Lung Cancer Survival Prediction using Machine Learning and Statistical Methods

Varun Nair

Master of Analytics

Auckland University of Technology

yfg2694@autuni.ac.nz

Supervisor: Dr. Victor Miranda

June 2025

Contents

List of Figures

List of Tables

Abstract

Lung cancer remains one of the leading causes of cancer-related mortality, underscoring the need for accurate survival prediction tools. This study investigates the effectiveness of statistical, machine learning, and dynamic survival models in predicting patient outcomes using clinical and post-treatment data. Using Cox Proportional Hazards and Random Survival Forests, we found that PFI time, age, gender, and residual tumor variables were key predictors, with RSF achieving a C-index of 0.86 and Cox Proportional Hazard 0.90. The findings highlight the importance of both baseline and post-treatment information in the context of building a dynamic survival model. Future work will focus on assumption testing, external validation, and hybrid modeling for enhanced clinical applicability.

Keywords- Lung Cancer Survival, Cox Proportional Hazards Model, Random Survival Forests, Dynamic Survival Modeling, Kaplan-Meier Analysis.

Chapter 1
Introduction

Lung cancer remains a leading cause of cancer-related mortality worldwide, accounting for a significant proportion of cancer diagnoses and deaths each year. Accurate survival prediction plays a vital role in guiding treatment decisions, resource allocation, and patient counseling.[7] Traditional prognostic models based on static variables such as age, gender, and tumor stage provide limited insight into individual patient outcomes.As a result, there is growing interest in creating prediction models that take into account data from both baseline and post-treatment to account for patients changing clinical status.

Among the most commonly used methods in survival analysis are the Kaplan-Meier estimator, the Cox proportional hazards model, and various machine learning techniques. A common non-parametric technique for comparing survival between groups is the Kaplan-Meier estimator, which calculates the likelihood of surviving over time. This is furthered by the Cox proportional hazards model, which estimates hazard ratios for various risk variables by adding covariates. [16] More recently, Random Survival Forests (RSF) have been used to analyze survival data as a machine learning technique that can handle high-dimensional data and intricate interactions while taking censoring into account. However, most existing models fail to integrate time-updated clinical variables, limiting their ability to adapt predictions to a patient's evolving condition. This presents a critical problem in survival prediction modeling that this study seeks to address.

By using a dynamic modeling technique, our study sets itself apart. In particular, we use PFI.time (Progression-Free Interval time) as a predictor, which records data regarding the course of the disease following initial therapy. The model may modify survival forecasts in response to modifications in a patient's clinical state thanks to this time-updated variable. The principles of dynamic survival modeling, in which predictions change as data accumulates, are consistent with this method. Recent research that incorporates time-varying imaging biomarkers or uses personalized, real-time prediction methods for the prognosis of lung cancer has validated this idea.

While some studies have explored time-varying data using imaging or biomarkers[6], there is limited research combining classical survival methods with machine learning on structured clinical datasets using dynamic predictors. In order to estimate patient survival outcomes in lung cancer, this study aims to assess and compare three popular survival analysis techniques: Random Survival Forests, Cox proportional hazards modeling, and Kaplan-Meier estimation. The study intends to evaluate each model's capacity to manage censored data, identify risk patterns, and produce significant diagnostic insights for individual patients by applying these techniques to a for treatment relevant dataset that includes baseline and follow-up data.In this study, overall survival time refers to the period from diagnosis to death or last follow-up, and censoring occurs when this outcome is unknown due to incomplete follow-up.

This report includes a review of the literature of related work, a statement of the research question, and a methodology section that describes the data set and the models used. Results are then presented and discussed, followed by conclusions, limitations, and suggestions for future research.

1.1 Motivation

This work is driven by a strong desire to apply analytical methods to practical healthcare problems, especially in oncology, where clinical decision-making can be directly influenced by predictive modeling. Because of its high mortality rate and intricate course, lung cancer is a crucial topic for studies aimed at predicting survival. From the standpoint of data science, survival analysis provides a rich fusion of contemporary machine learning methods with traditional statistical methods, opening up possibilities for the creation of powerful and interpretable models. This project is particularly interesting because it investigates dynamic modeling, an emerging field with substantial promise to enhance patient-specific diagnosis, in addition to static prediction. Clinicians looking for more sophisticated tools to direct treatment planning and follow-up care will find this study useful in addition to statisticians and machine learning practitioners.

Chapter 2
Literature Review

Researchers have developed a variety of survival prediction models in the past decades, ranging from traditional statistical methods to modern machine learning and deep learning approaches. Each methodology offers unique strengths, but also presents limitations. This section critically examines the evolution of survival modeling techniques in lung cancer, synthesizes key findings, highlights methodological limitations, and identifies unresolved gaps in the current research landscape.In particular, it explores how advancements in computational power and data availability have enabled more complex and dynamic models. Despite these innovations, challenges remain in model interpretability, generalizability across populations, and integration with clinical workflows.

2.1 Statistical Methods: Kaplan-Meier and Cox Proportional Hazards

Survival analysis in oncology has traditionally relied on well-established statistical models, most notably the Kaplan-Meier (KM) estimator and the Cox Proportional Hazards (CPH) model. These methods have formed the cornerstone of survival prediction and prognosis estimation in clinical research for decades. The Kaplan-Meier estimator, in particular, is a non-parametric method that is extensively used for estimating survival probabilities and generating survival curves. It provides an intuitive graphical representation of survival functions, allowing researchers and clinicians to easily compare survival experiences across different patient groups, such as treatment types, stages of disease, or biomarker levels. Its major strength lies in its ability to handle censored data that is, data from patients whose survival time is unknown due to loss to follow-up or the end of the study period. As a result, it has become a standard tool for exploratory survival analysis in clinical studies [1][2].

However, the KM estimator is fundamentally descriptive. It does not account for covariates or model the relationship between survival time and patient-specific features. This limitation becomes critical when attempting to build predictive models or understand how individual risk factors contribute to survival outcomes. The inability to include multiple explanatory variables simultaneously reduces its utility in multivariate settings or personalized medicine contexts.

To overcome these constraints, the Cox Proportional Hazards model was introduced as a semi-parametric alternative. Unlike KM, the CPH model enables the inclusion of multiple covariates, such as age, gender, tumor size, and clinical biomarkers, to estimate hazard ratios the relative risk of the event (e.g., death) occurring at a particular time point given the covariates. This approach offers both flexibility and interpretability, making it a widely used method in prognostic modeling. Studies such as [3] and [4] have demonstrated the effectiveness of the Cox model in predicting lung cancer outcomes, especially in the construction of clinical nomograms where variables like Eastern Cooperative Oncology Group (ECOG) status, tumor diameter, and serum biomarkers serve as significant predictors.

Despite its strengths, the CPH model makes a critical assumption: the proportionality of hazards over time. This means that the effect of a covariate is assumed to remain constant throughout the entire study period. In real world datasets especially in heterogeneous populations such as those with lung cancer this assumption may not hold. Disease progression, treatment effects, and patient response often vary over time, leading to violations of the proportional hazards assumption. Furthermore, the Cox model has limited capacity to capture non-linear relationships or complex interactions between variables. This restricts its effectiveness in high-dimensional or

dynamically changing clinical environments.

2.2 Machine Learning Methods: Random Survival Forests and Classification Models

Recent advances in machine learning (ML) have led to the development of more flexible survival prediction models, particularly those capable of capturing non-linear relationships, modeling interactions between variables, and handling high-dimensional data. Among these, Random Survival Forests (RSF) have gained prominence. RSF is an ensemble-based method that extends traditional decision tree algorithms to survival data. Instead of using a single tree, it constructs multiple survival trees based on bootstrapped samples and aggregates their predictions, enabling more robust and accurate results. This approach can accommodate complex covariate interactions, is naturally suited for handling missing values, and does not rely on proportional hazards or parametric assumptions.

Studies such as [5] have shown that RSF often outperforms traditional models like CPH in both short-term and long-term survival prediction, especially when working with heterogeneous datasets. Another study [4] applied RSF to a lung cancer cohort and reported a superior concordance index (C-index) compared to Cox models, reflecting better alignment between predicted and observed survival outcomes. RSF's ability to model non-linear effects is particularly valuable in clinical settings where variable interactions (e.g., between treatment type and genetic mutations) significantly affect patient outcomes.

In parallel, classification-based models have been used to simplify survival analysis by converting time-to-event data into categorical labels. For example, patients may be grouped into survival classes such as less than 6 months, 6–24 months, and more than 24 months. These models are generally easier to interpret in clinical decision-making, especially for triage or stratification purposes. When combined with regression approaches or probability calibration techniques, classification models can effectively segment patient populations and guide treatment pathways.

However, this simplification introduces significant information loss, as survival time is discretized rather than modeled as a continuous variable. The performance of classification models is also heavily influenced by the choice of time cutoffs, which may be arbitrarily selected or not clinically meaningful. Moreover, many classification models lack transparency and generalizability across different clinical settings.While machine learning methods offer substantial gains in predictive accuracy, a common criticism is their black-box nature.[14][12] Without rigorous post hoc interpretability tools (such as SHAP or LIME), it becomes difficult to explain why a model made a specific prediction a major barrier to adoption in evidence based medicine.

2.3 Deep Learning Models

Deep learning (DL) has recently gained traction in survival analysis due to its ability to model complex relationships and manage incomplete clinical data. In the study [7] a novel transformer-based architecture was proposed to predict overall survival in non-small cell lung cancer (NSCLC) patients. Unlike traditional approaches, this model directly handles missing values without requiring imputation by encoding temporal and semantic relationships within clinical features. The architecture integrates both static and time-dependent variables and learns latent representations that enhance survival prediction, even in the presence of incomplete or irregularly sampled patient data. This methodology reflects a shift toward more flexible and robust modeling frameworks in clinical prognostics.

Another promising approach is DeepHit, introduced in [6][15], which uses a deep neural architecture to estimate

survival probabilities directly. This model accounts for competing risks and generates a full survival distribution for each patient, improving flexibility and accuracy. Tools like LIME (Local Interpretable Model-agnostic Explanations) have also been applied to these architectures, providing local explanations for individual predictions and partially addressing the interpretability concern.

Despite these advancements, DL models face significant practical limitations. They require large volumes of high-quality, annotated training data, which are often unavailable in clinical practice due to privacy issues, inconsistent data standards, and sparse follow-up records. Furthermore, training deep models involves substantial computational resources and expertise, making their integration into clinical workflows difficult. Even with emerging explainability techniques, most deep learning models still struggle to achieve the transparency needed for clinical accountability. As a result, adoption remains limited to research settings or highly resourced institutions.

2.4 Dynamic Modeling Approaches

In recent years, dynamic survival modeling has emerged as a critical area of innovation. Unlike static models, which rely solely on baseline covariates, dynamic models incorporate time-updated information, reflecting the patient's evolving clinical trajectory. These models are particularly well-suited for diseases like lung cancer, where treatment response, biomarker levels, and functional status can change significantly over time.

Dynamic modeling techniques include methods that update survival probabilities as new information becomes available for instance, tumor recurrence, treatment changes, or follow-up imaging results. Studies like [4] and [7] have developed such models using time-series clinical data and recurrent neural networks, allowing real-time survival prediction and continuous risk stratification. These models align closely with the real-world nature of patient care, where clinicians adjust decisions based on a patient's latest condition rather than static baseline features.

However, deploying dynamic models in practice poses multiple operational and technical challenges. The most significant is the need for longitudinal data collection, which is not standardized across healthcare systems. Many electronic health records lack the consistency or structure required for effective time-series modeling. Moreover, real-time prediction systems must be continuously retrained and recalibrated as new data becomes available, demanding robust infrastructure and cross-functional collaboration between data scientists and clinicians. The cost and complexity of implementing such systems often outweigh perceived benefits, especially in resource-limited settings. Additionally, few studies have quantitatively measured the incremental benefit of incorporating updated variables over static models, raising questions about the return on investment for dynamic modeling pipelines.

2.5 Identified gaps in Literature

In summary, although there has been substantial progress in the development of statistical, machine learning, and deep learning models for lung cancer survival prediction, critical gaps remain. Many studies have evaluated these models in isolation, without systematic comparisons across methodological categories. Most modeling efforts are based on static datasets, which fail to capture the time-varying nature of disease progression and treatment response. While dynamic modeling frameworks offer promise, their adoption has been limited by a lack of standardized longitudinal data, the absence of robust validation frameworks, and significant implementation challenges.Moreover, very few studies have rigorously evaluated how much predictive performance improves with time-updated covariates such as treatment response, progression status, or biomarker trajectories. The lack of quantitative benchmarks for dynamic models creates uncertainty around their clinical value. This study aims to bridge these gaps by conducting a comprehensive evaluation of classical, machine learning, and dynamic survival models on a structured patient dataset that includes both baseline and follow-up variables. By doing so, we aim to assess not only the predictive accuracy of each modeling framework but also its practical feasibility and interpretability in a real-world clinical context.

Chapter 3
Research Questions and Scoping

The main research question guiding this study is:

How do statistical, machine learning, and dynamic survival models perform in predicting lung cancer patient outcomes using clinical data that includes both baseline and post-treatment information?

Sub-questions explored in this study include:

- Which survival modeling approach best handles censored data while providing meaningful insights for patient-level prognosis?
- What trade-offs exist between predictive accuracy and interpretability across these modeling approaches in a clinical context?

This question arises from critical gaps identified in the literature: while Kaplan-Meier and Cox Proportional Hazards models have been extensively used due to their interpretability and clinical familiarity [1][2][3], they assume proportional hazards and struggle to accommodate time-varying covariates. Random Survival Forests and classification models, as applied in [5] and [4], offer improved predictive performance but often lack transparency. More recent deep learning models such as DeepHit [6] and transformer-based networks [7] integrate time-series data but are resource-intensive and lack clinical interpretability.

Despite promising developments, few studies have evaluated these models in parallel using dynamic predictors such as PFI.time and Residual tumor which reflect a patient's evolving status following treatment. Furthermore, most existing work either focuses on static datasets or examines only a single modeling framework in isolation, limiting comparative insights and real-world applicability.

3.1 Aim of the Study:

This study aims to compare two key survival modeling approaches: the Cox proportional hazards model and the random survival forest to predict lung cancer patient outcomes using structured clinical data that include both baseline and follow-up variables. Kaplan-Meier analysis is also used to provide descriptive insights into group-wise survival trends. The evaluation focuses on model performance, clinical interpretability, and applicability in real-world settings where patient status evolves over time.

3.2 Scope of this study:

This study is confined to the use of structured clinical data derived from lung cancer patients, including both baseline and follow-up information such as demographic characteristics, tumor staging, and treatment outcomes. It deliberately excludes imaging data (e.g., CT or PET scans) and unstructured clinical text (e.g.,

physician notes or discharge summaries), which are often utilized in deep learning applications but require separate preprocessing pipelines and natural language processing techniques. Furthermore, the analysis does not incorporate high-dimensional biomarker or genomic datasets unless the variables are available in a tabular, structured format suitable for traditional modeling. Lastly, while the study evaluates models with the potential for dynamic updating, it does not involve real-time deployment or live data streaming from clinical systems due to infrastructural constraints. However, these aspects are discussed in the context of future research opportunities.

3.3 Contributions:

By combining and comparing multiple survival modeling approaches in dynamic clinical data, this study contributes the following:

Benchmark Evaluation: It provides a comprehensive evaluation of survival models specifically the Cox Proportional Hazards model, Kaplan-Meier estimation, and Random Survival Forests using real-world clinical data that includes both baseline and follow-up features. This benchmark enables objective comparison across statistical and machine learning methods in the context of lung cancer survival prediction.

Model Performance Insights: The study explores the strengths and limitations of each approach, offering insights into the trade-offs between interpretability, predictive performance, and their suitability for evolving clinical scenarios. Particular attention is given to how each method manages censored data, non-linearity, and variable interactions.

Temporal Adaptability and Clinical Relevance: By integrating time-updated clinical features, the analysis demonstrates how dynamic modeling techniques can be used to reflect patient progress after treatment. This is crucial in oncology, where survival risk may shift due to interventions or disease progression.

Practical Implementation Considerations: The findings highlight not only the methodological differences but also the practical challenges of applying these models in clinical settings such as data availability, infrastructure requirements, and ease of interpretation for healthcare professionals.

Guidance for Future Research and Application: Finally, this study identifies areas where current modeling strategies fall short and outlines avenues for future work, such as improving model explainability, standardizing longitudinal data collection, and enabling integration with clinical decision-support systems.

Chapter 4
Research Design/Methods

This study employs a structured, comparative modeling approach to evaluate the effectiveness of different survival analysis techniques in predicting patient outcomes in lung cancer. The primary goal is to assess how well statistical and machine learning models perform when applied to structured clinical datasets that include both baseline and post-treatment information. The research is driven by the central question: How do Cox Proportional Hazards and Random Survival Forest models compare in terms of predictive accuracy, interpretability, and adaptability to time-updated clinical data?

4.1 Research design

This study investigates the research question: 'How do statistical, machine learning, and dynamic survival models perform in predicting lung cancer patient outcomes using clinical data that includes both baseline and post-treatment information?' To address this, a comparative analysis was conducted using real-world clinical data from lung cancer patients. The aim was to evaluate and compare the predictive performance, interpretability, and clinical relevance of three survival modeling approaches: the Kaplan-Meier estimator, Cox Proportional Hazards (Cox PH) model, and Random Survival Forests (RSF).

The research follows a structured pipeline beginning with data acquisition from the TCGA LUAD dataset, followed by integration and preprocessing of clinical variables. Predictors were then selected based on clinical relevance and data completeness. Three survival models Kaplan-Meier, Cox Proportional Hazards, and Random Survival Forest were developed and evaluated. The final step involved comparing these models in terms of predictive accuracy, interpretability, and clinical applicability.

Data collection: The clinical data for this study were obtained from the publicly available TCGA LUAD (Lung Adenocarcinoma) cohort via the UCSC Xena Browser. Two key datasets were downloaded and the coding part of the project was done using Google colab:

LUNG survival.txt – containing survival-related variables, including overall survival (OS), OS time, progression-free interval (PFI), and vital status. LUNG_survival.txt dataset

LUAD clinicalMatrix.json – containing a range of clinical features such as age at diagnosis, gender, tumor stage, treatment response, and additional pathological and molecular indicators. LUAD_clinicalMatrix.json dataset.

The use of TCGA data ensures high reliability, standardization, and quality, as it is sourced from multiple leading cancer research centers. As the datasets are publicly available and fully de-identified, no additional ethical clearance is required for their use in this secondary analysis.

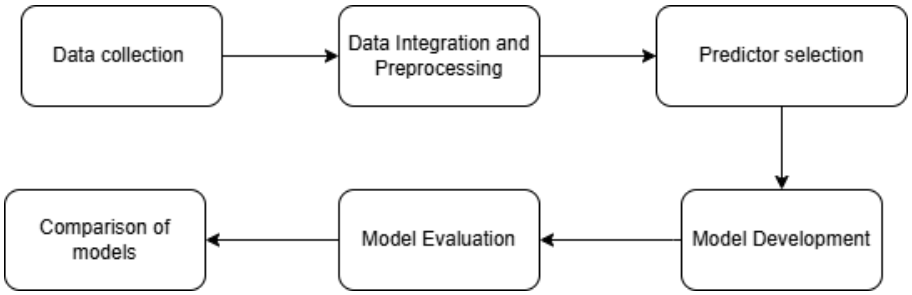

Fig. 4.1. Workflow diagram illustrating the structured research pipeline.

4.2 Data Preprocessing:

Preprocessing was a critical step in preparing the dataset for survival analysis and ensuring methodological rigor. The dataset was constructed by merging two publicly available sources from the TCGA LUAD cohort via the UCSC Xena Browser, using the sample ID as the common key. The resulting integrated dataset contained approximately 1,300 patient records and 146 variables, including demographic, clinical, and survival information. As per ethical standards, all data were de-identified and publicly accessible, eliminating concerns related to participant privacy or consent.

The preprocessing began by removing completely empty columns , followed by an initial inspection of missing data. While rows with missing values in critical target variables (OS.time and OS) were dropped to preserve survival label integrity, removing rows with missing predictor variables significantly reduced the dataset size. Therefore, missing values in predictor variables were imputed using median imputation via SimpleImputer from scikit-learn, preserving as much usable data as possible without making unrealistic assumptions about distribution.

Next, categorical variables were encoded to make them suitable for machine learning algorithms. The gender variable was label encoded, converting string categories into binary numerical format. The residual tumor variable, which had multiple categories, was one-hot encoded , allowing the model to treat each category independently without assuming any ordinal relationship. This encoding was particularly important for the Random Survival Forest, which is sensitive to data types.

A correlation heatmap was also generated to support variable selection. It helped identify features with stronger linear relationships to the target variable (OS.time) and ensured low multicollinearity between predictors. Based on both clinical relevance and correlation patterns, features such as PFI.time, age at initial pathologic diagnosis, days to new tumor event after initial treatment, gender, and residual tumor were retained. Variables with negligible correlation or insufficient clinical justification were excluded to improve model clarity and generalizability.

Predictor selection was guided by a combination of clinical relevance, data completeness, correlation with the target variable, and exploratory data analysis (EDA). Variables that are well-established in the oncology literature (e.g., age at diagnosis, tumor characteristics) were prioritized. Features that showed a meaningful statistical association with OS.time in the correlation matrix, and had minimal missing values, were retained. In contrast, variables with weak correlation, excessive missingness, or ambiguous clinical interpretation were excluded to enhance model stability and clarity.

Finally, a structured dataset with encoded, imputed, and filtered features was created. It was split into training and testing sets using an 80-20 stratification to evaluate model performance reliably. The complete preprocessing pipeline ensured consistency, minimized bias due to data loss, and aligned with the ethical and technical standards

expected for clinical survival modeling.

4.3 Model Development

To address the research objective of comparing statistical, machine learning, and dynamic survival models for predicting lung cancer outcomes, this study implements three modeling approaches: the Kaplan-Meier (KM) estimator, the Cox Proportional Hazards (CPH) model, and Random Survival Forests (RSF). Each method brings distinct strengths and assumptions to the analysis of time-to-event data.

4.3.1 Kaplan Meier Estimator

The Kaplan-Meier estimator is a non-parametric statistic used to estimate the survival function from observed survival times. It is particularly useful for visualizing survival probabilities and comparing different groups (e.g., based on gender or age). In this study, KM curves were generated for the overall dataset, and stratified analyses were conducted by gender and age group to explore subgroup survival patterns.

The survival probability at time t is denoted by S(t), is given by the product-limit formula

$$S(t) = \prod_{t_i \leq t} \left(1 - \frac{d_i}{n_i}\right), \qquad \text{——————} (1)$$

where:
- t_i is the time of the i-th event (e.g., death)
- d_i is the number of events at t_i
- n_i is the number of individuals at risk just prior to t_i.

This estimator accounts for right-censored data, where the exact time of the event is unknown for some individuals due to loss to follow-up or study end[1][2].

4.3.2 Cox Proportional Hazard Modelling:

The Cox Proportional Hazards model is a semi-parametric method that models the hazard function, allowing for the inclusion of multiple covariates. It assumes that the effect of each covariate is multiplicative with respect to the baseline hazard and remains constant over time (i.e., proportional hazards assumption).

The hazard function for subject i at time t is :

$$h_i(t) = h_0(t)\exp(\beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}), \qquad \text{——————} \quad (2)$$

where: - $h_i(t)$ is the hazard for individual i at time t
- $h_0(t)$ is the baseline hazard function,
- x_{ij} are the covariates for individual i,
- β_j are the coefficients corresponding to each covariate.

This model provides interpretable hazard ratios for each predictor and is widely used in clinical research due to

its simplicity and effectiveness in analyzing censored data[8].

4.3.3 Random Survival Forest:

Random Survival Forests (RSF) are an ensemble learning method that extends decision trees to time-to-event data. RSFs can handle high-dimensional datasets, non-linear relationships, and complex interactions without making assumptions about the underlying hazard distribution.

$$\hat{S}(t \mid X_i) = P(T > t \mid X = X_i), \qquad \text{——————} (3)$$

where:

T is the time-to-event (e.g., death, failure)

X i is the feature vector for individual

$\hat{S}(t \mid X_i)$ is the estimated survival probability at time t.

In this study, the RSF model was trained using the scikit-survival package, with survival labels encoded using Surv.from dataframe(). The model was optimized using parameters such as the number of trees, minimum samples per split, and maximum features. RSF provided risk predictions and a concordance index (C-index) was used to evaluate performance.

By integrating classical statistical techniques with machine learning-based RSF, this study allows for both interpretability and improved predictive power. The three models were subsequently evaluated and compared based on performance metrics and clinical interpretability.

To assess the relative performance of the survival models, this study employs a combination of evaluation metrics. The Concordance Index (C-index) serves as the primary metric, reflecting each model's ability to correctly rank patient survival times. Additionally, ROC (Receiver Operating Characteristic) curves are used to assess the discriminative ability of the models at specific time thresholds (1000 days), providing a visual and quantitative basis for comparison. These metrics are complemented by an analysis of model interpretability and underlying assumptions

While the primary focus of this study is on survival modeling using time-to-event data, preliminary experiments with classification-based approaches on a separate cancer dataset were conducted during the early phases of model exploration. These models, however, yielded poor results achieving around 50 percent accuracy with weak precision and recall highlighting their unsuitability for predicting censored survival outcomes. This outcome further reinforced the decision to adopt dedicated survival analysis techniques such as Cox Proportional Hazards and Random Survival Forests for this study, as aligned with the central research objective. All modeling steps and method selections were discussed in regular consultation with the research supervisor to refine the methodological choices and align them with the research question.

Chapter 5
Application of research method

5.1 Data Preprocessing

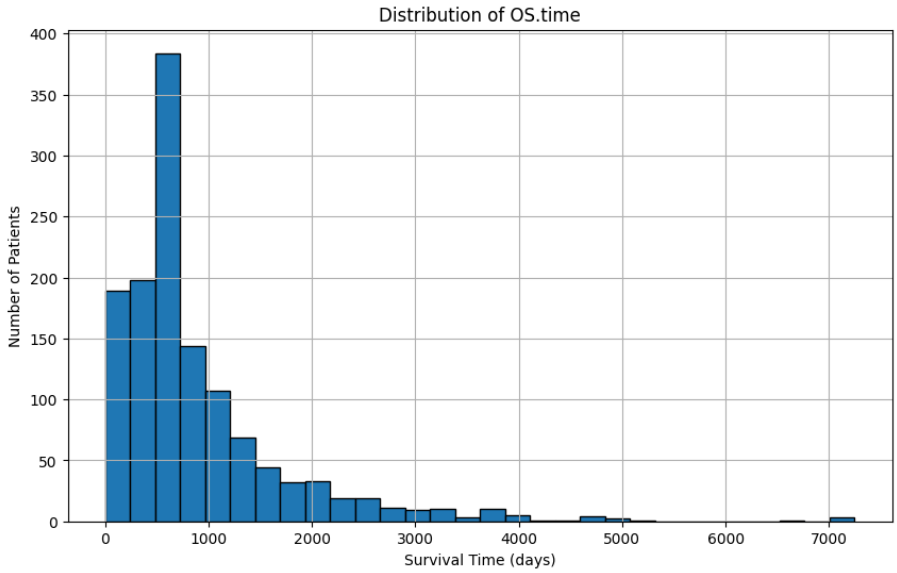

Fig. 5.1. Histogram of Overall Survival Time(Target Variable)

The histogram of OS.time (Figure 5.1) clearly shows a right-skewed, non-symmetric distribution with a long tail of high survival values. This shape deviates significantly from a normal (Gaussian) distribution, which would appear as a bell-shaped and symmetric curve [9]. The presence of skewness and outliers violates the assumption of normality, which underlies many traditional parametric models. Therefore, this justifies the use of non-parametric survival models (e.g., Kaplan-Meier) and semi-parametric models (e.g., Cox Proportional Hazards), which are better suited for handling non-normal, censored survival data.

The histogram distribution suggested the potential presence of outliers particularly patients with unusually long survival durations. To investigate this further, a boxplot was generated (Table 5.1), which confirmed the presence of multiple outliers. These are represented by individual data points lying beyond the upper whisker of the box, indicating survival times far exceeding the typical range defined by the interquartile range (IQR). The identification of these outliers is crucial, as they can affect the stability and interpretability of some statistical models. Their presence supports the use of robust, non-parametric methods such as the Kaplan-Meier estimator, and machine learning models like Random Survival Forests, which are better suited to handle non-normal, right skewed, and censored survival data.

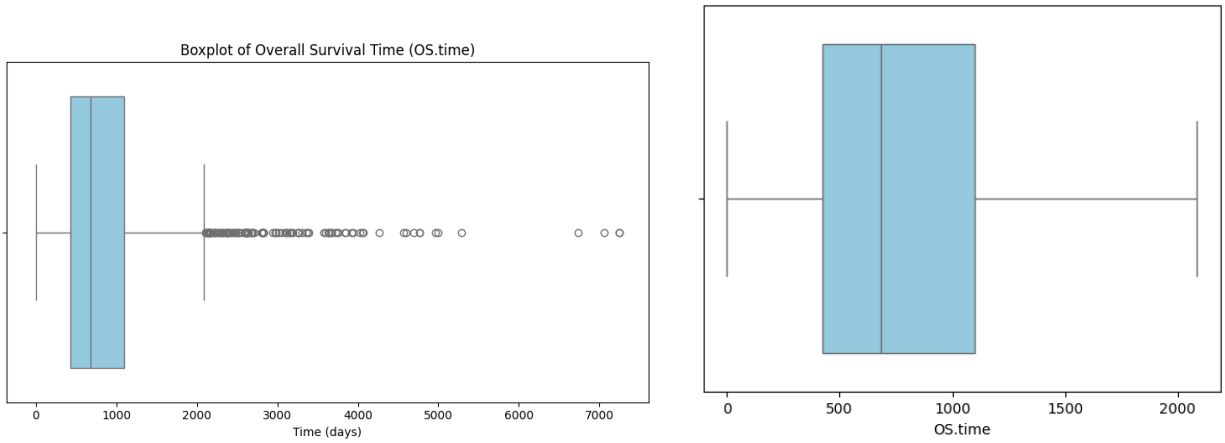

Table 5.1. Boxplot of Overall survival time before and after removing outliers

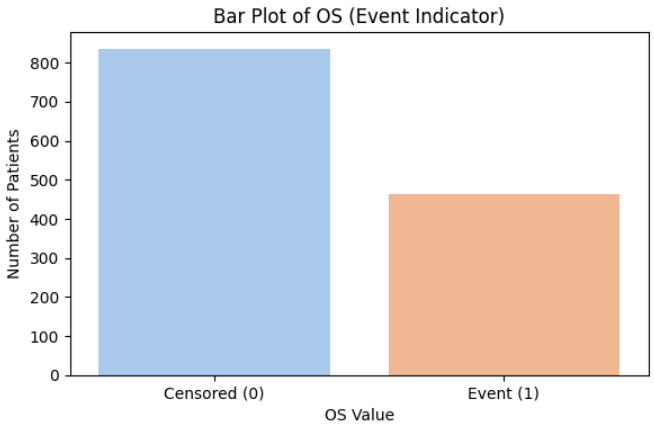

Fig. 5.2. Barplot of Overall Survival

After identifying extreme survival durations through the initial histogram and boxplot, outliers in the OS.time variable were removed using the Interquartile Range (IQR) method. Specifically, any values falling below Q1 1.5 × IQR or above Q3 + 1.5 × IQR were excluded. The resulting boxplot (Table 5.1) displays a more compact distribution of survival times, with no high-end outliers present. This transformation was essential to reduce the influence of extreme values that could potentially skew statistical analyses and inflate model error especially in regression-based approaches. Moreover, removing outliers helped stabilize estimates of central tendency (mean, median) and variance, thereby improving the reliability of modeling and interpretation. While advanced models like Random Survival Forests can tolerate outliers, this step was important for ensuring consistent preprocessing across all methods and highlighting how much of the original distribution was driven by a small number of extreme cases.

Figure 5.2 displays a bar plot of the binary event indicator OS, where a value of 1 represents patients who experienced the event (death) and 0 represents censored patients (i.e., those lost to follow-up or still alive at the time of analysis). The plot shows that a larger proportion of patients are censored, with approximately 825 patients (64 percent) censored and 470 patients (36 percent) having experienced the event. This level of censoring is typical in clinical survival datasets and must be properly accounted for in model development.

5.2 Feature Selection

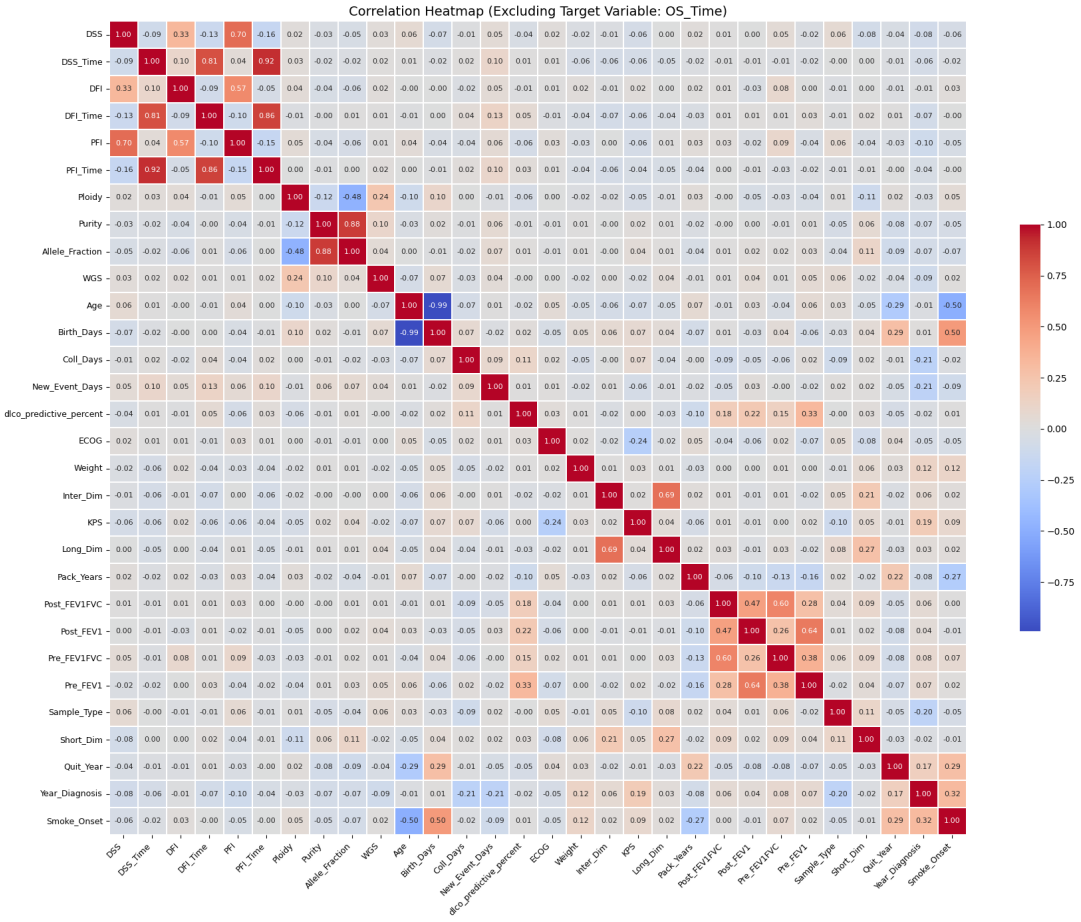

Fig. 5.3. Heatmap of correlation matrix

This imbalance is common in real-world clinical datasets, especially in longitudinal studies, where many patients may not yet have reached the event endpoint (death) by the time of data extraction [17]. Factors contributing to this include short follow-up duration for recent patients, improved treatment outcomes, or patients withdrawing from the study.Although this level of censoring is typical, it can affect model performance if not properly addressed. For this reason, the study employed censoring-aware survival models such as Kaplan-Meier, Cox Proportional Hazards, and Random Survival Forests, which are explicitly designed to handle incomplete outcome information without introducing bias. Figure 5.3 presents a full correlation heatmap generated to explore the linear relationships among numeric variables in the dataset, with a particular focus on identifying predictors that may be highly correlated. This visualization assists in assessing multicollinearity a critical consideration in model development as strong correlations between predictors can compromise the stability and interpretability of regression based survival models.The heatmap includes all numeric variable pairs without applying a correlation threshold, allowing a comprehensive overview of relationships across the dataset. Most variables exhibit low to moderate correlations, suggesting that multicollinearity is generally not a concern. This analysis also supported feature selection. Numeric predictors such as PFI.time(progression free interval time), age at initial pathologic diagnosis, and days to new tumor event after initial treatment were selected based on their relevance and lack of excessive correlation with other predictors, ensuring that the final model inputs were informative and independent.

5.3 Kaplan Meier Survival Results

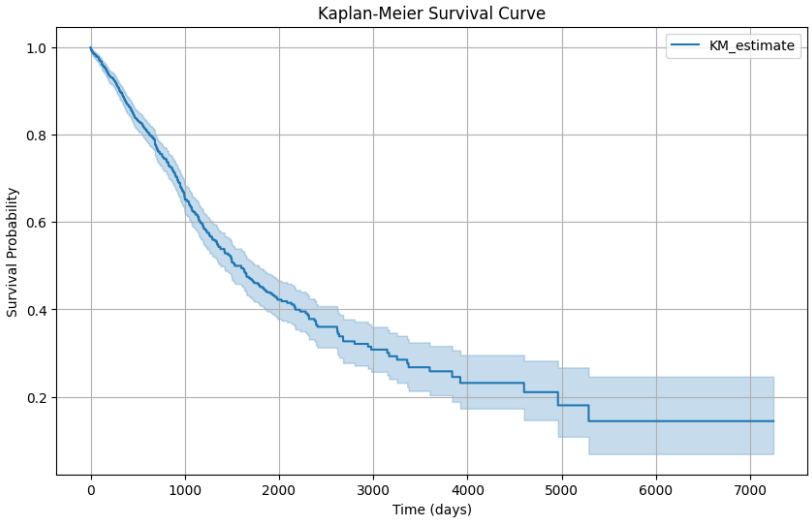

Fig. 5.4. Kaplan-Meier survival curve showing the estimated probability of overall survival over time

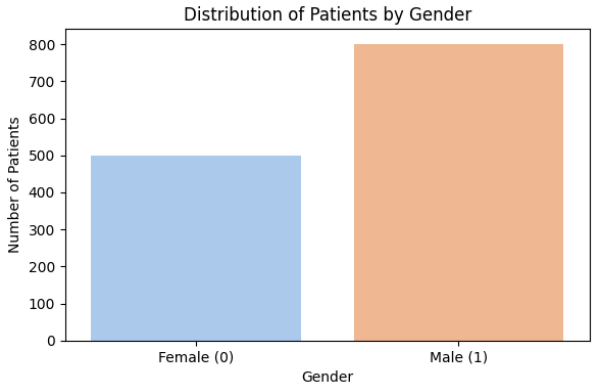

Fig. 5.5. Gender wise distrbution of population

The Kaplan-Meier survival curve (Fig 5.4), which illustrates the probability of patient survival over time. The curve starts at a survival probability of 1.0, indicating that all patients were alive at the beginning of the observation period. As time progresses, the survival probability declines in a stepwise manner, reflecting the timing and frequency of death events within the cohort. A steeper decline in the earlier part of the curve suggests a higher rate of mortality shortly after diagnosis or treatment, while the flattening of the curve at later time points indicates that a subset of patients experienced longer-term survival. By around 3000 days (8.2 years), the survival probability drops below 30 %, with a continued gradual decrease beyond that point.The shaded region around the curve represents the 95% confidence interval, which becomes wider at later time points due to the decreasing number of patients still under observation, a common occurrence in survival analysis. This widening reflects increased uncertainty in the survival estimates as fewer data points are available over time.Figure 5.5 shows the distribution of patients by gender. There are more male patients (801) compared to female patients (498) in the dataset. This imbalance is important to consider when interpreting survival analyses stratified by gender, as it may influence the shape and confidence intervals of survival curves.

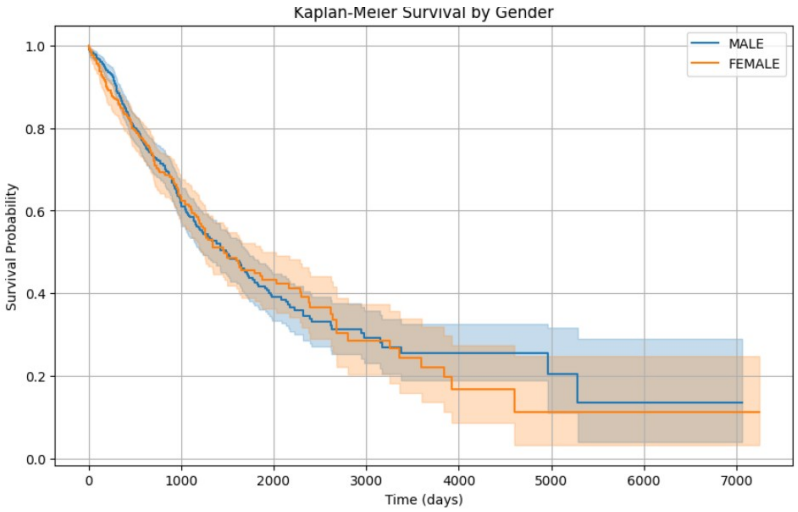

Fig. 5.6. Kaplan-Meier Survival graph by gender

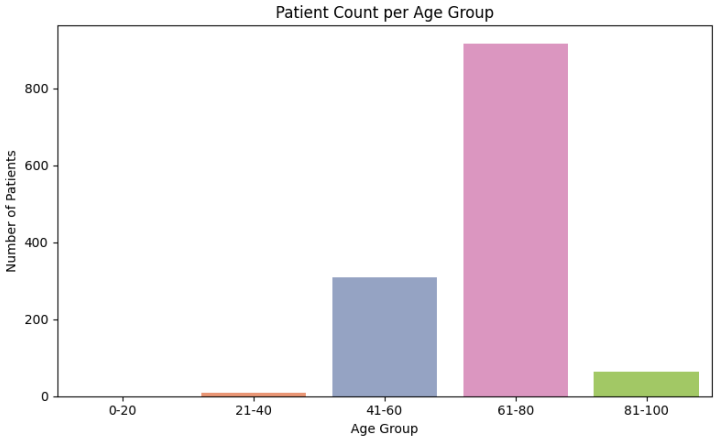

Fig. 5.7. Distribution of Age Group

The Kaplan-Meier survival curves stratified by gender (Fig 5.6), with males represented in blue and females in orange. The plot illustrates the estimated survival probabilities over time, starting from 100% and decreasing as death events occur.While both curves follow a similar declining pattern, male patients appear to exhibit slightly better long-term survival, particularly beyond 4000 days. This may be influenced, at least in part, by the unequal gender distribution observed in the dataset. As shown in the corresponding gender bar plot (Figure 5.5), there are more male patients (801) than female patients (498). This imbalance in group sizes can affect the survival estimates, especially in the tail of the distribution where fewer female patients remain under observation.

The shaded areas around each curve represent 95% confidence intervals, which grow wider as fewer patients remain at risk. Although minor differences in survival are observed between genders, the overall shape of the curves suggests broadly similar survival patterns, with the gender imbalance potentially contributing to the slight divergence seen in later stages.

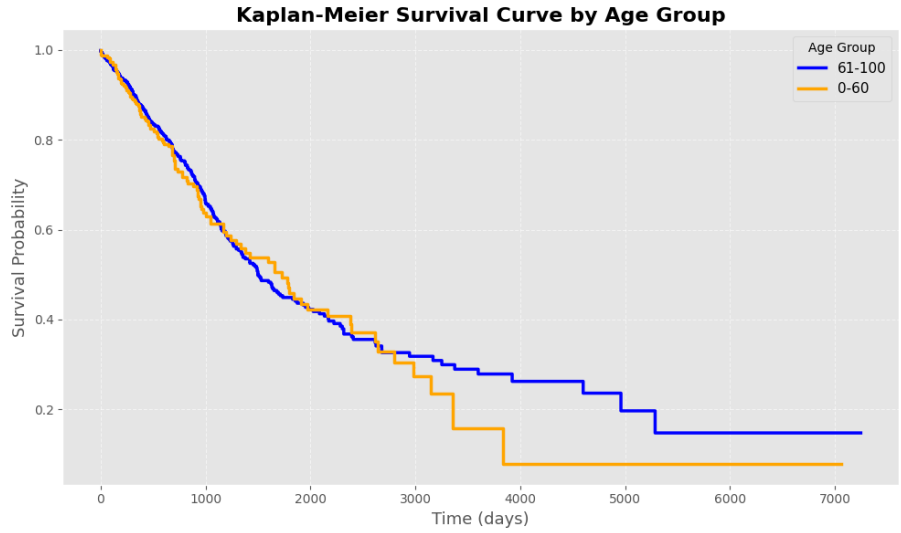

Fig. 5.8. Kaplan Meier Survival Graph by Age Group

Figure 5.7 shows the distribution of patients across predefined age groups. The majority of patients fall within the 61–80 age group, accounting for the highest number of cases (917 patients). This is followed by the 41–60 group, which includes 310 patients. The 81–100 group contributes 64 patients, while only 8 patients are in the 21–40 group, and none are in the 0–20 group.

This distribution reflects a skew toward older age groups, as the incidence of disease increases with age. The low representation in the younger age brackets suggests that age-stratified analyses should focus on splitting age group in 2 categories 0-60 and 61-100 for a better stratified graph, where the data is more concentrated and reliable for survival modeling.

The above (Figure 5.8) displays Kaplan-Meier survival curves for two age groups: 0–60 years (yellow) and 61–100 years (blue). These curves estimate the probability of survival over time (in days) for each group, starting at 100% and declining as events (deaths) occur.

In the initial stages, both groups show similar survival probabilities. However, beyond approximately 3000 days, a notable divergence appears: the older age group (61–100) maintains a higher survival probability than the younger group (0–60). By the end of the follow-up period, the survival curve for the younger group levels off at a lower point, indicating poorer long-term survival compared to older patients.

Although unusual, this trend may be influenced by factors such as disease severity, treatment differences, or stage at diagnosis that differ between age groups. Additionally, the larger number of patients in the 61–100 group may have contributed to the stability of their survival estimates. These results suggest that older patients in this cohort had better long-term survival, warranting further investigation into contributing clinical factors.

5.4 Cox Proportional Hazard Modelling Results

model	lifelines.CoxPHFitter
duration col	'OS.time'
event col	'OS'
baseline estimation	breslow
number of observations	1130
number of events observed	455
partial log-likelihood	-2536.49
time fit was run	2025-05-25 10:32:59 UTC

	coef	exp(coef)	se(coef)	coef lower 95%	coef upper 95%	exp(coef) lower 95%	exp(coef) upper 95%	cmp to	z	p	-log2(p)
PFI.time	-0.00	1.00	0.00	-0.00	-0.00	1.00	1.00	0.00	-18.33	<0.005	246.88
days_to_new_tumor_event_after_initial_treatment	0.00	1.00	0.00	-0.00	0.00	1.00	1.00	0.00	0.74	0.46	1.12
age_at_initial_pathologic_diagnosis	0.01	1.01	0.01	-0.00	0.02	1.00	1.02	0.00	1.26	0.21	2.28
gender_encoded	-0.23	0.80	0.08	-0.39	-0.07	0.68	0.94	0.00	-2.76	0.01	7.42
residual_tumor_R1	-0.29	0.75	0.29	-0.87	0.29	0.42	1.33	0.00	-0.99	0.32	1.63
residual_tumor_R2	0.46	1.58	0.45	-0.43	1.34	0.65	3.82	0.00	1.01	0.31	1.68
residual_tumor_RX	0.32	1.38	0.26	-0.18	0.82	0.83	2.28	0.00	1.25	0.21	2.25

Concordance	0.90
Partial AIC	5086.97
log-likelihood ratio test	528.72 on 7 df
-log2(p) of ll-ratio test	363.00

Fig. 5.9. Cox proportional hazard model output

Following the removal of outliers, a Cox Proportional Hazards model was developed using data from 1130 patients, among whom 455 experienced the event of interest (death). The model was built using the following predictors: PFI.time, days to new tumor event after initial treatment, age at initial pathologic diagnosis, gender (label encoded), and residual tumor categories (one-hot encoded: residual tumor R1, residual tumor R2, and residual tumor RX). Among these, PFI.time was found to be highly statistically significant ($p < 0.005$), although its hazard ratio was approximately 1.00, indicating a negligible change in risk per unit increase likely due to the large scale or limited variability in this variable.

The age at initial pathologic diagnosis showed a small, non-significant increase in hazard (HR = 1.01, $p = 0.21$), while gender was the only categorical predictor to reach statistical significance. Male patients (encoded as 1) had a 20% lower risk of death compared to females (HR = 0.80, $p = 0.01$), suggesting a survival advantage among males in this cohort. However, it is important to note that from Fig. 5.5, it was clearly evident that the number of male patients in the dataset was substantially higher than that of females, which may have influenced the stability and significance of this result. The observed survival benefit for males could, in part, reflect this sample size imbalance rather than a purely biological effect.

Other predictors, including days to new tumor event after initial treatment and the residual tumor categories, did not show statistically significant associations with survival. However, residual tumor R2 (HR = 1.58) and residual tumor RX (HR = 1.38) indicated a trend toward poorer survival outcomes relative to the baseline group (R0). Despite the lack of significance, these trends may warrant further investigation in larger samples. The overall model performance was strong, with a concordance index of 0.90, indicating excellent discriminative ability in predicting survival times.

The image(Fig 5.10) presents the Receiver Operating Characteristic (ROC) curve for the Random Survival Forest (RSF) model, evaluated at a prediction threshold of 1000 days. The ROC curve is a valuable diagnostic tool adapted for survival analysis to evaluate the model's ability to discriminate between patients who survive beyond a specific time point and those who do not. It plots the true positive rate (sensitivity) against the false positive rate (1 – specificity) across a range of classification thresholds, offering a comprehensive understanding of the model's binary decision performance at a fixed time horizon.

5.5 Random Survival Forest Result

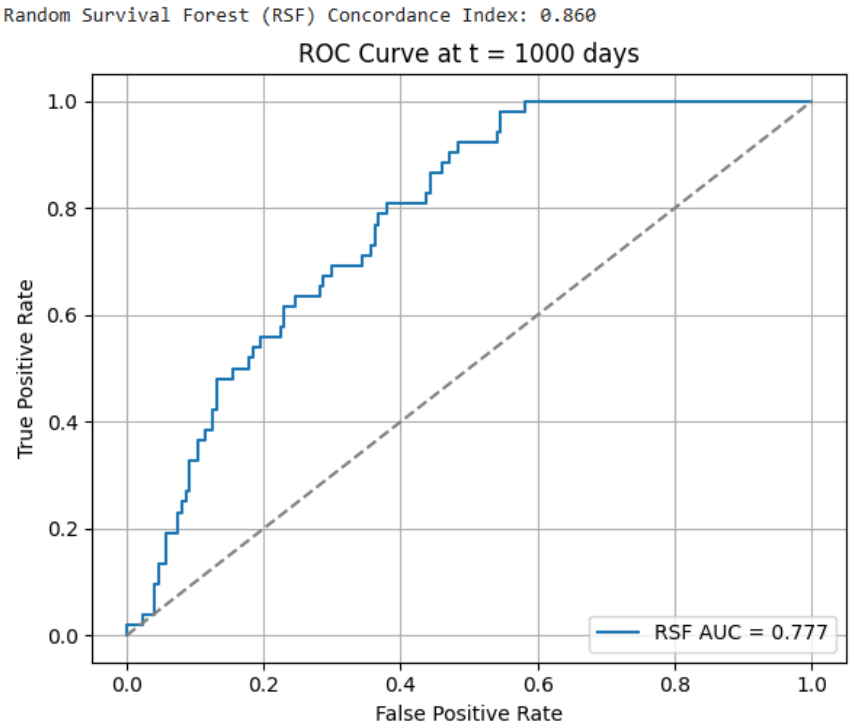

Fig. 5.10. Roc curve of RSF model and C-index

For consistency and comparability with the Cox Proportional Hazards model, the same set of predictors was used to train the RSF model. These include: PFI.time, days to new tumor event after initial treatment, age at initial pathologic diagnosis, gender encoded (label encoded), and the one-hot encoded residual tumor variables (R1, R2, RX), with R0 as the baseline. Using the same features ensures a fair evaluation of both models and allows direct comparison of their predictive performance.

The time threshold of 1000 days was carefully selected based on the distribution of overall survival times in the dataset. The histogram of OS.time revealed that the majority of patients had survival durations less than or close to 1000 days, making this cutoff a clinically relevant and data-driven choice for evaluating the model's discriminatory ability at a meaningful point in time.

The RSF model achieved an Area Under the Curve (AUC) of 0.777 at the 1000-day mark, which indicates good classification performance. An AUC close to 0.8 reflects the model's ability to correctly distinguish between high-risk and low-risk patients, outperforming a random classifier and approaching clinically useful prediction levels. Additionally, the concordance index (C-index) of 0.860, shown above the graph, further supports the model's robustness. The C-index evaluates the model's performance across the entire survival timeline, quantifying how well it can rank patients by predicted risk. A C-index of 0.860 signifies excellent agreement between predicted and observed survival outcomes.

5.6 Model Comparison and Evaluation

Model	C-index
Cox Proportional Hazards Model	0.90
Random Survival Forest (RSF) Model	0.860

Table 5.2. Comparison of Concordance Index (C-index) Between Models

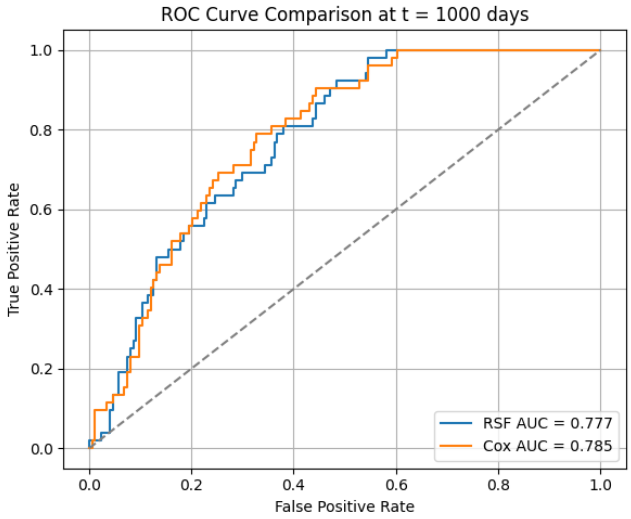

Fig. 5.11. ROC curve of both models

Table 5.2 compares the Concordance Index (C-index) of the Cox Proportional Hazards model and the Random Survival Forest (RSF) model. The C-index measures the model's ability to correctly rank patients based on predicted survival times. The Cox model achieved a C-index of 0.90, indicating excellent discriminatory performance. The RSF model, with a C-index of 0.860, also demonstrated strong predictive accuracy. Although the Cox model slightly outperformed the RSF in global ranking ability, both models show high reliability in distinguishing between low- and high-risk patients.

The graph (Fig 5.11) shows the ROC curve comparison between the Cox Proportional Hazards model and the Random Survival Forest (RSF) model, evaluated at a fixed time horizon of 1000 days. This time point was chosen based on the distribution of overall survival times in the dataset, which showed that the majority of patients had survival durations under or near 1000 days making it a clinically meaningful threshold for classification evaluation.

The Cox model achieved an AUC of 0.785, slightly outperforming the RSF model, which recorded an AUC of 0.777. Both models exhibit good discriminatory power at this time-specific evaluation point, as reflected by their ROC curves rising well above the diagonal (random chance) line.

These findings are consistent with earlier evaluations using the Concordance Index (C-index), where the Cox model also demonstrated stronger overall performance (C-index = 0.90) compared to the RSF model (C-index = 0.86). The similarity in both C-index and ROC AUC suggests that the Cox model consistently ranks patients more accurately and classifies survival outcomes more effectively at the 1000-day mark.Overall, this comparison reinforces the Cox model's strength in this dataset, offering both robust global ranking and reliable time-specific classification. While RSF remains a competitive non-linear alternative, the Cox model appears better suited for this particular survival prediction task.

Chapter 6
Evaluation of research method

6.1 Conclusion

This study implemented and compared multiple survival modeling approaches namely the Cox Proportional Hazards (CPH) model and the Random Survival Forest (RSF) to predict lung cancer patient outcomes using clinical data that included both baseline and post-treatment variables. By applying a combination of statistical and machine learning methodologies, the research comprehensively explored the potential of each model type in handling real-world survival prediction tasks.

The Cox Proportional Hazards model, a semi-parametric statistical method, provided interpretable insights into how individual predictors affected patient survival. It effectively identified variables such as gender, age at diagnosis, and residual tumor status as significant factors influencing the hazard of death. The model achieved a concordance index (C-index) of 0.900, which surpasses several published works such [3] study that reported a C-index of 0.78 .However, its reliance on the proportional hazards assumption and limited flexibility in modeling complex interactions posed constraints in scenarios where risks evolve over time. .

To overcome these limitations, a Random Survival Forest model was developed. The RSF model utilized both baseline and dynamic post-treatment variables to build a non-parametric, tree-based ensemble that captures nonlinear relationships and handles censored data without strict assumptions. It achieved a C-index of 0.860, aligning with results from [4] benchmark study emphasizing the strength of RSF for lung cancer survival prediction . Studies by [5] also highlighted comparative advantages of combining classical Cox models with machine learning methods. The model also incorporated progression-related features, enabling it to adapt better to changing risk over the follow-up period, especially for patients with recurrence events.

Post-treatment features like PFI.time and days to new tumor event were found to significantly contribute to predictive performance. Their inclusion enhanced the model's ability to stratify patients based on evolving risk. Also,[7] and [17] demonstrated that model augmentation with dynamic or knowledge-driven features can improve C-index, reinforcing this study's emphasis on incorporating post-treatment variables like PFI.time Notably, the exclusion of PFI.time resulted in a substantial drop in the C-index, highlighting the importance of dynamic data in improving survival estimates.

In summary, the study demonstrated that statistical models like CPH offer clear interpretability and clinical relevance, while machine learning models such as RSF provide strong predictive performance and flexibility in dealing with complex, real world datasets.[18] The integration of both baseline and follow-up clinical information proved essential in building accurate and practical survival models, suggesting that hybrid or complementary use of these approaches could yield optimal results in future applications.

6.2 Limitations

While this study provides meaningful insights into lung cancer survival prediction, several limitations must be

acknowledged to properly contextualize the findings. First, the dataset was relatively limited in both size and representativeness. Certain subgroups, such as younger patients and females, were underrepresented, which could bias model performance and reduce generalizability across a broader patient population. Moreover, the dataset exhibited imbalance in the outcome variable (OS), with a higher proportion of deceased patients. [20] This imbalance may have influenced the models to favor the majority class, potentially compromising sensitivity in predicting long-term survivors.

Second, although the Cox Proportional Hazards (CPH) model is a well-established statistical method, it assumes that hazard ratios remain constant over time an assumption that was not formally tested in this study. Violations of this assumption may affect the reliability of estimated survival probabilities. Additionally, the model's performance showed a strong dependency on the PFI.time (Progression-Free Interval) variable. During experimental trials, removal of this predictor led to a significant drop in C-index, indicating the model's heavy reliance on this single feature. This overdependence limits robustness, especially in scenarios where PFI data is missing or unavailable.

Finally, while internal validation was conducted through test-train splitting, no external validation was performed using an independent dataset.[19] As such, the models' predictive performance might not fully generalize to new patient populations or clinical settings.

6.3 Future Work

To strengthen the generalizability of the findings, future studies should include external validation using datasets from other hospitals or cancer registries. This would test whether the models perform consistently across different populations and settings. Additionally, expanding the dataset to include a more balanced representation across gender, age groups, and tumor characteristics would improve model fairness and reduce sampling bias. Stratified sampling or oversampling techniques might also be explored to counteract class imbalances observed in variables like gender or survival outcomes.

While this study evaluated model performance using a fixed 1000-day threshold in the ROC curve, future work should consider multiple clinically relevant survival time points. Evaluating performance across different durations (e.g., 2000 or 3000 days survival) would offer a more complete picture of how models perform over time and improve clinical utility for personalized prognosis.

The Cox Proportional Hazards model assumes that hazard ratios remain constant over time. While this assumption was not formally tested in the current study, future work should evaluate the proportional hazards assumption using diagnostics such as Schoenfeld residuals or time-dependent covariates[11]. This would ensure that statistical inferences drawn from the model remain valid and robust across the follow-up period.

The Random Survival Forest (RSF) model demonstrated strong performance but limited interpretability. Future studies should explore model interpretation techniques such as [12] [14] SHAP (SHapley Additive Explanations) or permutation importance to better understand feature contributions. These tools can provide clinicians with clearer explanations for risk predictions, enhancing the model's transparency and trustworthiness in a healthcare setting.

There is significant potential in developing hybrid survival models that combine the interpretability of Cox models with the flexibility and nonlinearity handling of machine learning approaches like RSF. Such hybrid systems can balance predictive power and clinical insight, offering both accuracy and transparency. Incorporating regularized Cox variants [13] (e.g., Elastic Net) or using ensemble techniques with interpretable learners could be promising directions.

References

[1] I. Etikan, S. C. Abubakar, and R. Alkassim, "The Kaplan-Meier estimator as a non-parametric technique," *Biom Biostat Int J.*, vol. 5, no. 4, pp. 1–4, 2017. Available: https://www.scirp.org/reference/referencespapers?referenceid=2744985

[2] M. K. Goel, P. Khanna, and J. Kishore, "Understanding survival analysis: Kaplan-Meier estimate," *International Journal of Ayurveda Research*, vol. 1, no. 4, pp. 274–278, 2010. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC3059453/

[3] J. Zhu, H. Shi, H. Ran, Q. Lai, Y. Shao, and Q. Wu, "Development and Validation of a Nomogram for Predicting Overall Survival in Patients with Second Primary Small Cell Lung Cancer After Non-Small Cell Lung Cancer: A SEER-Based Study," Int. J. Gen. Med., vol. 15, pp. 3613–3624, Apr. 2022, Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC8986201/pdf/ijgm-15-3613.pdf

[4] S. Khatua, "A Benchmark Study On the Comparative Advantage Of Random Survival Forest Analysis in Lung Cancer Survival Predictions" 2024. Available: https://www.academia.edu/116804882/A_BENCHMARK_STUDY_ON_THE_COMPARATIVE_ADVANTAGE_OF_RANDOM_SURVIVAL_FOREST_ANALYSIS_IN_LUNG_C
sb-sw=69037320&utm_

[5] J. A. Bartholomai and H. B. Frieboes, "Lung cancer survival prediction via machine learning regression, classification and statistical techniques," *Biomedical Engineering*, vol. 17, pp. 24–32, 2018. Available: https://pubmed.ncbi.nlm.nih.gov/31312809/

[6] C. Astley *et al.*, "Explainable deep learning-based survival prediction for non-small cell lung cancer patients undergoing radical radiotherapy," *Journal of Clinical Data Science*, vol. 12, pp. 115–126, 2023. Available: https://www.sciencedirect.com/science/article/pii/S0167814024000057

[7] C. M. Caruso, V. Guarrasi, S. Ramella, and P. Soda, "A deep learning approach for overall survival prediction in lung cancer with missing values," Comput. Methods Programs Biomed., vol. 254, p. 108308, 2024, Available:https://www.sciencedirect.com/science/article/pii/S016926072400302X

[8] L. Pu, R. Dhupar, and X. Meng, "Predicting Postoperative Lung Cancer Recurrence and Survival Using Cox Proportional Hazards Regression and Machine Learning," Cancers, vol. 17, no. 1, p. 33, 2024. Available: https://www.mdpi.com/2072-6694/17/1/33

[9] F.-Y. Dao, H. Lv, Y.-H. Yang, H. Zulfiqar, H. Gao, and H. Lin, "Computational identification of N6-methyladenosine sites in multiple tissues of mammals," Comput. Struct. Biotechnol. J., vol. 18, pp. 1084–1091, 2020, doi: Available: https://www.sciencedirect.com/science/article/pii/S2001037020302622

[10] M. Othus, B. Barlogie, M. L. LeBlanc, and J. J. Crowley, "Cure models as a useful statistical tool for analyzing survival," Clin. Cancer Res., vol. 18, no. 14, pp. 3731–3736, 2012. Available: https://pubmed.ncbi.nlm.nih.gov/22675175/

[11] Kuitunen, V. T. Ponkilainen, M. M. Uimonen, A. Eskelinen and A. Reito, "Testing the proportional hazards assumption in cox regression and dealing with possible non-proportionality in total joint arthroplasty research: methodological perspectives and review," BMC Musculoskeletal Disorders, vol. 22, no. 489, pp. 1–7, 2021. Available: https://bmcmusculoskeletdisord.biomedcentral.com/articles/10.1186/s12891-021-04379-2?

[12] L. Ter-Minassian, S. Ghalebikesabi, K. Diaz-Ordaz, and C. Holmes, "Explainable AI for survival analysis: a median-SHAP approach," arXiv preprint arXiv:2402.00072, 2024. [Online]. Available: https://arxiv.org/html/2402.00072v1?.

[13] A. J. Komorowski, M. Kiraly, A. Ren, and M. Z. Ding, "Interpretable machine learning for in-hospital survival of patients with COVID-19: a retrospective study of electronic health record data in the USA," *The Lancet Digital Health*, vol. 3, no. 7, pp. e447–e457, Jul. 2021, Available: https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-024-02525-z?

[14] Z. Shi, Y. Chen, A. Liu, J. Zeng, W. Xie, X. Lin, Y. Cheng, H. Xu, J. Zhou, S. Gao, C. Feng, H. Zhang, and Y. Sun, "Application of random survival forest to establish a nomogram combining clinlabomics-score and clinical data for predicting brain metastasis in primary lung cancer," Clin. Transl. Oncol., vol. 27, no. 4, pp. 1472–1483, 2025, Available: https://pubmed.ncbi.nlm.nih.gov/39225959/

[15] Y. H. Lai, W. N. Chen, T. C. Hsu, C. Lin, Y. Tsao, and S. Wu, "Overall survival prediction of non-small cell lung cancer by integrating microarray and clinical data with deep learning," Sci. Rep., vol. 10, no. 1, p. 4679, 2020, Available: https://pubmed.ncbi.nlm.nih.gov/32170141/

[16] Y. Liu, Z. Wang, X. Cao, M. Liu, and L. Zhong, "Machine learning models for predicting survival in lung cancer patients undergoing microwave ablation," Front. Med., vol. 12, 2025. Available: https://www.frontiersin.org/journals/medicine/articles/10.3389/fmed.2025.1561083/full?

[17] C. Fang, G. A. Arango Argoty, I. Kagiampakis, et al., "Integrating knowledge graphs into machine learning models for survival prediction and biomarker discovery in patients with non–small-cell lung cancer," J. Transl. Med., vol. 22, p. 726, 2024. Available: https://translational-medicine.biomedcentral.com/articles/10.1186/s12967-024-05509-9?

[18] M. R. Salmanpour, A. Gorji, A. Mousavi, A. F. Jouzdani, N. Sanati, M. Maghsudi, B. Leung, C. Ho, R. Yuan, and A. Rahmim, "Enhanced lung cancer survival prediction using semi-supervised pseudo-labeling and learning from diverse PET/CT datasets," Cancers, vol. 17, no. 2, p. 285, 2025. Available: https://www.mdpi.com/2072-6694/17/2/285?

[19] S. Salerno and Y. Li, "High-dimensional survival analysis: Methods and applications," Annu. Rev. Stat. Appl., vol. 10, no. 1, pp. 25–49, 2023. Available: https://www.annualreviews.org/content/journals/10.1146/annurev-statistics-032921-022127

[20] A. Groji, A. F. Jouzdani, N. Sanati, A. M. Ahmadzadeh, R. Yuan, A. Rahmim, and M. R. Salmanpour, "Censor-aware semi-supervised lung cancer survival time prediction using clinical and radiomics feature," arXiv preprint arXiv:2502.01661, 2025. Available: https://arxiv.org/abs/2502.01661