ML Based Tumor Classification Using Pan–Cancer RNA-Seq Data

Beck Schemenauer & Gavin Lynch

Context & Motivation

- **Tumor Identification:** Early detection is essential because finding a tumor sooner greatly improves treatment success and patient outcomes.
- **Tumor Distinction:** Different tumor types require different therapies, so identifying what kind of cancer it is guides proper treatment decisions.
- **Why Machine Learning:** Gene expression data is extremely high-dimensional, and ML is needed to discover patterns that humans cannot analyze manually.

Biological Definitions

Gene

A segment of DNA that contains the instructions for making proteins

Gene-Expression

A numerical measure of how active each gene is in a cell

PAN-Cancer

A dataset or analysis that compares multiple cancer types together.

RNA-Seq

A sequencing method used to measure gene expression levels.

Methodology

1 **EDA**

Explore class distribution and expression ranges

2 **PCA**

Visualize class separation

3 **Feature Selection**

Select most informative genes

4 **Prediction**

Train models to classify tumor type

5 **Baseline Comparison**

Evaluate against random features

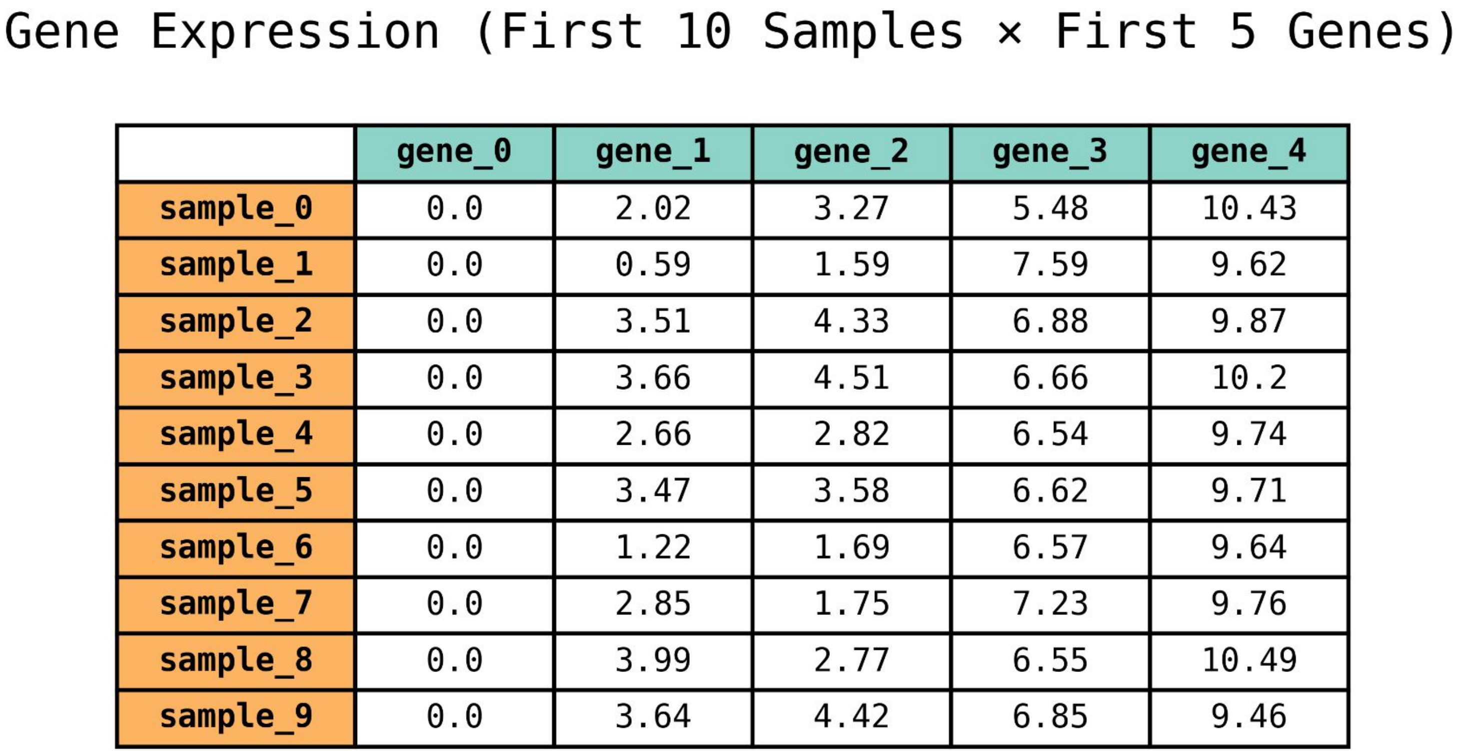
6 **Bio-Interpretation**

Explore gene-specific biological meaning

1

EDA

Gain an initial understanding of the dataset by examining its structure, class balance, and overall behavior.

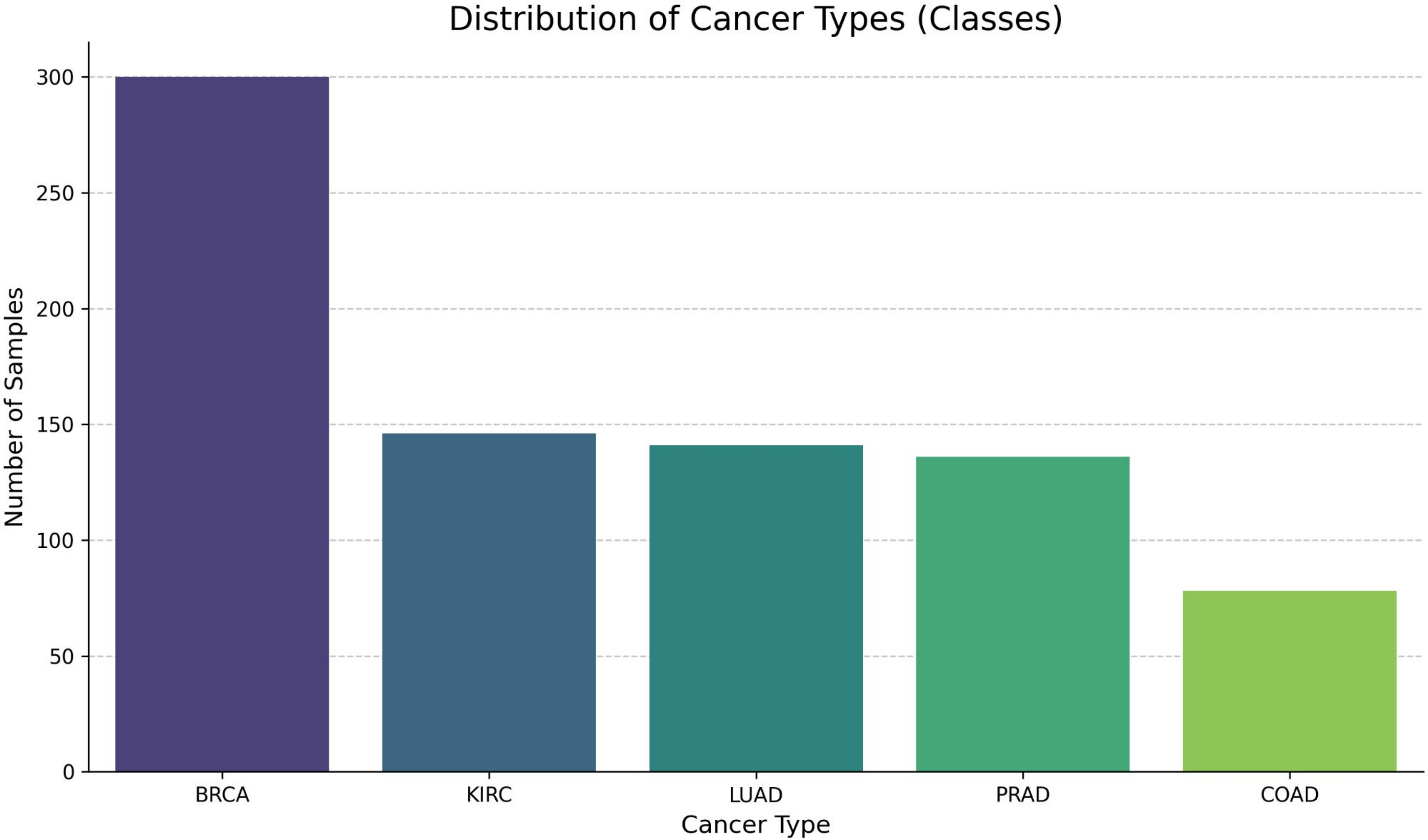


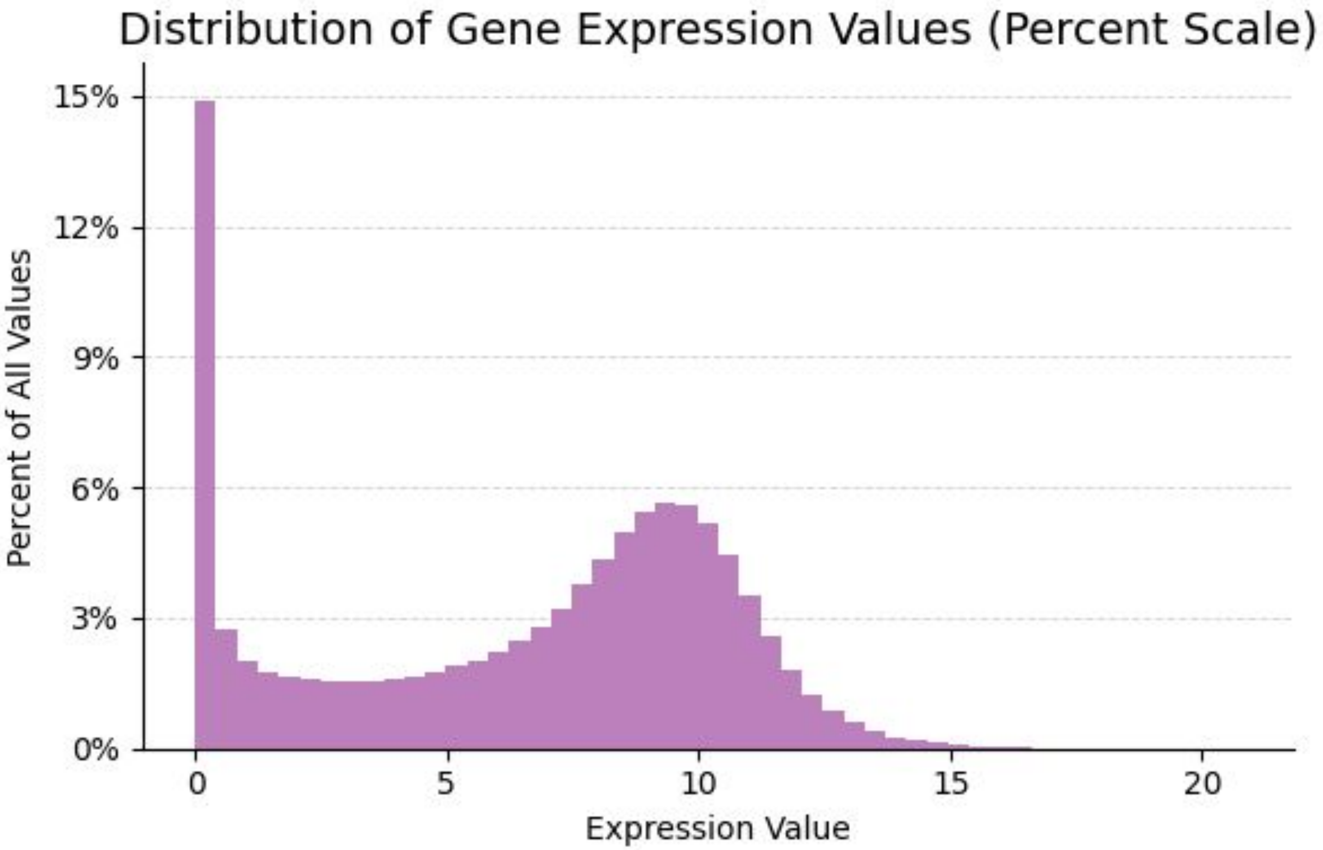
Gene Expression (First 10 Samples × First 5 Genes)

	gene_0	gene_1	gene_2	gene_3	gene_4
sample_0	0.0	2.02	3.27	5.48	10.43
sample_1	0.0	0.59	1.59	7.59	9.62
sample_2	0.0	3.51	4.33	6.88	9.87
sample_3	0.0	3.66	4.51	6.66	10.2
sample_4	0.0	2.66	2.82	6.54	9.74
sample_5	0.0	3.47	3.58	6.62	9.71
sample_6	0.0	1.22	1.69	6.57	9.64
sample_7	0.0	2.85	1.75	7.23	9.76
sample_8	0.0	3.99	2.77	6.55	10.49
sample_9	0.0	3.64	4.42	6.85	9.46

801 Total Samples

20,531 Total Features


Distribution of Cancer Types (Classes)
Number of Samples
0
50
100
150
200
250
300
BRCA
KIRC
LUAD
PRAD
COAD
Cancer Type

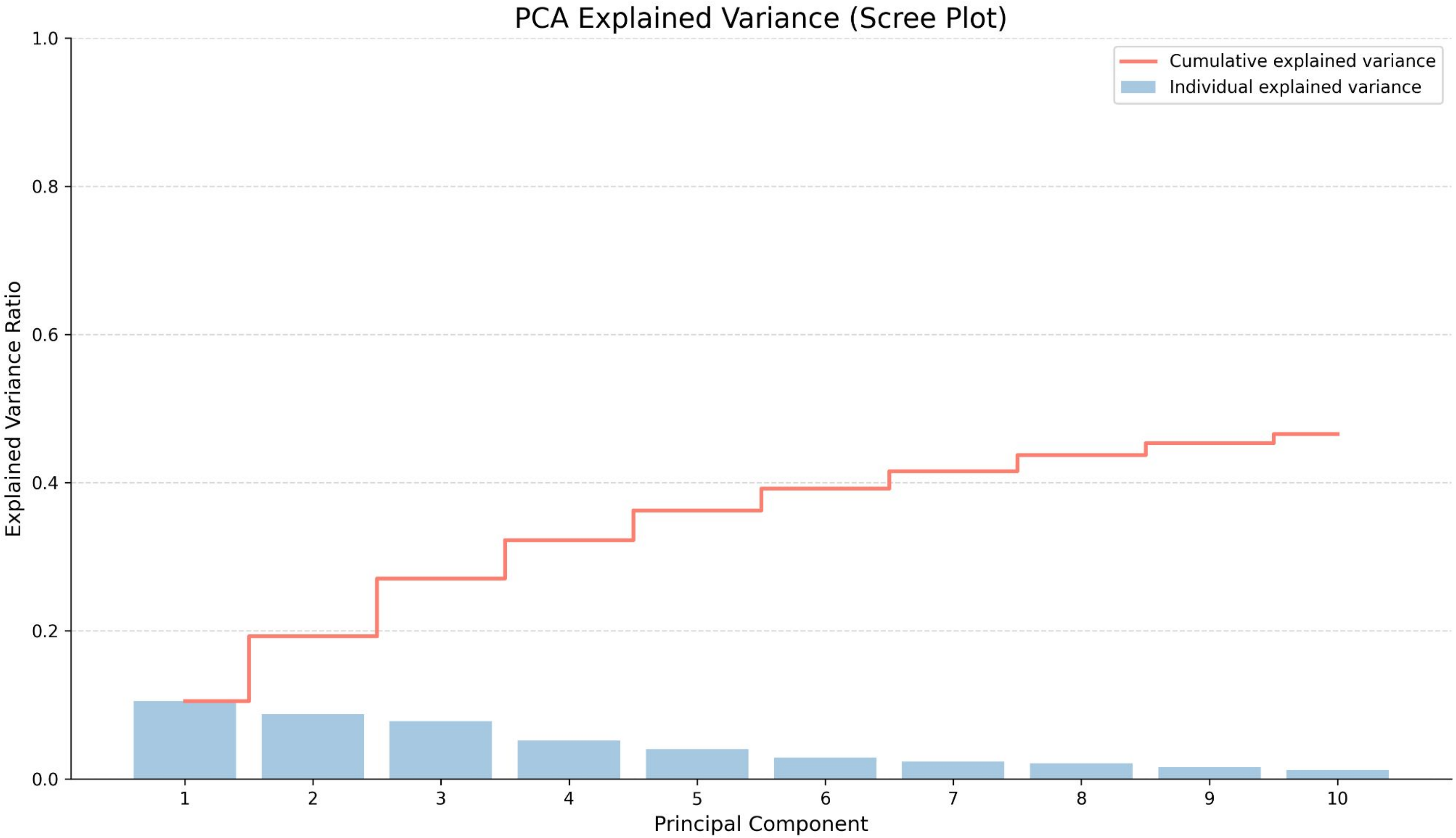


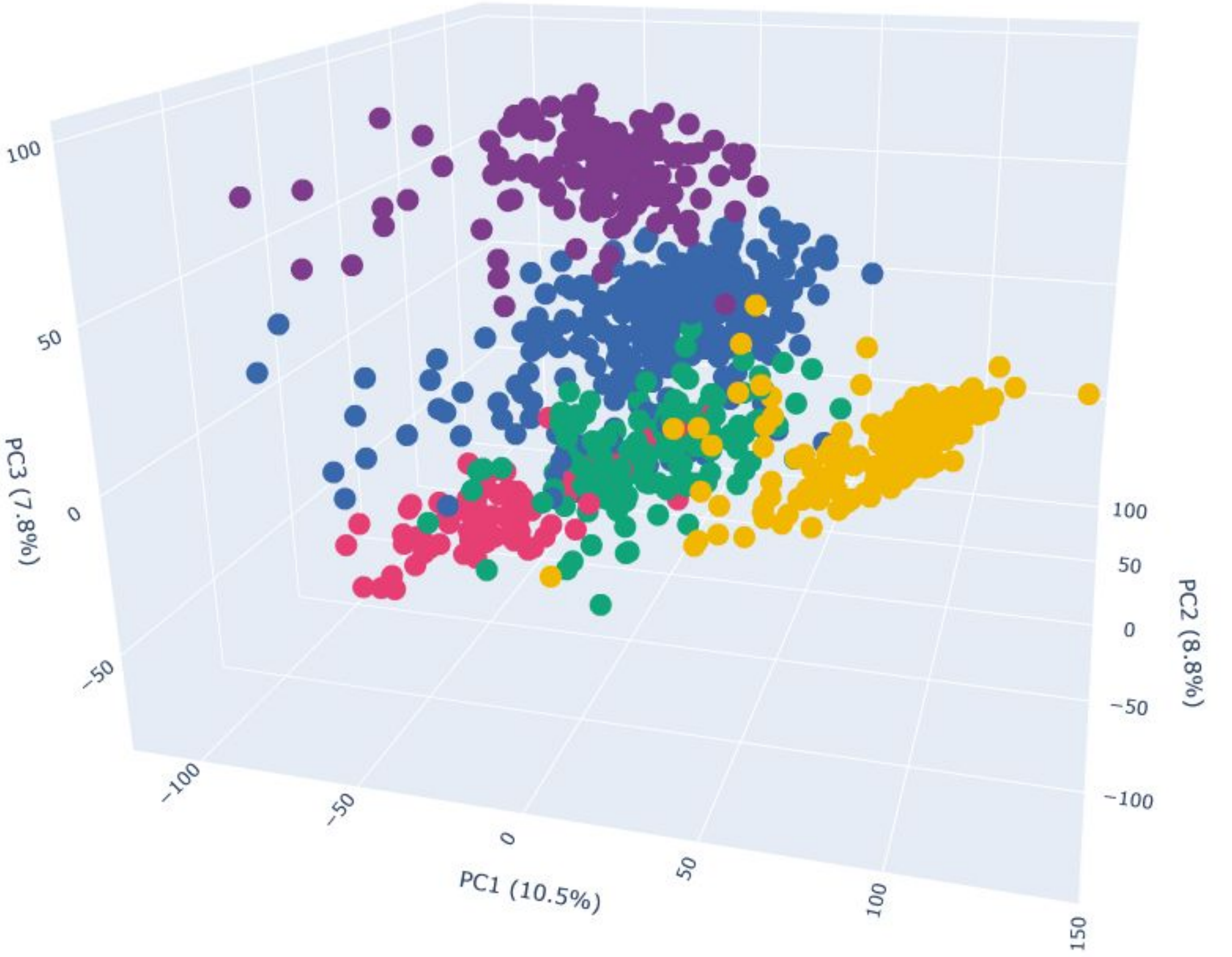
Distribution of Gene Expression Values (Percent Scale)
Percent of All Values
15%
12%
9%
6%
3%
0%
0
5
10
15
20
Expression Value

2

PCA

Project the data into lower dimensions to reveal visual patterns and highlight how well cancer types naturally separate.


PCA Explained Variance (Scree Plot)
Explained Variance Ratio
Principal Component
1.0
0.8
0.6
0.4
0.2
0.0
1
2
3
4
5
6
7
8
9
10
Cumulative explained variance
Individual explained variance


PC3 (7.8%)
100
50
0
−50
PC1 (10.5%)
−100
−50
0
50
100
150
PC2 (8.8%)
100
50
0
−50
−100

3

Feature Selection

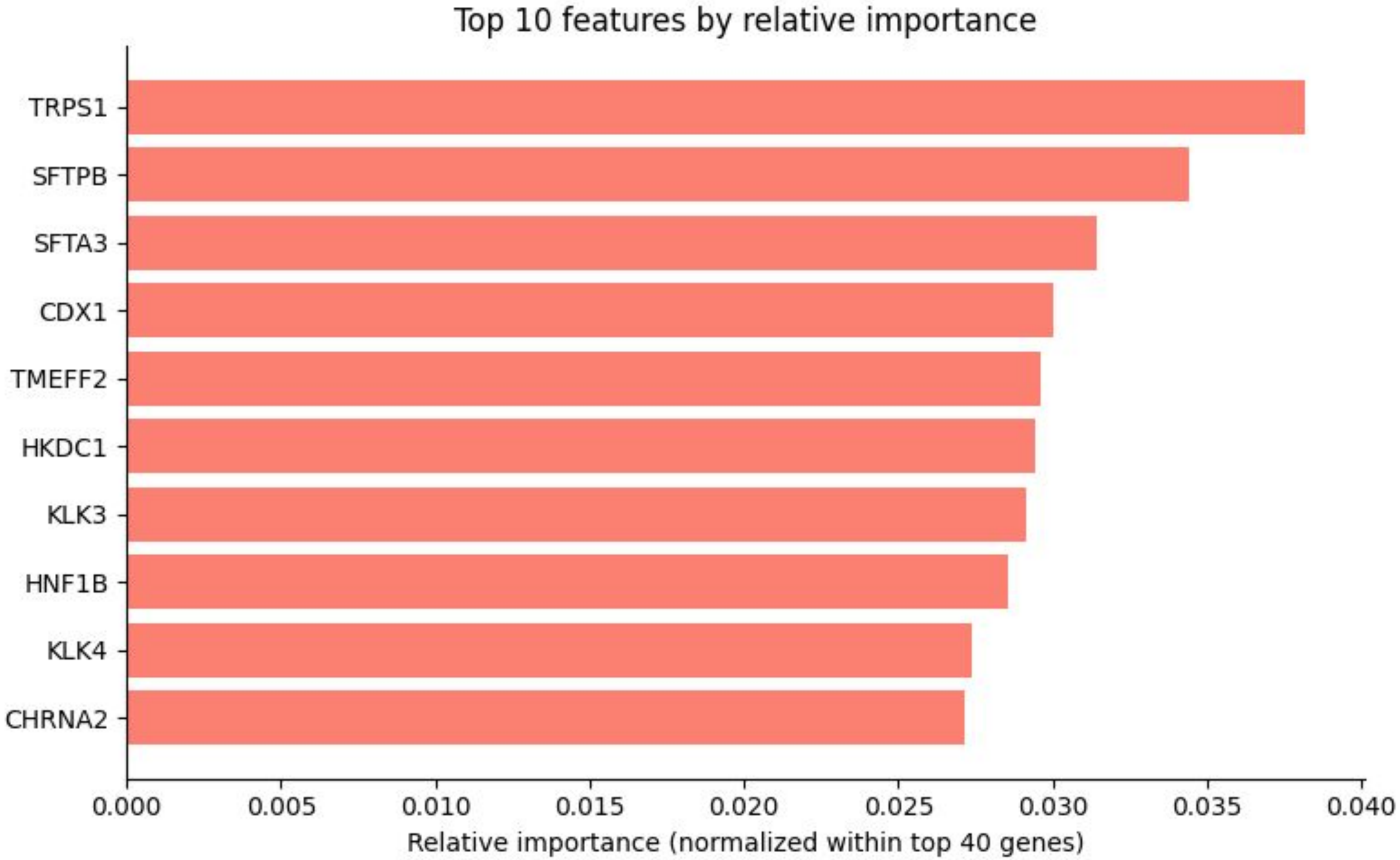
Identify the subset of genes that carry the strongest signal for distinguishing between cancer types.

Why Feature Selection?

- **Machine Learning:** Reduces noise and overfitting in high-dimensional data, improving model stability and lowering computational cost.
- **Biology:** Enables tracing gene functions and mapping selected genes to pathways, revealing meaningful biological mechanisms.
- **Practical Impact:** Supports low-cost, targeted sequencing panels and increases interpretability for doctors, with potential relevance for diagnostics or drug targets.

Feature Selection Process

- Start with full dataset (X, y)
- Perform outer 5-fold stratified cross-validation to create train/test splits
- Within each outer-train split, run an inner 5-fold CV
    - Train a Random Forest on each inner fold
    - Collect feature importances from each model
- Sum importances across inner folds to get an importance score per gene
- Rank all genes by their summed importance scores


Top 10 features by relative importance
TRPS1
SFTPB
SFTA3
CDX1
TMEFF2
HKDC1
KLK3
HNF1B
KLK4
CHRNA2
0.000
0.005
0.010
0.015
0.020
0.025
0.030
0.035
0.040
Relative importance (normalized within top 40 genes)

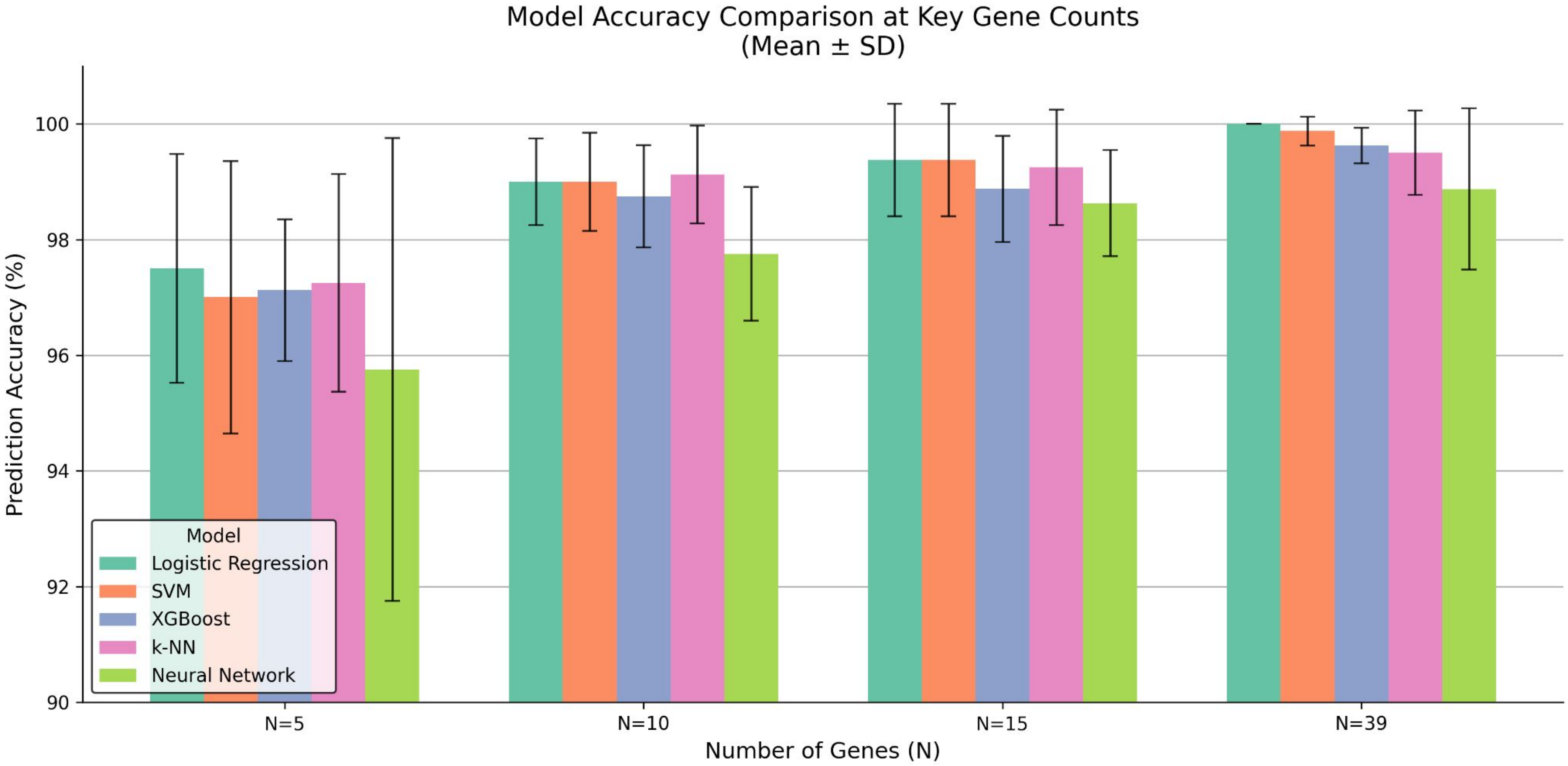
4

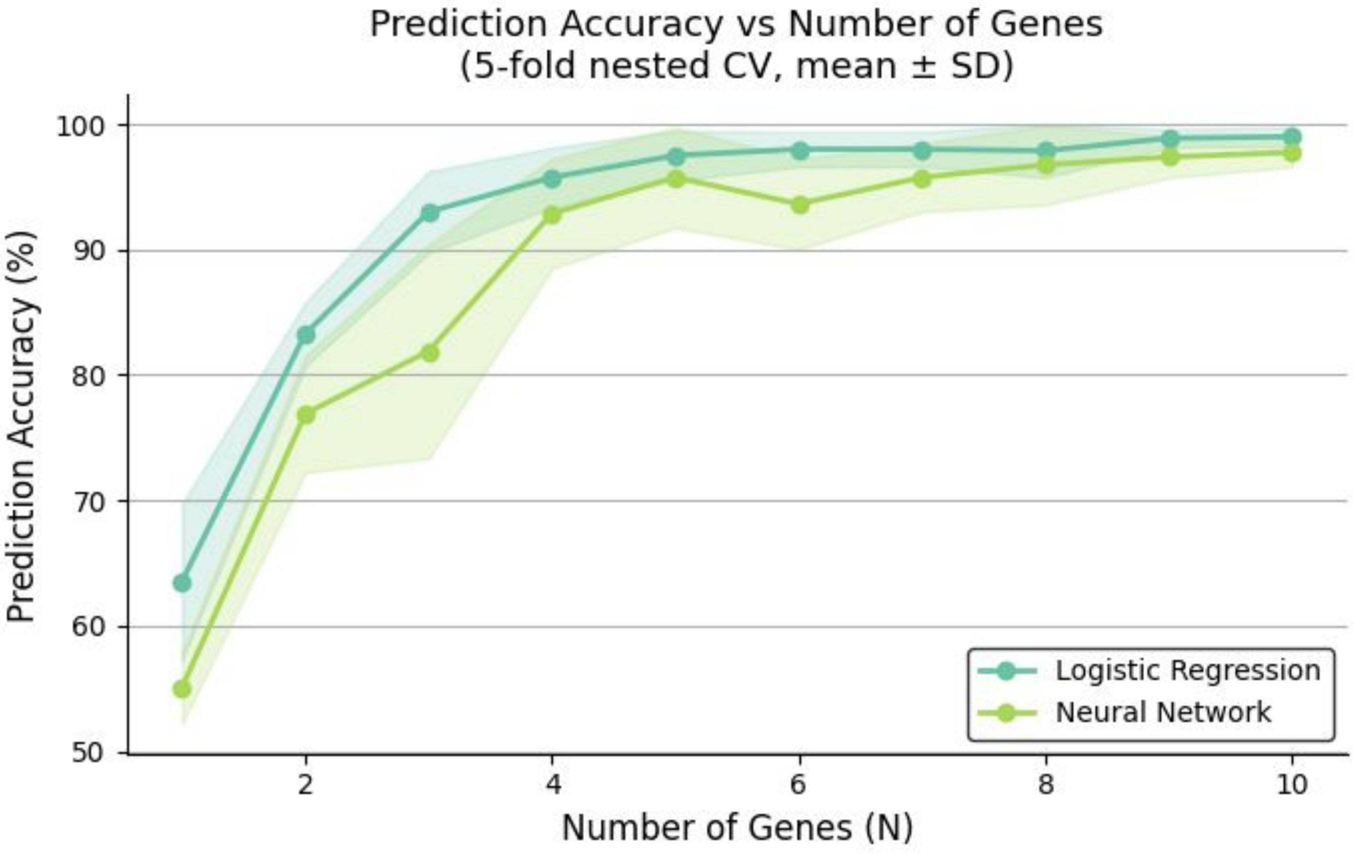
Prediction

Train machine-learning models to classify samples into tumor types based on selected gene features.

Models & Motives

- **Logistic Regression:** Tumor types can differ by strong, shifts in just a few key gene expression levels.
- **SVM:** PCA showed clear separation between tumor types with just 3 dimensions, hyperplanes could likely separate higher-dimensional data easily.
- **XGBoost:** Useful for modeling non-linear gene interactions that may define cancer subtypes more subtly.
- **kNN:** Samples cluster tightly by tumor type, so neighbors provide reliable labels.
- **Neural Network:** Helps capture multi-gene patterns or more abstract pathway signals that simpler models may miss.


Model Accuracy Comparison at Key Gene Counts
(Mean ± SD)
Prediction Accuracy (%)
100
98
96
94
92
90
N=5
N=10
N=15
N=39
Number of Genes (N)
Model
Logistic Regression
SVM
XGBoost
k-NN
Neural Network


Prediction Accuracy vs Number of Genes
(5-fold nested CV, mean ± SD)
Prediction Accuracy (%)
100
90
80
70
60
50
2
4
6
8
10
Number of Genes (N)
Logistic Regression
Neural Network

Prediction Summary

- Across all models prediction accuracy reached 99–100% with < 50 genes.
- Only 3-10 top-ranked genes were needed to achieve near-maximum performance.
- Accuracy curves flatten early, adding more genes provides minimal additional benefit.
- Performance was robust across models, indicating the signal is strong and model-independent.
- The 3D PCA plot already showed clear class separation, helping explain why models classify so accurately.

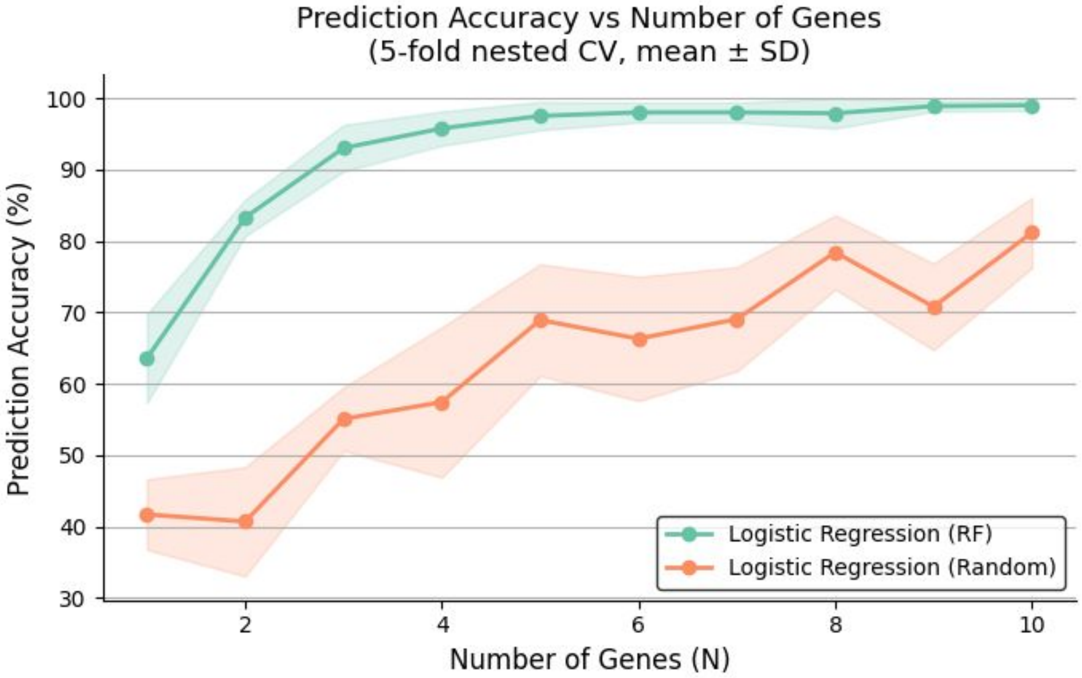
5

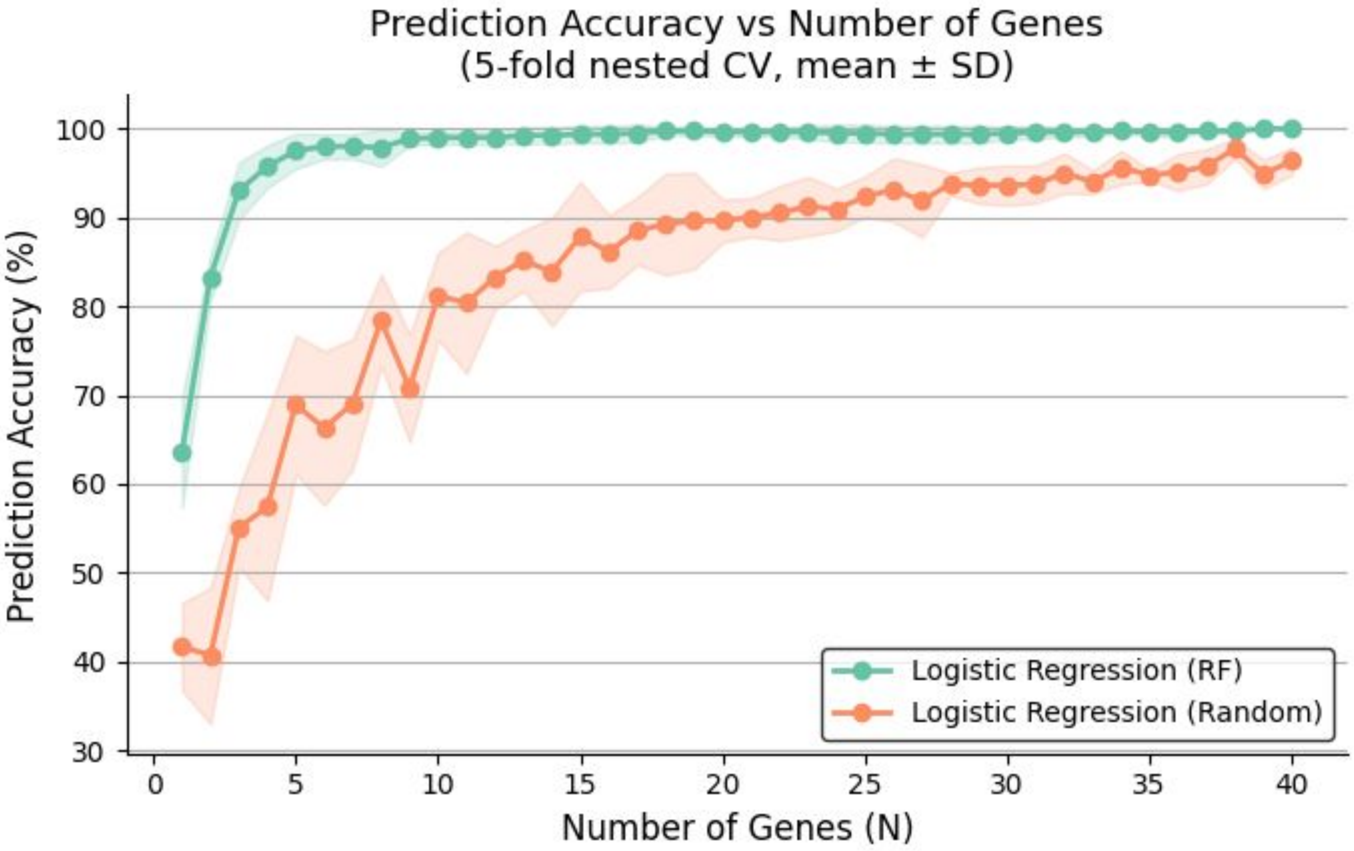
Baseline Comparison

Test whether RF-selected genes outperform randomly chosen genes in predicting tumor type.

Why Baseline Comparison?

- Show our selected genes add real predictive value.
- Check whether accuracy could occur just by chance.
- Validate that feature selection is actually helping.





Prediction Accuracy vs Number of Genes
(5-fold nested CV, mean ± SD)

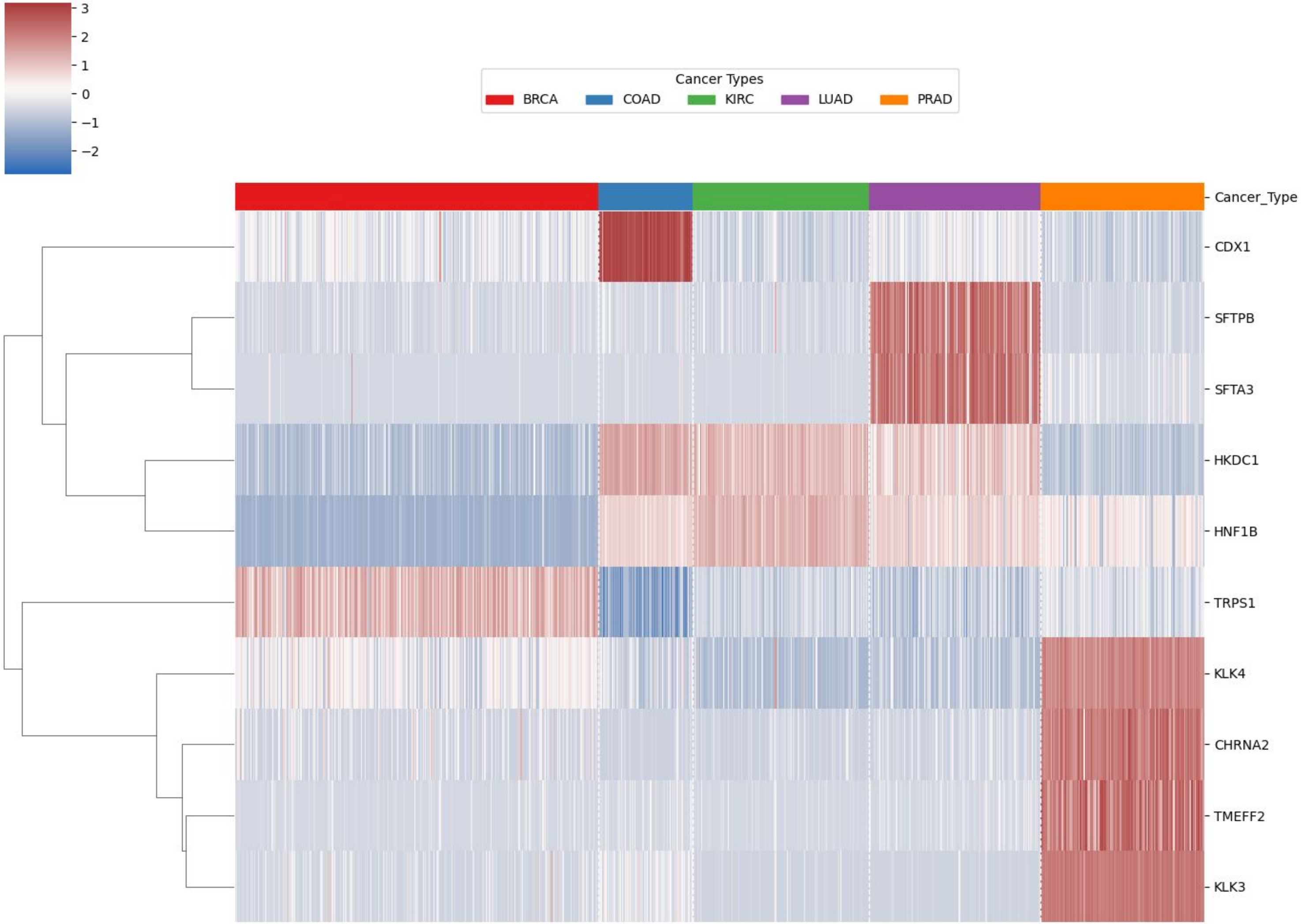
Comparison Summary

- High accuracy with random genes is expected when using many features, but unstable and unreliable.
- Feature selection achieves similar or higher accuracy with far fewer genes, proving it captures meaningful biological signal.
- Selected genes yield more stable performance across folds and a higher peak accuracy compared to random baselines.

6

Bio-Interpretation

Investigate the biological roles of top genes to uncover potential functional relevance or cancer-related mechanisms.


3
2
1
0
−1
−2
Cancer Types
BRCA
COAD
KIRC
LUAD
PRAD
Cancer_Type
CDX1
SFTPB
SFTA3
HKDC1
HNF1B
TRPS1
KLK4
CHRNA2
TMEFF2
KLK3

Gene Functional Analysis

Top 4 Genes	Function in the Body
CDX1	Codes for proteins involved in intestinal function, highly linked to colon cancer
SFTPB	Codes for lung lining proteins, linked to lung disease
TMEFF2	Tumor suppressing gene, highly linked to prostate cancer
TRPS1	Codes for connective tissue proteins, linked to breast cancer proliferation

Thanks