XTC	1.0
PAC	1.0
SVC	1.0
SGD	1.0
RF	1.0
BAG	1.0
KNN	1.0
NB	1.0
CAL	1.0
MLP	1.0
RCV	1.0
NC	1.0
ANN	1.0

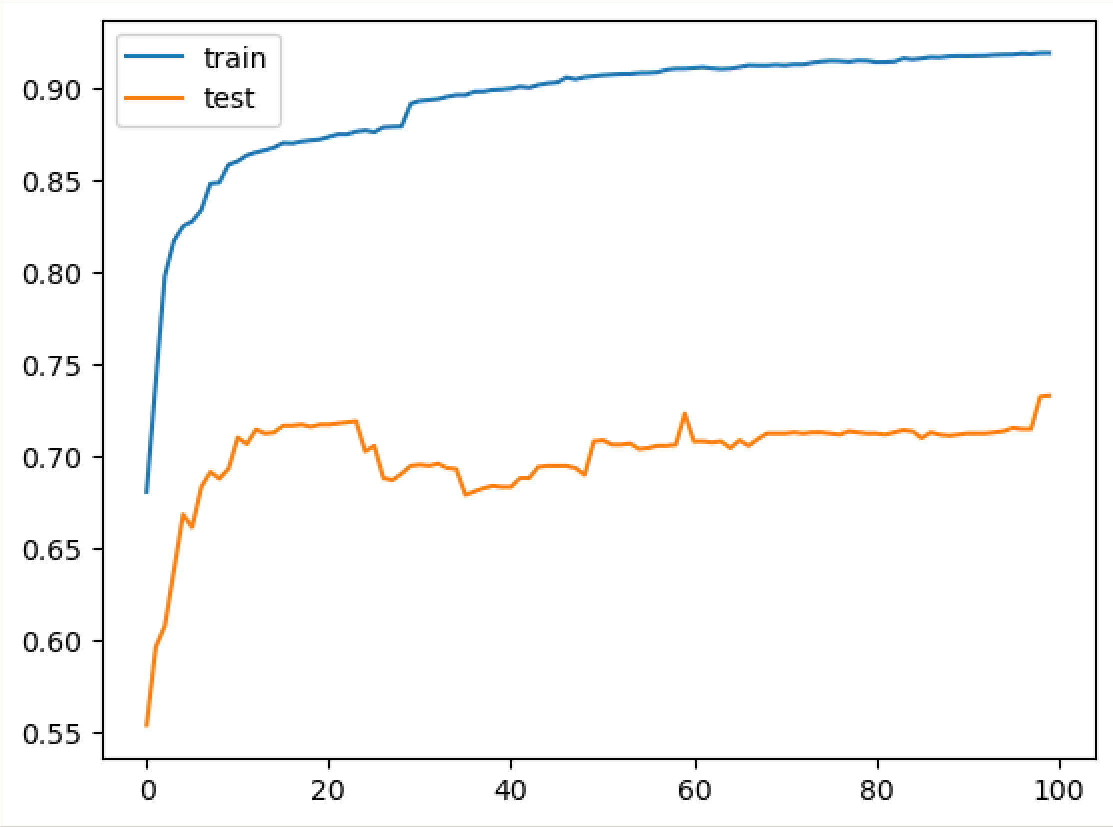
miRNA Dataset	
Algorithm	Acc
LR	0.942
LDA	0.950
CART	0.926
XGB	0.957
ADA	0.955
XTC	0.955
PAC	0.933
SVC	0.923
SGD	0.919
RF	0.955
BAG	0.942
KNN	0.955
CAL	0.901
MLP	0.956
RCV	0.953
NC	0.940
ANN	0.937

Combined Dataset	
Algorithm	Acc
LR	1.0
LDA	1.0
CART	0.999
XGB	0.999
ADA	1.0
XTC	1.0
PAC	1.0
SVC	1.0
SGD	1.0
RF	1.0
BAG	1.0
KNN	1.0
NB	1.0
CAL	1.0
MLP	1.0
RCV	1.0
NC	1.0
ANN	1.0

Gistic - AUC

Significant overfit

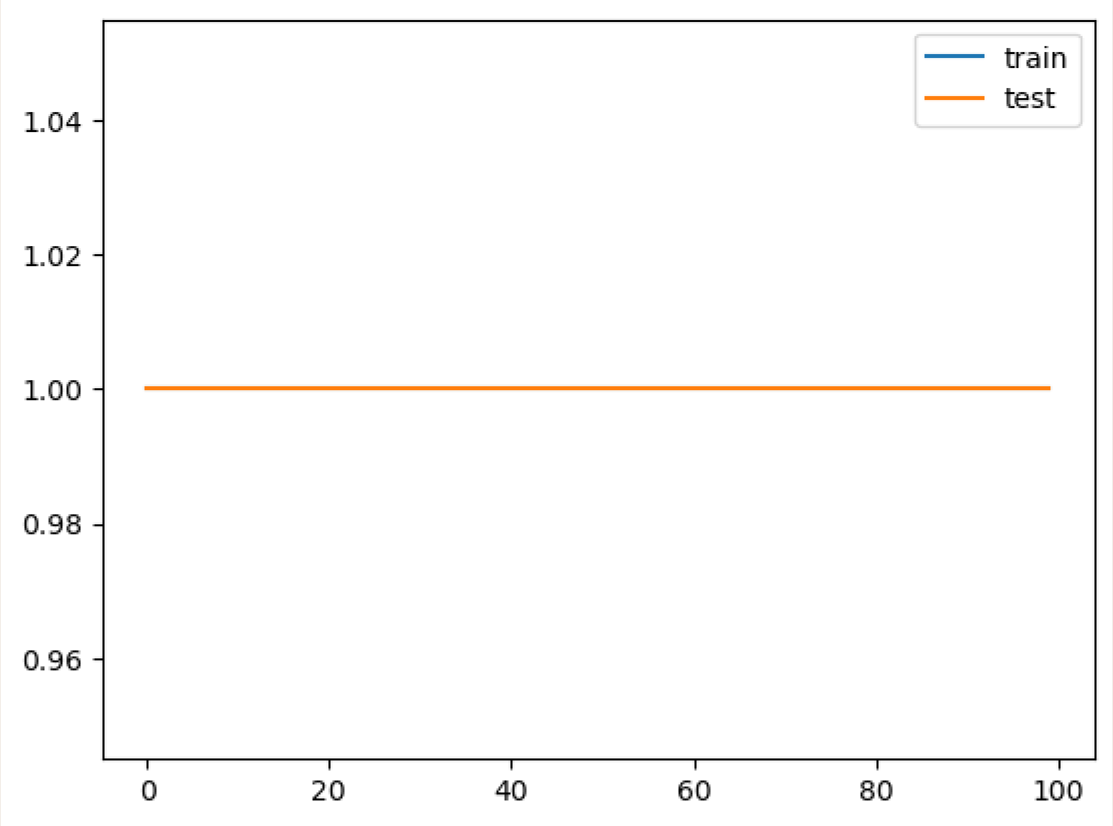
-need more data



Htseq – AUC

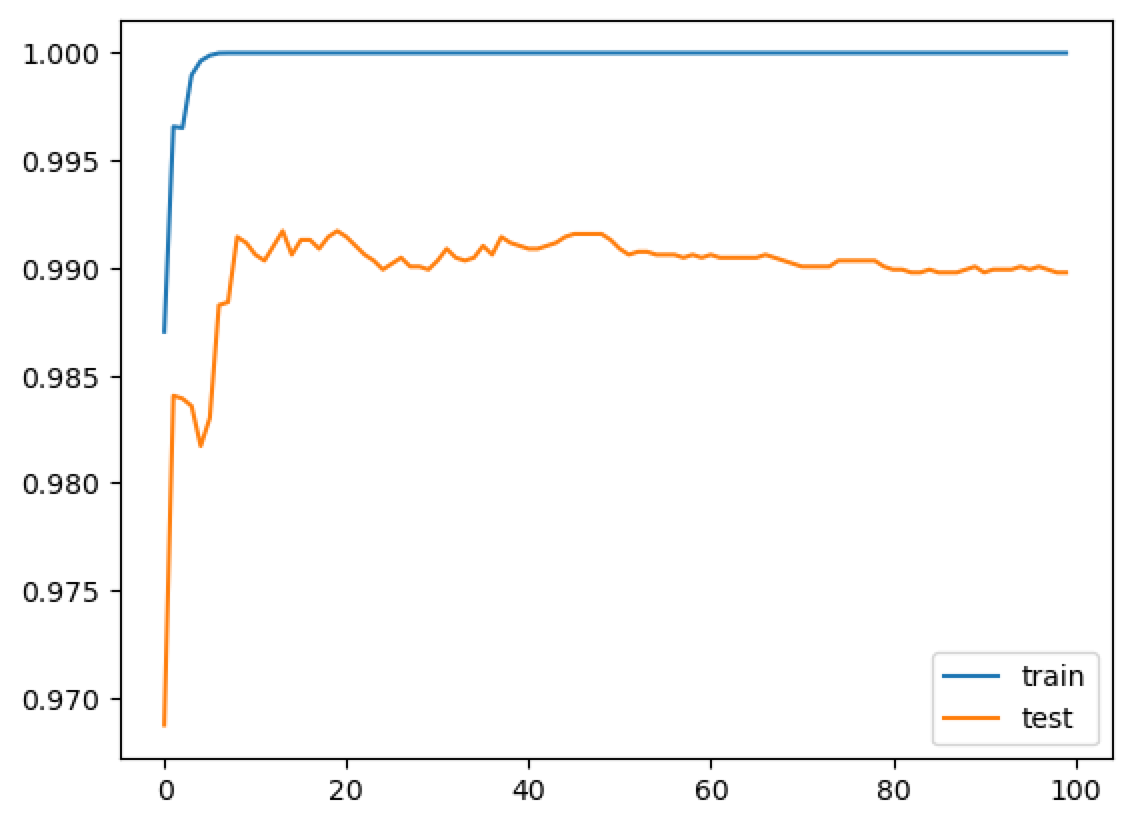
Perfect performance

-read counts significantly different for two cancers



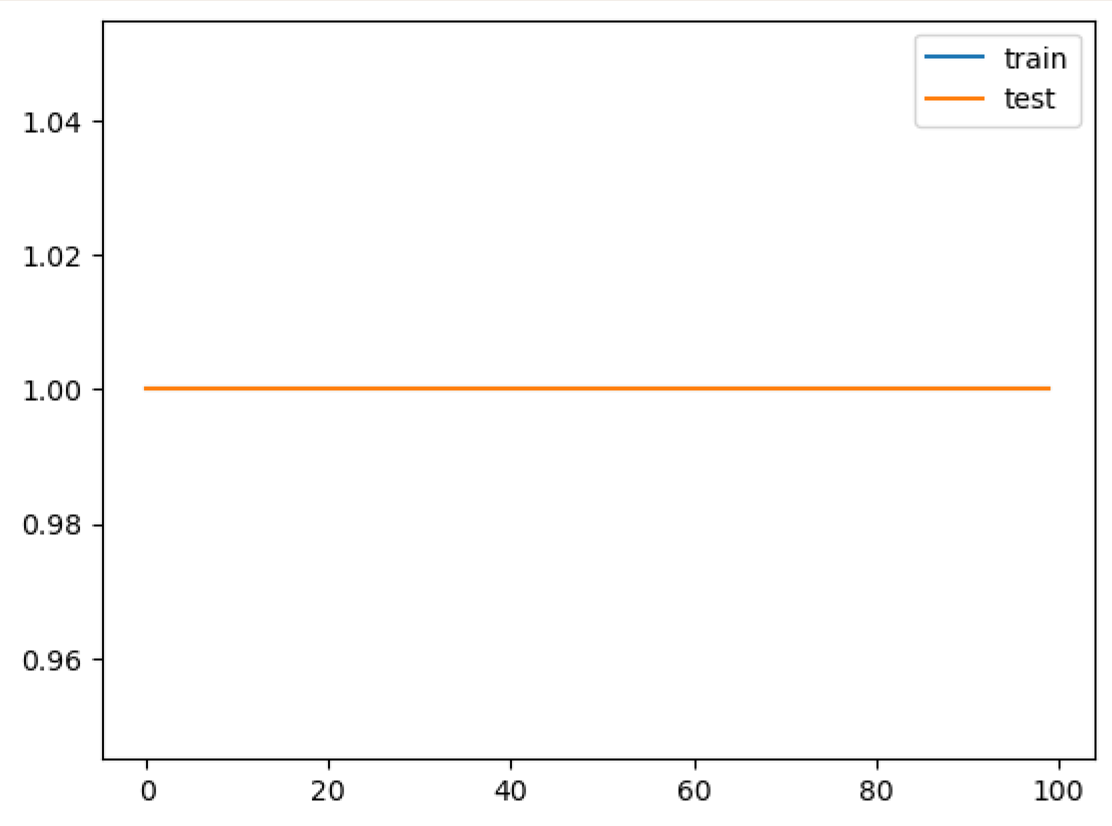
miRNA learning curve - AUC

Slight overfit but still performing well.



Combined learning curve

Also perfect performance



Important features

MOST SALIENT FEATURES FOUND FOR EACH ALGORITHM

SHAPLEY (SHAP) SCORES USED

-EXPLAINS HOW EACH FEATURE (PLAYER) CONTRIBUTES TO THE PREDICTION (TEAM)

GISTIC AND MIRNA = XGBOOST SHAP VALUES

HTSEQ & COMBINED = AVERAGED ACROSS ALL ALGORITHMS

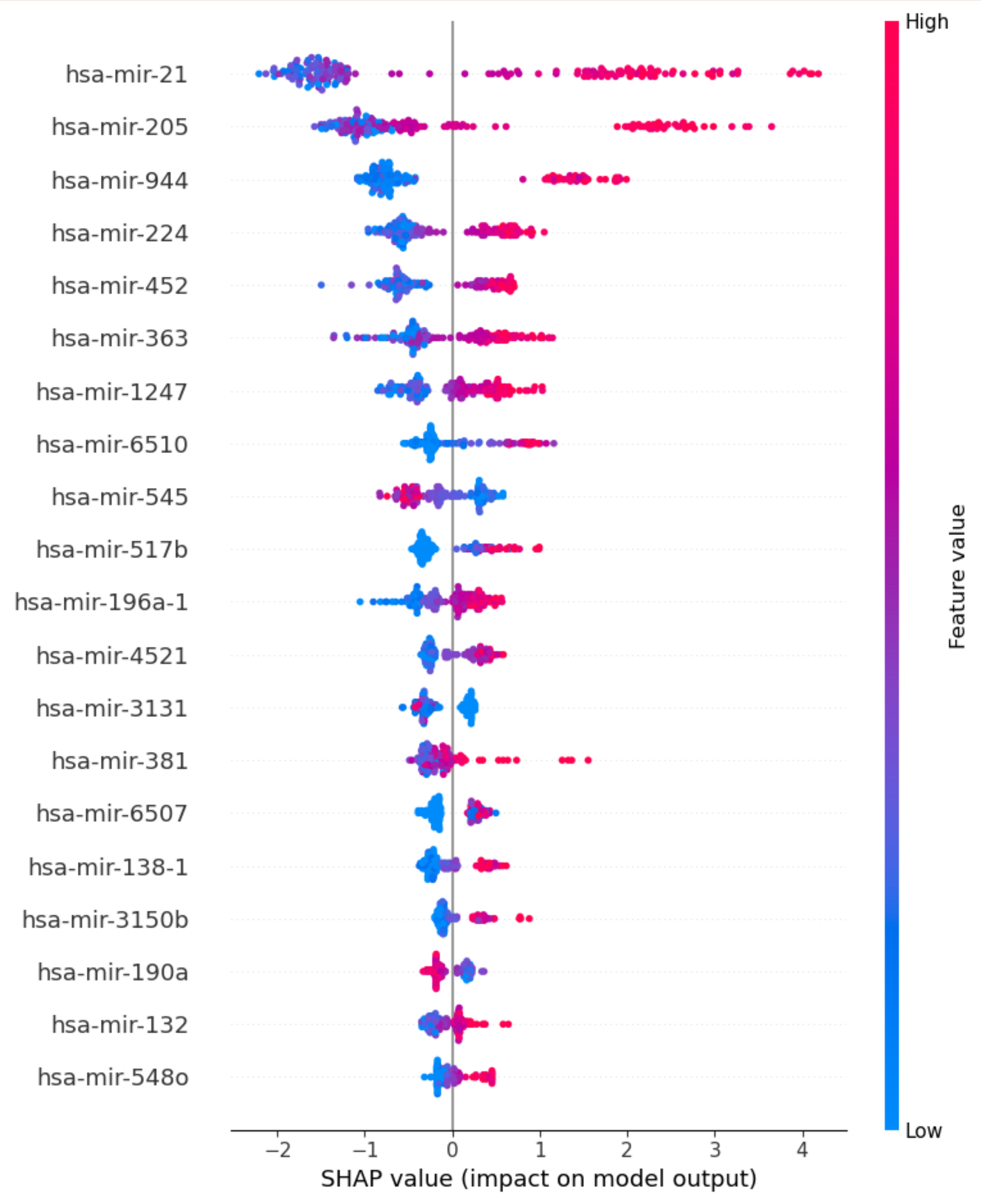
Source: Towardsdatascience

SHAP Directionality-miRNA data

- UCEC = 0
- CESC = 1

- Mir-944 & Mir-21
  - Low – UCEC
  - High – CESC

- Mir-190a
  - Low – CESC
  - High – UCEC



Medium

Gistic – XGB SHAP values

Gene Name	Cancer Specificity	Ave SHAP	Source
ENSG00000141905.16	Endometrial cancer	0.093	link
ENSG00000073282.11	cervical cancer, head and neck cancer, lung cancer, urothelial cancer	0.063	link
ENSG00000176009.3	None	0.062	link
ENSG00000028310.16	Detected in all	0.055	link
ENSG00000121858.9	Prostate cancer, Myeloma, Endometrial cancer, Breast cancer	0.054	link

endometrial biomarkers

Htseq – Average SHAP values

Gene Name	Cancer Specificity	Ave SHAP	Source
ENSG00000206630.1	endometrial cancer	0.061	link
ENSG00000182117.5	Diffuse large B cell lymphoma (DLBCL), bladder, & cervical	0.060	link
ENSG00000269899.1	Ovarian cancer	0.057	link
ENSG00000274501.1	Acute myeloid leukemia (LAML) & thymoma	0.044	link
ENSG00000215267.7	None	0.030	link

miRNA – XGBoost SHAP values

Gene Name	Cancer Specificity	SHAP val	Source
hsa-mir-21	Many cancers, incl. endometrial & cervical	0.154	link
hsa-mir-205	Endometrial, cervical, squamous cell carcinoma, colon cancer	0.110	Link link
hsa-mir-944	Endometrial, cervical, & breast cancers	0.082	Link link
hsa-mir-224	Hepatocellular carcinoma (HCC), Pancreatic ductal adenocarcinoma (PDAC), & Non-small cell lung cancer (NSCLC)	0.049	link
hsa-mir-452	Bladder cancer, uterine cancer	0.048	link

Combined – Average SHAP values

Gene Name	Cancer Specificity	Ave SHAP	Source
ENSG00000280231.1	Thymoma	0.064	link
ENSG00000215030.5	Ovarian and bladder cancer	0.061	Link
ENSG00000225131.2	Diffuse large B cell lymphoma (DLBCL) and Glioblastoma	0.058	link
ENSG00000128228.4	Diffuse large B cell lymphoma (DLBCL), Uterine, Bladder cancer	0.058	link
ENSG00000244268.1	endometrial cancer	0.034	link link

Discussion

MIRNA - HSA-MIR-224

XGBOOST SHAP VALUE = 0.049

NO CURRENT CANCER BIOMARKERS

HTSEQ - ENSG00000269899.1,
ENSG00000274501.1, &
ENSG00000215267.7
- SPECIAL INTEREST – HIGH ACC – NO
BIOMARKERS

COMBINED – TOP 4 WITH NO
BIOMARKERS FOR EITHER CANCER

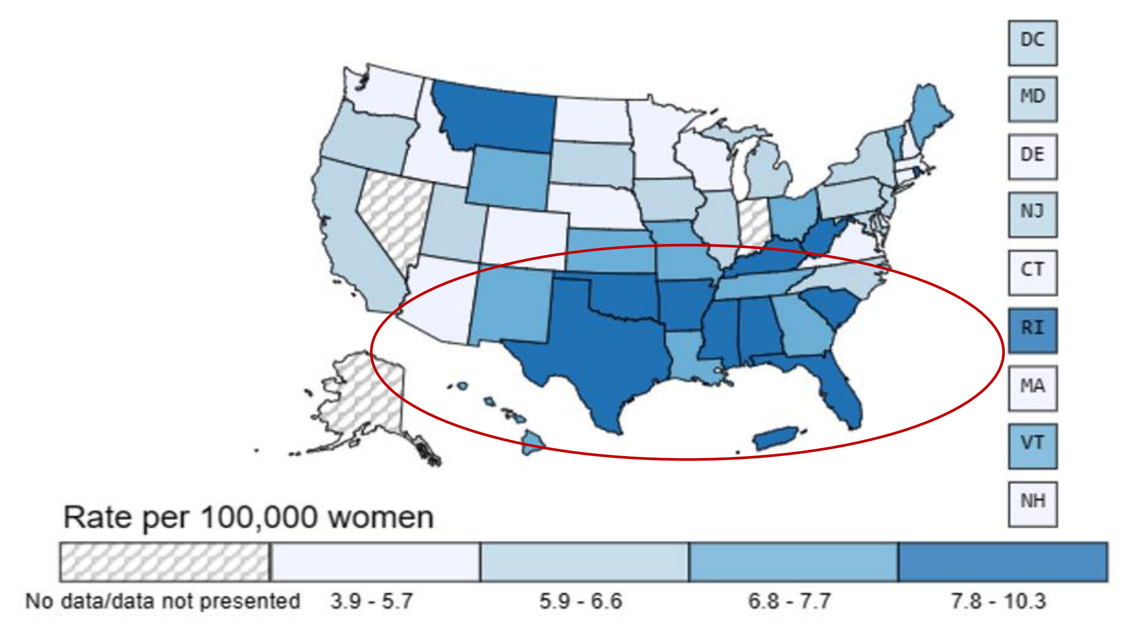
GISTIC (CAUTION) - ENSG00000176009.3

SHAP VALUE = 0.062

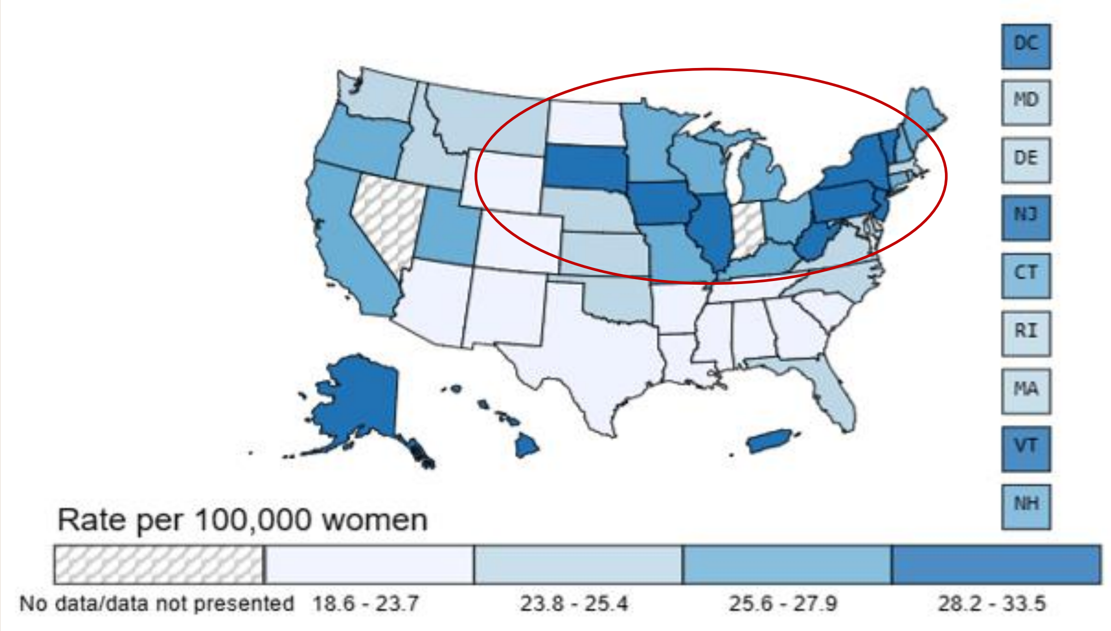
BUT LOW ACCURACY

Comparison Maps – US Rate of New Cancers

Cervical



Uterine



USCS Data Visualizations - CDC - 2016 to 2020

CONCLUSION

Biomarkers – applied to public health
+Genetic testing to predict disease
+Screening in high-prevalence areas

Future Work: Which combinations of biomarkers result in specific disease?

Acknowledgement

Special Thanks to the Machine Learning And Data Analytics Group (MLDAG) here at FIU

& my mom, Donna Ferguson for her endless support

Thank you

Genevieve Ferguson

gferg020@fiu.edu