Supervised ML Approaches for Predicting Survival Rate

Group 3
Darin, Darshan, David, Heng, Tina

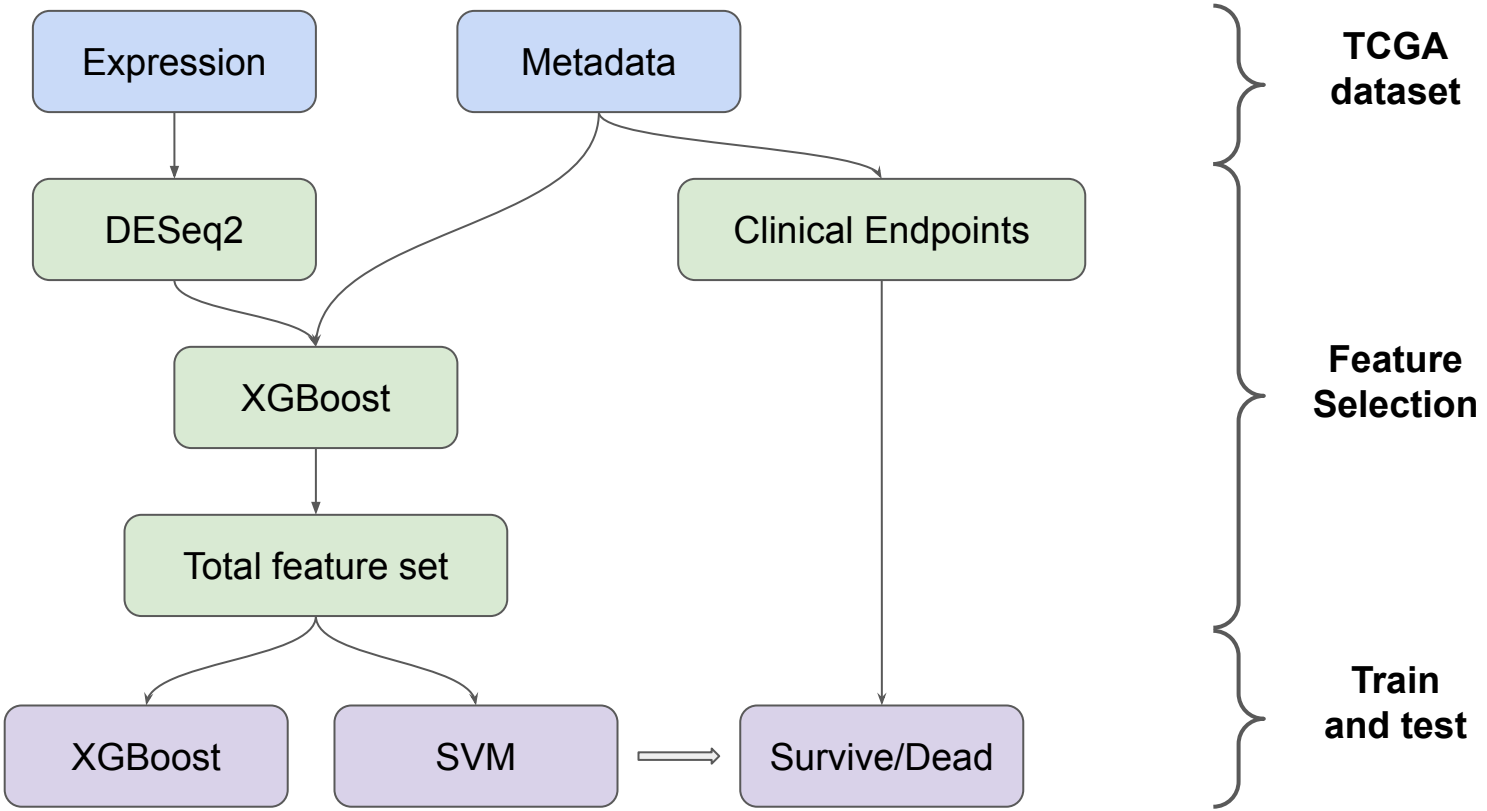
June 14, 2023

Background/Introduction

- **“How Long Have I Got?”**
- Most lung cancer patients worldwide [stage IV non small cell lung cancer (NSCLC)] have a poor survival: 25%–30% die <3 months.
- Of those surviving >3 months, 10%–15% (70,000–105,000 new patients worldwide per year) survive (very) long.
- **Goal:** develop and validate a survival prediction model in a cohort of different stages patients of lung cancer in TCGA.

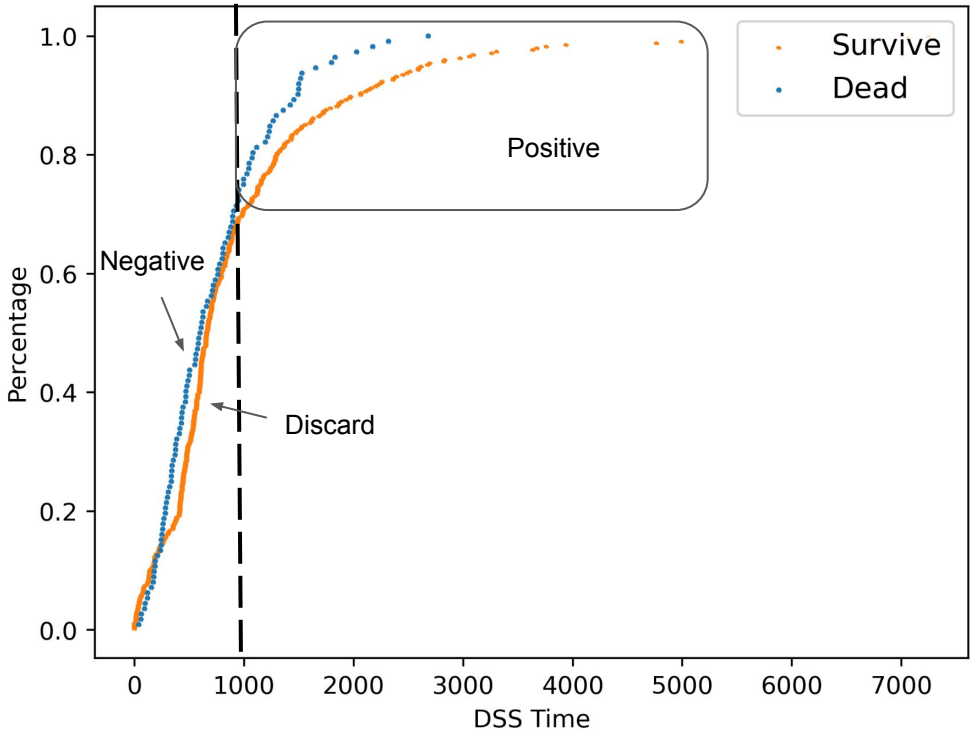
Guo et al.

Overview


Expression
Metadata
DESeq2
Clinical Endpoints
XGBoost
Total feature set
XGBoost
SVM
Survive/Dead
TCGA dataset
Feature Selection
Train and test

Clinical Endpoints

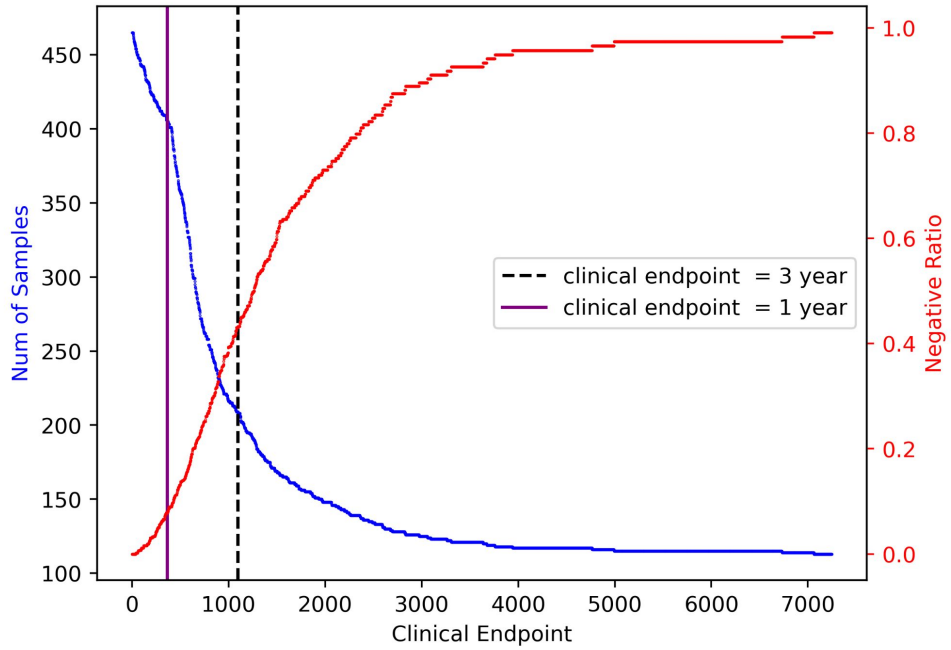
How to Define Negative/Positive Samples?



Clinical Endpoints

Trade-off between Sample Size and Sample Balance

Use two different clinical endpoints: 1 year and 3 year



Demographic Information

1 yr	N = 405	
Variable	**Living (N=373)**	**Deceased (N=32)**
Age (years)		
< 65(179)	163	16
≥65(226)	210	16
Sex		
Female (225)	210	15
Male(180)	163	17
Tumor stage		
Early(319)	293	26
Late(80)	74	6
Smoking status		
Smoker(256)	237	19
Non-smoker(149)	136	13

3 yr	N = 209	
Variable	**Living (N=119)**	**Deceased (N=90)**
Age (years)		
< 65(93)	49	44
≥65(116)	70	46
Sex		
Female(116)	66	50
Male(93)	53	40
Tumor stage		
Early(157)	99	58
Late(49)	18	31
Smoking status		
Smoker(134)	78	56
Non-smoker(75)	41	34

Data Preprocessing

Metadata

1. Age at diagnosis
2. Gender
3. Tumor stage:
   I, II as EARLY stage; III, IV as LATE stage
4. Smoking status:
   > 188 cigarettes/year as SMOKER
   ≤ 188 cigarettes/year as NON-SMOKER

Gene Expression

1. Remove all normal and recurrent samples
2. Log2 transformation

DESeq2 Result (Endpoint: 1 year)

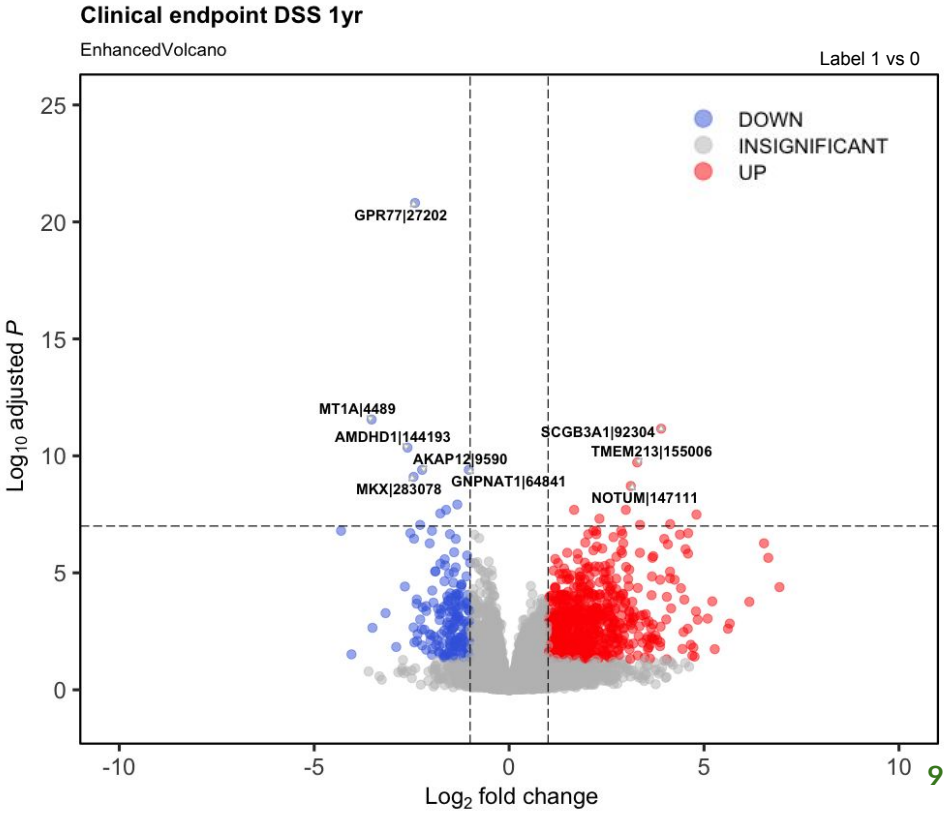
- **DESeq2 pre-filtering**:
  Remove genes with less than 10 counts (remove 757 genes out of 19382 with nonzero total read count, 3.9%)
- **3,226** genes (15.7%) being significant with padj < 0.05

GENE	log2FoldChange	pvalue	padj
GPR77	-2.41267	8.29E-26	1.54E-21
MT1A	-3.5278	3.01E-16	2.80E-12
SCGB3A1	3.90002	1.11E-15	6.92E-12
AMDHD1	-2.60538	9.55E-15	4.45E-11
TMEM213	3.2851	5.11E-14	1.90E-10

DESeq2 Result (Endpoint: 1 year)

- **DESeq2 pre-filtering**:
  Remove genes with less than 10 counts (remove 757 genes out of 19382 with nonzero total read count, 3.9%)
- **3,226** genes (15.7%) being significant with padj < 0.05

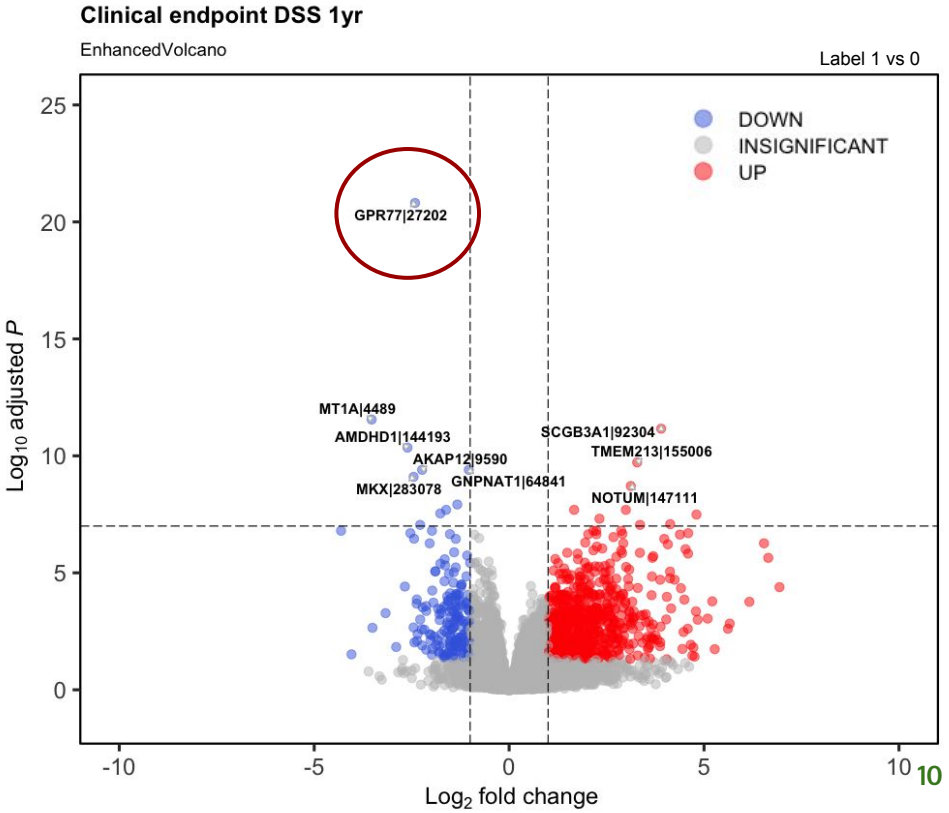
GENE	log2FoldChange	pvalue	padj
GPR77	-2.41267	8.29E-26	1.54E-21
MT1A	-3.5278	3.01E-16	2.80E-12
SCGB3A1	3.90002	1.11E-15	6.92E-12
AMDHD1	-2.60538	9.55E-15	4.45E-11
TMEM213	3.2851	5.11E-14	1.90E-10



DESeq2 Result (Endpoint: 1 year)

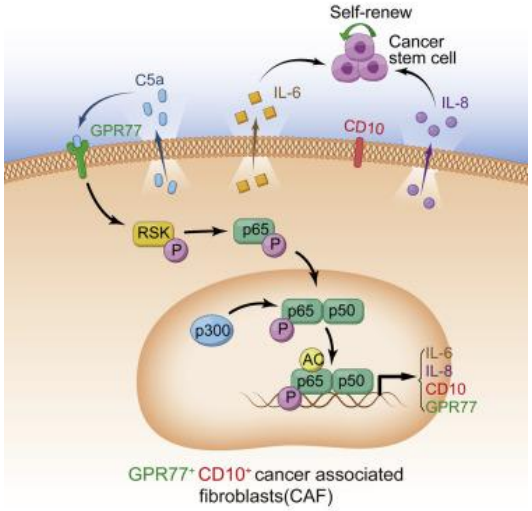
- **DESeq2 pre-filtering**:
  Remove genes with less than 10 counts (remove 757 genes out of 19382 with nonzero total read count, 3.9%)
- **3,226** genes (15.7%) being significant with padj < 0.05

GENE	log2FoldChange	pvalue	padj
GPR77	**-2.41267**	**8.29E-26**	**1.54E-21**
MT1A	-3.5278	3.01E-16	2.80E-12
SCGB3A1	3.90002	1.11E-15	6.92E-12
AMDHD1	-2.60538	9.55E-15	4.45E-11
TMEM213	3.2851	5.11E-14	1.90E-10

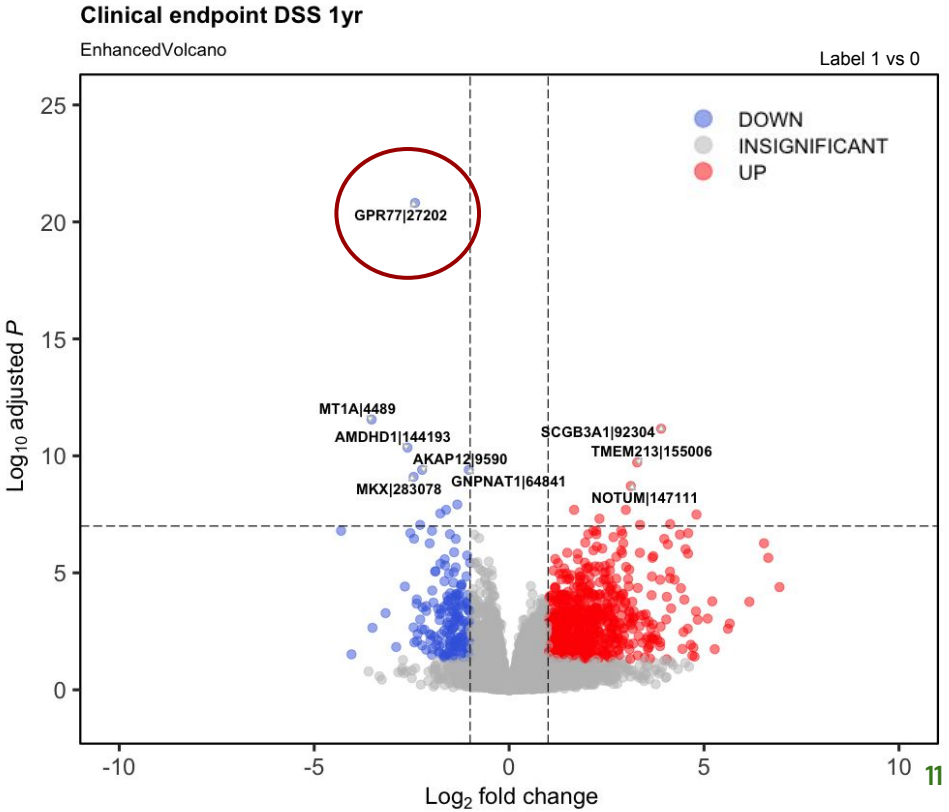


DESeq2 Result (Endpoint: 1 year)

- Role of GPR77 in lung cancer prognosis:
  GPR77+ carcinoma-associated fibroblasts (CAFs) sustain cancer stemness and promote tumor chemoresistance



Su et al., Cell (2018)



DESeq2 Result (Endpoint: 3 year)

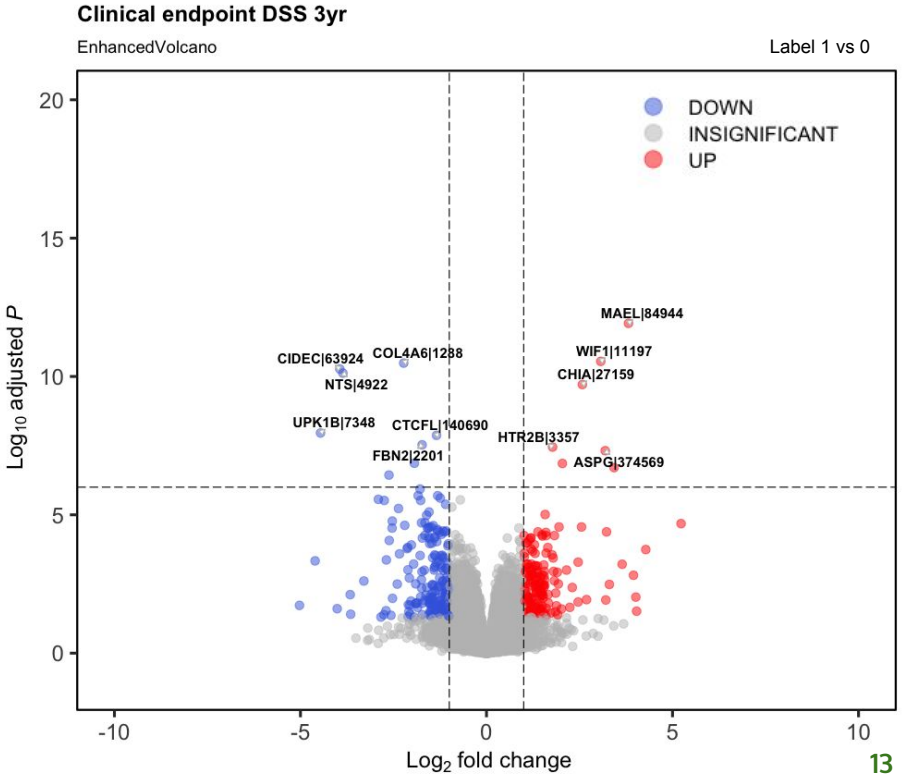
- **DESeq2 pre-filtering**:
  Remove genes with less than 10 counts (remove 1126 genes out of 19163 with nonzero total read count, 5.9%)
- **1606** genes (7.8%) being significant with padj < 0.05

GENE	log2FoldChange	pvalue	padj
MAEL	3.81582	6.64E-17	1.20E-12
WIF1	3.07129	3.15E-15	2.84E-11
COL4A6	-2.22539	5.45E-15	3.28E-11
CIDEC	-3.94203	1.19E-14	5.36E-11
NTS	-3.85991	2.10E-14	7.57E-11

DESeq2 Result (Endpoint: 3 year)

- **DESeq2 pre-filtering**:
  Remove genes with less than 10 counts (remove 1126 genes out of 19163 with nonzero total read count, 5.9%)
- **1606** genes (7.8%) being significant with padj < 0.05

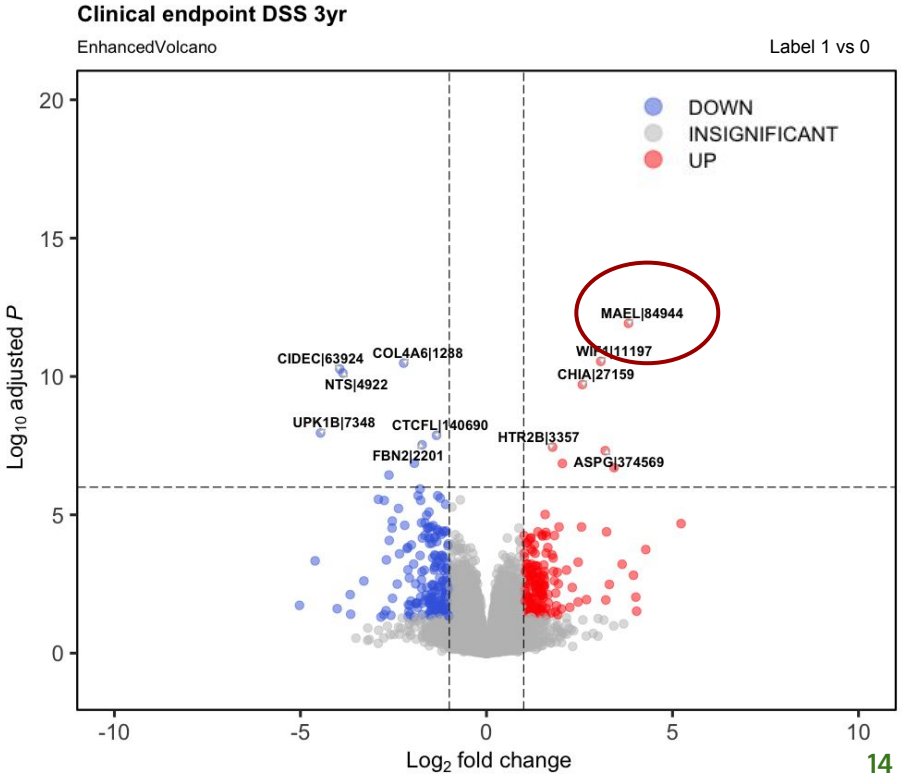
GENE	log2FoldChange	pvalue	padj
MAEL	3.81582	6.64E-17	1.20E-12
WIF1	3.07129	3.15E-15	2.84E-11
COL4A6	-2.22539	5.45E-15	3.28E-11
CIDEC	-3.94203	1.19E-14	5.36E-11
NTS	-3.85991	2.10E-14	7.57E-11



DESeq2 Result (Endpoint: 3 year)

- **DESeq2 pre-filtering**:
  Remove genes with less than 10 counts (remove 1126 genes out of 19163 with nonzero total read count, 5.9%)
- **1606** genes (7.8%) being significant with padj < 0.05

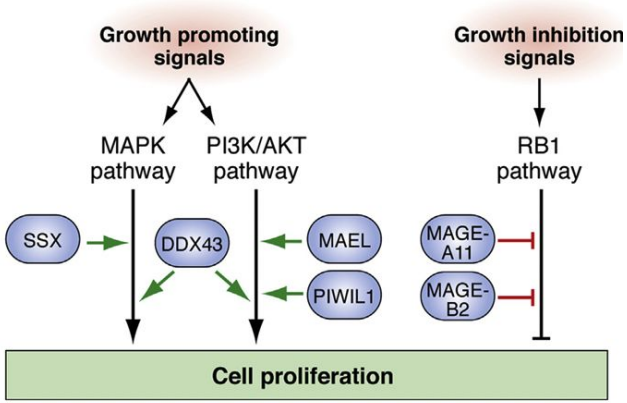
GENE	log2FoldChange	pvalue	padj
MAEL	**3.81582**	**6.64E-17**	**1.20E-12**
WIF1	3.07129	3.15E-15	2.84E-11
COL4A6	-2.22539	5.45E-15	3.28E-11
CIDEC	-3.94203	1.19E-14	5.36E-11
NTS	-3.85991	2.10E-14	7.57E-11



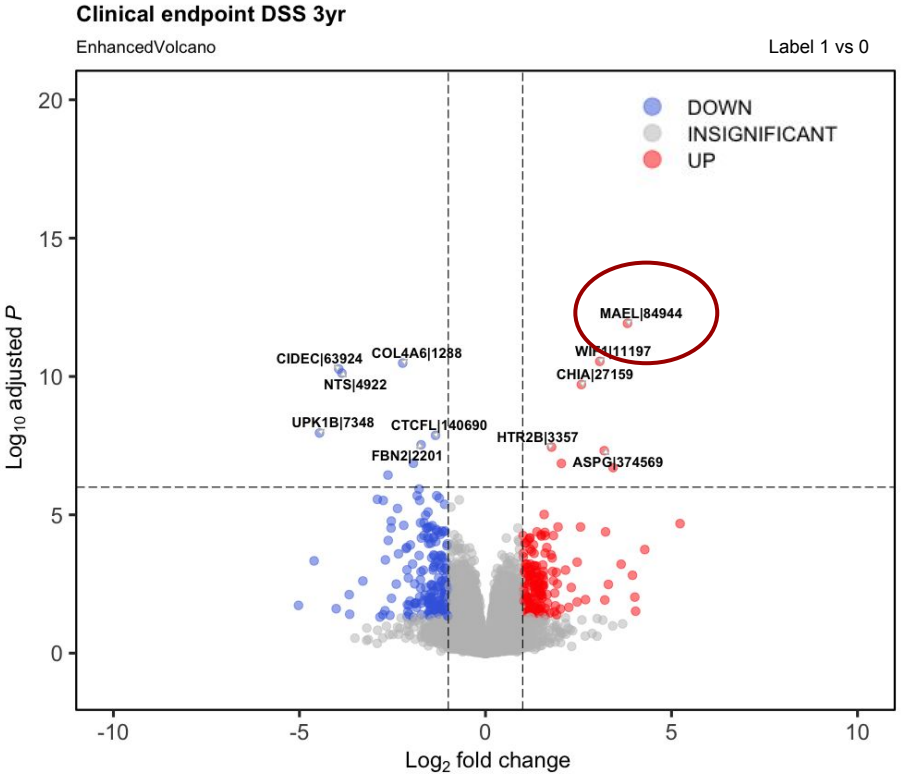
DESeq2 Result (Endpoint: 3 year)

- Roles of MAEL in lung cancer prognosis:

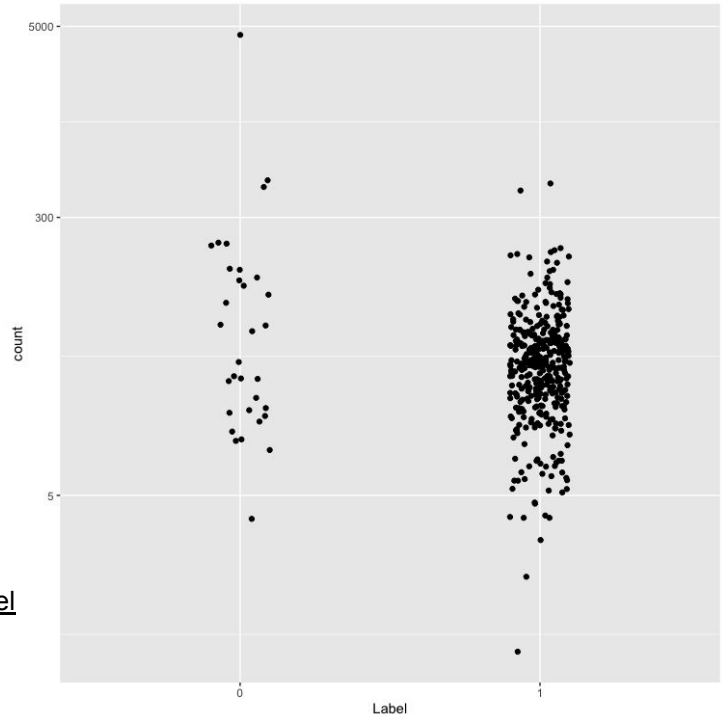
1. **Maelstrom (MAEL) is highly expressed in lung, liver, breast, bladder, and colorectal cancer**
2. **Correlated with epithelial-mesenchymal transition (EMT), tumor aggressiveness and poor prognosis outcome**



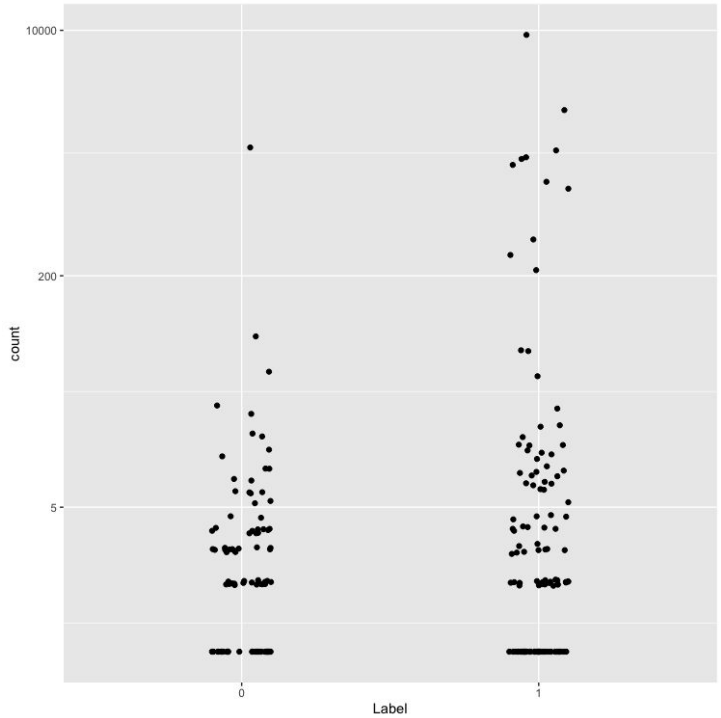
A.Van Tongelen et al. Cancer Letters (2017)



Read counts for the most significant genes



GPR77 (1 yr)



MAEL (3 yr)

Our classifier label
1 as living
0 as deceased

Feature Selection: XGBoost

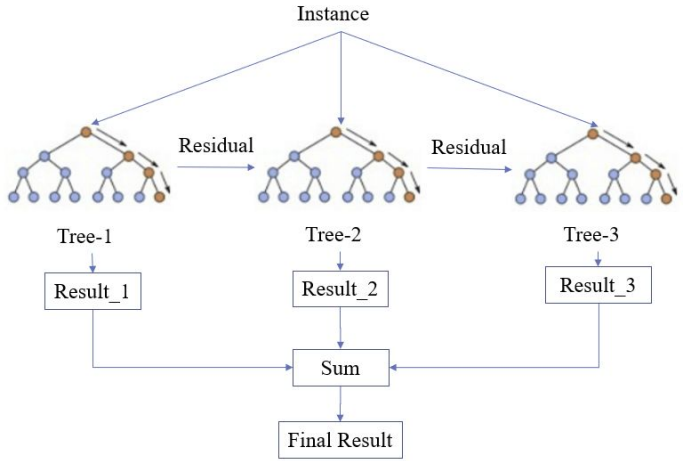
Selected Features for our models:

- Top 100 significant gene expression → DESeq2 Result
- Patient age → metadata dataset
- Tumor stage → metadata dataset
- Smoking status → metadata dataset
- Gender → metadata dataset
- Split our dataset into 80-20

Model: XGBoost

Combines weak predictive models (decision trees) to create a stronger predictive model

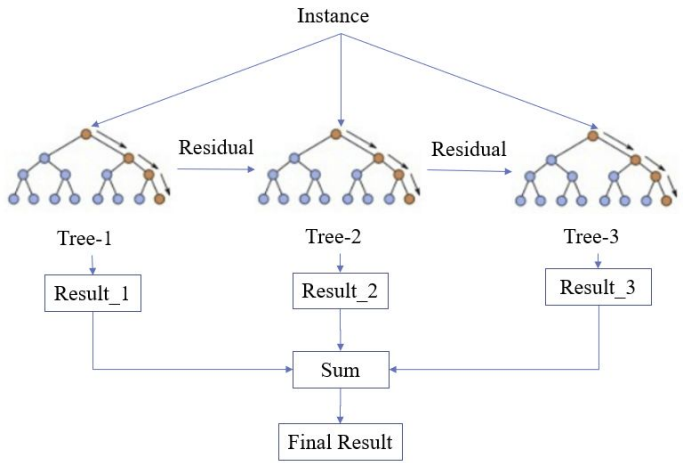
- Gradient boosting: each subsequent decision tree is trained to correct mistakes



Model: XGBoost

Combines weak predictive models (decision trees) to create a stronger predictive model

- Gradient boosting: each subsequent decision tree is trained to correct mistakes

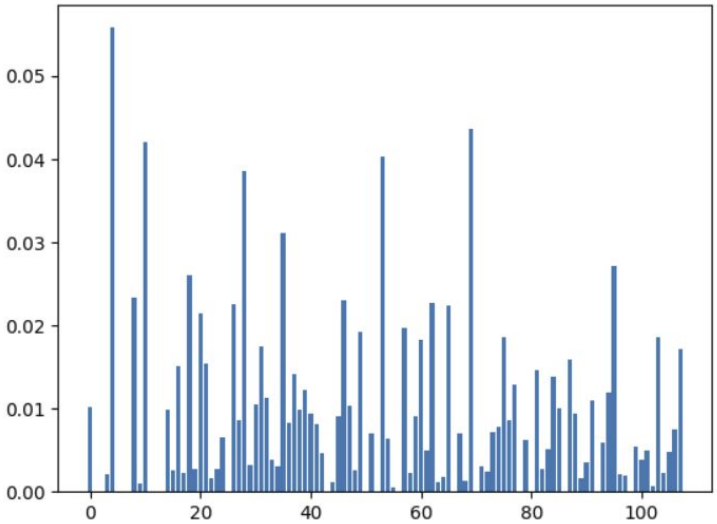


Parameters to vary:
- n_estimators
- max_depth
- learning_rate

Model: XGBoost

After training on the total feature set, XGBoost allows us to analyze the importance of the features using a significant threshold

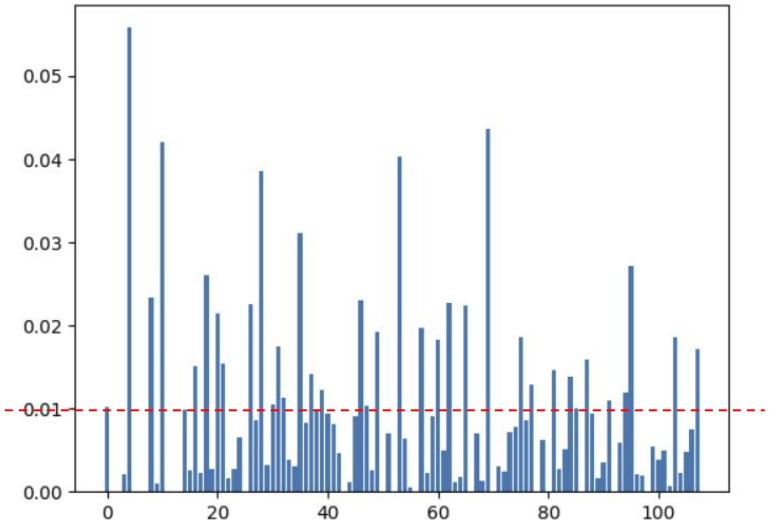
- Features above the threshold were used for classification



Model: XGBoost

After training on the total feature set, XGBoost allows us to analyze the importance of the features using a significant threshold

- Features above the threshold were used for classification

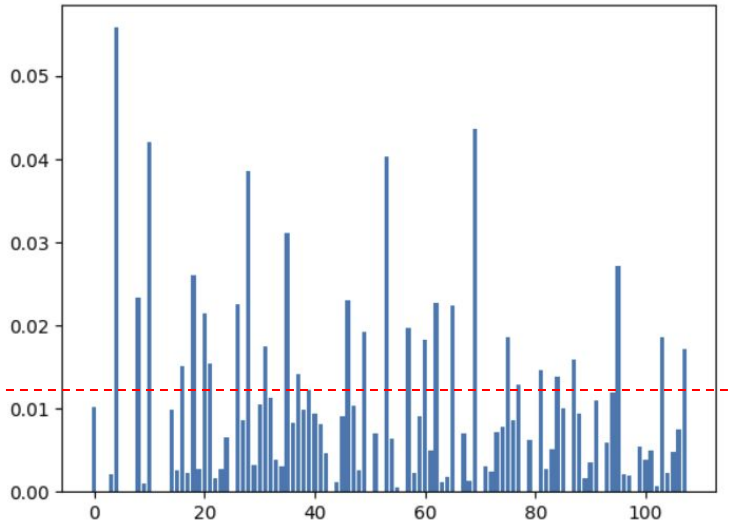


Iteration 1: threshold > 0.01

Model: XGBoost

After training on the total feature set, XGBoost allows us to analyze the importance of the features using a significant threshold

- Features above the threshold were used for classification

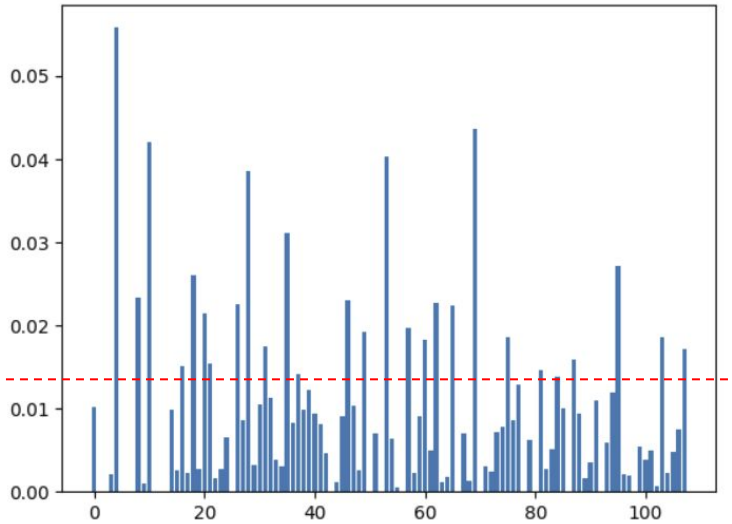


Iteration 2: threshold > 0.011

Model: XGBoost

After training on the total feature set, XGBoost allows us to analyze the importance of the features using a significant threshold

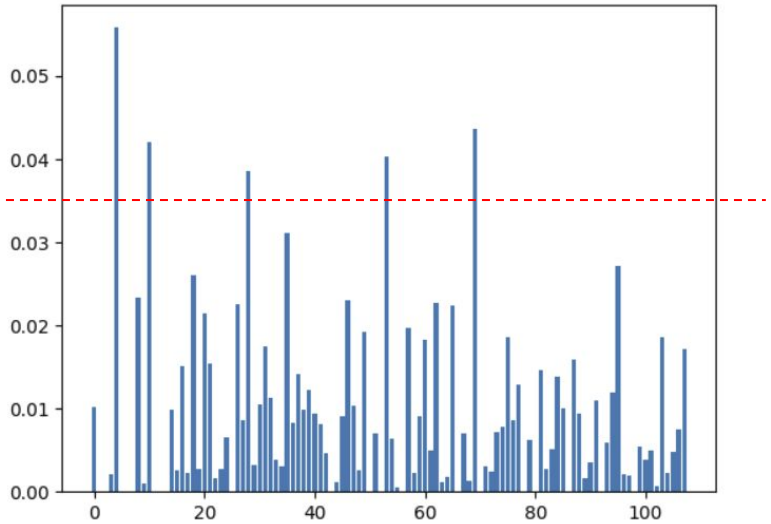
- Features above the threshold were used for classification



Model: XGBoost

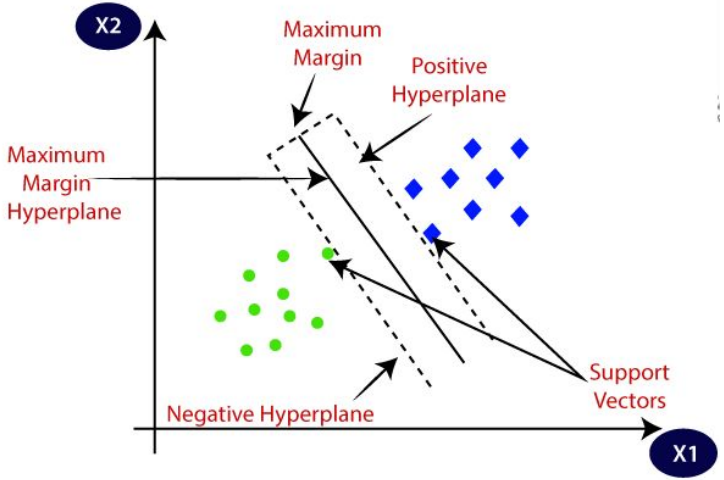
After training on the total feature set, XGBoost allows us to analyze the importance of the features using a significant threshold

- Features above the threshold were used for classification



Model: SVM

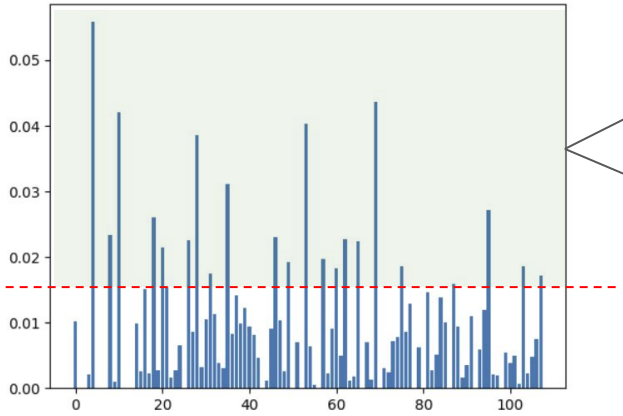
- Maps data to high-dimensional feature space to separate label data
- Looks to determine the optimal hyperplane that maximally separated the data points
  - Maximizing the margin
  - Maximizing the distance between the hyperplane and observations closest to the hyperplane
- Performed grid search varying the different parameter of SVM



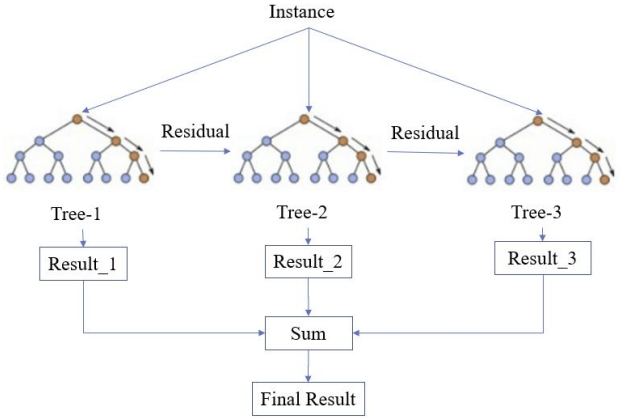
Parameters to Choose:
- Penalty
- kernel

Training/Testing

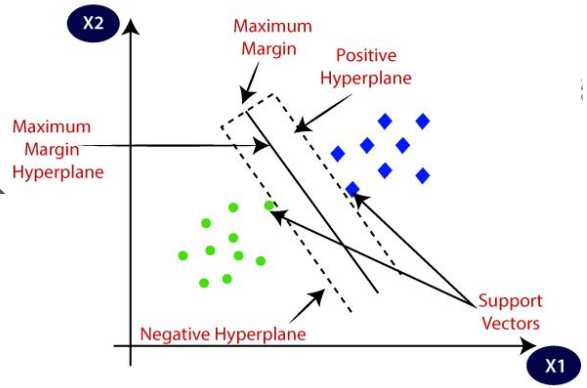
Feature Selection with XGBoost



XGBoost Classification



SVM Classification



Iterate over different significant thresholds + grid search over model parameters

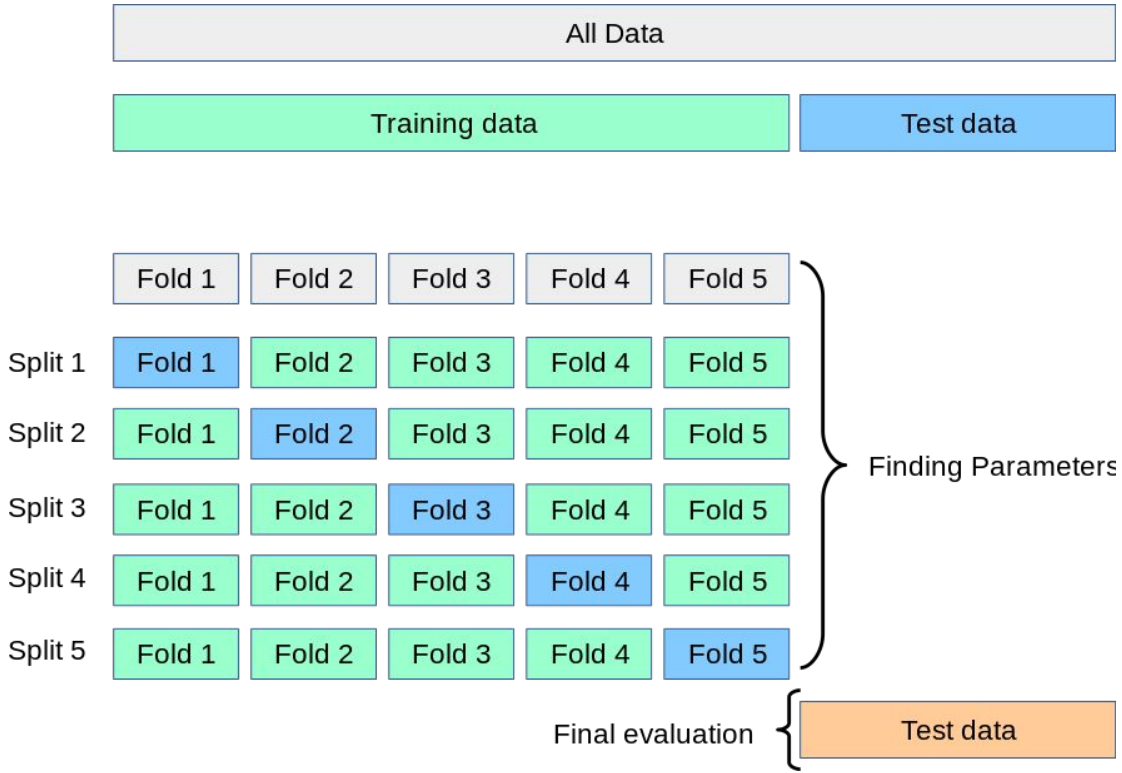
Training

Cross Validation

- 5 folds

Hyperparameter Tuning

- Grid search



Training

Cross Validation

- 5 folds

Hyperparameter Tuning

- Grid search

Parameters in XGBoost:
- N-estimators: 10, 50, 100
- Max_depth: 5, 10
- Learning_rate: 0.01, 0.1, 1, 10, 100

Parameters in SVM:
- C: 0.1, 0.5, 1, 2, 5, 10
- Kernel: Linear, RBF

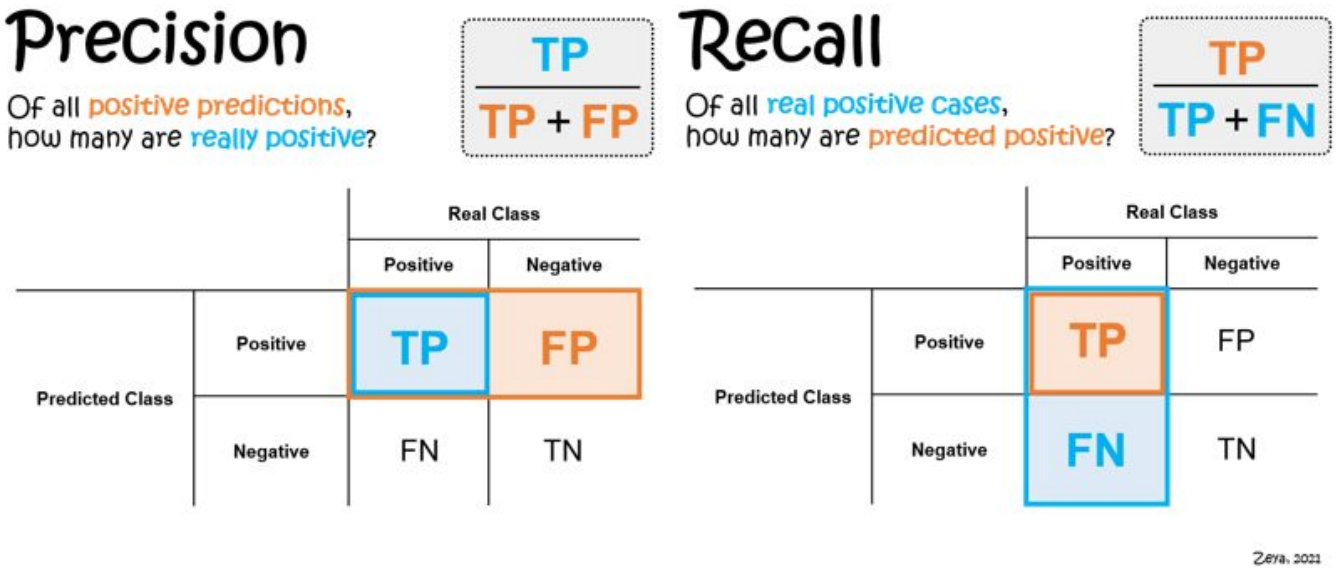
Evaluation

Accuracy:
0.85

Evaluation

Evaluation

- Accuracy
- F1-Score
  - **Imbalanced datasets**



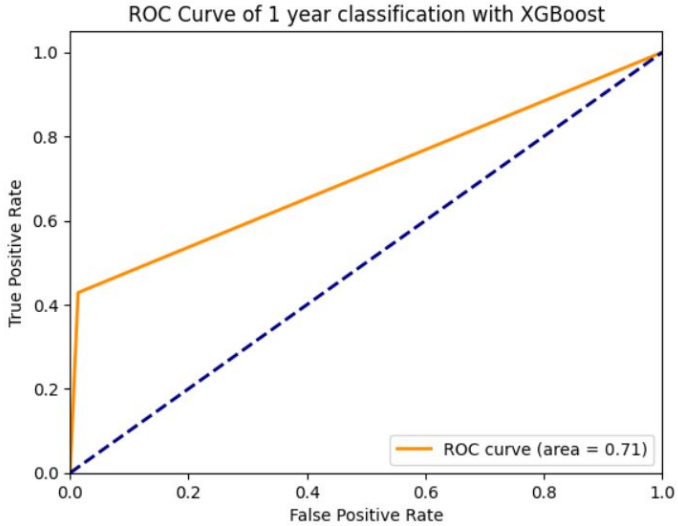
Evaluation

$$F1\ Score = 2 \times \frac{recall \times precision}{recall + precision}$$

RESULTS - 1 YEAR

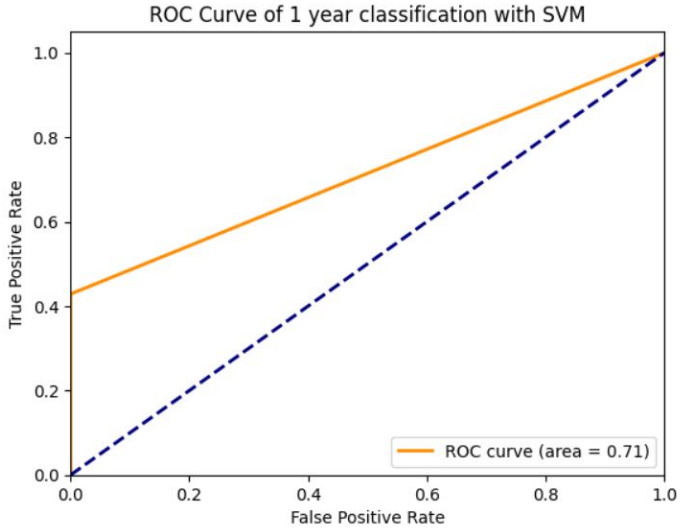
XGBoost

	Precision	Recall	F1
Alive	0.9481	0.9865	0.9669
Dead	0.7500	0.4286	0.5455
		Accuracy	0.9383



SVM

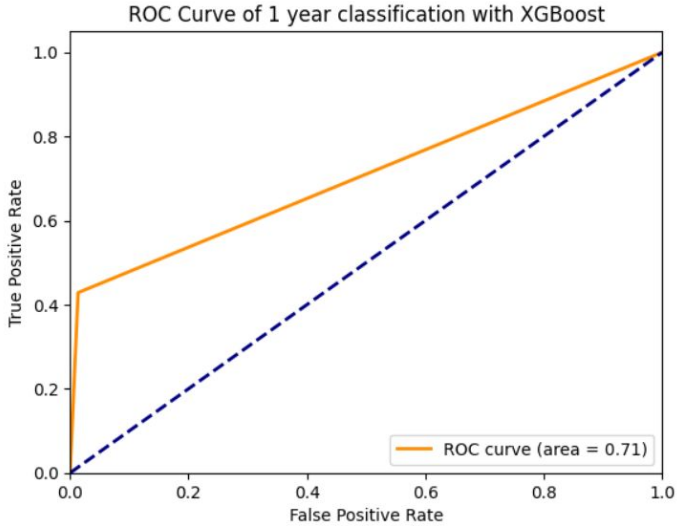
	Precision	Recall	F1
Alive	0.9487	1.0000	0.9737
Dead	1.0000	0.4286	0.6000
		Accuracy	0.9506



RESULTS - 1 YEAR

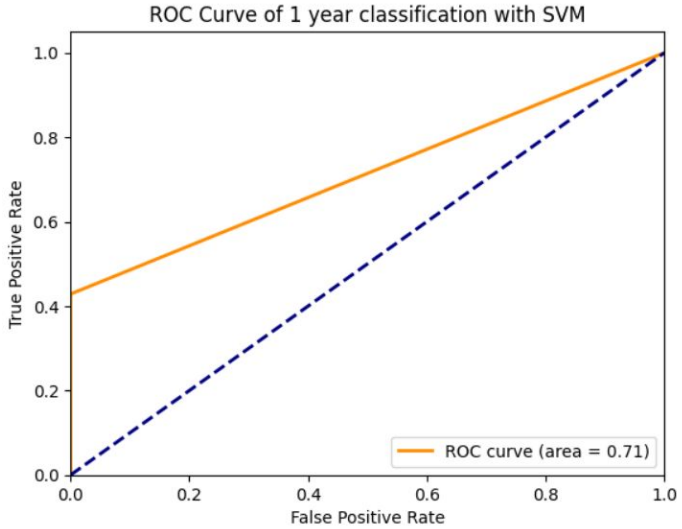
XGBoost

	Precision	Recall	F1
Alive	0.9481	0.9865	0.9669
Dead	0.7500	0.4286	0.5455
		Accuracy	0.9383



SVM

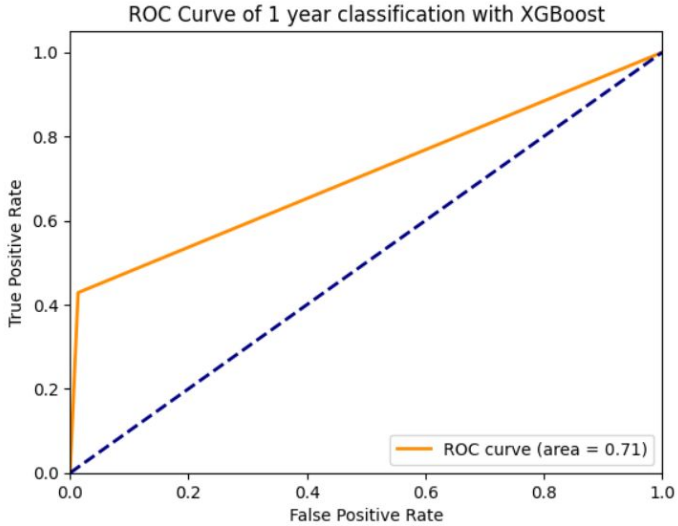
	Precision	Recall	F1
Alive	0.9487	1.0000	0.9737
Dead	1.0000	0.4286	0.6000
		Accuracy	0.9506



RESULTS - 1 YEAR

XGBoost

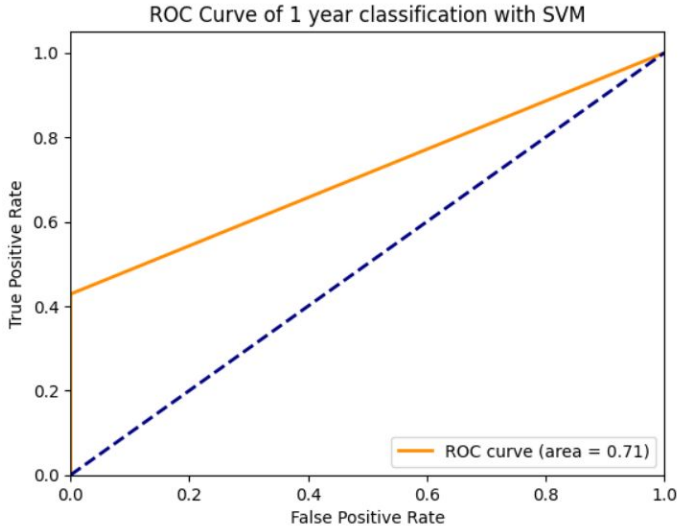
	Precision	Recall	F1
Alive	0.9481	0.9865	0.9669
Dead	0.7500	0.4286	0.5455
		Accuracy	0.9383



High accuracy expected, so we choose to focus on F1 score

SVM

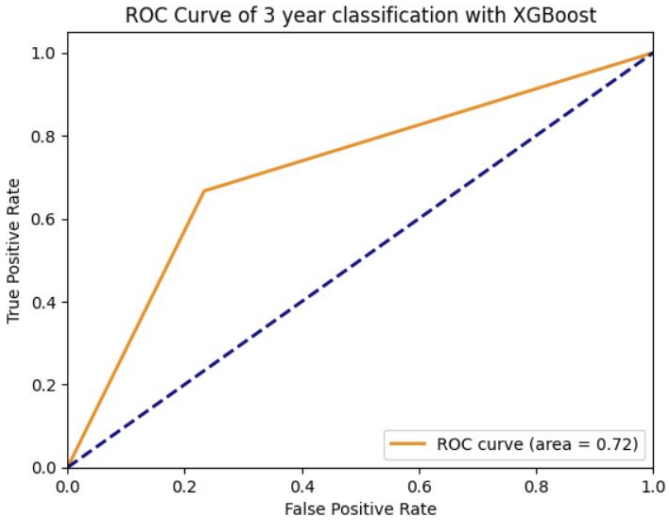
	Precision	Recall	F1
Alive	0.9487	1.0000	0.9737
Dead	1.0000	0.4286	0.6000
		Accuracy	0.9506



RESULTS - 3 YEAR

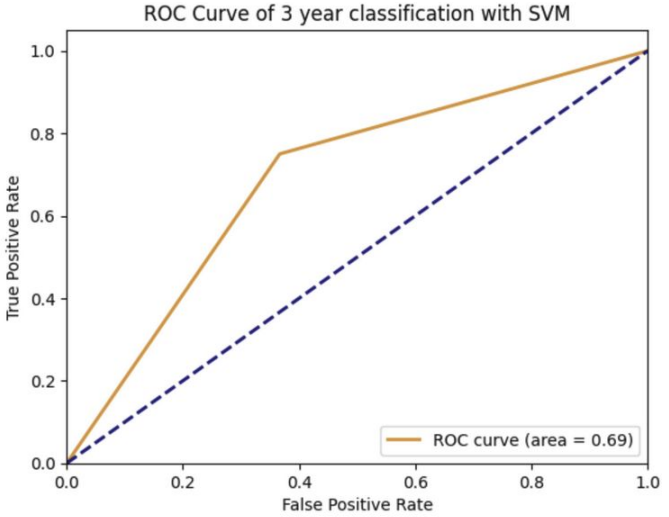
XGBoost

	Precision	Recall	F1
Alive	0.8148	0.7333	0.7719
Dead	0.4667	0.5833	0.5185
		Accuracy	0.6905



SVM

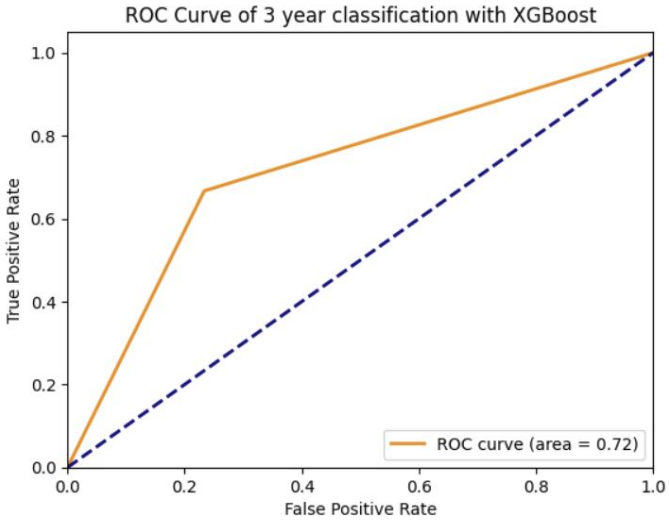
	Precision	Recall	F1
Alive	0.8696	0.6333	0.7308
Dead	0.4500	0.7500	0.5625
		Accuracy	0.6667



RESULTS - 3 YEAR

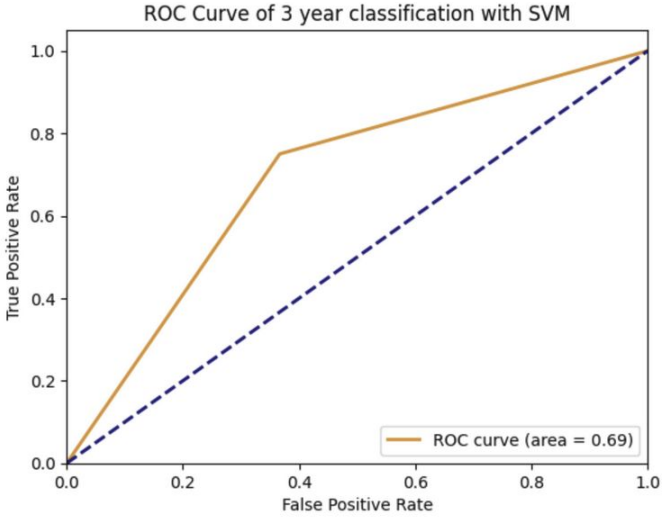
XGBoost

	Precision	Recall	F1
Alive	0.8148	0.7333	0.7719
Dead	0.4667	0.5833	0.5185
		Accuracy	0.6905



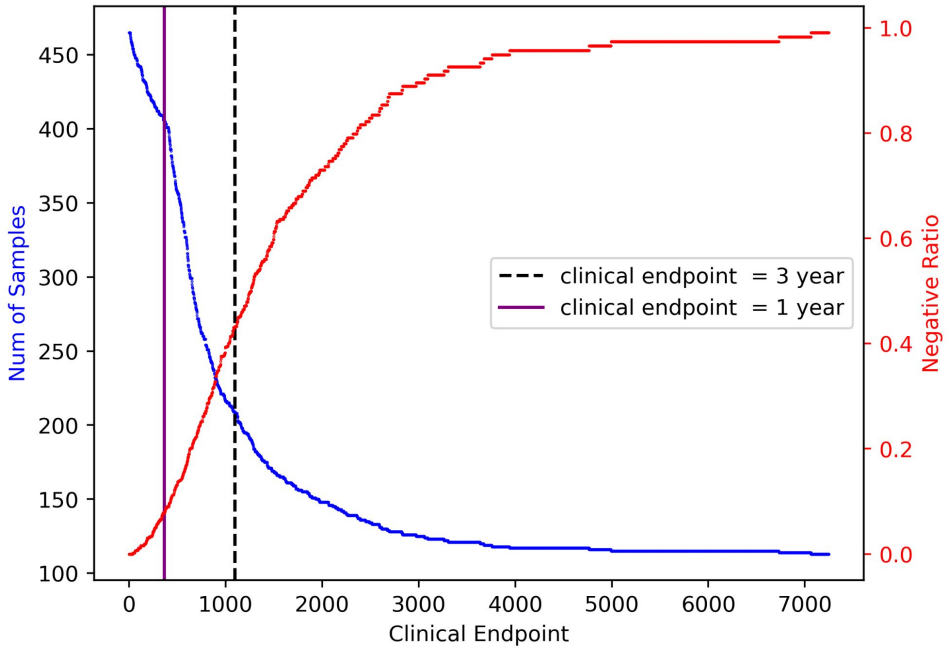
SVM

	Precision	Recall	F1
Alive	0.8696	0.6333	0.7308
Dead	0.4500	0.7500	0.5625
		Accuracy	0.6667



Discussion

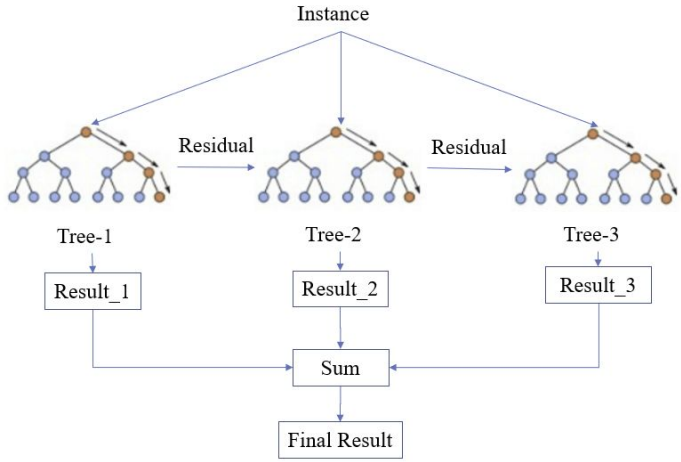
- Tradeoff between sample size and sample balance

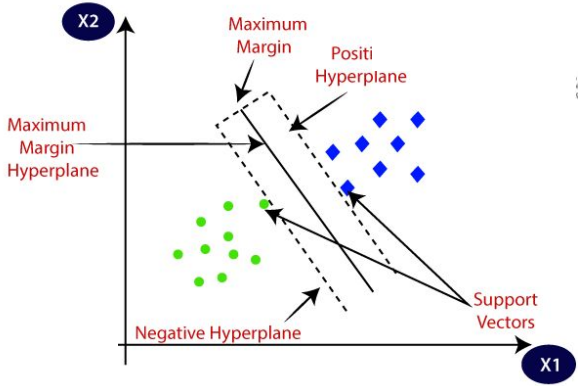


Discussion

- Trying different models
- Evaluation Metrics
  - Matthews Correlation Coefficient

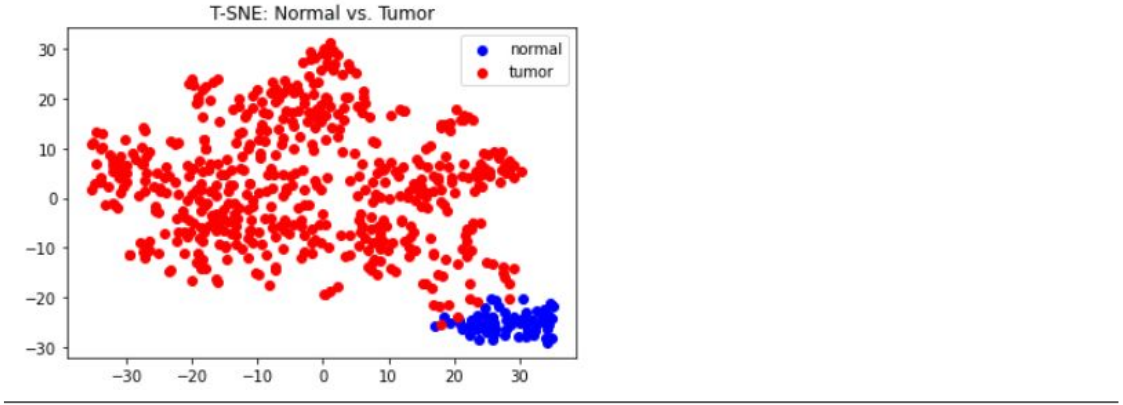
$$\mathrm{MCC} = \frac{\mathrm{TP} \times \mathrm{TN} - \mathrm{FP} \times \mathrm{FN}}{\sqrt{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}}$$

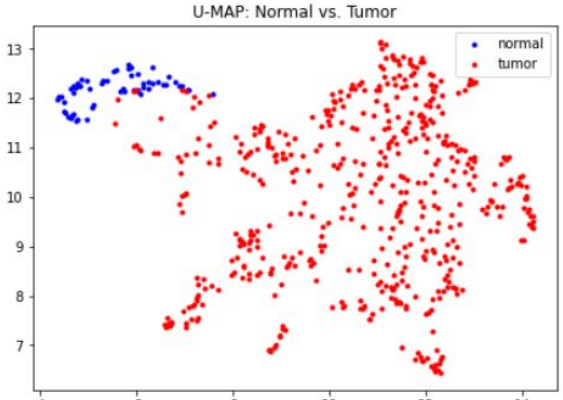




Discussion

- PCA, UMAP, t-SNE





Discussion

- Including mutation data
- Ran DESeq2 on training data vs. full

	patient_id	Hugo_Symbol	Entrez_Gene_Id	Center	NCBI_Build	Chromosome	Start_Position	End_Position	Strand	Variant_Classification
0	TCGA-05-4244	CPN1	0	.	GRCh37	10	101814119	101814119	+	Missense_Mutation
1	TCGA-05-4244	MKI67	0	.	GRCh37	10	129902901	129902901	+	Silent
2	TCGA-05-4244	NEBL	0	.	GRCh37	10	21104601	21104606	+	In_Frame_Del
3	TCGA-05-4244	RP11-445N18.7	0	.	GRCh37	10	45652518	45652518	+	RNA
4	TCGA-05-4244	ERCC6	0	.	GRCh37	10	50667200	50667200	+	Silent

THANK YOU!

- The coolest group, group 3