Survival Analysis of Cancer Patients Using Random Survival Forest

Arthur Liang

Introduction

This study analyzes survival outcomes across multiple cancer types using The Cancer Genome Atlas (TCGA) dataset. The primary objective was to develop a predictive model for patient mortality while accounting for various clinical and demographic factors. The dataset comprises 4,081 patients across seven cancer types, including clinical staging information, demographic data, and treatment indicators.

Methods

Exploratory Data Analysis and Data Preprocessing

Initial exploration of the dataset revealed several key characteristics:

- The dataset contains 4,081 patients with 12 variables including patient identifiers, staging information (TNM), demographics (race, sex), cancer subtype, and treatment indicators
- Missing data analysis showed:
    - Significant missing values in staging information (stage_m: 1,105, stage_t: 626, stage_n: 667)
    - Complete data for race, sex, subtype, and adjuvant radiotherapy status
    - Menopausal status was largely missing (3,108 missing values)
    - Event status (death/recurrence) had 2,654 missing values
    - Survival follow-up times ranged from 0 to 23.57 years (median: 2.08 years)
    - Age at diagnosis ranged from 26.57 to 90.00 years (mean: 63.37 years)
- Cancer subtypes were relatively balanced.

This initial exploration informed our preprocessing strategy, particularly the handling of missing values and the selection of relevant features for the survival model. We proceeded with data preprocessing as follows:

- TNM staging was converted to ordinal values using clinical staging hierarchy
- Missing values were handled using median imputation for numerical features (age, survival time) and mode imputation for categorical variables
- Categorical variables (race, sex, cancer subtype, treatment status) were encoded using one-hot encoding
- The target variable "death" was binary-encoded from the event column

Initial Modeling Approaches

Before settling on the Random Survival Forest, we first explored using a Cox Proportional Hazards (CoxPH) model:

- The initial CoxPH model was implemented using the lifelines package
- Features included TNM staging, demographic information, and treatment status
- The model showed:
    - Significant effects ($p < 0.005$) for stage_t, stage_n, sex, and years_at_dx
    - A C-index of 0.68 during training that didn't generalize to test where the C-index was 0.29
    - Strong association between age at diagnosis and survival (coefficient: 0.03, $p < 0.005$)
    - Significant impact of sex (coefficient: 0.42, $p < 0.005$)
    - No significant effect of adjuvant radiotherapy ($p = 0.86$)
- This preliminary analysis suggested that both clinical and demographic factors were important predictors and that the relationship between predictors and survival might be non-linear.

These findings suggested that a more flexible model could potentially capture complex interactions better leading to our decision to implement a Random Survival Forest model, which could better handle non-linear relationships and interactions between features.

Model Selection and Training

We implemented a Random Survival Forest model, chosen for its:

- Ability to handle censored data
- Capacity to capture non-linear relationships
- Robust performance with high-dimensional data
- Built-in handling of survival analysis requirements

The model was trained using:

- 80-20 train-test split, stratified by event status
- 100 trees in the forest
- Standard scaling of numerical features
- Final feature set of 23 variables after one-hot encoding

Results

Demographic and Clinical Characteristics

The patient population showed diverse characteristics:

- Age range: 26.6-90.0 years (median: 63.7 years)
- Gender distribution: 58.9% female, 41.1% male
- Treatment: Nearly equal distribution of adjuvant radiotherapy (49.3% received treatment)
- Racial composition: Predominantly White (65.8%), with significant representation of Black (9.6%) and Asian (3.8%) patients

Model Performance

The Random Survival Forest achieved a C-index of 0.719 on the test set, indicating good discriminative ability. The model's survival probability predictions showed:

- 1-year survival: 91.5% (mean probability)
- 3-year survival: 74.8% (mean probability)
- 5-year survival: 63.7% (mean probability)

Cancer-Specific Outcomes

Analysis revealed significant variation in mortality rates across cancer types:

- Highest mortality: Ovarian cancer (OV) with 59.45% event rate
- Lowest mortality: Prostate cancer (PRAD) with 2.00% event rate
- Median survival times ranged from 1.47 years (BLCA) to 2.74 years (OV)

Figure 1 shows the survival curve and distribution of survival times for different subtypes.

Conclusion

The Random Survival Forest model demonstrates strong predictive capability for cancer survival across multiple cancer types. The C-index of 0.719 indicates good discrimination ability, while the survival probability estimates provide clinically relevant temporal predictions. Additionally, the balanced performance across demographic groups suggests robust generalizability. The model's performance suggests potential utility in clinical decision-making, though careful consideration of cancer-specific factors is warranted. However, cancer type-specific survival patterns indicate the need for tailored prognostic approaches. Moreover, because of significant missing data in staging information (>15% for each TNM component), variable follow-up times across cancer types, and uneven distribution of events across cancer types further validation with more data is needed.

AI Coding

No AI was used for the code in Jupyter Notebook. Once the code was functional in main.py, to assist in writing the report, I used AI to edit the functions to write comments and output useful metrics and statistics (mostly in the run_analysis and visualize_results functions).

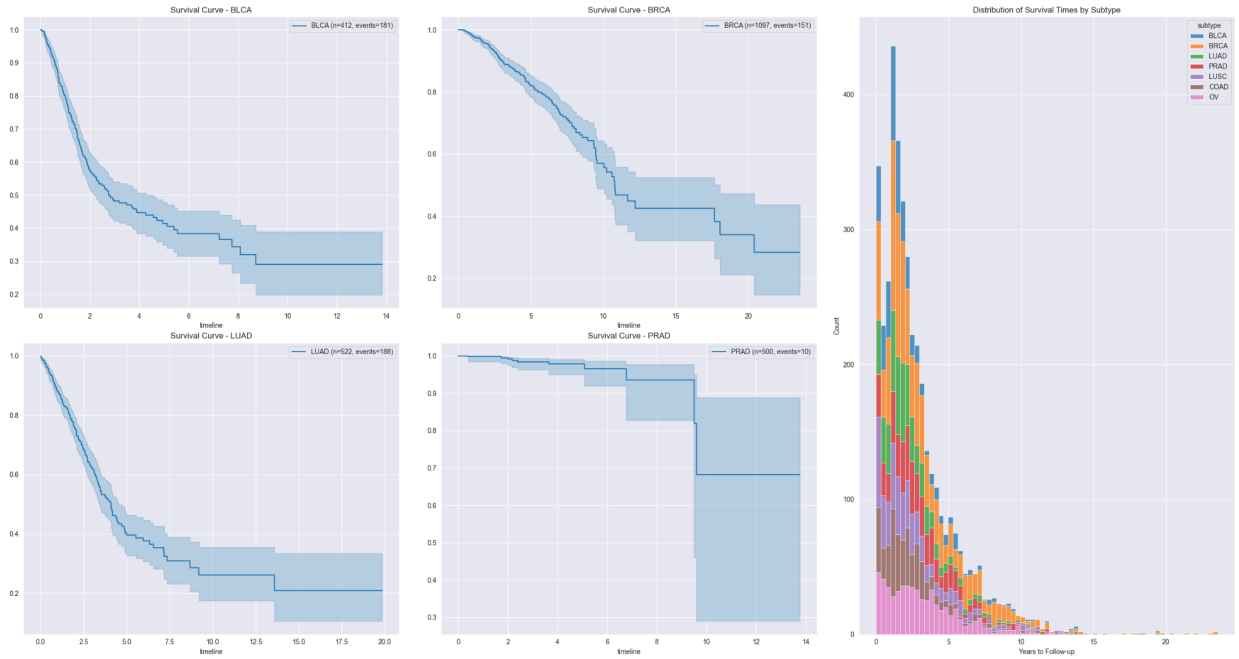

Figure 1: Survival Curves and Distribution of Survival Times for Different Subtypes