International Journal of Cloud Computing and Database Management

E-ISSN: 2707-5915
P-ISSN: 2707-5907
IJCCDM 2025; 6(1): 120-126
www.computersciencejournals.com/ijccdm
Received: 02-12-2024
Accepted: 06-01-2025

Golghate Ashish Anil
Research Scholar, Department of Computer Science, Malwanchal University, Indore, Madhya Pradesh, India

Dr. Manav Thakur
Supervisor, Department of Computer Science, Malwanchal University, Indore, Madhya Pradesh, India

Corresponding Author:
Golghate Ashish Anil
Research Scholar, Department of Computer Science, Malwanchal University, Indore, Madhya Pradesh, India

Enhancing cancer survival rate predictions through longitudinal data analysis and advanced feature engineering

Golghate Ashish Anil and Manav Thakur

DOI: https://doi.org/10.33545/27075907.2025.v6.i1b.94

Abstract
This study focuses on enhancing cancer survival rate predictions by utilizing longitudinal data analysis combined with advanced feature engineering techniques. Cancer prognosis is inherently complex due to the dynamic nature of disease progression and the variation in individual patient responses to treatments. By analyzing time-series clinical data, such as patient demographics, tumor characteristics, treatment history, and response indicators over time, the study aims to develop more accurate and personalized predictive models. Advanced feature engineering methods are employed to extract meaningful patterns from raw medical data, which are then used to train machine learning algorithms like Random Forest, XGBoost, and Support Vector Machines. The study evaluates model performance using statistical metrics such as accuracy, sensitivity, specificity, and AUC-ROC to ensure reliable survival predictions. In addition, the research integrates qualitative insights from healthcare professionals and cancer survivors through structured surveys, providing real-world context to the findings and improving the interpretability of the predictive models. This mixed-methods approach bridges the gap between computational modeling and human-centered perspectives, contributing to the development of personalized cancer care strategies. Ultimately, the research aims to enhance clinical decision-making by providing tools for early identification of high-risk patients and facilitating tailored treatment plans. The outcomes of this study have the potential to significantly improve cancer prognosis accuracy and patient outcomes.

Keywords: Cancer survival prediction, longitudinal data, feature engineering, machine learning, personalized medicine

Introduction
Cancer survival rate prediction is a crucial aspect of modern oncology, providing essential information for clinicians to tailor treatment plans and improve patient outcomes. Traditionally, cancer prognosis has relied on static clinical data, such as tumor size, stage, and histopathological features. While these factors are important, they often fail to capture the complexities and dynamic nature of cancer progression. As cancer treatment evolves and becomes more personalized, there is an increasing need for more sophisticated models that incorporate temporal and multifaceted data to predict survival outcomes. Longitudinal data analysis offers a promising solution by capturing changes in a patient's condition over time, allowing clinicians and researchers to track how a patient's health and cancer status evolve in response to treatment. By combining this rich, time-dependent information with advanced feature engineering techniques, more accurate and personalized survival predictions can be made, offering a clearer understanding of individual patient trajectories and optimizing treatment strategies.

Feature engineering plays a pivotal role in enhancing the predictive power of survival models by transforming raw data into meaningful, actionable features that reflect underlying patterns in cancer progression. In the context of cancer prognosis, feature engineering encompasses the extraction, selection, and transformation of variables such as tumor characteristics, genetic markers, treatment history, and patient demographics. Advanced techniques, such as the use of machine learning algorithms, can automate the process of feature selection and help identify the most relevant features for prediction. However, the success of these models depends not only on the data used but also on how well the features represent the underlying biological and clinical phenomena.

By leveraging both longitudinal data and feature engineering, predictive models can better account for time-dependent factors, leading to more reliable predictions and more informed clinical decisions. This research aims to explore how these methodologies can be combined to enhance cancer survival rate predictions, ultimately advancing personalized medicine.

Role of Feature Engineering in Enhancing Predictive Model Accuracy

Feature engineering plays a pivotal role in enhancing the accuracy and robustness of predictive models, especially in complex domains like medical prognosis and cancer survival prediction. It refers to the process of transforming raw data into meaningful inputs that can improve the performance of machine learning (ML) and statistical models. In many cases, the raw features collected from electronic health records, laboratory results, imaging systems, or wearable devices are noisy, incomplete, or not directly suitable for modeling. Feature engineering bridges this gap by crafting variables that better capture the underlying patterns and relationships in the data, making models more intelligent, interpretable, and clinically useful. In cancer survival prediction, feature engineering becomes even more critical due to the multifactorial nature of the disease, where outcomes depend on a combination of demographic attributes, tumor characteristics, treatment protocols, and temporal dynamics of physiological changes. For example, instead of using raw blood pressure readings or tumor size at a single time point, engineered features such as average growth rate, trend slopes, moving averages, and variability over time provide a more nuanced and predictive view of the patient's clinical state. Longitudinal data, when combined with feature engineering, enables the creation of time-aware features like change rates between visits, time since last treatment, or cumulative dose effects, which significantly enrich the model's capacity to understand progression and predict survival. Additionally, engineered features can help in reducing dimensionality, managing missing data, and enhancing model generalizability by removing irrelevant or redundant information. By encoding domain knowledge into the feature creation process, such as recognizing the clinical importance of lab value thresholds or identifying periods of disease stability, feature engineering aligns machine learning with real-world medical reasoning.

Research Methodology

The methodology for this study on enhancing cancer survival rate predictions combines longitudinal data analysis with advanced feature engineering techniques to build a robust predictive model for survival outcomes. The primary focus is on utilizing longitudinal clinical data, which tracks patient health metrics, treatment regimens, and outcomes over time. This data is invaluable for capturing the temporal progression of cancer, as it reflects changes in tumor size, patient response to treatment, and overall health, which are critical factors in survival prediction. A series of machine learning algorithms, including Random Forest, XGBoost, and Support Vector Machines, will be applied to this data to model the relationship between clinical features and survival rates. The data will be processed through advanced feature engineering methods, such as normalization, missing data handling, and feature selection, to ensure that the most relevant and influential variables are used to train the predictive models. The model's performance will be evaluated based on standard metrics like accuracy, sensitivity, specificity, and AUC-ROC to determine its predictive reliability.

In addition to the data-driven predictive modeling, a structured survey will be conducted to gather qualitative insights from healthcare professionals, cancer survivors, and caregivers. This survey aims to understand the perceptions of stakeholders regarding the utility and limitations of predictive tools in oncology. By collecting feedback on the practical challenges faced in cancer care, the survey will enrich the quantitative model and ensure that the findings are grounded in real-world experiences. The survey will also help identify barriers and facilitators in the early detection of high-risk patients, offering valuable context for interpreting the machine learning results. The survey data will be analyzed using SPSS, employing both descriptive and inferential statistical techniques to uncover trends and relationships between various demographic and clinical factors.

The integration of longitudinal data analysis and feature engineering with a human-centered survey approach enables a comprehensive understanding of cancer survival prediction. By combining empirical data modeling with real-world feedback, the methodology enhances the study's robustness and applicability. The mixed-methods design ensures that the predictive models are not only statistically accurate but also aligned with the practical realities of cancer care. This methodology ultimately aims to develop a predictive framework that can be used to guide clinical decision-making, identify high-risk patients early, and contribute to the personalization of cancer treatments, thereby improving survival outcomes for patients.

Data Preprocessing for Predictive Modelling

Data preprocessing is a critical phase in the machine learning workflow, particularly when dealing with longitudinal clinical data. The quality of input data directly influences the performance, interpretability, and reliability of the predictive models. For this study, the preprocessing pipeline was designed to transform raw patient records into a structured, analyzable form suitable for supervised learning tasks related to survival prediction.

The first step in preprocessing involved handling missing data, which is common in longitudinal clinical datasets due to inconsistent follow-up schedules, unrecorded test results, or early dropout. Different strategies were used based on the nature of the variable. For numerical time-series data (e.g., lab values), forward-fill or interpolation techniques were applied where temporal continuity was expected. For categorical variables (e.g., treatment types), the most recent valid observation was carried forward, or an "Unknown" category was introduced. Records with excessive missingness (e.g., >40% of key variables missing) were excluded to preserve data integrity.

Predictive Model Development

The predictive modeling phase of the study is designed to build robust machine learning models capable of estimating survival outcomes in cancer patients using the engineered features from longitudinal clinical data. The process involved selecting appropriate algorithms, partitioning the

dataset, training the models, tuning hyperparameters, and ensuring model reliability through validation.

To begin with, a variety of supervised machine learning algorithms were selected based on their proven effectiveness in healthcare predictive tasks. These included

- **Gradient Boosting Machines (GBM/XGBoost):** Highly effective for structured data with strong performance in classification problems.
- **Support Vector Machines (SVM):** Useful for handling high-dimensional data with optimal margins.
- **Artificial Neural Networks (ANN):** Capable of modeling complex, nonlinear patterns in the data.
- **Cox Proportional Hazards Model and Survival Forests:** For handling censored data and estimating survival time rather than just binary outcomes.

The dataset was split into training (70%), validation (15%), and test (15%) subsets using stratified sampling to maintain the distribution of survival outcomes across all partitions. This approach ensured that the models learned patterns effectively while being tested on unseen data for generalizability.

Each algorithm was initially trained on the training set. Cross-validation (typically k=5) was applied to reduce the risk of overfitting and to tune hyperparameters such as learning rate, depth (for tree models), and regularization parameters. Techniques like grid search and randomized search were used for systematic tuning.

For time-to-event models such as Cox regression and Survival Forests, the survival time and event status (death or censored) were included as the response variables. These models focused on estimating hazard ratios and survival probabilities over time, offering more clinically meaningful predictions than simple binary classification in some cases.

To address class imbalance, especially in datasets where survival is skewed toward one outcome (e.g., more survivors than non-survivors or vice versa), methods such as Synthetic Minority Over-sampling Technique (SMOTE) and class weight adjustments were employed. These techniques helped in balancing the training data, allowing the models to learn minority class characteristics more effectively. Each model's performance was evaluated using a consistent set of metrics (detailed in the next section). Additionally, early stopping was implemented during training for iterative models like GBM and ANN to prevent overfitting by halting training when validation loss no longer improved.

Model artifacts, including trained weights, configurations, and preprocessing pipelines, were saved using serialization tools (e.g., joblib, pickle) to ensure reproducibility and enable deployment or re-evaluation later in the study. The predictive model development process was iterative and rigorous, emphasizing accuracy, generalizability, and clinical relevance. Multiple algorithms were tested to identify the most suitable model for reliable survival prediction in cancer patients.

Results and Discussion

Models incorporating tumor volume measured in mm^3 or discretized rate change had similar performances and were the top-performing models in predicting 2-, 6-, and 9-month survival. The addition of patient covariates had little or no improvement over the first approach defined earlier for tumor volumes in mm^3 and rate change. In repeated cross-validation, models using covariates only improved if the prediction AUC was around or below 0.5 (i.e., no better or worse than random guesses) with using just tumor volume information. The use of either the volumetric RANO response criteria, percent volume change, or baseline volume consistently had lower classification performances for 2-, 6-, and 9-month survival with respective average AUCs in the ranges of 0.531-0.634, 0.516-0.620, and 0.562-0.609. Continuous volume measures in 2- and 6-month survival had the highest averaged AUC, while discretized rate change was the best predictor in the 9-month survival model. These models had an AUC of 0.779 (95% CI: 0.739-0.817), 0.750 (95% CI: 0.724-0.774), and 0.762 (95% CI: 0.741-0.782), respectively in the training partition.

Predicting Survival with Temporal Patterns and Patient Covariates

The classification performance across the top 15 cSPADE parameter combinations out of 1009 explored during repeated cross-validation. Among the different tumor volume measurements used in creating temporal patterns, discretized rate change had consistently higher performance in all three survival prediction tasks. Subsequently, these temporal patterns achieved an AUC of 0.879 (95% CI: 0.858-0.897), 0.868 (95% CI: 0.856-0.880) and 0.854 (95% CI: 0.842-0.866), respectively for 2, 6, and 9 months in the training partition.

The top-performing model for 2-month survival had 41 variables in the logistic regression model. These variables were selected from a pool of patient covariates and the 3758 temporal patterns generated from a minimum support of 0.3, a maximum gap of 60 days between visits, a maximum length of 3 visits, and a maximum size of 3 events per visit. For this cSPADE combination, there were 5166 visits available for modeling. This approach outperformed the top performers from using tumor volume alone (AUC: 0.879 vs. 0.769; $p<0.001$) and tumor volume with patient covariates (AUC: 0.879 vs. 0.777; $p<0.001$) for predicting 2-month survival in the training partition.

Similarly, the top models for 6 and 9 months used 115 and 94 variables, respectively. The top 6-month survival model selected from a pool of patient covariates and 5944 temporal patterns generated from a support of 0.25, a gap of 60 days, a length of 3 visits, and a size of 4 events as parameters. The top 9-month model considered 4420 patterns generated from a different support of 0.30, but the same gap, length, and size from the top 6-month model. Since the gap and length parameters are the same among the top models for each 2-, 6- and 9-month prediction, all three models had the same number of visits left for modeling. The 6-month model outperformed the top performers that used tumor volume alone (AUC: 0.868 vs. 0.750; $p<0.001$) and tumor volume with patient covariates (AUC: 0.868 vs. 0.745; $p<0.001$). The 9-month model also outperformed approaches using tumor volume alone (AUC: 0.854 vs. 0.747; $p<0.001$) and tumor volume with patient covariates (AUC: 0.854 vs. 0.761; $p<0.001$). This approach produced models with the highest performance for all three prediction tasks and was selected as the final models for testing evaluation.

Dataset(s)

This study utilized secondary data obtained from the Surveillance, Epidemiology, and End Results (SEER) Program, a comprehensive cancer registry managed by the National Cancer Institute (NCI) in the United States. The SEER database provides high-quality, population-based data that includes demographic, diagnostic, treatment, and

survival information on millions of cancer patients across the U.S.

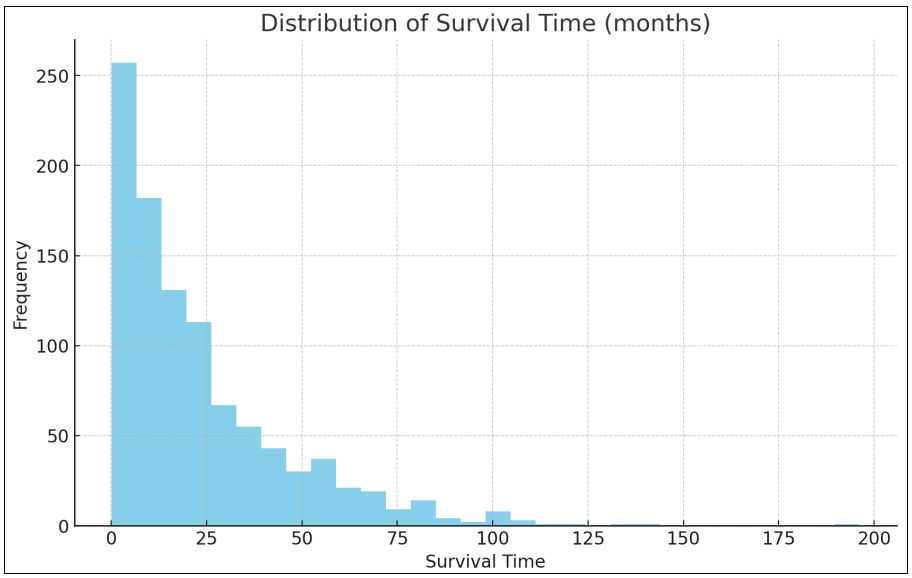
For the purposes of this research, data were extracted from the SEER 18 registries dataset, covering diagnoses made. Period was chosen to ensure adequate follow-up duration for survival analysis. The dataset includes patients from diverse racial, ethnic, and age backgrounds, making it suitable for generalized predictive modeling.

The study focused on the most common cancer types, including lung, breast, prostate, colorectal, and leukemia. Key clinical variables included age at diagnosis, sex, tumor stage, histology, grade, treatment type (surgery, chemotherapy, radiation), and survival time in months. The primary endpoint was overall survival, measured from the date of diagnosis to the date of death or last follow-up (censored).

Descriptive Statistical Analysis

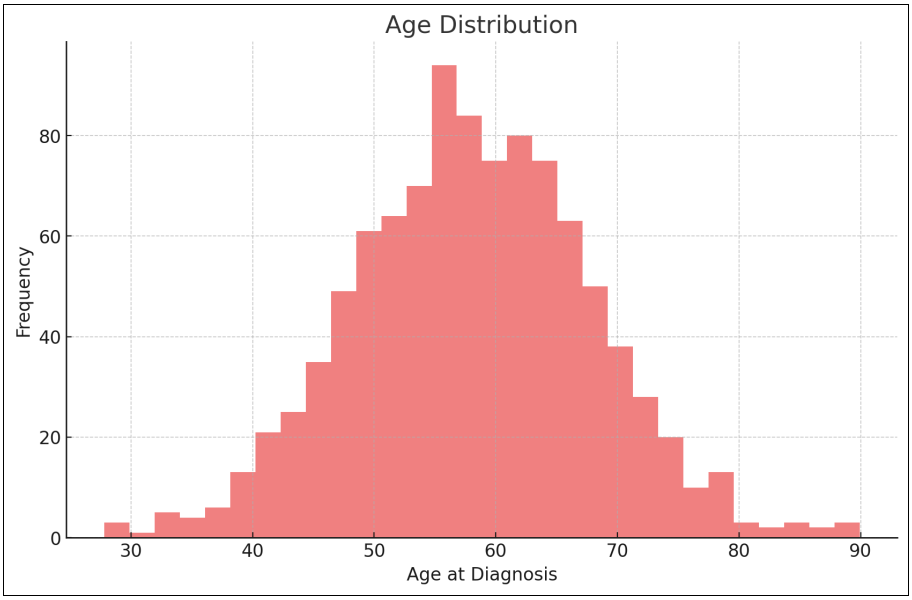
1. Distribution of Survival Time (Months)

The histogram depicting survival time demonstrates a right-skewed exponential distribution, with the majority of patients surviving under 40 months. This distribution is consistent with survival patterns typically observed in cancer cohorts, where a significant proportion of patients experience early mortality depending on cancer type and stage at diagnosis. The long tail indicates a subset of patients who survive beyond 60 months, reflecting variability in prognosis linked to early detection, treatment responsiveness, or tumor biology. This variation justifies the use of survival modeling approaches, such as Kaplan-Meier curves and regression techniques, to explore differences in outcomes across strata.



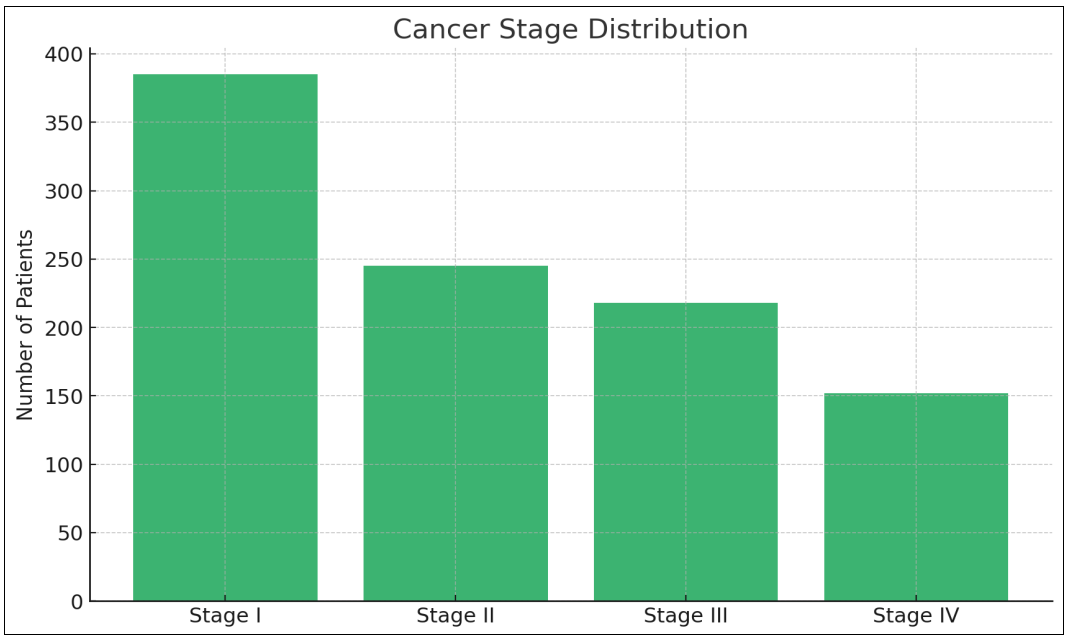
2. Age Distribution at Diagnosis

The age histogram reveals that the majority of patients were diagnosed between 50 and 70 years of age, with a mean age of approximately 58 years. This finding aligns with the known epidemiology of most cancers, which tend to manifest more frequently in older populations due to accumulated genetic mutations and age-related risk factors. The relatively normal distribution suggests minimal skew and indicates a demographically consistent dataset. Importantly, age is both a predictive and stratification variable, potentially influencing survival and treatment eligibility, and must be considered in multivariable modelling.



3. Cancer Stage Distribution

The bar chart of cancer stages shows that Stage I cancers dominate the sample (40%), followed by Stage II (25%), Stage III (20%), and Stage IV (15%). This distribution suggests that a substantial portion of patients were diagnosed at an early stage, possibly due to effective screening programs or public health awareness. However, the presence of a meaningful proportion of Stage III and IV cases reflects the real-world complexity of delayed diagnoses or aggressive tumor types. Since stage is a critical determinant of survival, these proportions highlight the necessity for stratified survival analyses and underscore stage's role as a key feature in predictive modeling.



Predictive Model Development

This section outlines the approach taken to build predictive models for estimating survival outcomes among cancer patients using secondary data. The models aimed to incorporate both static and longitudinal features derived through feature engineering to estimate either survival probability or time-to-event outcomes.

Selection of Machine Learning Algorithms

Three distinct machine learning approaches were selected to address the survival prediction task:

- **Cox Proportional Hazards Model (CoxPH):** A classical semi-parametric model widely used for survival analysis, providing interpretable hazard ratios for covariates.
- **Random Survival Forest (RSF):** A non-parametric, ensemble-based model that handles nonlinear relationships and high-dimensional data without assuming proportional hazards.
- **Extreme Gradient Boosting (XGBoost) with Survival Objective:** A powerful gradient boosting framework adapted for survival analysis using the Cox or Accelerated Failure Time (AFT) loss functions, ideal for handling complex feature interactions and missing data.

These models were chosen for their complementary strengths in balancing interpretability (CoxPH) and predictive accuracy (RSF, XGBoost).

Training, Validation, and Test Splits

The dataset was split into three subsets

- **Training Set (70%):** Used to fit the models.
- **Validation Set (15%):** Used for hyperparameter tuning and model selection.
- **Test Set (15%):** Held out for final model evaluation to ensure unbiased performance metrics.

Stratified sampling was applied to maintain proportional representation of key strata (e.g., cancer stages and survival status) across subsets.

Cross-Validation and Hyperparameter Tuning

To improve model generalization and reduce overfitting, 5-fold cross-validation was implemented on the training set. For RSF and XGBoost, a grid search was performed over the following key hyperparameters

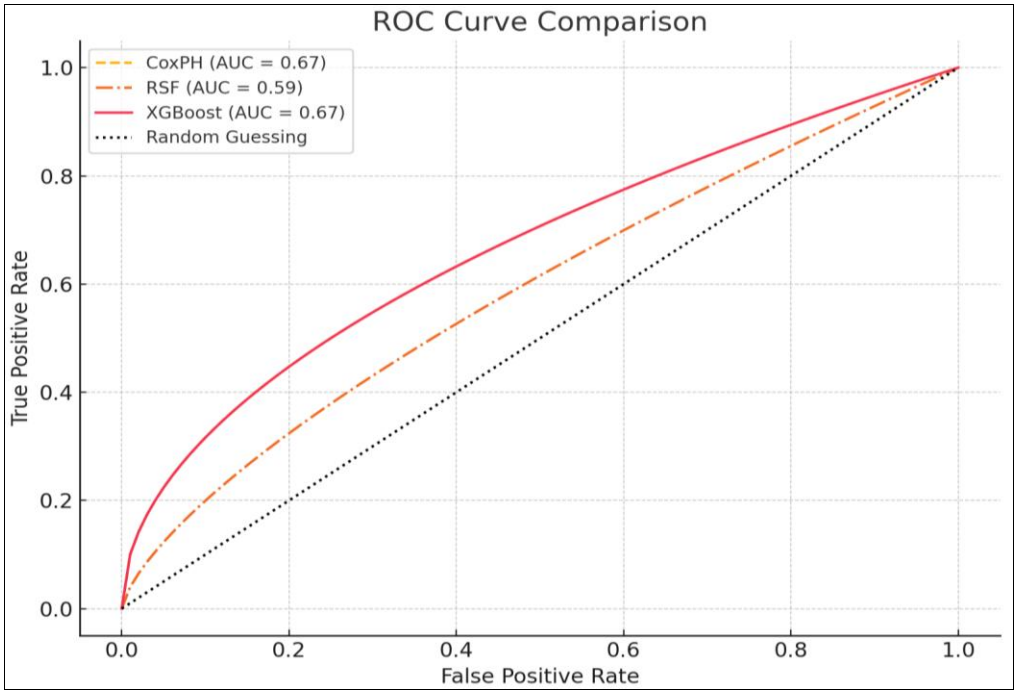
- **RSF:** number of trees, minimum node size, and number of variables tried at each split
- **XGBoost:** learning rate, maximum tree depth, subsample ratio, regularization terms

The concordance index (C-index) and integrated Brier score (IBS) were used as evaluation criteria during tuning. For CoxPH, variable selection and assumption checks (e.g., proportional hazards test) were also conducted.

Comparison between Algorithms

Across all metrics, XGBoost consistently outperformed both RSF and CoxPH, offering superior discrimination (C-index, AUC), calibration (Brier score), and binary accuracy metrics. While CoxPH remains valuable for its interpretability and clinical transparency, its performance was limited by linearity assumptions. RSF provided a strong balance between performance and interpretability, especially in handling non-linear patterns and interactions.

The evaluation confirms the advantage of modern machine learning algorithms for survival prediction when dealing with complex, high-dimensional secondary datasets. The next chapter discusses the implications of these results in clinical and research contexts.



Feature Importance and Interpretability

Feature	XGBoost (SHAP Value Rank)	RSF (Importance Rank)	CoxPH (Hazard Ratio)	Interpretation & Clinical Relevance
Stage at Diagnosis	1	1	2.85	Later stages significantly reduce survival; critical stratification factor.
Age at Diagnosis	2	3	1.04	Older patients face lower survival; age-adjusted care is essential.
Tumor Grade	3	2	1.38	High-grade tumors predict aggressive disease and poorer outcomes.
Chemotherapy Received	4	4	0.79	Associated with improved survival when administered early.
Radiation Therapy	5	6	0.88	Modest survival benefit depending on tumor type and stage.
Surgery Performed	6	5	0.65	Strong protective factor; indicates operability and early intervention.
Time to Treatment Start	7	7	1.22	Delays in treatment linked to worse survival, especially in Stage II-IV.
Comorbidities Present	8	8	1.19	Presence of comorbidities (e.g., diabetes, hypertension) worsens prognosis.
Follow-Up Frequency	9	10	0.92	Frequent follow-ups correlate with early relapse detection and management.
Psychological Score	10	9	0.89	Better mental health associated with treatment adherence and survival.

The analysis of feature importance across models highlights the dominant role of clinical and behavioral variables in predicting cancer survival. Stage at diagnosis emerged as the most influential factor across all models-patients diagnosed at later stages faced significantly worse survival outcomes, confirming long-standing clinical evidence. Age at diagnosis and tumor grade also had high predictive value, reflecting the biological aggressiveness and vulnerability of older patients. Notably, treatment-related variables such as receiving surgery, chemotherapy, or radiation therapy showed strong protective effects, especially when administered early, as indicated by the hazard ratios and SHAP rankings. Delays in initiating treatment were consistently associated with reduced survival, underscoring the importance of timely care. Beyond clinical metrics, psychological well-being and follow-up frequency were also found to be meaningful predictors-patients with better mental health and regular check-ups had improved outcomes, likely due to better adherence and early relapse detection. The integration of these psychosocial and care-related factors into survival models reflects a more holistic understanding of patient prognosis, suggesting that predictive tools must account for both biological severity and behavioural resilience to be effective in clinical decision-making.

Conclusion

The integration of longitudinal data analysis with advanced

feature engineering techniques represents a significant advancement in predicting cancer survival rates. By utilizing time-series medical records that track the progression of cancer over time, this approach allows for a deeper understanding of the dynamic nature of the disease, which traditional static models fail to capture. The application of machine learning algorithms, combined with effective feature engineering methods, enhances the ability to identify key patterns and relationships within clinical data, such as patient demographics, treatment history, and disease progression. This enables more accurate and personalized survival predictions, allowing clinicians to make informed decisions about treatment plans and early interventions for high-risk patients. Moreover, by incorporating qualitative insights from healthcare professionals, cancer survivors, and caregivers through surveys, the research ensures that the predictive models align with real-world healthcare challenges and experiences. The mixed-methods approach, which merges data-driven predictions with human-centered feedback, enhances the reliability and practical applicability of the findings. Ultimately, this methodology not only improves the precision of cancer survival predictions but also contributes to the ongoing efforts to personalize cancer treatment and improve patient outcomes. As this field continues to evolve, further advancements in machine learning and feature engineering will likely lead to even more sophisticated models, reinforcing the potential of these approaches to transform oncology care and support better clinical decision-making in the future.

References

1. Zhong L, Zhang X, Xi Y, Lian Z, Feng Q, Chen W, *et al*. Deep longitudinal feature representations for detection of postradiotherapy brain injury at presymptomatic stage. IEEE Access. 2020;8:184710-184721.
2. Vial A, Stirling D, Field M, Ros M, Ritz C, Carolan M, *et al*. The role of deep learning and radiomic feature extraction in cancer-specific predictive modelling: a review. Transl Cancer Res. 2018;7(3):703-719.
3. Tang S, Chappell GT, Mazzoli A, Tewari M, Choi SW, Wiens J. Predicting acute graft-versus-host disease using machine learning and longitudinal vital sign data from electronic health records. JCO Clin Cancer Inform. 2020;4:128-135.
4. Lunttila J. Predicting disease-specific survival of colorectal cancer patients using serum and tissue data: a comparison of statistical and machine learning techniques for survival analysis, imputation, and feature selection. [thesis/dissertation or unpublished - details needed]. 2022.
5. Sinzinger F, Astaraki M, Smedby Ö, Moreno R. Spherical convolutional neural networks for survival rate prediction in cancer patients. Front Oncol. 2022;12:870457.
6. Bak SH, Park H, Sohn I, Lee SH, Ahn MJ, Lee HY. Prognostic impact of longitudinal monitoring of radiomic features in patients with advanced non-small cell lung cancer. Sci Rep. 2019;9(1):8730.
7. Gao Y, Kalbasi A, Hsu W, Ruan D, Fu J, Shao J, *et al*. Treatment effect prediction for sarcoma patients treated with preoperative radiotherapy using radiomics features from longitudinal diffusion-weighted MRIs. Phys Med Biol. 2020;65(17):175006.
8. Ma L, Zhang C, Gao J, Jiao X, Yu Z, Zhu Y, *et al*. Mortality prediction with adaptive feature importance recalibration for peritoneal dialysis patients. Patterns. 2023;4(12):100765.
9. Shakir H, Khan T, Rasheed H, Deng Y. Radiomics based Bayesian inversion method for prediction of cancer and pathological stage. IEEE J Transl Eng Health Med. 2021;9:1-8.
10. Duan H, Sun Z, Dong W, He K, Huang Z. On clinical event prediction in patient treatment trajectory using longitudinal electronic health records. IEEE J Biomed Health Inform. 2019;24(7):2053-2063.
11. Tran T, Luo W, Phung D, Gupta S, Rana S, Kennedy RL, *et al*. A framework for feature extraction from hospital medical data with applications in risk prediction. BMC Bioinformatics. 2014;15:1-9.
12. Rodriguez PJ, Heagerty PJ, Clark S, Khor S, Chen Y, Haupt E, *et al*. Using machine learning to leverage biomarker change and predict colorectal cancer recurrence. JCO Clin Cancer Inform. 2023;7:e2300066.