

From Gene Expression to Clinical Outcomes: A Comparative Analysis of Statistical and Machine Learning Models in Pan-Cancer Survival Prediction

Github Repository:

<https://github.com/rotemfisher/ML-and-statistics-in-medical-fields.git>

Abstract

Background:

Cancer prognosis is traditionally guided by clinical staging systems; however, substantial inter-patient variability in survival persists within identical stages. This heterogeneity reflects underlying molecular diversity that is not captured by conventional clinicopathologic features. Advances in transcriptome-wide RNA sequencing enable systematic interrogation of tumor biology and create opportunities to integrate molecular signals with machine learning-based survival modeling.

Methods:

We analyzed the TCGA Pan-Cancer Atlas, comprising approximately 11,000 patients with matched whole-transcriptome RNA-Seq profiles and clinical survival annotations across multiple cancer types. Given the high dimensionality, multicollinearity, and censoring inherent to genomic survival data, we designed a structured machine learning pipeline tailored to survival analysis.

Transcriptomic features were standardized and compressed using principal component analysis (PCA), preserving 95% of total variance while mitigating correlated gene effects. Clinical and genomic representations were jointly modeled using Random Survival Forests (RSF), enabling non-linear interactions without parametric assumptions. Model hyperparameters were optimized using Optuna-based hyperparameter tuning, and permutation-based feature importance was employed to quantify the contribution of individual molecular and clinical components. A refined survival model was subsequently trained using the most informative features. Prognostic relevance was evaluated using the concordance index (C-index) and biologically interpretable risk stratification through Kaplan-Meier analysis and Log-rank statistical testing.

Results:

The refined RSF model achieved a C-index of 0.7827, outperforming the optimized baseline genomic model (C-index = 0.7427) and substantially exceeding linear Cox-based baselines (C-index = 0.6633). Permutation importance analysis identified both established clinical determinants (age, tumor stage) and multiple transcriptome-derived components as dominant

contributors to survival prediction, indicating that molecular expression programs encode complementary prognostic information. Risk stratification based on model-derived hazard scores yielded a pronounced separation between high-risk and low-risk patient groups, confirmed by a highly significant Log-rank test ($\chi^2 = 400.88$, $p = 3.55 \times 10^{-89}$).

Conclusions:

By integrating dimensionality-aware transcriptomic representations with non-linear survival learning algorithms, this study demonstrates how machine learning can extract biologically meaningful survival signals from high-dimensional genomic data. The results support a paradigm in which algorithmic modeling and tumor biology jointly inform prognosis, advancing molecularly driven precision oncology beyond traditional stage-based frameworks.

Introduction

Cancer prognosis is central to clinical decision-making, guiding treatment selection, follow-up strategies, and patient counseling. Conventional prognostic frameworks rely primarily on clinicopathologic variables such as tumor stage, grade, and patient age. While effective at the population level, these factors often fail to explain the substantial variability in survival outcomes observed among patients with identical clinical staging.

This limitation reflects the molecular heterogeneity of cancer. Tumors that appear clinically similar may differ markedly at the genomic and transcriptomic levels, leading to divergent disease trajectories and treatment responses. Large-scale cancer genomics initiatives, most notably The Cancer Genome Atlas (TCGA), have enabled systematic characterization of tumor molecular profiles across diverse cancer types, providing new opportunities to refine prognostic modeling beyond traditional staging systems.

Transcriptome-wide RNA sequencing (RNA-Seq) captures coordinated gene expression programs associated with tumor biology, pathway activity, and microenvironmental interactions. However, leveraging RNA-Seq data for survival prediction presents substantial challenges, including high dimensionality, strong feature correlation, and right-censored outcomes. Classical statistical survival models such as Cox Proportional Hazards (CoxPH) often struggle under these conditions due to linearity and proportional hazard assumptions.

Machine learning-based survival models offer a flexible alternative. In particular, tree-based methods such as Random Survival Forests (RSF) can model non-linear relationships and complex feature interactions without parametric assumptions. When combined with dimensionality reduction and feature selection, these approaches are well suited for high-dimensional genomic survival data.

In this study, we evaluate whether integrating transcriptome-derived representations with machine learning survival models improves prognostic performance in a pan-cancer setting. Using the TCGA Pan-Cancer Atlas, we compare classical Cox-based modeling with Random Survival Forests and assess whether a reduced subset of molecular and clinical features can meaningfully stratify patients into distinct survival risk groups.

Results

3.1 Baseline Survival Modeling

The Cox Proportional Hazards model achieved a C-index of 0.6633, indicating limited discriminatory ability when applied to high-dimensional clinical and transcriptomic features.

3.2 Machine Learning Survival Modeling

A naïve Random Survival Forest model trained with default hyperparameters achieved a C-index of 0.7246, demonstrating a substantial improvement over the Cox baseline. After Optuna-based hyperparameter optimization, RSF performance further improved to a C-index of 0.7427.

3.3 Feature Importance and Model Refinement

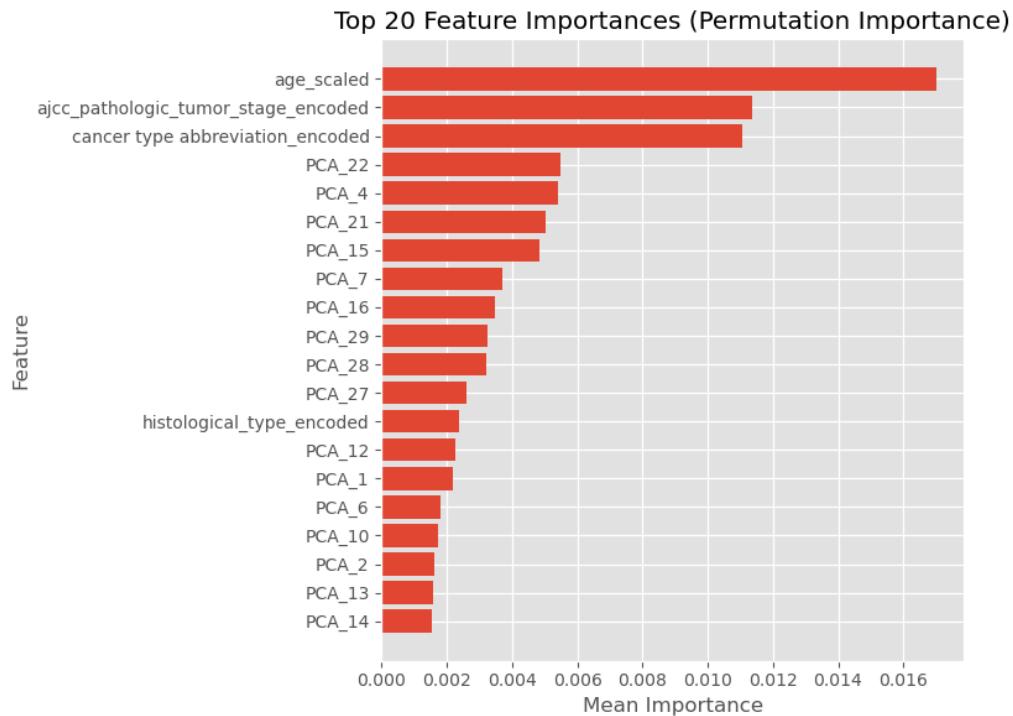
Permutation-based feature importance analysis revealed that both established clinical variables (age, tumor stage) and multiple transcriptome-derived PCA components ranked among the most influential predictors of survival. Based on importance rankings, the top 100 features were selected, reducing dimensionality from over 3,000 variables to a compact subset.

The refined RSF model trained on this reduced feature space achieved a C-index of 0.7827, representing a performance gain of $\Delta = 0.0400$ relative to the optimized RSF model.

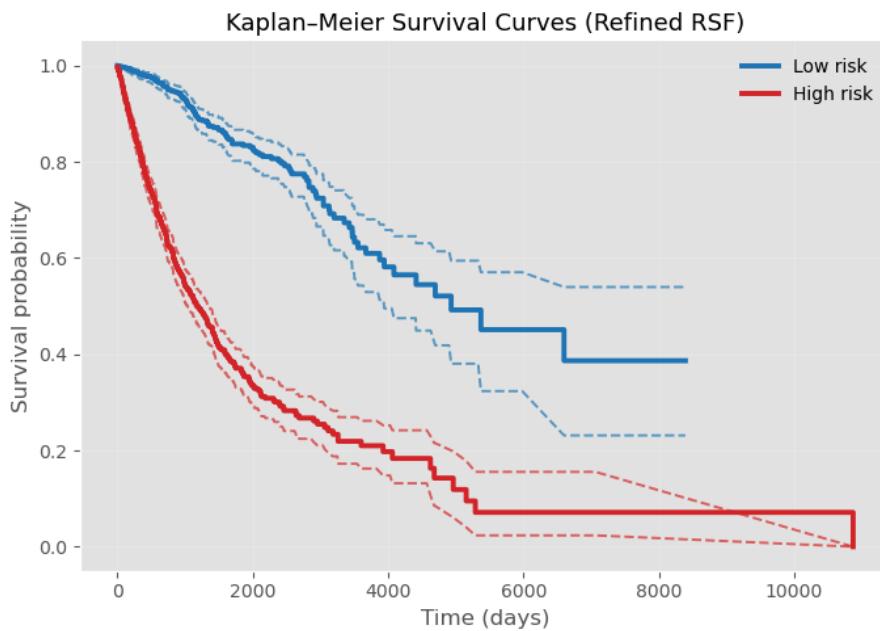
3.4 Risk Stratification

Patients were stratified into high-risk and low-risk groups using refined RSF risk scores. Kaplan–Meier survival analysis demonstrated a pronounced separation between groups, confirmed by a highly significant Log-rank test ($\chi^2 = 400.88$, $p = 3.55 \times 10^{-89}$), indicating clinically meaningful survival discrimination.

4.1 Figure 1: Feature importance



4.2 Figure 2: Kaplan–Meier survival curves for high- and low-risk groups derived from the refined RSF model.



4.3 Figure 3: Model comparison

Data Element	Value (C- index)
CoxPH	0.6633
Naive RSF	0.7246
Optimize RSF	0.7427
Refined RSF	0.7827
Performance improvement RSF	0.0581
C-index Improvement over CoxPH	0.1194

Methods

5.1 Study Design

We conducted a comparative survival modeling study using clinical and transcriptomic data from the TCGA Pan-Cancer Atlas. The analytical pipeline included preprocessing and feature integration, baseline statistical modeling, machine learning-based survival modeling, feature importance–driven refinement, and risk stratification.

5.2 Data Preprocessing and Feature Integration

Clinical features included age at diagnosis, cancer type, AJCC tumor stage, histological subtype, gender, and race.

Age was treated as a continuous variable, median-imputed, and z-score normalized. Categorical variables were label-encoded, with missing values assigned to an “Unknown” category.

RNA-Seq gene expression features were:

- Imputed with zero for missing values
- Standardized using z-score normalization

5.3 Dimensionality Reduction

To address high dimensionality and multicollinearity, Principal Component Analysis (PCA) was applied exclusively to transcriptomic features. Components explaining 95% of total variance were retained and concatenated with clinical variables to form the final feature set.

5.4 Survival Modeling

- **Baseline model:** Cox Proportional Hazards model with L2 regularization
- **Machine learning model:** Random Survival Forest (RSF)

An initial RSF model was trained using default hyperparameters. Hyperparameter optimization was subsequently performed using Optuna, with the concordance index (C-index) on a validation set as the optimization objective.

5.5 Feature Importance and Model Refinement

Permutation-based feature importance was computed on the optimized RSF model. Based on importance rankings, the top 100 features (clinical and PCA-derived transcriptomic components) were selected. A refined RSF model was retrained using this reduced feature set.

5.6 Evaluation and Risk Stratification

All models were evaluated using the concordance index (C-index). For interpretability, refined RSF risk scores were used to stratify patients into high- and low-risk groups using a median split. Survival differences were assessed using Kaplan–Meier curves and the Log-rank test.

Discussion

6.1 Principal Findings

In this study, we evaluated the contribution of transcriptome-derived features to survival prediction in a large pan-cancer cohort using a structured comparison between classical statistical models and machine learning-based survival approaches. Our results demonstrate a clear and consistent improvement in predictive performance when transitioning from linear Cox-based modeling to non-linear, tree-based survival models.

Specifically, the Random Survival Forest (RSF) framework substantially outperformed the Cox Proportional Hazards baseline, and further gains were achieved through targeted hyperparameter tuning and permutation-based feature selection. The refined RSF model achieved the highest concordance index (0.7827) and produced clinically interpretable risk stratification with highly significant separation between patient groups.

6.2 Why Non-Linear Survival Models Outperform Cox-Based Approaches

The comparatively limited performance of the Cox Proportional Hazards model observed in this study is consistent with known methodological constraints of linear survival models when applied to high-dimensional molecular data. Cox models rely on proportional hazards and linear associations between covariates and risk, assumptions that are often violated in heterogeneous cancer populations.

In contrast, Random Survival Forests are capable of modeling complex, non-linear interactions between clinical and transcriptomic features without requiring parametric assumptions. This flexibility is particularly advantageous in pan-cancer settings, where baseline hazard functions, molecular drivers, and survival trajectories vary substantially across tumor types. The observed performance gains suggest that survival-relevant molecular information is encoded in interaction patterns that are not adequately captured by linear models.

6.3 Interpretation of Transcriptomic Contributions

Permutation-based feature importance analysis revealed that transcriptome-derived components consistently ranked among the most influential predictors of survival, alongside established clinical determinants such as age and tumor stage. Importantly, the transcriptomic features used in this study were derived from principal component analysis rather than individual gene-level measurements.

This finding suggests that survival-associated molecular information may be better represented at the level of aggregated expression programs, potentially reflecting coordinated pathway activity, cellular composition, or tumor microenvironmental states. By compressing gene expression into orthogonal components, PCA enabled the model to leverage biologically meaningful variation while mitigating noise and redundancy inherent to high-dimensional transcriptomic data.

6.4 Impact of Feature Selection on Model Performance

An important observation of this study is that reducing the feature space from thousands of variables to a focused subset of 100 highly informative features resulted in a notable improvement in predictive accuracy. This improvement indicates that survival-relevant signals are concentrated within a limited subset of clinical and molecular dimensions, while many features contribute little or no prognostic value.

Feature selection via permutation importance likely reduced overfitting and enhanced model stability by focusing learning capacity on the most relevant inputs. These results highlight the importance of integrating feature importance analysis into survival modeling pipelines, particularly when working with high-dimensional genomic data.

6.5 Clinical Interpretability and Risk Stratification

Beyond predictive accuracy, the refined RSF model demonstrated strong clinical interpretability through risk-based patient stratification. Kaplan–Meier analysis revealed a pronounced and statistically robust separation between high-risk and low-risk patient groups derived from model predictions.

While this stratification does not imply immediate clinical applicability, it illustrates the potential of transcriptome-informed models to complement traditional staging systems by capturing molecular heterogeneity that influences survival outcomes. Such models may serve as decision-support tools for hypothesis generation or retrospective risk assessment in future translational studies.

6.6 Limitations

Several limitations of this study should be acknowledged. First, the pan-cancer design introduces substantial biological heterogeneity, and the model does not explicitly account for cancer-type-specific baseline hazards. Second, transcriptomic features were represented using PCA components, which, while effective for dimensionality reduction, limit direct biological interpretability at the individual gene or pathway level.

Additionally, missing gene expression values were imputed with zeros, an assumption that may not fully capture the complexity of RNA-Seq measurement noise. Finally, model evaluation was performed on a single dataset without external validation, and performance estimates may therefore reflect dataset-specific characteristics.