Created by: Rotem Fisher &
Orian Aziz

FROM GENE EXPRESSION TO CLINICAL OUTCOMES

GitHub

BACKGROUND & MOTIVATION: THE NEED FOR PRECISION ONCOLOGY

THE CLINICAL CHALLENGE

- Traditional cancer prognosis relies mainly on clinical stages.
- 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.

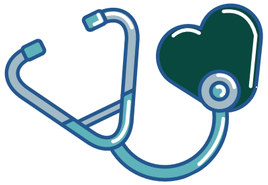
THE OPPORTUNITY

Machine Learning enables modeling of high-dimensional, non-linear genomic patterns that may improve survival prediction.

RESEARCH QUESTION

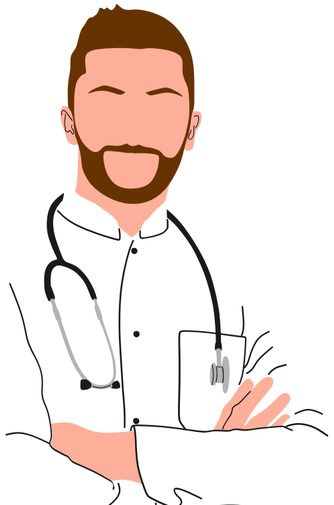
- **Primary Research Question:** Can genomic features improve survival prediction beyond traditional clinical variables in Pan-Cancer patients?
- **Secondary Question:** Does a non-linear Machine Learning model (Random Survival Forest) outperform a traditional statistical survival model (Cox Proportional Hazards)?
- **Hypothesis:** High-dimensional genomic expression patterns contain prognostic information that is not fully captured by clinical variables alone.

THE DATASET (TCGA PAN-CANCER ATLAS)



GENOMIC FEATURES TABLE:

- Content: Whole-transcriptome RNA-Seq gene expression data.
- Raw Scale: ~20,000 genes across ~11,000 patients.
- After preprocessing: 3,399 genomic features used for modeling



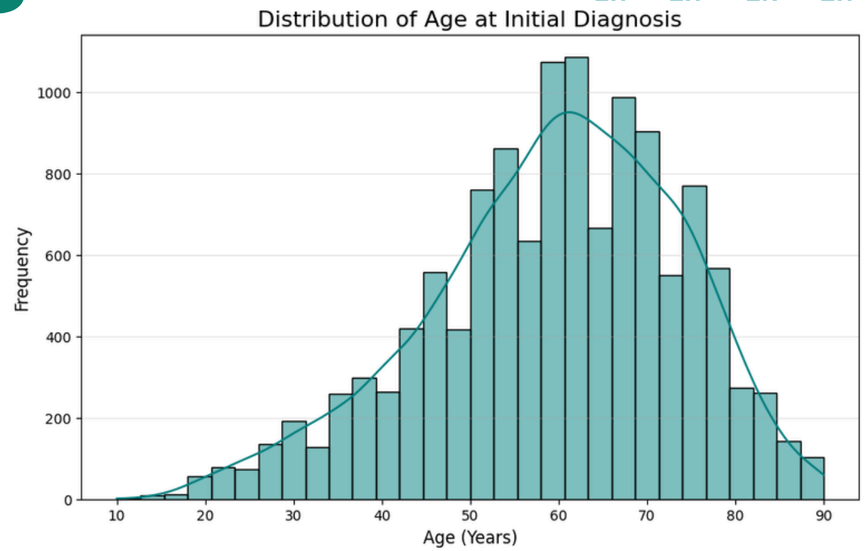
CLINICAL & SURVIVAL ANNOTATION TABLE:

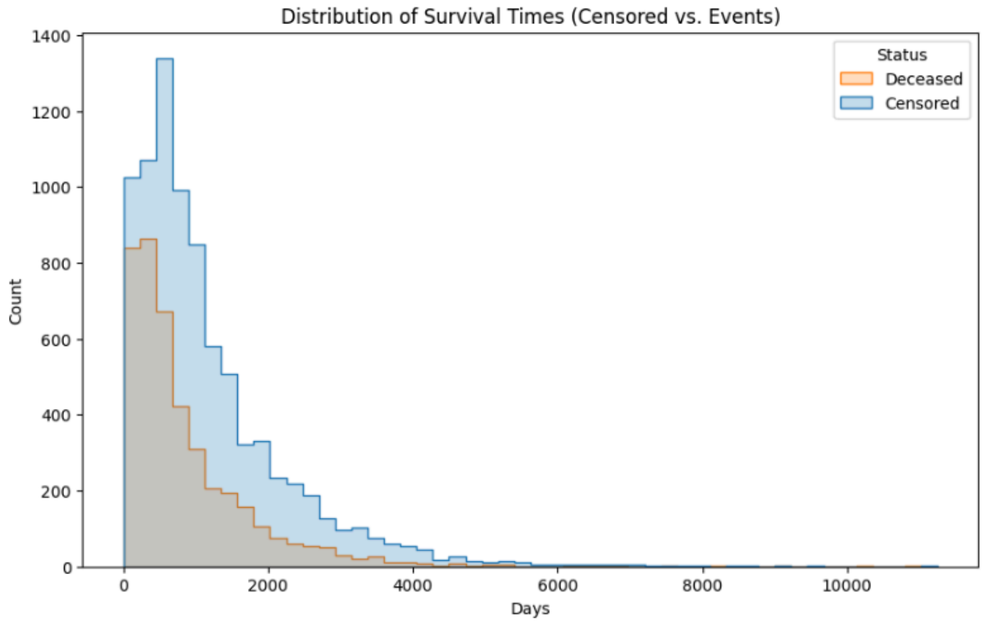
- Scope: Patient-level clinical, demographic, and survival outcomes.
- Key Variables: Age, Cancer Type, Tumor Stage, Histological Type.
- Survival Endpoint: Time-to-event (days) and censoring indicator.
- Final Cohort Size: 10,952 patients.

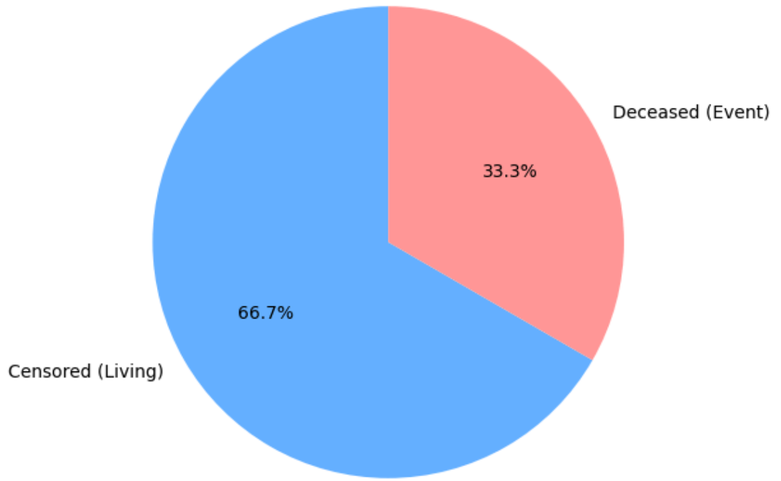
EXPLORATORY DATA ANALYSIS

KEY INSIGHTS FROM EDA

- Approximately 67% of observations are right-censored, supporting the use of survival analysis methods.
- The age distribution is centered around ~60 years, consistent with cancer epidemiology.
- Survival times exhibit a long right tail, indicating substantial heterogeneity in patient outcomes







THE CHALLENGES

1. The Curse of Dimensionality

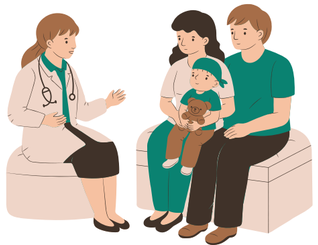
- We have ~20,000 genomic features but only ~10,952 samples
- High dimensionality increases overfitting risk and reduces model generalizability.

2. Statistical Assumptions & Multicollinearit

- Gene expression features are highly correlated due to shared biological pathways.
- This may destabilize linear survival models such as Cox PH.

3. Data Heterogeneity & Censoring

- 67% right-censored observations.
- Substantial survival variability across cancer types.



MITIGATION STRATEGIES

1. Tackling Dimensionality

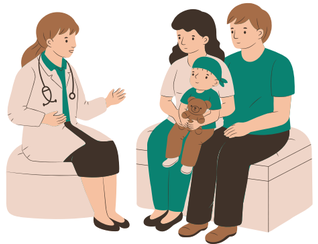
- Variance filtering + univariate Cox screening to reduce noise.
- PCA (95% explained variance) for dimensionality reduction.
- Final Top-K Feature Selection (100 features) to improve generalization.
- Regularization (L1 / L2) to prevent overfitting.

2. Handling Multicollinearity

- PCA transformation to generate uncorrelated components.
- Random Survival Forest (RSF) to model correlated genomic features without linearity assumptions.

3. Addressing Censoring & Heterogeneity

- Survival-specific models accounting for right-censoring.
- Inclusion of Cancer Type as a covariate to handle Pan-Cancer variability.
- Stratified train/test split to ensure robust evaluation.



METHODOLOGY - ANALYSIS PIPELINE

DATA PREPARATION

- Clinical & RNA-Seq integration
- Variance filtering + Cox screening
- Z-score normalization
- PCA (95% explained variance)

BASELINE MODEL

Regularized Cox PH (Clinical variables)

- Statistical benchmark
- Ridge penalty

ADVANCED MODEL

Random Survival Forest (Genomic + Clinical)

- Non-linear survival modeling
- Hyperparameter tuning (Optuna)
- Final Top-100 Feature Selection

MODEL VALIDATION

- Final evaluation on held-out test set
- C-index comparison
- Kaplan–Meier risk stratification
- Log-rank statistical test

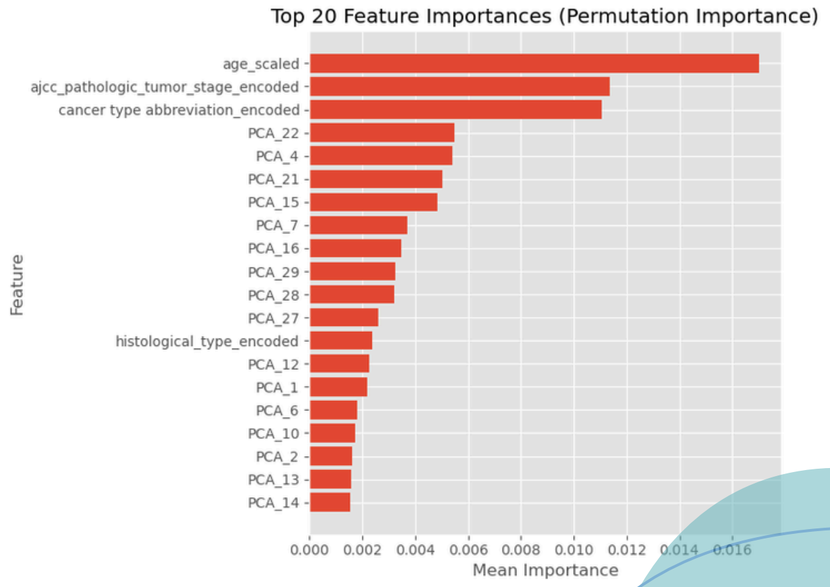
PREDICTIVE PERFORMANCE

Model	C-Index
Clinical Cox	0.6667
Genomic Cox	0.6633
RSF (All Features)	0.7427
RSF (Top-100 Features)	**0.7827**

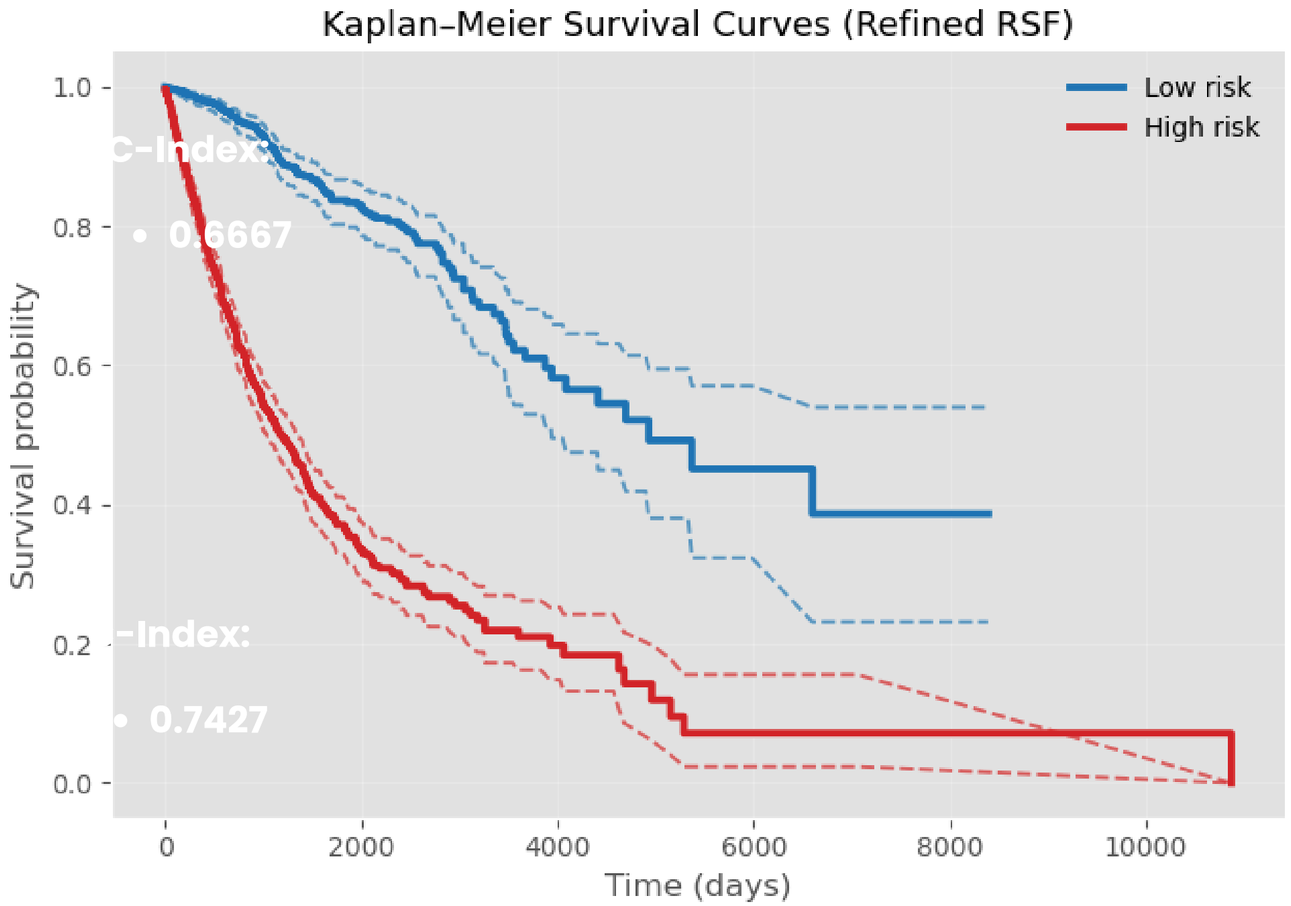
Δ Improvement (Cox → Refined RSF): +0.116

CONCLUSIONS

- Linear genomic modeling did not improve over clinical baseline.
- Non-linear modeling substantially enhanced predictive performance.
- Feature refinement improved generalization (+0.04 C-index).



RISK STRATIFICATION USING REFINED RSF



Risk Stratification Using Refined RSF Top 100

- Patients were stratified into high- and low-risk groups based on RSF risk scores.
- Clear separation between survival curves is observed.
- Log-rank test confirms statistically significant difference ($p = 3.55 \times 10^{-89}$).

FINAL CONCLUSIONS

- Clinical variables alone provide moderate predictive performance (C-index ≈ 0.66).
- Linear genomic modeling (Cox) does not improve survival prediction.
- Non-linear modeling (RSF) substantially enhances performance.
- Feature refinement further improves generalization (C-index = 0.7827).
- The integration of genomic and clinical data enables clinically meaningful risk stratification.

THANK YOU FOR YOUR ATTENTION