Created by: Rotem Fisher & Orian Aziz

FROM GENE EXPRESSION TO CLINICAL OUTCOMES

GITHUB REPO

BACKGROUND & MOTIVATION: THE NEED FOR PRECISION ONCOLOGY

THE CLINICAL CHALLENGE

- Traditional cancer prognosis relies mainly on clinical stages (I-IV).
- However, patients with the same stage often have drastically different survival

THE BIOLOGICAL INSIGHT

- Cancer is a disease of the genome.
- Gene expression profiles (RNA-Seq) capture the tumor's molecular activity.

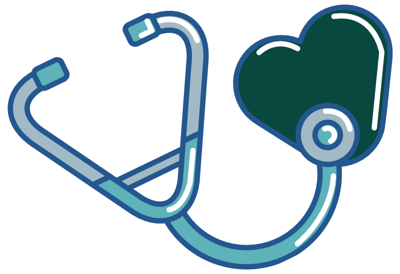
THE OPPORTUNITY

Using Machine Learning to identify complex genomic patterns that predict patient survival better than random chance.

RESEARCH QUESTION

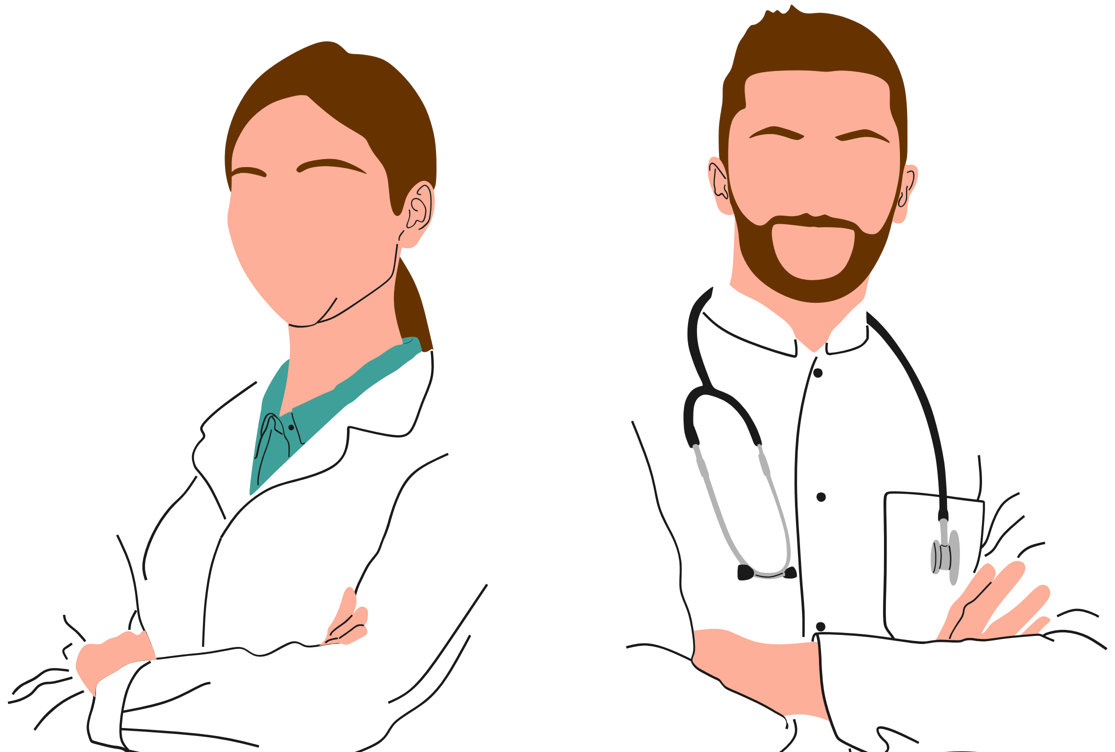
- **Primary Goal:** To identify a minimal set of genomic signatures that can accurately stratify patients into high-risk and low-risk survival groups.
- **Key Hypothesis:** Specific gene expression patterns are strongly correlated with patient prognosis, independent of clinical stage.
- **Comparison:** Can Machine Learning models (e.g., Random Forest) outperform traditional cancer prognosis ?

THE DATASET (TCGA PAN-CANCER ATLAS)



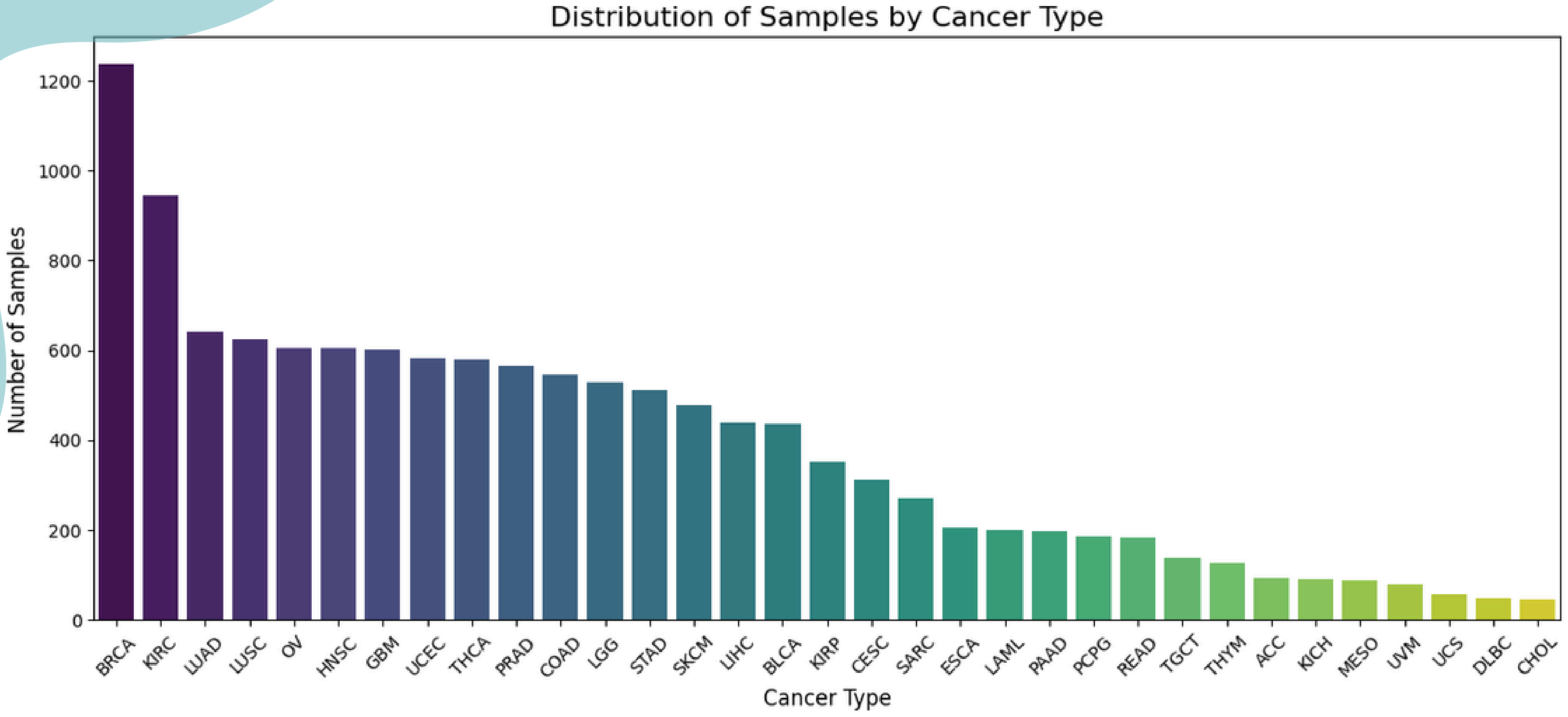
GENOMIC FEATURES TABLE:

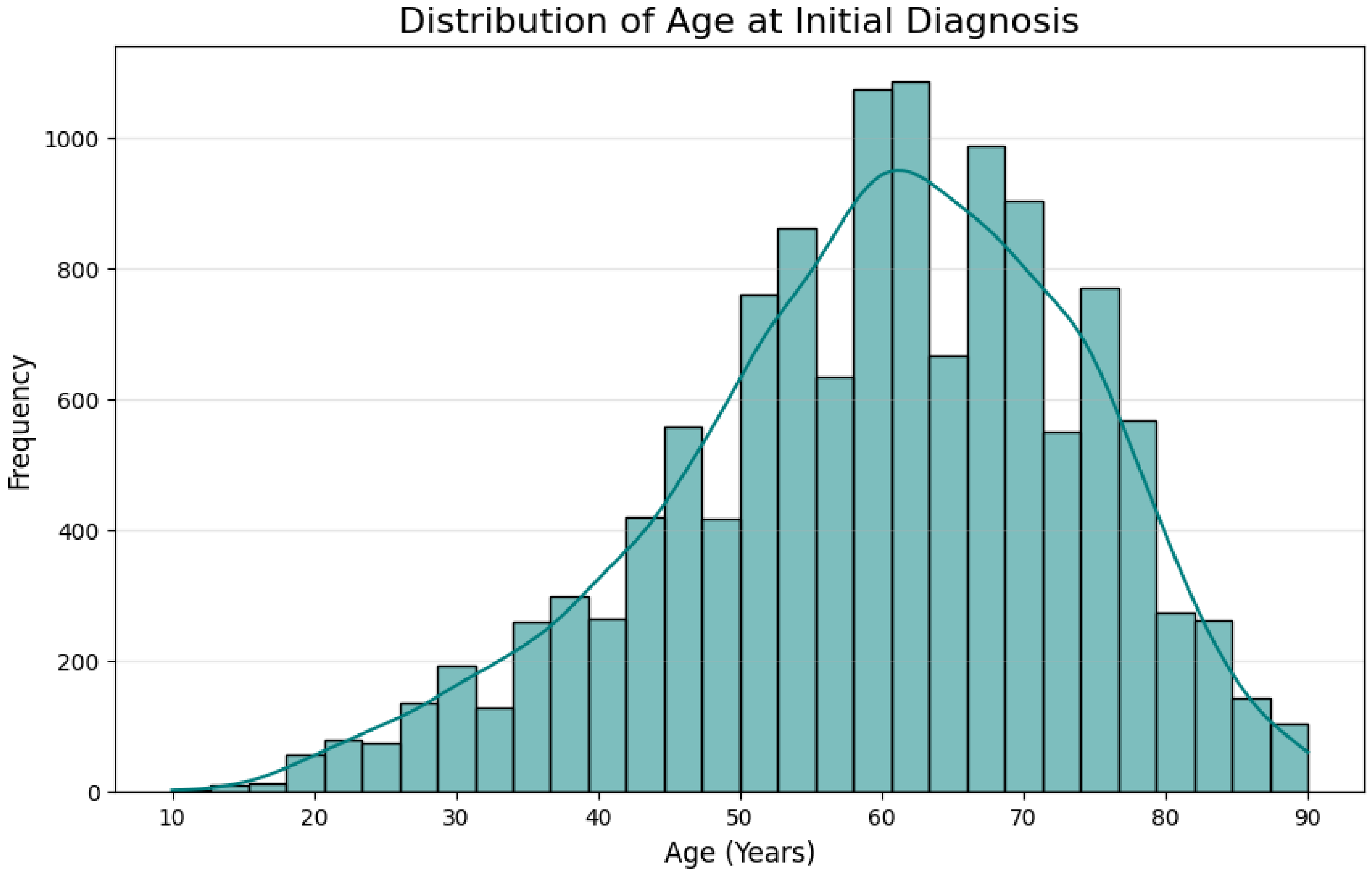
- Content: Whole-transcriptome RNA-Seq gene expression data.
- Scale: High-dimensional matrix containing ~20,000 genes across ~11,000 samples.

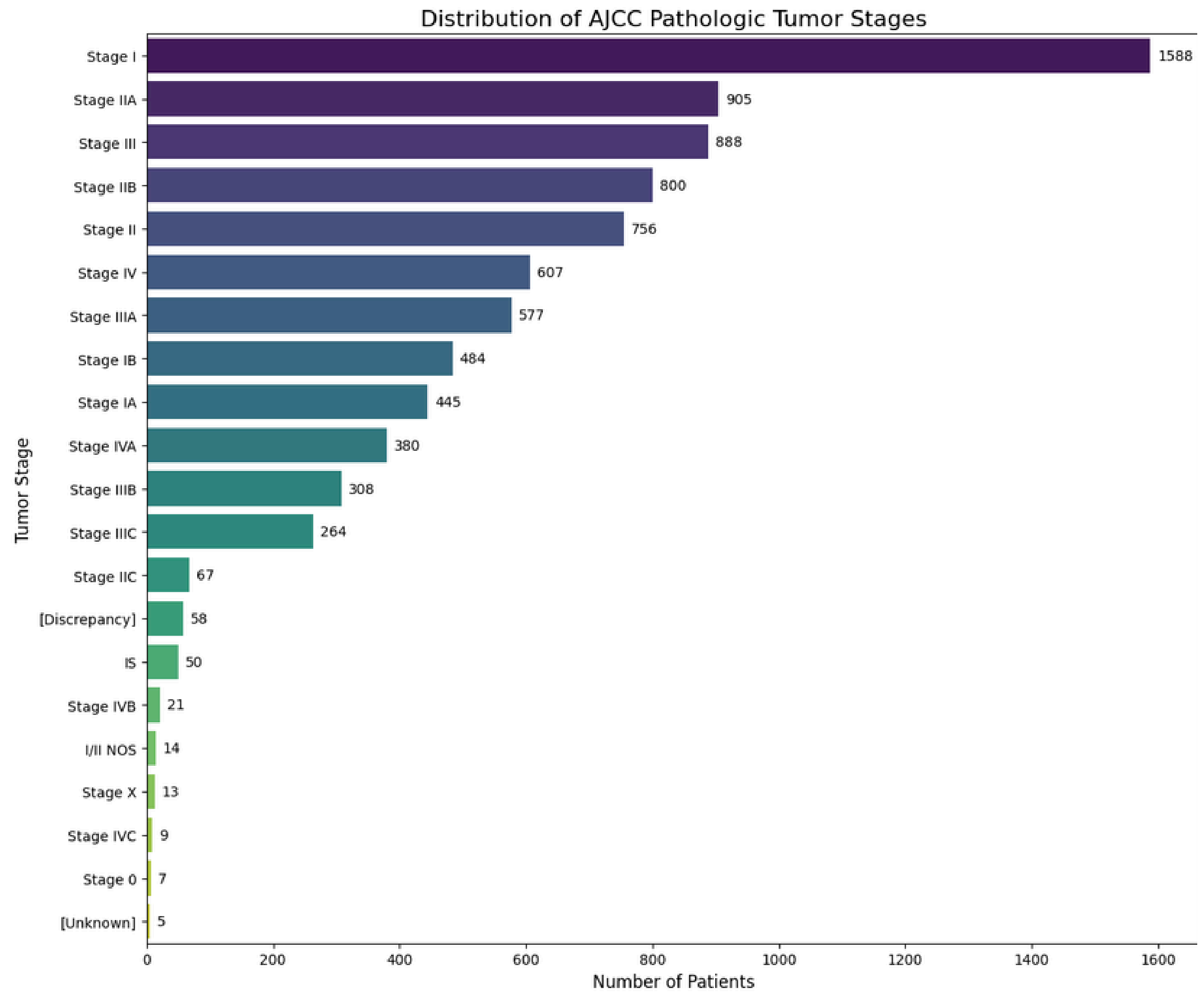


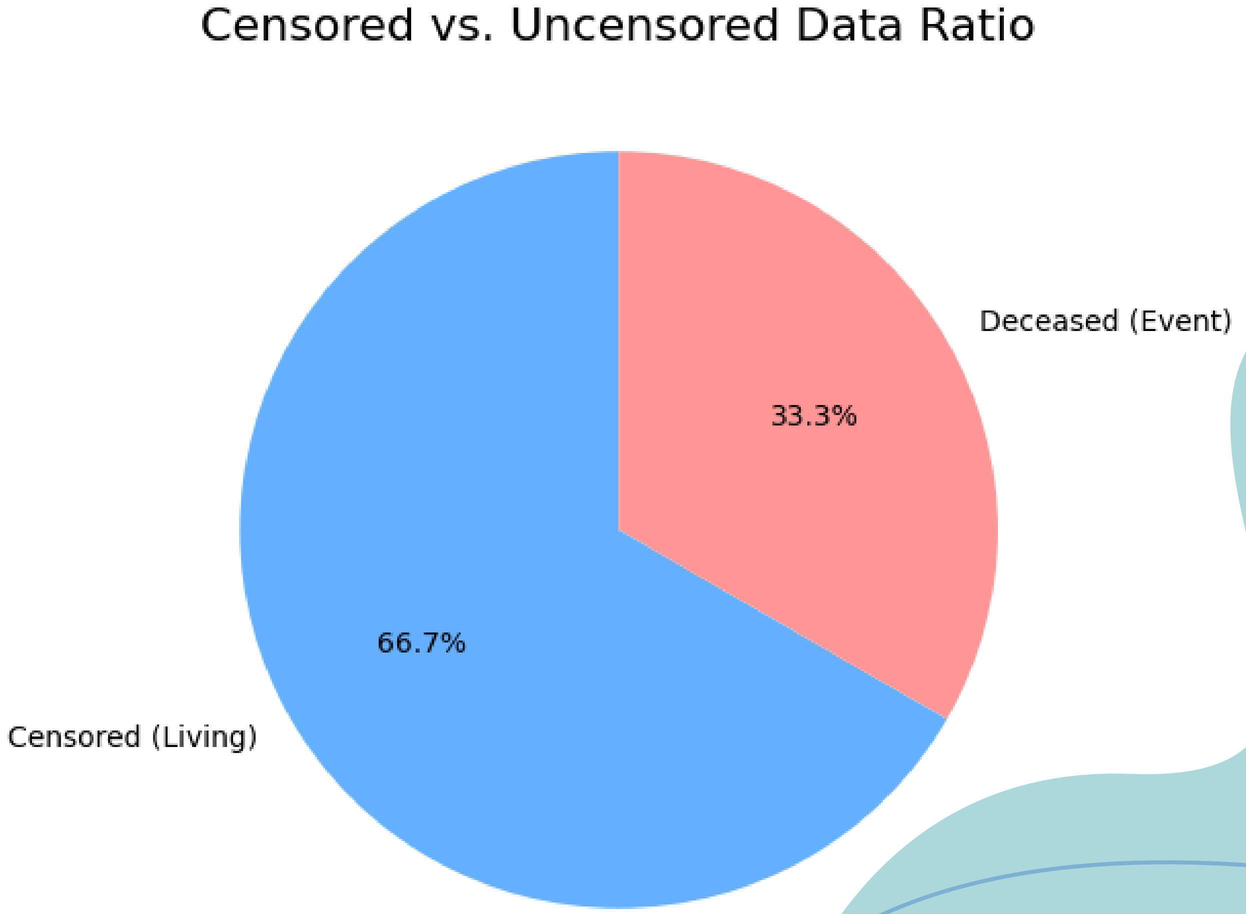
CLINICAL & SURVIVAL ANNOTATION TABLE:

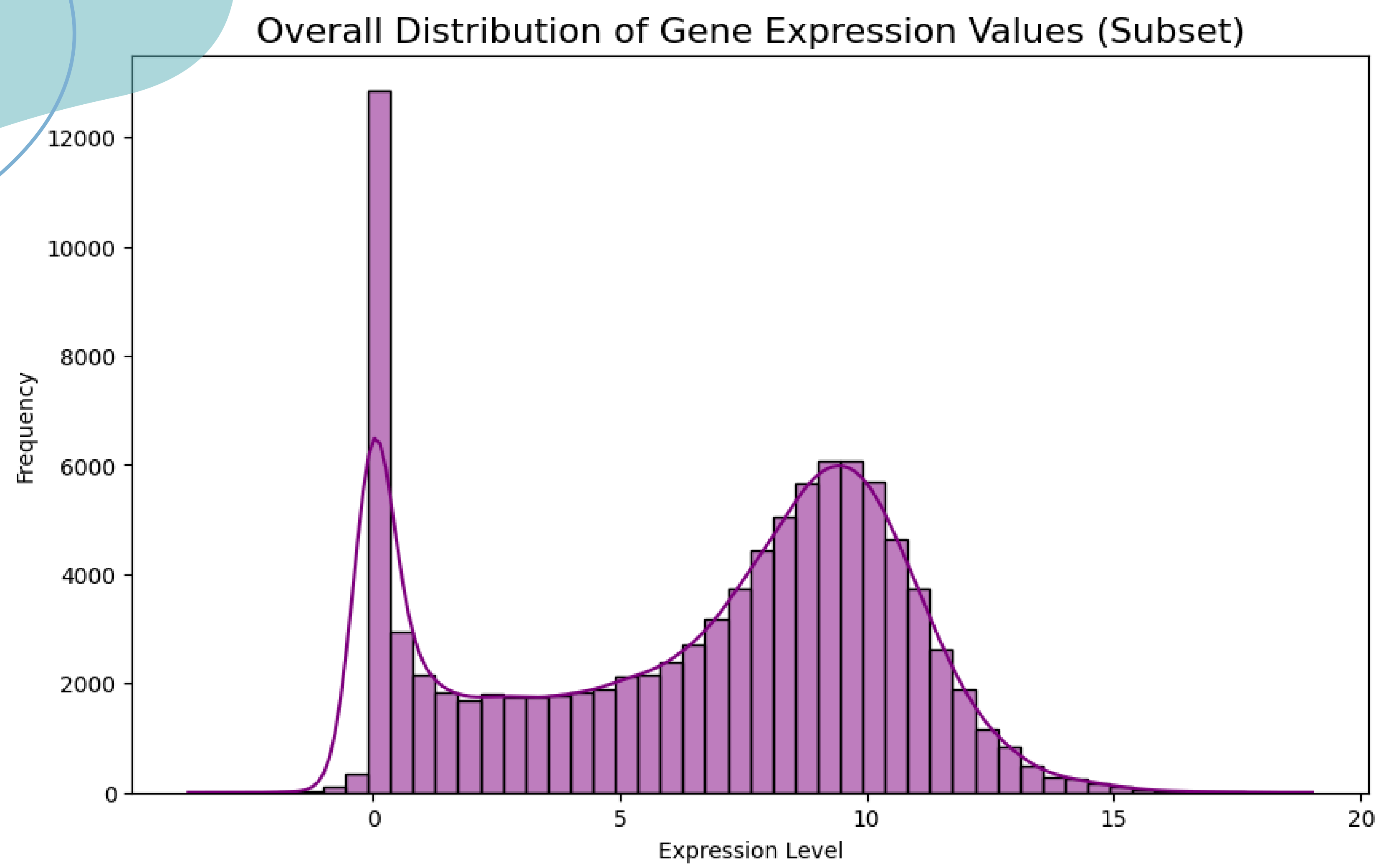
- Scope: Patient-level clinical, demographic, and survival outcomes across the Pan-Cancer cohort.
- Key Features: Patient metadata (Age, Gender), tumor characteristics (Cancer Type, Stage, Grade), and treatment info.
- Outcomes: Standardized survival endpoints (Time & Status) used as ground truth for model training.

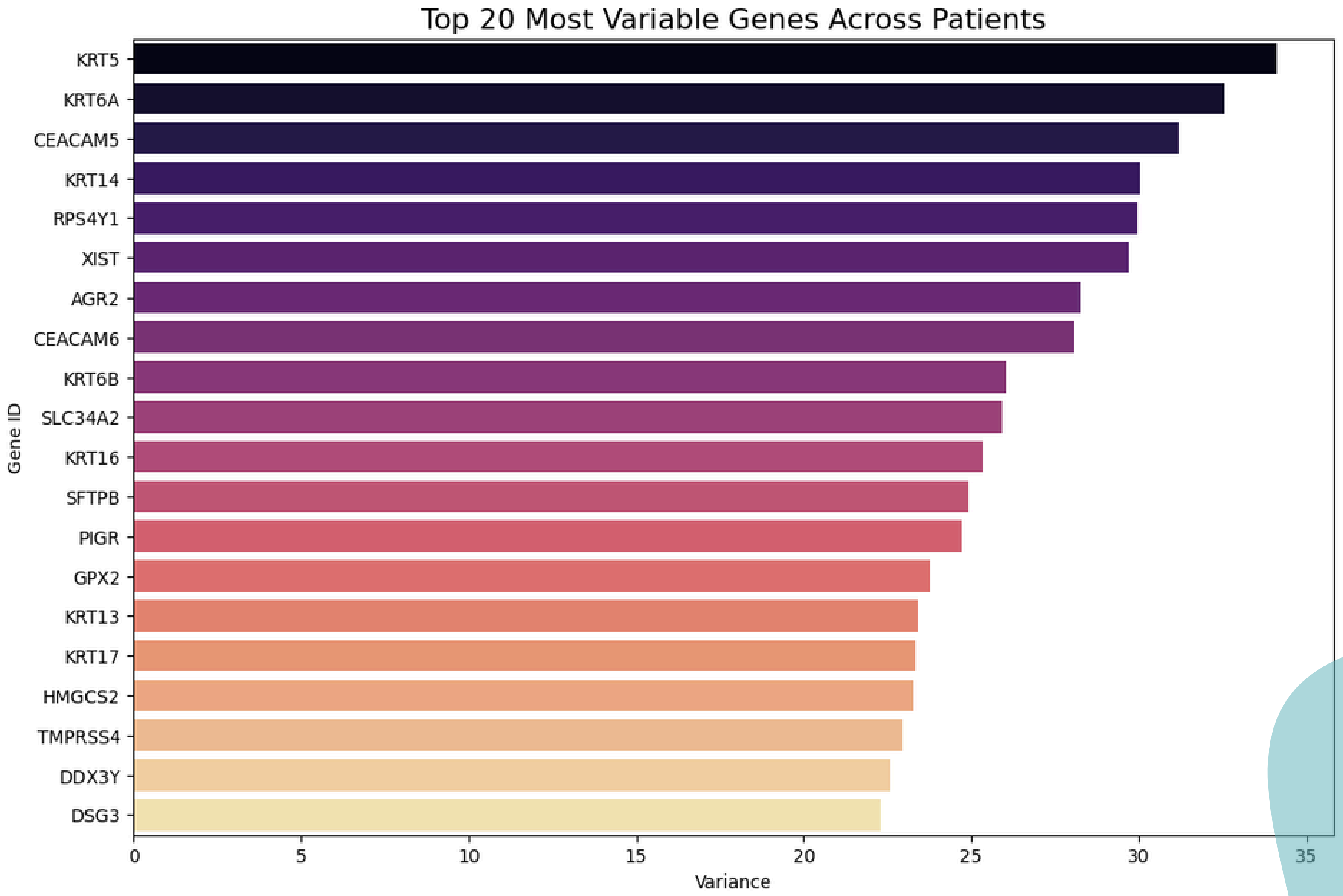

Distribution of Samples by Cancer Type
Number of Samples
0
200
400
600
800
1000
1200
BRCA
KIRC
LUAD
LUSC
OV
HNSC
GBM
UCEC
THCA
PRAD
COAD
LGG
STAD
SKCM
LIHC
BLCA
KIRP
CESC
SARC
ESCA
LAML
PAAD
PCPG
READ
TGCT
THYM
ACC
KICH
MESO
UVM
UCS
DLBC
CHOL
Cancer Type


Distribution of Age at Initial Diagnosis
Frequency
0
200
400
600
800
1000
10
20
30
40
50
60
70
80
90
Age (Years)


Distribution of AJCC Pathologic Tumor Stages
Tumor Stage
Stage I 1588
Stage IIA 905
Stage III 888
Stage IIB 800
Stage II 756
Stage IV 607
Stage IIIA 577
Stage IB 484
Stage IA 445
Stage IVA 380
Stage IIIB 308
Stage IIIC 264
Stage IIC 67
[Discrepancy] 58
IS 50
Stage IVB 21
I/II NOS 14
Stage X 13
Stage IVC 9
Stage 0 7
[Unknown] 5
0
200
400
600
800
1000
1200
1400
1600
Number of Patients


Censored vs. Uncensored Data Ratio
Deceased (Event)
33.3%
66.7%
Censored (Living)


Overall Distribution of Gene Expression Values (Subset)
Frequency
Expression Level
0
2000
4000
6000
8000
10000
12000
0
5
10
15
20


Top 20 Most Variable Genes Across Patients
Gene ID
Variance
KRT5
KRT6A
CEACAM5
KRT14
RPS4Y1
XIST
AGR2
CEACAM6
KRT6B
SLC34A2
KRT16
SFTPB
PIGR
GPX2
KRT13
KRT17
HMGCS2
TMPRSS4
DDX3Y
DSG3
0
5
10
15
20
25
30
35

METHODOLOGY: THE ANALYSIS PIPELINE

DATA PREPROCESSING & INTEGRATION

Handling high-dimensional RNA-Seq matrices and clinical metadata. The process includes data integration, normalization (e.g., Log2, Z-score standardization), and dimensionality reduction using techniques like PCA and Variance Thresholding.

MODELING STRATEGY: CLINICAL VS. GENOMIC

- Baseline Model (Statistical): Cox Proportional Hazards using standard clinical attributes (Stage, Age) to establish a performance benchmark.
- Genomic Model (Machine Learning): Advanced survival algorithms capable of handling high-dimensional data, such as Random Survival Forests (RSF) or XGBOOST.

MODEL EVALUATION & COMPARISON

- Benchmarking: Comparing the C-Index of the Genomic Model against the Clinical Baseline (Stage-based).
- Risk Stratification: Testing if the genomic signature can separate survival curves significantly better than standard staging (Log-Rank P-value).

ANTICIPATED CHALLENGES & MITIGATION STRATEGIES

1. The Curse of Dimensionality

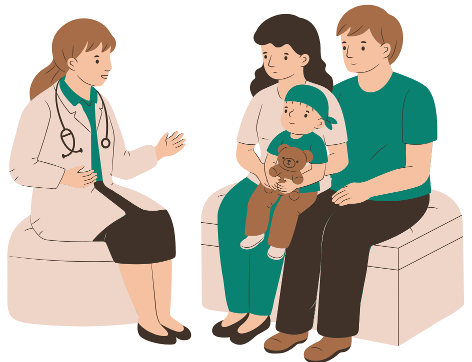
We have ~20,000 genomic features but only ~11,000 samples. This creates a severe risk of Overfitting – the model might learn noise instead of true biological signals.

2. Statistical Assumptions & Multicollinearit

Genes operate in biological pathways and are highly correlated (violating the Independence assumption of standard regression models). High Multicollinearity can lead to unstable coefficient estimates.

3. Data Heterogeneity & Censoring

- Right-Censoring: High rate of censored data (living patients) limits exact survival information.
- Pan-Cancer Variability: Diverse survival baselines across cancer types create complex distribution shifts.



PROPOSED MITIGATION STRATEGIES

1. Tackling Dimensionality

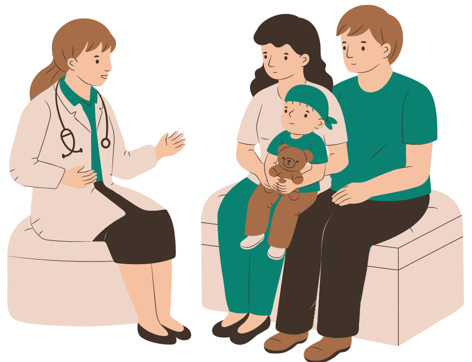
- Feature Selection: Using Variance Thresholding to remove noise and Cox Screening to select top predictive genes.
- Regularization: Applying Lasso (L1) / Ridge (L2) penalties to shrink coefficients and prevent overfitting.

2. Handling Multicollinearity

- Tree-Based Models: Utilizing Random Survival Forests which naturally handle correlated features without assuming independence.
- Dimensionality Reduction: Using PCA to transform correlated genes into a smaller set of uncorrelated components.

3. Addressing Censoring & Heterogeneity

- account for censored survival times.
- Stratification: Including Cancer Type as a covariate in the model or stratifying the analysis to account for different baseline survival rates across the Pan-Cancer cohort.



OUR PROCESS:

THE PROGRESS SO FAR

- **Data Pipeline:** Integrated clinical and genomic data, utilizing PCA to reduce ~20,000 gene features while retaining 95% variance.
- **Baseline Modeling:** Implemented Cox Proportional Hazards with Ridge regularization to handle high dimensionality.
- **Advanced Modeling:** Developed a Random Survival Forest (RSF) to capture non-linear relationships.
- **Optimization:** Maximized model performance (C-Index) using Optuna for hyperparameter tuning.
- **Score:**
  - Best C-Index Found: 0.7368
  - Best Parameters: n_estimators: 250, max_depth: 17, min_samples_split: 17, min_samples_leaf: 30, max_features': sqrt

WORK TO BE DONE

- **Advanced Evaluation:** Analyze feature importance to identify specific genes and clinical factors driving survival predictions.
- **Model Expansion:** Benchmark performance against Gradient Boosting (e.g., XGBoost) and Deep Learning (e.g., DeepSurv) models.
- **Refined Dimensionality Reduction:** Implement alternative techniques (e.g., MIPMLP) to enable interpretable SHAP and LIME analysis on gene features.
- **Clinical Validation:** Stratify patients into risk groups to assess the practical clinical utility of the model's risk scores.

THANK YOU FOR YOUR ATTENTION

GITHUB REPO