Ihona Maria Correa de Cabo

Predicting Lung Cancer Survival: A Comparative Study of Statistical and Machine Learning Methods

MASTER'S THESIS

Supervisors: Dr. Dídac Florensa (Universitat de Lleida) and Dr. Jordi Vilaplana (Universitat de Lleida)

Tutor: Dr. Jordi Mateo (Universitat de Lleida)

Master's Degree in Health Data Science























Barcelona, 2025

Abstract

Survival analysis is a statistical method used to predict the time to an event such as death or disease progression, among others. A distinguishable aspect of survival analysis is the presence of censored data, which indicates that the event of interest has not occurred (at least) during the study period. Traditionally, purely statistical methods have been widely used in the community for the analysis of censored data. These models are valued for their interpretability and their extensive clinical validation. However, traditional models rely on many assumptions and have been mostly designed for small data sets. In addition, they may fail to capture the complexities inherent in medical data.
Machine Learning models, in contrast, offer more flexibility in modeling and handling non-linear patterns, complex interactions and high-dimensional data.

This study aims to bridge the gap between traditional statistical methods and Machine Learning in survival analysis by revealing which techniques are more effective in predicting the survival and key risk factors in lung cancer patients.
Uncovering and understanding strong predictors that influence survival times in patients can be beneficial to prevent the disease and influence on clinical decision-making.

First, an extensive traditional survival analysis is performed on a dataset from “Hospital Universitari Arnau de Vilanova” (Lleida), which contains patients diagnosed with lung cancer from 2012 to 2021.
Using the same dataset, this study also compares the performance of Cox and Weibull models against Machine Learning based models that include Random Survival Forests, Gradient Boosting, and neural-network techniques. This comparison is assessed through the Concordance Index, the Integrated Brier Score and feature importance plots.
The findings reveal that Machine Learning models, particularly Random Survival Forests, slightly outperform traditional statistical methods in terms of predictive accuracy. However, no groundbreaking differences were found, suggesting that for a survival dataset with a small number of predictors, statistical models should be used first, since they are more interpretable and their coefficients can be explained in terms of hazard ratios, providing valuable insights.
Further research on Machine Learning model interpretability is needed, especially for neural-network based models, which could have a direct impact in clinical applicability.

Keywords: Survival Analysis, Lung Cancer, Machine Learning, survival prediction, Hazard Ratios

Glossary

AF – Acceleration Factor

AI – Artificial Intelligence

AIC – Akaike Information Criterion

C-index – Concordance Index

CI – Confidence Interval

CSV – Comma Separated Values

DL – Deep Learning

GBSA – Gradient Boosting Survival Analysis

h(t) – Hazard Function

HR – Hazard Ratio

IBS – Integrated Brier Score

ICD – International Classification of Diseases

ML – Machine Learning

NSCL – Non–Small-Cell Lung Cancer

OS – Overall Survival

PH – Proportional Hazards

RR – Crude Risk Ratio

RSF – Random Survival Forest

S(t) – Survival Function

SCLC – Small Cell Lung Cancer

WHO – World Health Organization

Table of Contents

Abstract **1**
Glossary **2**
1. Introduction **6**
1.1. Motivation 6
1.2. Objectives 6
2. Background and State of the Art **7**
2.1. Lung Cancer 7
2.2. Diagnosis and treatment 8
2.3. Traditional Survival Analysis 9
2.4. Machine Learning for Survival Analysis 13
3. Methodology **16**
3.1. Study design and dataset 16
3.2. Tools 17
3.3. Data preprocessing 18
3.3.1. Data transformations 19
3.3.2. Exploratory Data Analysis 22
3.4. Traditional Survival Analysis 24
3.5. Machine Learning Survival Analysis 30
3.6. Performance evaluation and comparison 31
4. Results **35**
4.1. Bivariate Analysis 35
4.2. Kaplan-Meier Curves 36
4.3. Cox Regression Models 40
4.3.1. General Cox model 40
4.3.2. Adding interaction terms 42
4.3.3. Stratifying by sex 43
4.4. Weibull model 44
4.5. Comparisons 45
5. Discussion and conclusions **48**
5.1. Results interpretation 49
5.2. Assumptions, limitations and future work 51
5.3. Conclusions 52
6. Ethical-social impact, sustainability, and diversity **53**
6.1. Ethical-social impact 53
6.2. Sustainability 54
6.3. Diversity 54
7. References **55**
Annex **62**
A1. Code availability 62
A2. Proportional Hazard assumption test 62
A3. Hyperparameter ranges of the ML models 63
A4. Project organization 64

List of Figures

1. Types of lung cancer with its relative statistics and incidence.
2. Distribution of survival times for lung cancer patients.
3. Age distribution of patients categorized by death status.
4. Death status by age group.
5. Death status by tobacco consumption.
6. Overall Kaplan-Meier Survival curve in months.
7. Survival curve stratified by sex.
8. Survival curve stratified by smoking status.
9. Survival curve stratified by age groups.
10. Survival curve stratified by tumor morphology.
11. Forest plot of HRs and their 95% CIs for the general Cox regression analysis.
12. Log-log plot for confirming the Weibull distribution assumption.
13. Feature importance plot for Cox model.
14. Feature importance plot for Weibull model.
15. Permutation-importance feature plot for RSF.
16. Permutation-importance feature plot for GBSA.
17. Gantt diagram chart for project organization planning.

List of Tables

1. Morphology column values included in the dataset.
2. Age group categories included in the dataset.
3. Descriptive table containing the total patients and the number of patients that died or survived in the end of the study.
4. Output of a Cox model in R.
5. Output of a Weibull model in R.
6. Bivariate analysis.
7. Log-rank test results of stratified Kaplan-Meier curves.
8. Multivariate analysis—Cox regression. Reference groups are marked as -.
9. p-values of the Schoenfeld residuals.
10. Cox model results of the interaction term sex*tobacco.
11. Cox regression stratified by sex (Men, Women). Reference groups are marked as -.
12. Weibull results. Reference groups are marked as -.
13. Traditional survival models (Weibull, general Cox and Cox with interaction term) are compared through the AIC and ordered from the best model (lowest AIC) to the worst model (highest AIC).
14. Traditional models (Cox, Weibull) are compared against ML models (RSF, GBSA, DeepSurv) through the C-index and IBS metrics.
15. Schoenfeld residuals test results for the general Cox model.
16. Schoenfeld residuals test results for the Cox model with interaction term sex*tobacco.
17. Schoenfeld residuals test results for the sex stratified Cox model (Men and Women).
18. Hyperparameter ranges and selected value for the RSF model.
19. Hyperparameter ranges and selected value for the GBSA model.

1. Introduction

1.1. Motivation

Lung cancer is known as one of the most common and serious types of cancer, accounting for the highest mortality rates among both men and women [1].

Despite advancements in diagnosis and treatment, this type of cancer still has a 5-year survival rate below 20%, and more than 80% of the patients are diagnosed in advanced stages [2].
Late diagnosis and the inherent variability in the imaging characteristics and histopathology of lung cancer present a barrier for clinicians when determining the optimal treatment approach, which are now the main limits to the optimization of the treatment [3].

These challenges highlight how personalized and data-driven approaches could support and improve clinical decision-making. For instance, by modeling time-to-event data, it is possible to estimate the survival of patients and to detect risk patterns that could help to assess the effectiveness of treatments and provide insights for clinical practice.

The main motivation behind this thesis was to contribute to the research field with knowledge that could have a positive impact on lung cancer patients. I have always been fascinated by the intersection of Data Science and healthcare, and this was an excellent opportunity to challenge myself and immerse in a real-world situation.

1.2. Objectives

The presented thesis is an observational, retrospective and longitudinal study that aims to contribute to the research community and ultimately improve the survival rate and early detection of lung cancer.

The primary objectives are:

- To compare the performance of traditional statistical models (Cox proportional hazards and Weibull models) with Machine Learning-based approaches (Random Survival Forest, Gradient Boosting, and neural network-based models) in the context of lung cancer survival analysis.

- To evaluate each model based on its predictive accuracy and clinical utility using metrics such as the Concordance Index, the Integrated Brier Score (IBS), and feature importance plots.

- To identify and analyze the key predictors that influence survival times in lung cancer patients.

- To assess whether Machine Learning models provide substantial advantages over traditional statistical methods in terms of predictive performance and risk factor identification.

2. Background and State of the Art

2.1. Lung Cancer

Lung cancer is a type of cancer that starts when abnormal cells grow in an uncontrolled way in the lungs [1] and can afterwards spread to other parts of the body. The two major forms of lung cancer are non–small-cell lung cancer, which account for 85% of all lung cancers, and small-cell lung cancer, representing about 15% of the cases [4].

Small Cell Lung Cancer (SCLC) is strongly related to smoking and is characterized by a rapid growth and early metastasis, making it more aggressive.

Non–small-cell lung cancer (NSCL), on the other hand, has a slower growth rate and can be divided into three major histologic subtypes (*Figure 1*):

- Squamous-cell carcinoma: Originates from squamous epithelial cells, often located in central airways, and is closely associated with smoking.
- Adenocarcinoma: Most common type in patients who have never smoked. It arises from glandular epithelial cells, often found in peripheral lung regions.
- Large-cell lung cancer: It is poorly differentiated and lacks histological features. It is often diagnosed by default through the exclusion of other possibilities, and sometimes begins in the central part of the lungs.

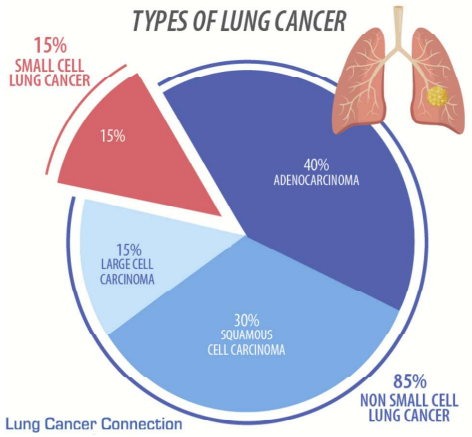


Figure 1.- Types of lung cancer with its relative statistics and incidence. (Source: Lung cancer connection)

Smoking is the leading risk factor of lung cancer, causing at least 80% of lung cancer deaths [5]. However, other factors such as radiation, environmental and occupational exposure to certain

pollutants, can increase the risk of developing this condition. Certain gene mutations and chromosomes have also been linked to an increased risk of lung cancer.

2.2. Diagnosis and treatment

An early and accurate diagnosis of lung cancer is crucial for better patient treatment and survival.
Current diagnosis of lung cancer includes different types of imaging that are complemented with pathological biopsies [6].

High-risk patients that present symptoms such as blood cough, trouble breathing, unexplained weight loss and fatigue, among others, typically undergo a chest radiography in order to discover possible abnormal areas in the lungs. Nevertheless, since chest X-rays cannot give a definitive diagnosis because they often cannot distinguish between cancer and other conditions [7], a Computed tomography (CT) scan is typically performed afterward. Basically, a CT scan uses x-rays to make cross-sectional images, taking many images and combining them computationally to show a lung slice, in this case. By using a CT scan, professionals can evaluate the exact tumor location, size, potential metastasis and possible lymph node involvement.

If the CT scan results confirm that the patient has cancer, other imaging techniques such as Magnetic resonance imaging (MRI) and Positron emission tomography (PET) scans are performed to discover whether the cancer has spread to other parts of the body. These imaging scans also help in determining the cancer staging and treatment planning.

The stages of lung cancer typically range from 1 to 4, with lower numbers indicating that the cancer is small and contained in the lung, and higher numbers that the cancer grows larger or spreads outside of the lungs. For instance, a stage 4 lung cancer has spread to other areas of the body, and therefore it is more aggressive and has a poorer prognosis. [8]

In order to distinguish between different histological types of lung cancer, a biopsy is needed to sample and analyze the tumor under a microscope. Furthermore, with the advancement of genetics and biomarkers testing, specific mutations have been identified with the aim of targeting the treatment for individual patients.

Depending on the staging and the type of lung cancer, patients are eligible for certain treatments.
In early stages, for instance, where the cancer is limited to a single and well localized tumor, a surgery is preferred to remove the cancer cells.
In contrast, for more advanced stages or when surgery is not feasible, chemotherapy, radiation therapy or immunotherapy might apply. Nevertheless, these treatments are very aggressive and can also destroy healthy tissue, harming the patient and producing many negative side-effects.

The limitations and adverse effects of traditional cancer treatment have promoted the development of targeted therapies, revolutionizing the field of NSCLC treatment. Targeted therapies are more used in NSCLC rather than SCLC because of its biological characteristics, which include the presence of

genetic mutations that cause tumor growth. Therefore, these therapies are designed to inhibit specific mutations, offering a more precise treatment approach towards cancer cells.

Targeted therapies focus on molecular alterations (typically identified through biomarker testing), and are often consumed either orally or intravenously. Usually, targeted therapies are applied in advanced cancer stages, where the tumor has spread through the body and traditional methods fail to target the malignant cells correctly. In addition, targeted therapies can be used either alone or combined with chemotherapy, for example.

These therapies offer significant advantages, particularly in patients with an advanced NSCLC, achieving higher response rates and longer progression-free survival for those patients that have actionable mutations. In summary, while traditional treatments are suitable for the broader population (regardless of their molecular profile), they are less precise and the side effects can be more severe and systemic.

2.3. Traditional Survival Analysis

Survival analysis is a well-studied field in health statistics research that aims to predict the time until the occurrence of an event of interest (the tumor recurrence or the death after some treatment intervention, for example). The time between the beginning of follow-up and the occurrence of the event is called *survival time* or *failure time* [10], and is a critical variable to consider in many medical studies, particularly in oncology.

Moreover, in survival analysis, one important issue to be considered is the censoring problem, which happens when the survival time is not observed for certain individuals. This phenomenon, also known as right-censoring, may arise when [11]

- The patient has not (yet) experienced the event of interest (such as relapse or death) by the time of the end of the study.
- The patient is lost to follow-up during the study.
- The patient experiences a different event that makes further follow-up impossible.

Censored survival times underestimate the true (but unknown) time to event of the subjects, so it is important to take into consideration that survival analysis normally underestimate the population's survival.
Besides right censoring, there are other types of censoring. Nevertheless, for this study, only right censoring is observed and taken into consideration.

Since the exact survival times of the censored subjects are unknown, and they account for a large portion of the data, standard statistical methods such as linear regression are not suitable for survival time data [10].

Survival data are generally described and modelled in terms of two related probabilities, namely survival and hazard [11]. The survival probability or survival function S(t) is the probability that an individual survives from the time origin (start of the study) to a specified future time "t". Survival probabilities for different values of t are crucial for survival analysis since they directly describe the survival experience of a study cohort.

On the other hand, the hazard probability or hazard function h(t) is the probability that an individual who is under observation at a time t has an event at that time. In other words, it represents the instantaneous event rate or risk for an individual who has already survived to time t [11].

It is interesting to observe that, while the survival function focuses on the "non-occurrence" of the event (the patient has not died; the disease has not progressed), the hazard function focuses more on the "occurrence" of that event.

In traditional survival analysis, there are mainly three statistical methods to assess these probabilities:

a) NON-PARAMETRIC METHODS

The **Kaplan-Meier estimator** is a popular and highly used non-parametric method for estimating the survival function S(t) from right censored data.

The estimator is defined as:

$$\widehat{S}(t) = \Pi_{t_i \leq t}\left(1 - \frac{d_i}{n_i}\right) \; [Eq.\,1]$$

Where $\widehat{S}(t)$ is the estimated probability of survival beyond time point t, t_i the observed event times, d_i the number of events and n_i the number of individuals at risk just before t_i that are alive and uncensored.

The Kaplan-Meier estimator handles censored data by updating the survival probability only at observed event times. Moreover, it produces a step function for S(t) that provides confidence intervals for survival probabilities. This stepwise calculation results in a survival curve that decreases at each event time and remains flat between events.
The advantage of this method is that it accommodates right-censored data, ensuring accurate survival estimates. Moreover, since it is non-parametric, it makes no assumptions about the underlying distribution of the survival times. Another benefit is that the intuitive survival curves are easy to interpret and compare across groups.

Nevertheless, this estimator can not model the impact of covariates (for example age or treatment type) on survival probabilities and can not directly model hazard ratios, which limits its use. As a result, the Kaplan-Meier estimator is more suitable for univariate analysis, where the goal is to estimate and compare survival across predefined groups (for example treated vs. untreated, smokers vs. non-smokers) rather than to model an individualized risk.

b) SEMI-PARAMETRIC METHODS

Semi-parametric methods combine the flexibility of non-parametric techniques with the ability to model the covariates' effects on survival and hazard functions.

The **Cox proportional hazards model** emerges as a semi-parametric method to understand the relationship between different factors (covariates) and the hazard of an event (death, disease progression) happening over time.
This model assumes that the effect of covariates on the hazard rate is proportional over time, which suggests that if one group has twice the risk of another group at one point in time, this ratio will remain constant throughout the whole study.

The Cox model is used to model the hazard function [Eq.2] which is the instantaneous risk of an event happening at time t, given that it has not happened yet.

$$h(t|X) = h_0(t) \cdot e^{\Sigma \beta_i \cdot X_i} \quad [Eq.2]$$

From [Eq.2] , $h(t|X)$ is the hazard function at time t given a set of covariates X, $h_0(t)$ the baseline hazard (which is not directly estimated in the standard Cox model), β the coefficient estimated for each covariate and $e^{\beta X}$ the hazard ratio.

The term inside the exponent $(\Sigma \beta_i \cdot X_i)$ is known as the log-risk score. It is a linear combination of the covariates and their corresponding coefficients, and it reflects the logarithm of the relative hazard associated. A higher log-risk score basically indicates a higher predicted hazard.

In summary, the Cox model calculates the hazard ratio for each covariate, representing how a one-unit increase in a covariate changes the hazard. For instance:

- Hazard ratio > 1: The covariate increases the risk of the event.
- Hazard ratio < 1: The covariate decreases the risk of the event.
- Hazard ratio = 1: The covariate has no effect on the event.

Since the Hazard ratio exemplifies how the risk compares between two individuals or groups, it is a relative Hazard, not an absolute measure.

The Cox model is very popular because it does not require to know the baseline hazard, (the hazard when all covariates are zero), since it only estimates the relative hazard between groups.

Moreover, while it does not directly model the survival function, it can be derived using the estimated hazard and the baseline survival curve. In addition, it can include data from censored individuals.
Nevertheless, the strong assumption of proportional hazards must be tested, as violating this assumption can lead to incorrect conclusions. Also, the Cox model struggles with time-dependent covariates, but other extensions such as time-dependent Cox models have been developed for these cases.

c) PARAMETRIC METHODS

Parametric methods assume that survival times follow a specific probability distribution, such as exponential, Weibull, log-normal, or Gompertz. Therefore, these methods are very specific, and they allow for a direct and precise estimation of both survival and hazard functions.

Each distribution offers a different flexibility when it comes to modeling the hazard function. For example, the exponential distribution assumes a constant hazard rate over time, while the Weibull can model increasing or decreasing hazard rates, providing greater flexibility to handle real data where risks may change over time.

One of the main advantages of parametric methods is that they can predict survival probabilities and hazard rates beyond the observed data, which can be useful for long-term extrapolation such as estimating the future survival of patients after treatment. Moreover, these models allow incorporating covariates through regression techniques so that the survival and hazard functions estimate their effects.

However, there are some limitations, because if the chosen distribution does not properly fit the data, the estimates may be biased and misleading. To address this, there are model selection techniques such as goodness-of-fit tests and comparisons using information criteria (AIC, BIC) that are used to ensure that the chosen model aligns with the observed survival patterns.

It is worth mentioning that the Cox proportional hazards model is the most common survival prediction model for cancer prognosis. It is frequently used to evaluate the impact of various covariates in survival times and can offer insights into the relationship between patient characteristics and outcomes.

However, despite its popularity, the Cox model is not designed to handle high-dimensional data. In fact, a large number of features in the dataset can cause the model to overfit the data and reduce the accuracy of the results [14].

Other traditional approaches researchers may use in contemporary cancer survival analysis include the Kaplan–Meier estimator, used to estimate survival probabilities over time. Parametric models, on

the other hand, are not as commonly used as non-parametric and semi-parametric methods since they are not as flexible and make a lot of assumptions on the data distribution. Moreover, in real-world clinical datasets, survival times can have complex patterns that may not align well with specific parametric distributions.

Nevertheless, parametric models are still valuable in certain contexts, especially in cost-effectiveness analysis or when clinical trials wish to estimate outcomes and risks beyond the follow-up period.

In recent years, there has been a shift towards integrating Machine Learning (ML) techniques into survival analysis, revealing a bright new direction in this field. These ML techniques aim to cope with the limitations that traditional survival analysis models present, such as the inability to capture nonlinear effects in the data and handle high-dimensional data, as well as the amount of assumptions to hold.

2.4. Machine Learning for Survival Analysis

Artificial Intelligence (AI) has revolutionized the management of NSCLC by improving different aspects, including staging, prognosis assessment, treatment prediction, response evaluation, recurrence prediction, and personalized prognostic assessment [15]. For instance, AI algorithms may classify NSCLC stages using ML techniques and deep imaging data analysis, facilitating personalized treatment decisions.

One notable example is the Watson for Oncology (WFO) system. WFO is a decision-support tool that analyzes patient's clinical data, including tumor type, stage, biomarkers and comorbidities, among others, to suggest individual treatment options, among others. [16].

Similarly to WFO, many learning systems have leveraged ML and clinical data to support personalized medicine. However, when it comes to predicting time-to-event data, which is a common scenario in clinical practice, the options have been less explored [17].

Despite that, the increasing number of data and variables, as well as the required assumptions many classic survival analysis methods imply, have started to propel the use of ML methods to analyze survival data.

It is worth mentioning that ML methods have not originally been proposed to handle censored data, making traditional ML models not appropriate for survival analysis, as they may lead to inaccurate or biased results [18].
Fortunately, ML techniques that inherently handle high-dimensional data have been adapted to also handle censored data, allowing ML to offer more flexible alternatives for analyzing high-dimensional, right-censored, heterogeneous data [14]. Moreover, ML methods can deal with data that violates common assumptions.

Research has shown that ML can provide more accurate alternatives to purely statistical approaches such as the Cox model for high-dimensional survival data [14]. In fact, reviews compared ML and traditional logistic regression for binary outcome prediction, highlighting better performance in ML algorithms such as Random Forest, Gradient Boosting, and Neural Networks [17].

The selection of the most suitable ML algorithm for survival analyses should be based on the particular research question as well as the characteristics of the underlying datasets, for example, how large the sample size is, how many variables are available, and how balanced the datasets are [17].

It is worth mentioning that, while ML models offer more flexibility in modeling complex non-linear patterns and may potentially have greater accuracy (especially in high dimensional data), they can be less interpretable and can be difficult to explain how predictions are derived [19].

In this thesis, the ML methods recommended for the analysis of survival data that will be examined are the following:

a) RANDOM SURVIVAL FOREST

Random Forest is a supervised ML algorithm introduced by Breiman (2001) that is widely used for classification and regression tasks. In essence, Random Forest grows multiple decision trees which are merged or ensembled together for a more accurate prediction [20].

Each tree is trained on a bootstrap sample (random sampling with replacement) of the training data. At each node, a random subset of features is considered for splitting. In the case of classification, trees vote for the most common class, whereas for regression, trees output a prediction. This method is highly robust to overfitting and scalable to large datasets.

Random Survival Forest (RSF) (Ishwaran et al., 2008) is an extension of Random Forests especially designed for survival analysis, where the goal is to model time-to-event data with possible censoring. It uses the same ensemble strategy as normal Random Forest, but it modifies the training and prediction process to estimate survival functions, cumulative hazard functions, or risk scores instead of classes or continuous values. In contrast to the normal Random Forest, where the splitting at each node is guided by a measure of impurity, here the splitting is guided by survival-specific criteria such as the log-rank test, and the final predictions are obtained by aggregating survival estimates across trees. This makes RSF a very powerful non-parametric method for analyzing censored data, as it does not require the proportional hazards assumption.

b) GRADIENT BOOSTING SURVIVAL MODEL

A gradient boosted model is a powerful ensemble learning technique introduced by Friedman (2001). It is similar to a Random Survival Forest, in the sense that it relies on multiple base learners (trees) to produce an overall prediction, but differs in how those are combined.

Random Forest uses many independent trees trained in parallel on different subsets of the data (bagging) , whereas Gradient Boosting builds trees sequentially, with each tree trying to fix the mistakes of the previous one (boosting), and is focused on bias reduction. This gives the model the ability to learn more complex patterns, although with a higher risk of overfitting.

Gradient Boosting Survival Analysis (GBSA) [24] extends the original gradient boosting method to handle time-to-event data with censoring, and it builds upon scikit-learn's gradient boosting framework to support survival loss functions.
In contrast to the RSF, GBSA does not directly predict the survival function, so it does not represent time, probabilities or Hazard Ratios directly. Instead, it outputs a relative risk score derived from the Cox proportional hazards model. This score can be used to rank individuals by risk, since higher values indicate higher hazard (and therefore shorter expected survival). Basically, GBSA models log-risk score like Cox does, but using boosting instead of linear regression.

c) DEEPSURV

DeepSurv is a deep feed-forward neural network for survival analysis that was introduced by Katzman et al. (2018) [25]. It extends the Cox proportional hazards model by replacing the linear risk function of Cox with a neural network, which enables the model to learn more complex and non-linear relationships between the covariates and the survival risk.

In DeepSurv, this feed-forward network takes the patient covariates as input and then outputs a single log-risk score (log HR) that is used to compare the relative hazard between individuals.

In essence, Deepsurv is a Deep Learning (DL) generalization of the Cox model, but instead of using a linear function of covariates, it learns a nonlinear function through a neural network.

While DeepSurv maintains the proportional hazards assumption, it enables greater flexibility and accuracy by modeling the risk function non-linearly.

DeepSurv has shown greater predictive performance in various real-world applications that include treatment recommendation, cancer prognosis, and personalized risk stratification.

3. Methodology

The aim of this section is to outline the approach and techniques used during the study. It also focuses on the study design, as well as the data and the tools used throughout the development of the thesis.

3.1. Study design and dataset

The present study is an observational study that is focused on assessing associations between different factors and survival outcomes. The primary outcome for this study is the Overall survival (OS) of lung cancer patients, defined as the time from diagnosis to disease death.
Moreover, the study can also be categorized as retrospective and longitudinal, since it examines historical hospital data where patients were followed over a period of time in order to observe changes in disease progression and mortality patterns.

This work uses patient data from *Hospital Universitari Arnau de Vilanova and Hospital Universitari Santa Maria,* Lleida, which contains patients that were diagnosed with lung cancer from 2012 to 2021. Specifically, cancer cases were extracted from the Lleida Population-based Cancer Registry using eight consecutive years of incidence data (2012–2019). These cases were validated by checking medical records, and hospital and pathology records were used as the main sources of information. Cancers were identified following the definitions of the International Association of Cancer Registries, the International Agency for Research on Cancer, and the European Network of Cancer Registries. Additionally, information regarding the cause of death was extracted from the Mortality Registry of Catalonia (2012–2021), provided by the Health Department.
In addition, to ensure a disease-specific survival analysis, patients who died from causes unrelated to lung cancer were excluded from the cohort.

The data extraction period extended to 2021 to ensure at least two years of follow-up for cases diagnosed in 2019.

3.2. Tools

The development of this thesis required various software environments and programming tools to handle the data processing, statistical modelling and survival analysis.

The exploratory data analysis, as well as the data processing and preparation, was performed in a Google Collab environment using Python programming language. Python was chosen due to its flexibility and built-in libraries, which include:

- *Pandas* [27]: Essential for loading the data and manipulate it.
- *Numpy* [28]: Used for array operations on the dataset.
- *Matplotlib* [29]: Used to generate static visualizations for data exploration.
- *Seaborn* [30]: Offers advanced plotting and better aesthetics.

Moreover, the Machine Learning models for survival analysis were also implemented in Python using the following libraries:

- *Scikit-learn* [31]: Standard library for Machine Learning workflows. Provides tools for preprocessing data, building models, evaluating them and performing cross-validation.
- *XGBoost* [32]: An optimized gradient boosting framework for better performance and speed.
- *Scikit-survival* [33]: It is an extension of the scikit-learn library but designed for survival analysis. Includes models like Random Survival Forests and Gradient Boosting Survival Analysis. It also includes evaluation metrics such as Concordance Index and Brier score.
- *PyTorc*h [34]: Deep learning framework used for building and training neural networks. In this context, it is used as a backend for survival models implemented with PyCox.
- *PyCox* [35]: A library for survival analysis with neural networks. It supports models like Cox Proportional Hazards and is built on top of PyTorch.
- *Lifelines* [36]: Survival analysis library that provides statistical models and evaluation tools such as the concordance index.

In order to perform the traditional survival analysis, the processed and prepared dataset was downloaded in CSV format and then loaded into Rstudio. R was chosen because it is well-suited for classic survival analysis and provides powerful specialized libraries explicitly designed for it. The key libraries used during the analysis are:

- *Survival* [37]: It is the core package for survival analysis. Provides functions for Kaplan-Meier estimations, Cox models and parametric survival models.
- *Flexsurv* [38]: Used for flexible parametric survival models.
- *Survcomp* [39]: Provides tools for comparing and evaluating survival models such as the concordance index.
- *Survex* [40]: Helps to interpret survival models, for example with feature importance.
- *Pec* [41]: Used to obtain prediction error curves, particularly for computing the Brier scores and integrated Brier scores.

- *RiskRegression* [42]: Offers methods for risk prediction and model comparison.
- Survminer [43]: Complementary package that offers visualization capabilities for survival curves and diagnostic plots.
- *Ggplot2* [44]: Data visualization package.
- *Gtsummary* [45]: Used for creating summary tables and reporting statistical outputs in a structured format.
- *Gt* [46]: Simplifies the creation of aesthetic tables.
- *Caret* [47]: Interface for model training, resampling, and preprocessing.
- *Car* [48]: Includes various statistical functions useful for regression diagnostics and model validation.
- *magrittr* [49]: Provides the pipe operator for writing readable code.

3.3. Data preprocessing

The provided dataset was processed using Python to ensure its quality and suitability for the posterior survival analysis. The main objectives of this step were to explore the dataset, clean the data, apply necessary transformations, and basically structure the information for the survival modeling.

The original raw dataset, stored in Excel format, contained a total of 1,696 patient records (identified by the ID number) and 27 variables extracted from the hospital database. After the processing steps, the final dataset contained 1,451 observations and 10 columns.

The first preprocessing step involved excluding irrelevant columns and those with an excessive proportion of missing values.

For instance, two variables that were rejected from the study were the diagnostic method and the cancer staging. The diagnostic method variable included seven possible diagnostic methods used by doctors to assess lung cancer, with biopsy being the most referred one. However, this variable could introduce bias, as the choice of the diagnostic method is often influenced by factors such as the patient's age, the symptom severity and the overall health condition.

On the other hand, while cancer staging is expected to have a strong association with survival (patients diagnosed at higher stages tend to have a poorer prognosis), this data was only available for 137 patients. This limited its applicability and unfortunately was excluded from the primary study.

3.3.1. Data transformations

Several transformations were applied to standardize and obtain relevant variables.

First of all, the death day, month and year were combined to form a single column indicating the official and real death date of the patient. The dataset also contained a death date from the hospital, but only the official date issued by the authorities was considered.

The dataset recorded both lung cancer diagnoses and causes of death using the International Classification of Diseases (ICD or CIE, in Spanish) [50], which is a globally recognized system maintained by the World Health Organization (WHO) that standardizes disease classification. Note that malignant tumor of the trachea, bronchi and lungs is categorized with the ICD code C34 and C33.

Similarly, there is a column that references the morphology of the tumor by using the ICD-O-3 (International Classification of Diseases for Oncology, 3rd Edition) , also developed by WHO [51]. This classification is widely used in cancer registries to categorize neoplasms (tumors) based on their histology and behavior (benign, malignant, metastatic, etc). In the dataset, there were 46 different morphology classification codes and, in order to simplify it, they have been grouped in the following histological categories:

ICD-O-3 Code	Histology Type	Description	Dataset proportion
M-81403	Adenocarcinoma	Most common subtype of lung cancer, often found in non-smokers and located in the outer lung regions.	37.5%
M-80703	Squamous	Strongly linked to smoking and it usually originates in the bronchi.	19.16%
M-80413	Small cell carcinoma	Highly aggressive cancer, strongly associated with smoking.	9.26%
M-80713	Epidermoid	A subtype of squamous cell carcinoma.	7.49%
M-80003	Blastoma	Rare lung cancer type.	6.67%
M-80463	Non-small cell carcinoma	General lung cancer category that has a slower growth rate.	6.54%
Others	Several	Other morphologies not mentioned above.	13.4%

Table 1.- Morphology column values included in the dataset.

Moreover, the age of the patient at the diagnosis was computed and used to create a column that categorized patients into different age groups:

Category	Age group	Dataset proportion
1	< 50 years	4.9%
2	50–59 years	17.40%
3	60–69 years	29.42%
4	70–79 years	29.07%
5	≥ 80 years	19.22%

Table 2.- Age group categories included in the dataset.

The data was also explored to ensure there were no inconsistencies. For instance, it has been confirmed that there were no alive patients with a death code, and vice versa.

It is important to mention that, although the extracted dataset contained only patients diagnosed with lung cancer, the registered deaths may not always be attributed to this condition. In fact, 81% of the registered deaths were caused by lung cancer, while the rest were attributed to other causes, which may be comorbidities of the disease. For instance, the second major cause of death observed in the dataset besides lung cancer was Chronic Obstructive Pulmonary Disease (ICD code J44), followed by Malignant Neoplasm with an unspecified primary site (ICD code C80), and Chronic Ischemic Heart Disease (ICD code I25).

Patients with a death code different from lung cancer were filtered out to ensure that only patients that died from the interested condition were analyzed.

A new column named "death" was defined as a boolean indicator that specifies if the patient died (1) or if the patient was alive (0) at the end of the study. A patient is considered death if it has a registered death date assigned, while if the death date is null, it is assumed that the patient was still alive (censored) at the end of the follow-up period.

From the 1452 patients that were diagnosed with lung cancer, 1081 of them died of the disease and only 371 survived.

The following descriptive table (*Table 3*) contains a summary of the selected variables of interest, accounting for the total counts and the number of deaths:

	Total counts	Death	Alive	Survival %
Sex				
Man	1122	864	258	22.99
Female	329	216	113	34.35
Morphology				
Non-small cell carcinoma	91	74	17	18.68
Adenocarcinoma	571	385	186	32.57
Blastoma	86	81	5	5.81
Epidermoid	117	87	30	25.64
Others	183	125	58	31.69
Small cell carcinoma	143	130	13	9.09
Squamous	260	198	62	23.84
Tobacco				
No	566	413	153	27.03
Yes	885	667	218	24.63
Age group				
[0-49]	81	51	30	37.04
[50- 59]	269	183	86	31.97
[60-69]	438	305	133	30.37
[70-79]	411	314	97	23.60
[80 - ~)	252	227	25	9.92

Table 3.- Descriptive table containing the total patients and the number of patients that died or survived in the end of the study.

Finally, the data was prepared for the posterior survival analysis.

The death column already indicates if the patient died or if the data is censored, either because the patient is alive or due to lost of follow-up. Nevertheless, in order to perform a survival analysis, it is necessary to represent the time to the event. This column is calculated by assessing the difference in

months between the diagnosis date and the death date. For censored patients, the time to event column contains the time difference between the diagnosis date and the end of the study, which is 31/12/2021.
In summary, this continuous variable "time to event month" indicates the number of months a patient survived from the date of diagnosis, representing the primary outcome of interest in the study.

Figure 2 shows the distribution of survival times (in months) for lung cancer patients in the dataset. The histogram is highly right-skewed, highlighting that most patients die relatively early (within the first few months).

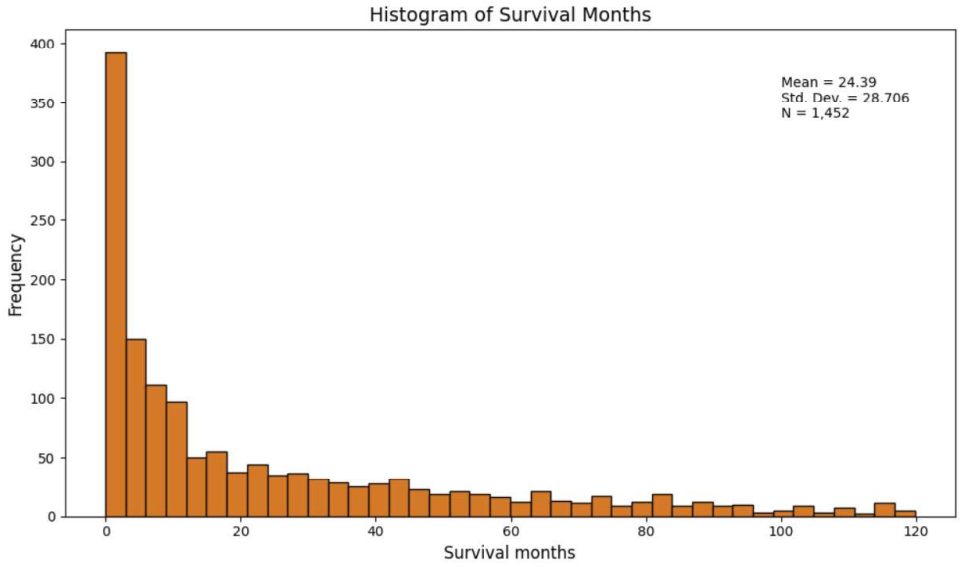


Figure 2.- Distribution of survival times for lung cancer patients.

3.3.2. Exploratory Data Analysis

The variables were inspected further through an Exploratory Data Analysis to observe patterns and trends.

It has been observed that 77% of the patients from the dataset are male. This disparity is supported by the literature, which suggests that historically, there has been higher smoking rates among men and more occupational exposures to carcinogens such as industrial chemicals, for instance. The combination of these environmental factors, along with biological influences, may help explain the higher prevalence of lung cancer in men [52] [53].

Moreover, most of the patients are between 60 and 80 years old, being a condition which is more common in the older population. This also aligns with the fact that lung cancer develops over time due to a prolonged exposure to risk factors such as smoking, environmental pollutants, and occupational carcinogens. While rare, lung cancer can also occur in younger individuals, especially

due to genetic mutations in this case. In the dataset, the minimum age of the patient at the diagnosis was 19, and the older 97.

Figure 3 illustrates the age distribution of patients categorized by death status. The peak of the distribution for surviving patients (blue) occurs at a younger age compared to the deceased group (orange), which indicates that patients who were alive at the end of the study tended to be younger. This also suggests that mortality is more common in older patients. Moreover, the substantial overlap between both distributions indicates that age alone is not the only determinant of survival.

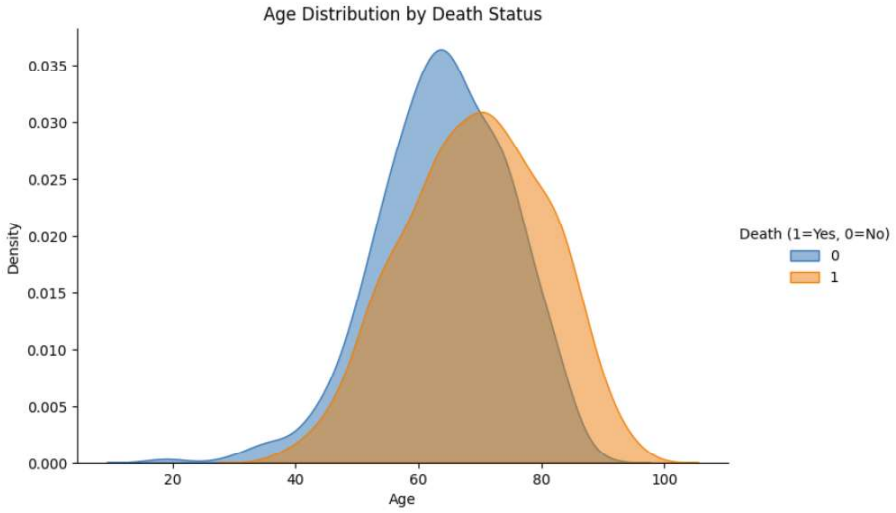


Figure 3.- Age distribution of patients categorized by death status.

Taking into account the age group, it is possible to observe that patients from group 4 (which are between 70 and 79), account for the most deaths *(Figure 4)*:

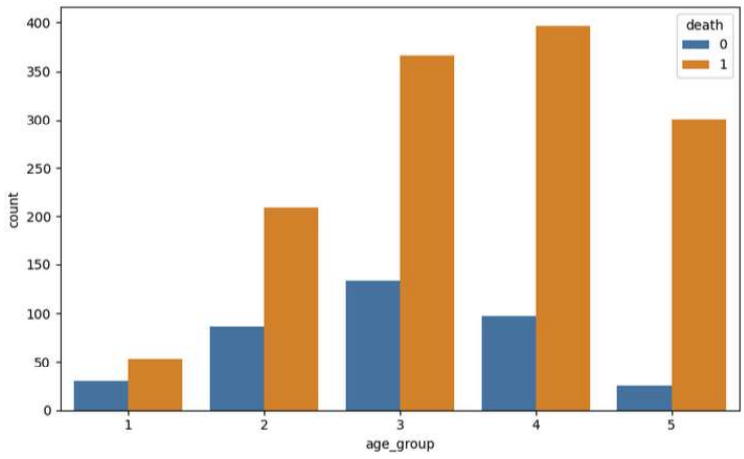


Figure 4.- Death status by age group.

In addition, 61% of the patients identified themselves as smokers. *Figure 5* shows the smoking habits categorized by death status. At first sight, it is possible to observe few differences between the survival of smokers versus non-smokers.

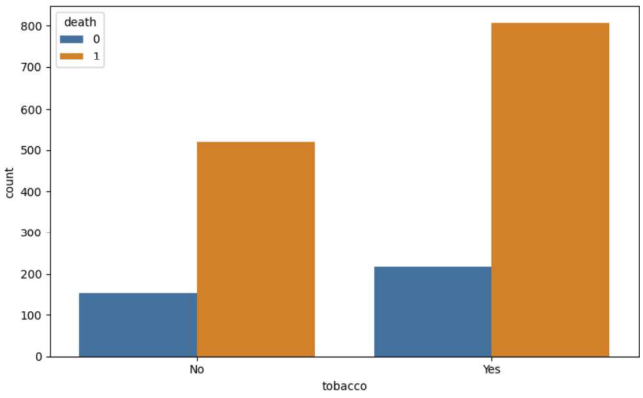


Figure 5.- Death status by tobacco consumption.

3.4. Traditional Survival Analysis

This study applies traditional survival analysis techniques to evaluate the survival of lung cancer patients by using past hospital data.

The primary outcome is overall survival (OS), defined as the time in months from the diagnosis until death (event of interest) or last follow-up (censoring).

Moreover, to conduct a comprehensive survival analysis, two additional sub-versions of the dataset were created for a deeper exploration of survival trends. The objective was to identify patterns that might be masked when analyzing the dataset as a whole.

The survival analysis was performed on the following datasets:

1. Full Lung Cancer Dataset – Includes all patients, providing an overall survival estimate.
2. Male-Only Subset – Analyzes survival probabilities for men.
3. Female-Only Subset – Analyzes survival probabilities for women.

By doing a separate survival analysis for male and female patients, the confounding effects are reduced, as it is ensured that gender-specific survival outcomes are not influenced by differences in smoking rates, treatment access, or tumor biology.

The statistical approaches that have been employed to analyze patient survival are: a bivariate analysis, the Kaplan-Meier estimator, the Cox Proportional Hazards Model and the Weibull regression model.

As a first step, a bivariate analysis was conducted to explore potential associations between categorical variables (sex, age group, tobacco use, and tumor morphology) and patient survival. This analysis is also referred to as “crude”, because it does not control for the influence of other variables.

Crude ratios (with 95% confidence intervals and p-values) have been calculated for each category of the variable. The crude risk ratio (RR) is the unadjusted ratio of the probability (or risk) of death in a specific group compared to the reference group.
[Eq. 3] shows how the RR are calculated. A and B represent the number of deaths and survivors in the group of interest, while C and D represent the same but in the reference group.

$$RR = \frac{A/(A+B)}{C/(C+D)} \quad [Eq. 3]$$

The reference group, by definition, has a crude ratio of 1.0. If the crude ratio of a given category is greater than 1, it means there is a higher risk of death compared to the reference group, while if it is less than 1, it indicates a lower risk.

Performing a bivariate analysis is particularly useful in the early stages of the analysis, since it helps to explore the data and it allows identifying potential associations between individual variables and the outcome binary variable. Nevertheless, since it does not account for censoring and it does not adjust for other variables, the results should be complemented with more complex, multivariable models.

Subsequently, the Kaplan-Meier estimator was used to compute and visualize the survival probabilities over time. In contrast to the bivariate analysis, Kaplan-Meier incorporates the time dimension and accounts for censoring.
A *survfit()* function was applied to estimate the overall survival curve, which represents the probability of patients surviving beyond a certain time point.

In order to explore differences in survival across subgroups, patients were stratified based on targeted clinical and demographic variables, including:

- Sex (Male vs. Female)
- Tumor Morphology (Different histological subtypes)
- Smoking Status (Smoker vs. Non-Smoker)
- Age Group (Categorized age intervals)

For each subgroup, a separate Kaplan-Meier survival curve was generated, allowing for a visual comparison of survival patterns among them. A steeper decline in the survival curve indicates a higher mortality of patients belonging to that particular subgroup.

Besides the survival curves, the Kaplan-Meier estimator provides with the number of individuals in each group, the number of events (deaths), the median survival times, and the confidence intervals (CI).

To statistically assess whether survival distributions differed significantly between groups, the log-rank test was employed. The log-rank test [54] is a hypothesis test that compares the entire survival curves of two or more groups over time. It consists of a null hypothesis that assumes both

groups have identical distribution curves and no difference in survival, and an alternative hypothesis that states that at least one group has a different survival experience.

In summary, for each time point where an event (death) occurs, the test calculates the observed deaths (O) and the expected deaths (E), which are the number of deaths expected if there was no difference in survival, for "i" groups [Eq. 4]:

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i} \quad [Eq. 4]$$

The Chi-Square (χ^2) measures how much the observed events variate with the expected ones, with the assumption that no difference between the groups takes place. Then, the p-value is computed using the Chi-Square cumulative distribution function (CDF), and taking into account the degrees of freedom.

If the p-value is less than the threshold value of 0.05, the null hypothesis is rejected, which indicates that there is a statistically significant difference in survival between the compared groups.

This statistical test assumes the following: that the survival times of individuals in each group are independent to each other, that the probability of censoring is the same for all individuals within each group, and finally that the hazard rates (the risk of an event occurring) for the compared groups is consistent over time.

While Kaplan-Meier survival curves provide visual comparisons and median survival times, they analyze only one variable at the time and do not adjust for confounding factors. This means that the observed survival differences could be due to the influence of other variables rather than the variable itself.

Therefore, in order to perform a robust survival analysis, Cox proportional hazards model (also known as Cox regression model) is applied afterwards. Contrary to the Kaplan-Meier, the Cox model allows for a simultaneous analysis of multiple covariates, adjusting for potential confounders. It can determine whether a variable still influences on survival after accounting for the others.
In addition, it quantifies the impact each predictor has on the survival through hazard ratios (HRs), providing the relative risk of an event occurring in one group compared to another.
It can also incorporate interactions between predictors, making it an even more powerful tool.
The Cox regression model was also implemented in Rstudio, leveraging its built-in functions to handle censoring, multivariate survival analysis, and statistical validation. R automatically computes the hazard ratios, confidence intervals, and model significance tests.

Table 4 summarizes the output of a Cox proportional hazards model computed using R:

Column in R	Description
coef (β)	Represents the log HR. Negative values indicate a protective effect (lower risk of death), while positive values indicate a higher risk of death.
exp(coef)	It is the HR, representing the relative risk of the event occurring for one unit increase in the variable.
se(coef)	The standard error, measures the variability of the coefficient estimate.
z-score	Tests whether the coefficient is significantly different from zero.
Pr(>|z|)	P-value to asses statistical significance.
exp(-coef)	Represents the relative decrease in hazard.
lower .95 and upper .95	Represent the lower and upper bound of the 95% confidence interval for the HR.

Table 4.- Output of a Cox model in R.

Additionally, the model provides overall statistics such as:

- Concordance Index: Measures how well the model predicts the survival times.
- Likelihood Ratio Test: A significant p-value suggests the model is better than a null model.
- Wald test: Assesses whether the coefficients of the predictors are significantly different from zero.
- Log-rank test: A test for the overall effect of the predictors on survival. A significant p-value suggests the predictors have a strong effect on survival.

In order to calculate the percentage of increase or decrease in hazard, the following formula is used:

$$\% = (HR - 1)\, x\, 100 \; [Eq. 5]$$

If the HR is higher than 1, the % will be positive and this implies a % increase in risk when compared with the reference group. On the other hand, if the HR is smaller than 1, the factor has a protective effect and the % will be negative, suggesting a % decrease of risk.

In a multivariable Cox regression model, where multiple covariates are included, HRs are adjusted hazard ratios, since they account for the influence of all the other variables in the model.

Before interpreting the results, there are some assumptions that need to be met in order to trust and interpret them correctly. The most important ones are listed below:

- Proportional Hazards (PH) Assumption: It states that the hazard ratio remains constant over time, meaning that the effect of the predictor variables on the hazard function is assumed to be constant over time.

- Independence assumption: Survival times should be independent to each other, so the survival of one individual should not influence the survival of another individual.
- Linearity assumption: The relationship between the predictor variables and the log of the hazard rate is assumed to be linear.
- No multicollinearity among the predictor variables.

From the listed assumptions, the PH is the central assumption of the Cox regression and therefore the most crucial to test. If this assumption is violated, the hazard ratios may be biased or misleading [55].
To validate this critical assumption, several statistical and graphical methods have been used, including the Schoenfeld residuals test. This test examines whether there is a significant relationship between time and the residuals of the model.
In a valid Cox model, the Schoenfeld residuals should be randomly scattered over time, indicating that the effect of a predictor is constant. On the other hand, if a trend is observed, it suggests that the hazard ratio of a variable changes over time and consequently violates the PH assumption. In addition, the test outputs a p-value. If this p-value is large enough ($p > 0.05$), it implies that there is no significant relationship and that the PH holds. In contrast, if the p-value is small, there is a relationship and the PH does not hold.

For cases when the PH assumption does not hold, other approaches such as parametric models may provide a better fit. In this study, the Weibull model in particular has been applied, assuming a specific mathematical form for how the hazard function changes over time.

The Weibull regression model follows a log-linear relationship for survival time, estimating:

$$log(T) = \beta_0 + \beta_1 \cdot X_1 + \ldots + \beta_n \cdot X_n + \sigma \cdot W \quad [Eq. 6]$$

Where T is the survival time, β_n the coefficients , X_n the predictor variables, σ the scale parameter and W a noise term.

In contrast to the Cox models, the β coefficients for the Weibull model represent the change in the log survival time rather than a change in the log HRs. While Weibull model reports negative coefficients for risk factors associated with shorter survival, the Cox model reports Hazard Ratios (HRs) > 1 for factors that increase risk.

In this context:

- If $\beta > 0$, the predictor is associated with longer log survival.
- If $\beta < 0$, the predictor is associated with shorter log survival.
- If $\beta = 0$, there's no effect.

If β coefficients are exponentiated [Eq. 7], it is possible to interpret the results in acceleration factor (AF) terms, where:

- If AF > 1, the survival time is longer.
- If AF < 1, the survival time is shorter.

$$AF = e^{\beta} \; [Eq.7]$$

AF represents the multiplicative effect of a predictor on survival time, and can be easier to interpret than β coefficients, which are on a log scale.

Moreover, the scale parameter σ specifies how long survival times are, on average, and it is very useful in this context to extract the shape parameter γ:

$$\gamma = \frac{1}{\sigma} \; [Eq.8]$$

The shape parameter allows distinguishing between different survival patterns:

- γ=1 : The hazard function is constant over time.
- γ>1 : The hazard function increases over time.
- γ<1 : The hazard function decreases over time.

Thus, the Weibull model is very flexible and can represent increasing and decreasing hazard rates, unlike the Cox model, which assumes proportional hazards.

In the output of the Weibull model in R, each row corresponds to a regression coefficient for a predictor in the model. The first row, however, corresponds to the intercept, which is the baseline log survival time when all covariates are at their reference level.

Columns from the output are described in the following table *(Table 5)* [57]:

Column in R	Description
Value (β)	Estimated coefficient for each predictor. Quantifies the effect of a predictor in the log of survival time.
Std. Error	Standard error of the coefficient estimate.
z	Tests whether the coefficient is significantly different from zero.
p	P-value to assess statistical significance of the predictor.

Table 5.- Output of a Weibull model in R.

On the other hand, it is important to ensure that the data follows a Weibull distribution tendency, since it is its main assumption.
In this study, this was visually verified by plotting the log-log plot. The idea behind this plot is that, if a natural logarithm is applied on both sides of the Weibull survival function, the result is a straight-line equation [57].
Therefore, if the points in the plot form a straight line, it means the Weibull distribution is a good fit. Otherwise, another survival model such as the log-normal or the log-logistic may be more appropriate.

Finally, the different models obtained in this section were compared by using the Akaike Information Criterion (AIC) [58], which is a measure used for model selection. AIC helps to compare different models by balancing goodness of fit and model complexity. The key idea is that it penalizes overly complex models to avoid overfitting and that a lower AIC indicates a better balance between fit and simplicity. Moreover, since different survival models have different assumptions, comparing them using AIC can help to determine the best approach for the data.

3.5. Machine Learning Survival Analysis

Machine Learning models designed for survival analysis have been applied to the processed data in order to compare the performance against traditional and purely statistical methods. These models, which have been only been applied to the Full Lung Cancer Dataset, can be grouped into three major categories based on their learning strategies and model architectures:

1. Ensemble Tree-Based Methods - Random Survival Forest (RSF)
2. Boosted Survival Models - Gradient Boosting Survival Analysis (GBSA)
3. Neural Network-Based Survival Models - Deepsurv architecture

Although the data was already cleaned and prepared, some additional processing was undertaken to appropriately prepare the data for the ML algorithms.
The relevant features were the same as in the traditional analysis : “age group”, “morphology”, “sex”, “tobacco”, “death” indicator and “time to event” columns.

In the case of the ensemble and boosted models, the processing steps followed were the same. First of all, the survival target variable was constructed using the *Surv.from_dataframe* utility from the *sksurv* library. This variable created a structured array that encodes the time-to-event variable along with the event indicator (death), which is the binary variable indicating if the event occurred or if the observation was censored. This target format is required for training survival analysis models, as it enables the model to account for right-censored data and predict time-to-event outcomes.
Next, categorical covariates (age, morphology, sex and tobacco) were converted into binary indicator columns by using one-hot encoding, making them suitable for tree-based ML algorithms, which expect numerical inputs.

Afterwards, the dataset was split into training and testing subsets and a 5-fold nested cross validation was employed for both model selection and performance assessment.

The inner loop was responsible for the hyperparameter tuning by using a randomized search with 25 iterations, while the outer loop was used to evaluate the performance of the best-found model on unseen data. This strategy ensures that each step of the model tuning is repeated on independent training or test splits, therefore reducing the risk of overfitting and ensuring robust evaluation.

The tuning process was applied to both RSF and GBSA models, although the hyperparameters explored slightly differed (see Annex A3). Moreover, the evaluation metric chosen for model selection was the concordance index, commonly used to evaluate risk models in survival analysis.
The outer loop collected the concordance index (C-index) of the test fold after each iteration and returned the mean concordance index together with the standard deviation.

For the Neural Network based model, a Deepsurv-inspired architecture was built using the CoxPH module from the *pycox* library, where a neural network learns risk scores by optimizing the Cox partial likelihood. Before training the Network, the survival time in months and the event indicator (death) were separated from the input features. Then, the categorical variable age group was treated as ordinal and transformed using label encoding. The rest of the categorical variables, on the other hand, were one-hot encoded. Afterwards, all input features were standardized using z-score normalization (*StandardScaler*). The input features (x) were separated from the survival time (y) and the death indicator (event). Finally, a 5-fold cross-validation also using the C-index as the evaluation metric was performed.

The DeepSurv neural network was trained using the training folds and evaluated on the corresponding test fold. The architecture consisted of two hidden layers with 32 units, ReLU activations, and batch normalization, followed by a single output neuron representing the log-risk function. After training, the risk scores were predicted for the test fold, and the final performance was reported as the mean and standard deviation of the C-index across the five folds.

3.6. Performance evaluation and comparison

The performance of traditional vs. ML-based survival models has been evaluated through the C-index and the Integrated Brier Score (IBS) metrics.

Harrell's **C-index** (Harrell et al., 1982) is a goodness of fit measure for models which produce risk scores. A higher C-index denotes better model discrimination, which means that is capable of distinguish patients with varying survival times based on their estimated risk scores [19].

The intuition behind this index is the following: the patient with the higher risk score should have a shorter time-to-event [59].

The index is measured on a scale of 0.5 to 1, with higher values indicating greater ability and 0.5 indicating no discrimination at all [19].

The C-index was computed using 5-fold stratified cross-validation, as mentioned before, to ensure an unbiased estimation of how well each model can rank patients. During this process, the dataset is randomly split into five parts of equal size (called folds).

Each fold serves once as the test set, while the remaining four form the training set.

Then, the model is trained on the training set and evaluated on the corresponding test set, and this is repeated five times so that every fold is used in the testing exactly once. The C-index from each fold is stored and, at the end, the mean ± standard deviation is reported.

For the Cox and Weibull models, *concordance.index()* function from R was used to calculate the C-index scores from predicted risk scores. It is worth mentioning that, in the Cox model, risk scores were directly derived from the linear predictor. However, for the Weibull model, the expected survival time was predicted and negated to obtain risk scores (since lower predicted survival time implies a higher risk).

For the ML-based survival models, the C-index was computed in Python. In the case of RSF and GBSA, the C-index was evaluated using a nested 5-fold cross-validation loop, where hyperparameter tuning was performed in the inner folds, and the evaluation using *sksurv's* C-index function. For DeepSurv, a neural network was evaluated using a 5-fold cross-validation and the *lifelines* C-index function.

The **Brier Score**, on the other hand, measures the overall performance, incorporating both the discrimination and calibration aspects of a model.

The Brier Score is calculated as follows:

$$Brier\ Score\ =\ \frac{1}{n}\Sigma_{i=1}^{n}(p_i - O_i)^2 \quad [Eq.9]$$

where n is the number of subjects, p_i is the probability of event predicted by the model for the *i-th* subject, and O_i is the observed outcome in the *i-th* subject (1 for event or 0 for non-event) [60].

The Brier score is calculated separately for each follow-up time point, but there is an extension (the IBS) that integrates the multiple Brier score values obtained at all follow-up times. The **IBS** calculates the mean squared difference between predicted probabilities and actual outcomes, providing an overall performance measure of how well predicted survival curves match actual outcomes on average. In this context, a score closer to 0 indicates a better predictive performance.

Unlike the C-Index, which is a single summary statistic, the Brier Score is time-dependent and requires survival probabilities at specific time points to compare them with the actual survival status at those same points. Therefore, to compute the IBS integral, a time grid is needed to provide the discrete time points to compare.

To define a fair time grid, the distribution of observed survival times in the dataset was examined. In the end, the time grid was set from 1 to 70 months, being a time window that covers up to the 90th percentile of survival times. This ensures that the IBS is computed in a time range where there are sufficient events and uncensored patients.

The IBS for the Cox and Weibull models was computed using the *pec* package in R, which evaluates time-dependent prediction error across a specified time grid.

To ensure robustness and to minimize possible overfitting bias, the IBS was computed using the .632 bootstrap method. This method samples the dataset with replacement to generate 100 bootstrap samples. In each iteration, a survival model was trained on the bootstrap sample, and the Brier Score was calculated on the observations that were not included in the bootstrap sample (out-of-bag). The .632 bootstrap is an improvement on plain bootstrapping where, on average, 63.2% of the original samples are selected in each bootstrap dataset and the remaining 36.8% are out-of-bag (not seen during training). Therefore, in each iteration, there is a natural train/test split. In the end, the apparent IBS (on the training set) and the out-of-bag IBS (on the test set) are combined using the following formula:

$$IBS_{.632} = 0.368 \cdot IBS_{Apparent} + 0.632 \cdot IBS_{Boostrap} \; [Eq. 10]$$

The final IBS values were extracted using the *crps()* function, which returns the integrated Brier score for each model in the predefined time grid. Moreover, by using all iterations of IBS values, confidence intervals or standard deviations can be computed.

This approach provides a more accurate generalization performance than evaluating on the training set alone, since it simulates training/testing on many datasets.

For ML-based survival models, the IBS score was implemented in Python [63] using a similar bootstrap-based evaluation for consistent comparison.

For RSF and GBSA, the *sksurv.metrics.integrated_brier_score()* function was used directly from the returned survival curves calculated with the *predict_survival_function()* method.

In contrast, DeepSurv, being a neural extension of the Cox Proportional Hazards model, outputs a relative risk score and does not directly provide survival probabilities. Therefore, the log-risk score needs to be transformed into survival functions using the estimated baseline hazard. Then, *model.predict_surv_df()* is used to generate the survival probabilities for each patient at each time point and the Brier scores are calculated. Finally, IBS is computed manually by integrating the Brier scores over the time grid using *numpy.trapezoid()*.

While performance metrics quantify the predictive ability and calibration of survival models, they do not provide insights into which features or variables influence the most on patient survival. Therefore, in order to complement these global metrics, feature importance plots were generated as a way to visualize and interpret how individual covariates contribute to the predictions of the model. These plots can help to compare models and they offer interpretability and transparency, especially for the ML models.

For the Cox model, feature importance was computed using the model coefficients, which represent the log HRs, where a positive ratio indicates that a variable increases the hazard, whereas a negative one indicates a protective effect. In contrast, the coefficients of the Weibull model represent changes in log survival time, so to make them comparable to HRs, they were transformed into log HRs using the formula:

$$log(HR) = -\frac{coefficient}{scale} \quad [Eq. 11]$$

Unlike the Cox or Weibull models, tree-based machine learning models do not provide direct coefficients. Instead, the permutation importance was used to assess how much each feature contributes to model performance. Basically, permutation importance works by randomly shuffling multiple times the values of a single feature (breaking its relationship with the target) and measuring how much the performance of the model gets degraded as a result. If shuffling a feature reduces the performance, it means that the model relied on that feature to make predictions. In contrast, if performance remains unchanged, the feature is likely unimportant.

The permutation importance was calculated in Python using the *sklearn.inspection.permutation_importance function*, and the performance metric used was the C-index.

DeepSurv, on the other hand, transforms input features through multiple hidden layers and therefore does not produce simple, interpretable coefficients or feature splits that can be directly mapped to an effect on survival. Because of these characteristics, DeepSurv can often be considered as a "black-box" model, since its predictions are the result of complex, non-linear interactions that are not easy to decompose and even understand. Its interpretation is not as straightforward and would require the use of advanced post-hoc techniques (such as SHAP values), which are out of the scope of this thesis.

4. Results

4.1. Bivariate Analysis

According to the bivariate analysis *(Table 6)*, female patients showed a significantly higher survival rate (34.4%) compared to male patients (23.0%), with a crude risk ratio of 0.8 (95% CI: 0.8–0.9, p < .001), indicating a 20% lower unadjusted risk of death compared to men.

Patients with blastoma and small cell carcinoma, in contrast, had significantly lower survival rates and high crude risk ratios of 7.0 (95% CI: 2.8–17.76, p < .001) and 4.3 (95% CI: 2.4–7.9, p < .001), respectively, when compared to adenocarcinoma.

On the other hand, the crude risk ratio for tobacco use was 1.03 (95% CI: 1.0–1.1, p = 0.15), indicating a non- statistically significant difference.

In terms of age, the survival appears to decrease when age increases. Patients 80 or older had a significantly lower survival rate (9.9%) and also the highest crude risk ratio (5.3 with 95% CI: 2.8–9.9, p < .001), when compared to the youngest age group (0–49 years, 33.3%). Intermediate age groups ([50–79]) showed a moderate increase in risk, but their crude risk ratios were not statistically significant.

	Total counts		Death		Alive		Survival	Bivariate Analysis		
	N	%	N	%	N	%	%	Crude ratio	CI 95%	p-value
Sex										
Man	1122	77,3	864	80	258	69,5	22.99	Ref. Group		
Female	329	22,7	216	20	113	30,5	34.35	0.8	0.8-0.9	<.001
Morphology										
Adenocarcinoma	571	39,4	385	35,6	168	45,3	32.57	Ref. Group		
Non-small cell carcinoma	91	6,3	74	6,9	17	4,6	18.68	1.8	1.0-3.3	.009
Blastoma	86	5,9	81	7,5	5	1,3	5.81	7.0	2.8-17.76	<.001
Epidermoid	117	8,1	87	8,1	30	8,1	25.64	1.2	0.8-2.0	0.15
Others	183	12,6	125	11,6	58	15,6	31.69	0.9	0.6-1.3	0.4
Small cell carcinoma	143	9,9	130	12,0	13	3,5	9.09	4.3	2.4-7.9	<.001

Squamous	260	17,9	198	18,3	62	16,7	23.84	1.4	1.0-1.9	.02
Tobacco										
No	566	39,0	413	38,2	153	41,2	27.03	Ref. Group		
Yes	885	61,0	667	61,8	218	58,8	24.63	1.03	1.0-1.1	0.15
Age group										
[0-49]	81	5,6	51	4,7	30	8,1	33.33	Ref. Group		
[50- 59]	269	18,5	183	16,9	86	23,2	31.97	1.2	0.7-2.1	0.2
[60-69]	438	30,2	305	28,2	133	35,8	30.37	1.3	0.8-2.2	0.12
[70-79]	411	28,3	314	29,1	97	26,1	23.60	1.3	0.8-2.2	0.12
[80 - ~)	252	17,4	227	21,0	25	6,7	9.92	5.3	2.8-9.9	<.001

Table 6.- Bivariate analysis.

4.2. Kaplan-Meier Curves

The following Kaplan-Meier survival curve *(Figure 6)* represents the OS, estimating the probability of patients surviving beyond specific time points that are defined in months. The dashed blue lines represent the 95% confidence interval (CI) for the estimated survival probability.

It can be observed how the curve starts at 1.0 (100% survival probability), but rapidly declines within the first few months. This indicates that a significant proportion of patients experience mortality early during the follow-up period. In fact, the median overall survival in the cohort was estimated at 10.94 months (95% CI: 9.99–12.75 months), suggesting that half of the study population experienced the event (death) within approximately 11 months of follow-up.

After the initial drop, the decrease in the survival probability slows down over time, suggesting that those patients who have survived the first critical months tend to have better prognosis.

Moreover, the CI is narrower at the beginning, highlighting how the uncertainty increases as fewer patients remain in the study.

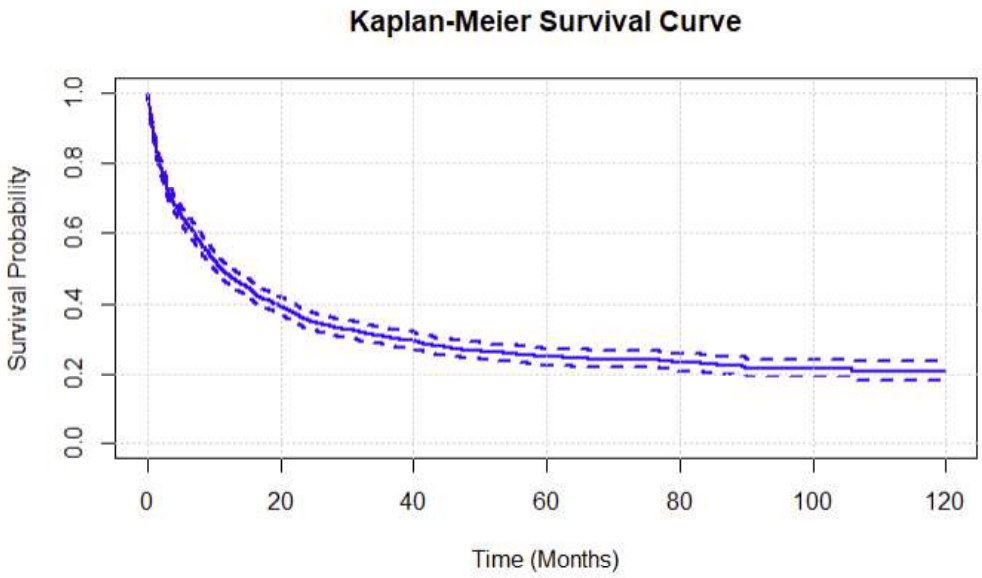


Figure 6.- Overall Kaplan-Meier Survival curve in months.

The following figures show the Kaplan-Meier survival curves stratified by sex, smoking habits, age group and morphology.

From *Figure 7*, it is possible to see that females have better survival outcomes when compared to males, since their survival probability remains higher during the complete follow-up period. The median OS for man is estimated at 9.54 months (95% CI: 8.41 – 10.71 months), while for women was 22.54 months (95% CI: 15.84 – 28.58 months). Since the confidence intervals do not overlap, it is suggested that females have a statistically significant survival advantage compared to males.

Nevertheless, there are far more males (1122) than females (329) in the study, so the overall survival curve is influenced by the male survival pattern. While this makes the male survival estimate more precise, it produces wider confidence intervals in the female, suggesting a greater uncertainty and variability.

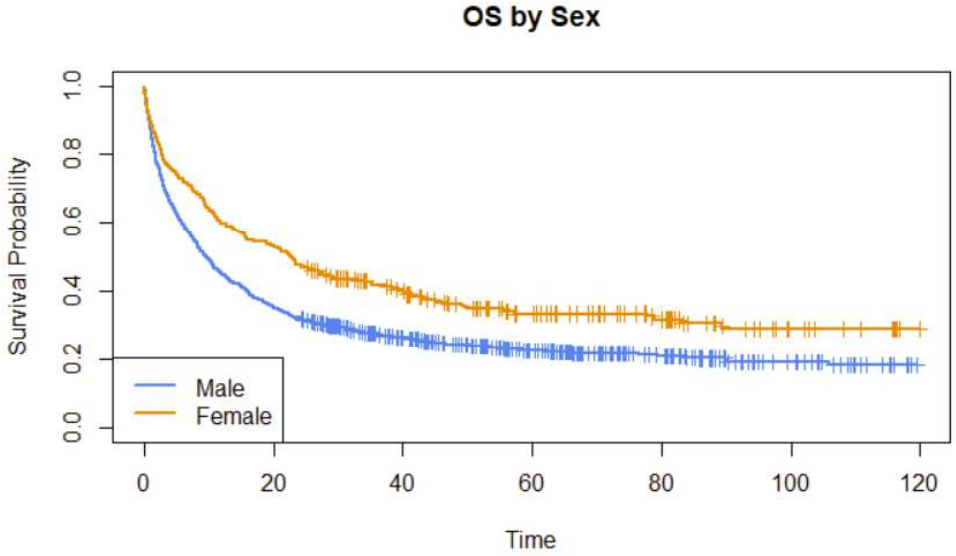


Figure 7.- Survival curve stratified by sex.

Surprisingly, the survival curve for smokers and non-smokers (*Figure 8)* shows no clear advantage for one group over the other, since the survival probability decreases at a similar rate for both. This may suggest that smoking status alone may not be a strong determinant of survival in this particular dataset.

The median OS for non-smokers is 10.91 months (95% CI: 9.30 – 13.83 months), while for smokers is 10.97 months (95% CI: 9.76 – 13.24 months). Nevertheless, in this case, the confidence intervals do overlap, which indicates that the difference is not significant and that additional statistical testing is needed.

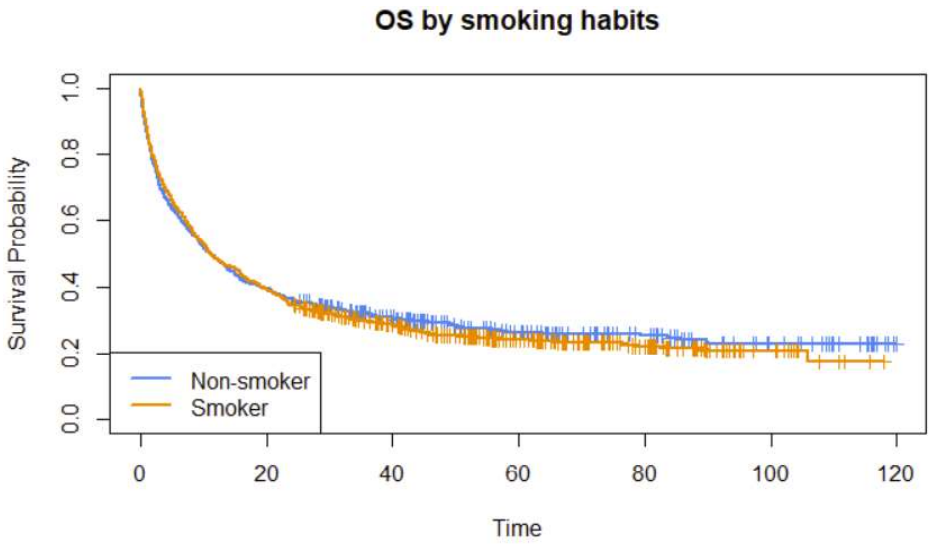


Figure 8.- Survival curve stratified by smoking status.

The survival curve that is stratified by age groups (*Figure 9*), highlights that younger patients have better survival outcomes.

For instance, patients that are under 50 have the highest median survival probability, which is 23.10 months (95% CI: 13.93 – 46.59 months), and their survival curve declines more gradually, meaning they have a lower early mortality.

In contrast, the age group of over 80 years old has the steepest drop in the survival curve, and the lowest median survival, which is around 4 months (95% CI: 3.19 – 5.85) This indicates higher early mortality in older patients.

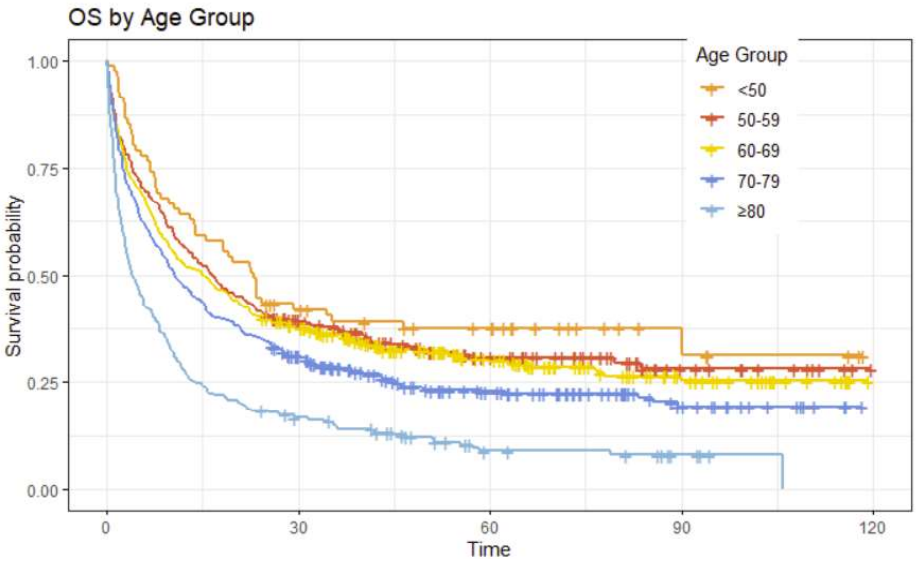


Figure 9.- Survival curve stratified by age groups.

Next, the Kaplan-Meier survival curve was stratified by the tumor morphology *(Figure 10)* in order to provide insights into how different histological subtypes of lung cancer impact on the OS.
There is a clear stratification of survival probabilities, which indicates that the tumor morphology is indeed a strong determinant of survival.

Adenocarcinoma (known as the most common NSCLC subtype) has the highest median survival time, 19.02 months (95% CI: 15.61 – 23.03), and has a slower decline in the survival curve.

In contrast, Blastoma appears to be the most aggressive tumor type, with a median survival time of only 1.84 months (95% CI: 1.28 – 3.45) and a faster decline in the survival curve.

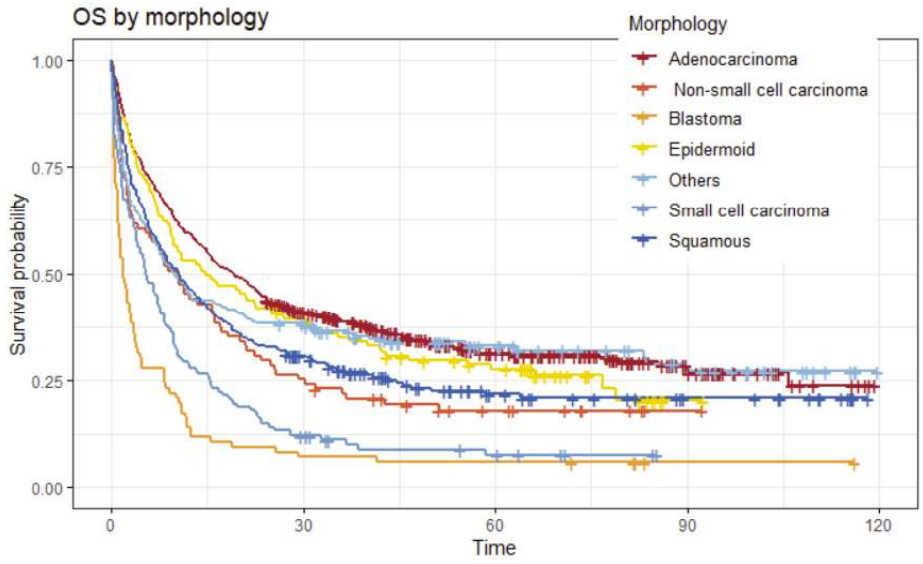


Figure 10.- Survival curve stratified by tumor morphology.

Table 7 contains a summary of the log-rank test results of the previous survival curves:

Variable	p-value	Result
Sex	1e-06	Significant
Morphology	<2e-16	Significant
Tobacco	0.6	Non significant
Age group	<2e-16	Significant

Table 7.- Log-rank test results of stratified Kaplan-Meier curves.

4.3. Cox Regression Models

4.3.1. General Cox model

The first Cox regression model applied included as predictors the sex, age group, smoking status and morphology. *Table 8* contains a summary of the obtained results:

Characteristic	HR	95% CI	p-value
Sex			
Male	—	—	
Female	0.74	0.63-0.86	<0.001
Age Group			
<50	—	—	
50-59	1.18	0.87-1.62	0.3
60-69	1.21	0.90-1.63	0.2
70-79	1.45	1.08-1.96	0.014
>80	2.26	1.66-3.08	<0.001
Tobacco			
No	—	—	
Yes	1.00	0.88-1.13	>0.9
Tumor Morphology			
Adenocarcinoma	—	—	
Non-small cell carcinoma	1.47	1.14-1.88	0.003
Blastoma	2.42	1.88-3.13	<0.001
Epidermoid	01.02	0.81-1.29	0.9
Others	1.18	0.96-1.44	0.11
Small cell carcinoma	2.16	1.77-2.65	<0.001
Squamous	1.21	1.01-1.45	0.034

Table 8.- Multivariate analysis—Cox regression. Reference groups are marked as -.

It is possible to observe that female patients had a significantly lower risk of death, more specifically of 26%, and this difference is significant (95% CI: 0.63–0.86, p < 0.001).

Regarding the age group, it seems that the older the patient, the higher the HR, which implies that older age groups have a higher mortality. Patients with ages between 70 and 79 have 1.45 times more risk of death (HR: 1.45, 95% CI: 1.08–1.96, p = 0.014) when compared with the reference group, which are patients younger than 50 years. Similarly, patients over 80 years have a 2.26 times more risk of death (HR: 2.26, 95% CI: 1.66–3.08, p < 0.001) compared with the reference group. In the rest of the age groups, due to the lack of significance, it is not possible to state that they have a real and non-random effect on patient survival.
After adjusting for the other variables, tobacco use does not appear to have a significant impact survival (HR: 1.00, 95% CI: 0.88–1.13, p > 0.9).

Blastoma and Small-cell carcinoma had the worst survival, with HRs of 2.42 and 2.16 respectively (both p < 0.001), when compared to the reference category (Adenocarcinoma). Non-small cell carcinoma had a 47% higher risk of death (HR: 1.47, 95% CI: 1.14–1.88, p = 0.003) and Squamous a 21% (HR: 1.21, p = 0.034). Moreover, there are no strong evidences supporting that Epidermoid and other morphologies affect the survival differently when compared to the reference category.

Figure 11 illustrates the HRs and 95% confidence intervals for the variables from the previous Cox model. Statistically significant variables do not have an HR equal to 1 and the CIs do not include 1.

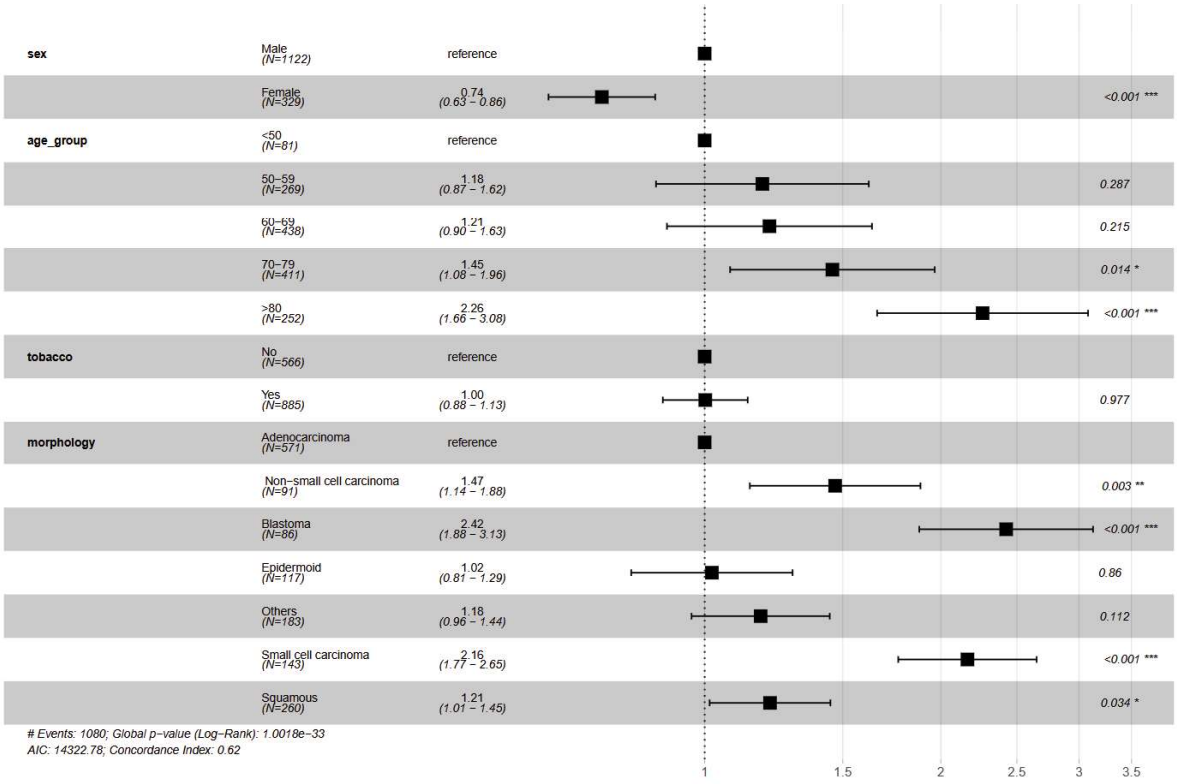


Figure 11.- Forest plot of HRs and their 95% CIs for the general Cox regression analysis.

Next, the PH assumption was checked by using the *cox.zph()* function in R. This function tests whether the Schoenfeld residuals are correlated with time and computes a p-value.
The p-values of the test are summarized in *Table 9,* whereas the complete table can be found in Annex A2.

Variable	p-value	Result
Sex	0.65	Assumption holds
Age group	0.66	Assumption holds
Tobacco	0.066	Assumption holds
Morphology	0.00046	Assumption does not hold

Table 9.- p-values of the Schoenfeld residuals.

It is possible to observe that the morphology variable violates the PH assumption, suggesting that the effect of morphology on survival is not constant over time.

4.3.2. Adding interaction terms

Instead of assuming that smoking affects men and women in the same way, an interaction term (sex* tobacco) is added to test whether the effect of tobacco on survival differs by sex.

Table 10 contains the summary result table after adding the interaction term. Note that age group and morphology have not been included in the table, as these were already analyzed in the previous Cox model.

Variable	HR	95 % CI	p-value
TobaccoYes	0.99	(0.86 – 1.13)	0.85
SexFemale	0.71	(0.55 – 0.91)	0.0061
SexFemale:TobaccoYes	1.077	(0.79 – 1.47)	0.64

Table 10.- Cox model results of the interaction term sex*tobacco.

While sex remains significant after the interaction term, tobacco alone is still not significant (p=0.85), as well as the interaction term (p=0.64). This means that tobacco effect does not significantly differ between men and women in this dataset.
In addition, the morphology variable also violates the PH assumption in this Cox model (see Annex A2).

4.3.3. Stratifying by sex

Table 11 contains a summary of the results of both Cox models, one containing only men and the other only women.

It is observed that age generally has a stronger impact on older men, whereas it only becomes significant for women at over 80 (HR: 2.17, 95% CI: 1.27–3.69, p = 0.004).

Smoking does not significantly affect the survival in either sex (HR men: 0.99, p = 0.9, HR women: 0.99, p > 0.9).

Regarding the tumor morphology, small cell carcinoma is associated with the worst prognosis, particularly for women (HR = 3.75, p < 0.001) when compared to men (HR = 2.09, p < 0.001). In addition, Blastoma also significantly increases mortality risk in both sexes. Interestingly, the "Others" tumor category shows a higher mortality risk in men (HR = 1.67, p < 0.001) but a protective effect in women (HR = 0.39, p < 0.001), which may suggest a possible classification bias or biological differences.

	Men			Women		
Characteristic	HR	95% CI	p-value	HR	95% CI	p-value
Age Group						
<50	—	—		—	—	
50-59	1.11	0.75-1.64	0.6	1.17	0.68-2.01	0.6
60-69	1.13	0.78-1.65	0.5	1.26	0.74-2.17	0.4
70-79	1.48	1.02-2.16	0.039	1.21	0.70-2.09	0.5
>80	2.22	1.50-3.27	<0.001	2.17	1.27-3.69	0.004
Tobacco						
No	—	—		—	—	
Yes	0.99	0.86-1.14	0.9	0.99	0.74-1.32	>0.9
Tumor Morphology						
Adenocarcinoma	—	—		—	—	
Non-small cell carcinoma	1.54	1.17-2.04	0.002	1.33	0.70- 2.51	0.4
Blastoma	2.51	1.88-3.35	<0.001	2.18	1.27-3.77	0.005
Epidermoid	1.04	0.81- 1.34	0.7	1.20	0.56-2.58	0.6
Others	1.67	1.34-2.10	<0.001	0.39	0.24-0.66	<0.001
Small cell carcinoma	2.09	1.67-2.61	<0.001	3.75	2.26- 6.20	<0.001
Squamous	1.23	1.01- 1.49	0.037	1.38	0.85-2.23	0.2

Table 11.- Cox regression stratified by sex (Men, Women). Reference groups are marked as -.

The PH assumption was tested and held in both Cox models separated by sex (see Annex A2), validating the interpretation and results.

4.4. Weibull model

Table 12 summarizes the results of the Weibull model. It is possible to see an alignment between the predictors of survival chosen by Cox and Weibull, highlighting sex and age as strong survival predictors. Female patients had a β coefficient of 0.55, which corresponds to an acceleration factor (AF) of 1.73 and therefore indicates that the survival time is 1.73 times longer in the female group. In essence, this implies a 73% longer survival time when compared to males ($p < 0.001$).

In addition, compared to patients younger than 50, those that were aged 70–79 had a 50% shorter survival time (β of -0.69 and AF = 0.50, $p = 0.007$), while those over 80 had a 78% shorter survival time (β of -1.5 and AF = 0.22p , < 0.001).

Again, tobacco use was not significantly associated with survival (β = -0.03, $p = 0.8$).

Finally, tumor morphology also had an impact on survival outcomes. When compared to adenocarcinoma, patients with non-small cell carcinoma (β = -0.70, AF = 0.50, $p = 0.001$), blastoma (β = -1.5, AF = 0.22, $p < 0.001$), and small cell carcinoma (β = -1.4, AF = 0.25, $p < 0.001$) experienced significantly shorter survival times, while the rest did not show any statistically significant difference at all.

Characteristic	Beta	95% CI	p-value
Sex			
Male	—	—	
Female	0.55	0.29, 0.81	<0.001
Age Group			
<50	—	—	
50-59	-0.31	-0.83, 0.21	0.2
60-69	-0.35	-0.86, 0.15	0.2
70-79	-0.69	-1.2, -0.18	0.007
>80	-1.5	-2.0, -0.97	<0.001
Tobacco			
No	—	—	
Yes	-0.03	-0.24, 0.18	0.8
Tumor Morphology			
Adenocarcinoma	—	—	
Non-small cell carcinoma	-0.70	-1.1, -0.28	0.001
Blastoma	-1.5	-1.9, -1.1	<0.001
Epidermoid	-0.01	-0.41, 0.39	>0.9
Others	-0.22	-0.56, 0.12	0.2
Small cell carcinoma	-1.4	-1.8, -1.1	<0.001
Squamous	-0.34	-0.64, -0.04	0.025

Table 12.- Weibull results. Reference groups are marked as -.

The shape parameter is approximately 0.598, and since $\gamma < 1$, it is possible to state that hazard decreases over time.

The Weibull model assumption is checked with the log-log plot *(Figure 12)*. The plot shows a line tendency, which means that the assumption for the Weibull distribution is valid.

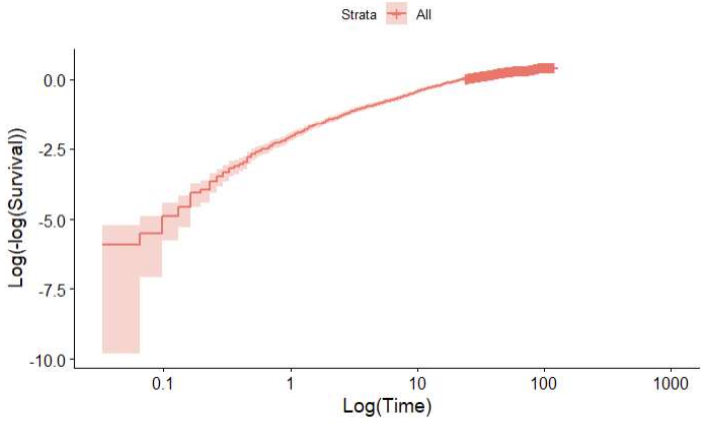


Figure 12.- Log-log plot for confirming the Weibull distribution assumption.

4.5. Comparisons

First, traditional survival models are compared between them through the AIC in order to determine which model best describes the present survival data without overfitting.

Table 13 contains the AIC results of the different models that have been compared. Note that AIC assumes that the models compared are fit to the same dataset, so the two Cox models fitted to the male data and female data separately have been excluded from the comparison.

It is possible to observe that the Weibull model has the lowest AIC (8819.91), which confirms that the data follows a specific distribution. Therefore, using this model can be more effective and can provide a more interpretable survival function.

Model	AIC
Weibull model	8819.91
Cox model without interaction	14322.78
Cox model with sex*tobacco interaction	14324.56

Table 13.- Traditional survival models (Weibull, general Cox and Cox with interaction term) are compared through the AIC and ordered from the best model (lowest AIC) to the worst model (highest AIC).

Next, traditional survival analysis models (Cox, Weibull) are compared against ML-based survival analysis models (RSF, GBSA and Deepsurv). By comparing the C-index and IBS (*Table 14*), it is possible to evaluate each model's predictive accuracy and power.

Model	C-index	IBS
Cox	0.616 ± 0.027	0.189
Weibull	0.615 ± 0.026	0.183
RSF	0.622 ± 0.016	0.199 ± 0.007
GBSA	0.620 ± 0.014	0.198 ± 0.007
DeepSurv	0.603 ± 0.012	0.226 ± 0.009

Table 14.- Traditional models (Cox, Weibull) are compared against ML models (RSF, GBSA, DeepSurv) through the C-index and IBS metrics.

Figures 13 and *14* display the feature importance of each predictor derived from Cox and Weibull models, respectively. *Figures 15* and *16*, on the other hand, show permutation-based feature importance for the RSF and GBSA models.

In the four figures, longer bars indicate a higher importance, suggesting that the corresponding variable has a more influential role in determining the patient survival.

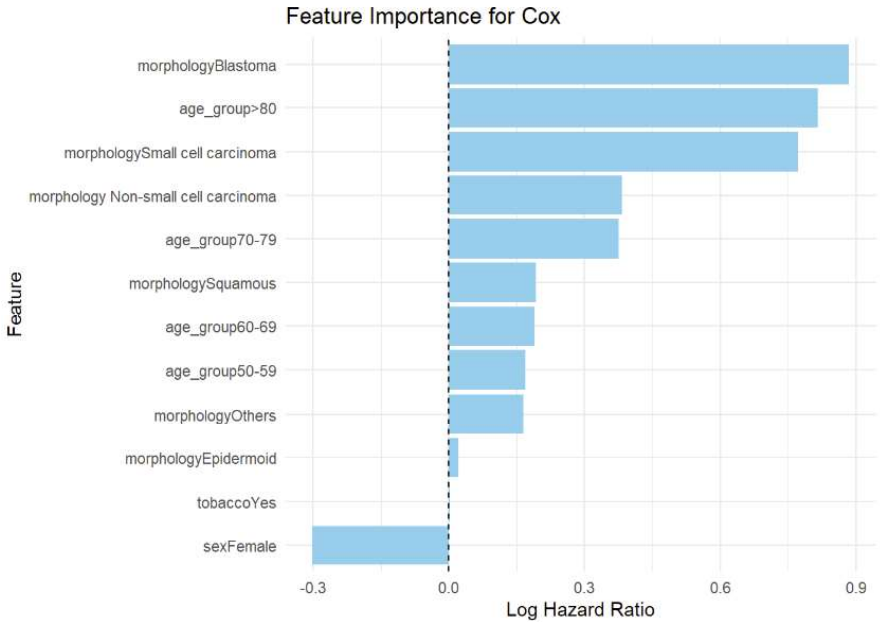


Figure 13.- Feature importance plot for Cox model.

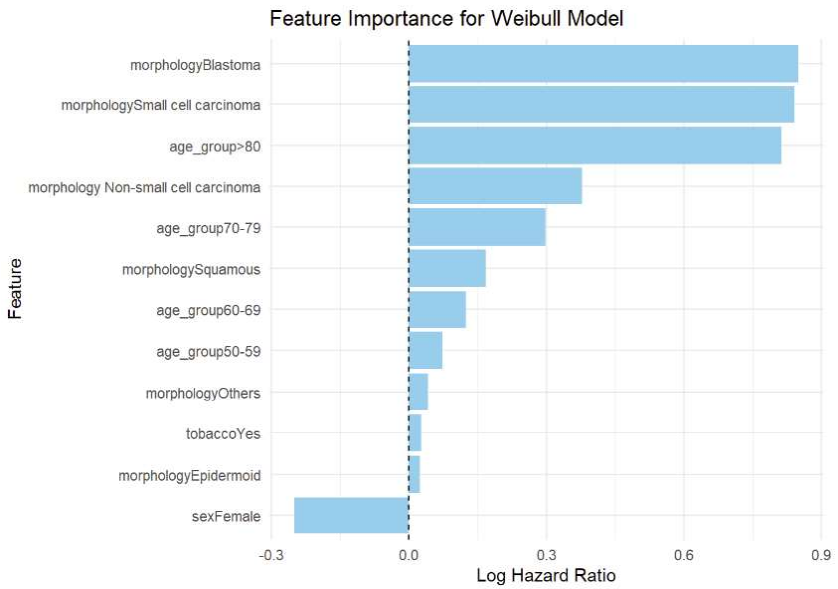


Figure 14.- Feature importance plot for Weibull model.

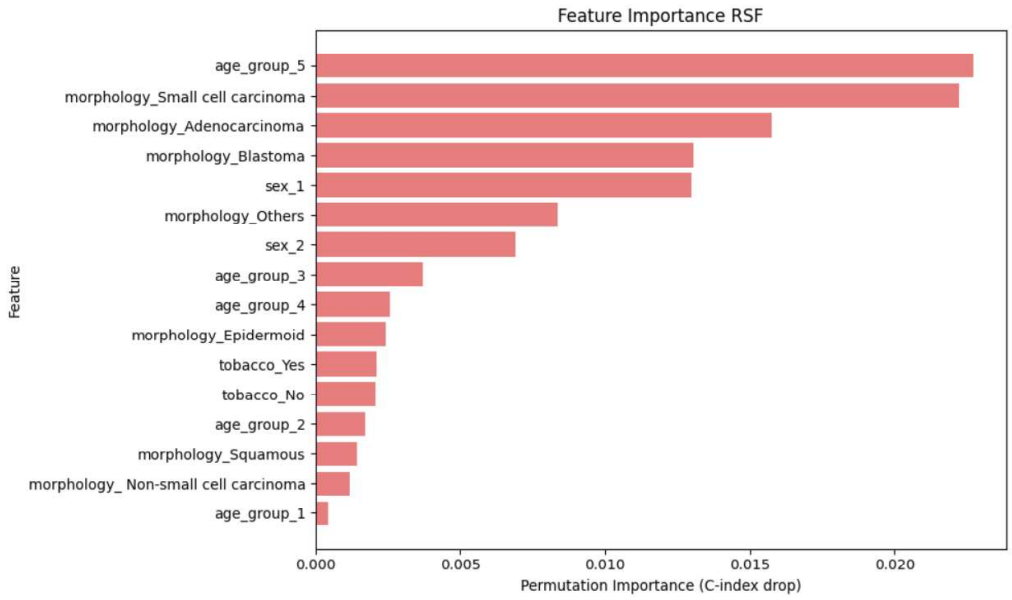


Figure 15.- Permutation-importance feature plot for RSF.

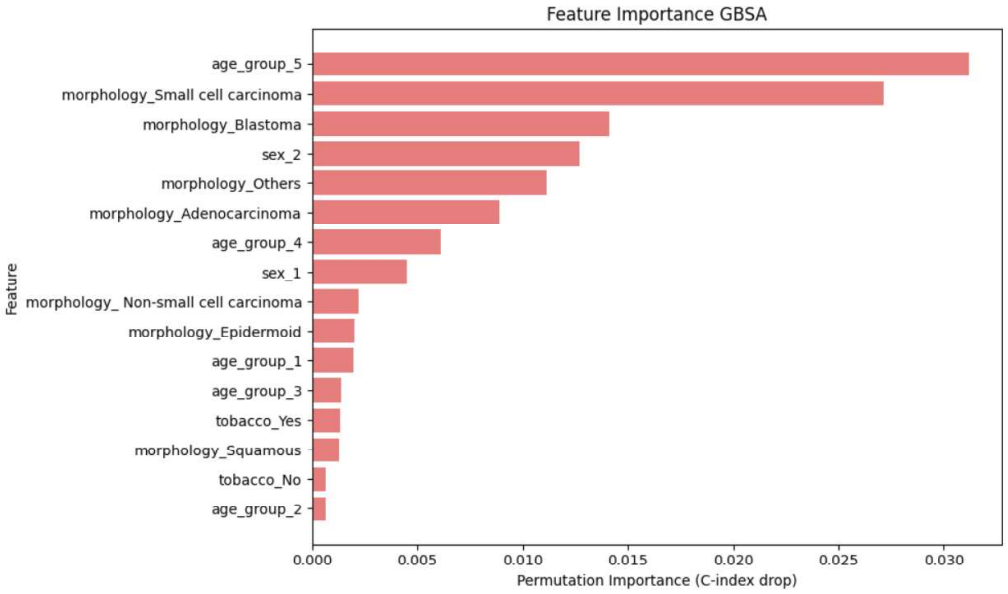


Figure 16.- Permutation-importance feature plot for GBSA.

5. Discussion and conclusions

The findings of this thesis provide some valuable insights into the factors that impact the survival of lung cancer patients. Moreover, two types of survival analysis have been performed and compared: traditional survival analysis, and ML-based survival analysis.
This section aims to analyze, discuss and present the findings of the study in a coherent and organized way. Moreover, the challenges, assumptions and limitations faced are also commented to give a direction for improvements and future work.

The predictors being evaluated were sex, age group, smoking condition and tumor morphology.

First of all, the traditional survival analysis was performed in depth.
Survival estimates were visually represented through the Kaplan-Meier curves. Then, several Cox proportional hazard models were applied in order to estimate the hazard ratios of each predictor.
The first Cox model included all variables, without any interaction term or stratification. Although the results seemed reasonable, the morphology variable violated the PH assumption, which suggested that the effect of morphology in survival was not constant over time.
A violation of the PH assumption does not mean that Cox regression can not be used, but it suggests the need to slightly modify it to account for non-proportional hazards. This is typically done by including an interaction between a predictor and time or by stratifying by a predictor.
An interaction term between sex and tobacco consumption was included in the Cox model in order to test whether the effect of tobacco on the survival depended on the sex of the patient. Unfortunately, the included interaction was not significant and, in addition, the PH assumption for morphology did not hold either.

The complete lung cancer dataset was then stratified by sex, obtaining a dataset only containing man and a dataset only containing woman. The idea was to avoid a possible bias towards male-dominated patters due to the dataset imbalance, and to observe any distinct pattern on the tobacco significance. No major differences were revealed after developing the stratified model, since the key predictors (age, morphology) remained significant while smoking continued to appear non-significant in both sexes. Nevertheless, this model was useful to compare in depth the effects of a particular predictor by sex.

By visualizing the log-log plot, it was tested that the data could follow a Weibull distribution, so this parametric model was applied towards the end of the traditional survival analysis. The Weibull model provided a solution for the first Cox model violating the PH assumption and gave interesting insights on the hazard progression.
The performance and comparison of the traditional models was evaluated using the AIC.

Afterwards, several ML-models (RSF, GBSA and DeepSurv) were explored to see if they could outperform and improve the results of purely statistical methods. The models were compared against traditional ones throughout the C-index, IBS score and feature importance plots, revealing interesting insights that will be commented during this section.

5.1. Results interpretation

The results obtained from the Cox proportional hazards models and the Weibull model provide consistent insights into the factors that influence survival in lung cancer patients.

Across all models, sex emerges as a significant variable, with females exhibiting a lower risk of mortality compared to males. In the Cox models, the HRs suggest that being female has a protective effect against mortality. The Weibull model reinforces this trend with a positive beta coefficient that implies longer survival times for females.

Age is another crucial determinant of survival, with an increasing risk observed in older patients, especially relevant in patients over 80. It is important to take into consideration that older adults tend to have a physiological decline, are less tolerant to certain treatments, or present other comorbidities that can lead to an increased mortality compared to younger patients.

Tumor morphology, on the other hand, while it is also an important predictor, it shows a substantial variation in its impact on survival. What is clear is that, across all models, blastoma and small cell carcinoma are associated with the highest risk of mortality and lowest survival. Literature confirms that SCLC has a very poor prognosis characterized by a rapid progression and limited long-term survivors [65], which is consistent with the observed results from the models and survival curves.
In contrast, epidermoid carcinoma and other histological subtypes demonstrate more modest and non-significant effects in many cases, suggesting a weaker or inconsistent associations with survival.
Moreover, it was observed in the Kaplan-Meier curves that Adenocarcinoma had the higher survival. This lung cancer subtype is often diagnosed earlier, particularly in non-smokers who benefit from targeted therapies [66], which may improve the survival of patients.

The stratified Cox model further explores this distinction between morphology subtypes, showing that some morphologies behave differently in men and women, particularly small cell carcinoma, which appears to have a higher relative risk for women.

Interestingly, the smoking status does not appear to have a significant impact on survival on any of the models, or the survival curve, even though literature demonstrated this association [67] [68]. These findings, although controversial, suggest that once lung cancer has developed and the treatment has begun, the role of the smoking in survival outcomes may be negligible, at least in the studied cohort. For this particular study, the reason behind this discrepancy may be due to the sample size, the data quality and possible confounding factors. This column was built based on self-reported data obtained from the primary care assistance, which may introduce misclassification bias. Moreover, not accounting for factors such as the duration of smoking cessation, the exposure to second hand smoking and the quantity of smoking can hide the relationship between smoking and survival. Finally, differences in treatment approaches and access to healthcare can also influence the survival outcome.

The Weibull model also suggests that the hazard is decreasing over time. This means that, while early deaths are more likely, those patients who survive initially tend to have a lower risk of death later. This result aligns with the Kaplan-Meier survival curve (*Figure 6*) , which shows a steeper drop in the beginning.
In general, while minor discrepancies exist between the Cox and Weibull models, particularly regarding certain histological subtypes, the overall patterns align well, reinforcing the robustness of these findings.
It is worth mentioning that, although the Cox model did not meet the proportional hazards assumption completely, this was not reflected in the performance results.

Regarding the comparison between traditional and ML models, it has been observed that tree-based models (RSF and GBSA) achieved the highest C-index values (0.622 and 0.620, respectively), slightly outperforming the Cox and Weibull models (0.616 and 0.615). The better performance of tree-based methods can be due to their ability to capture non-linear interactions and handle complex patterns in the data without specific assumptions about the hazard function.
DeepSurv, despite being a neural network, showed a slightly lower C-index (0.603).

For the IBS score, the Weibull model achieved the lowest IBS (0.183), followed closely by Cox (0.189), indicating that these models had better predictive accuracy and overall calibration compared to the ML-based models. DeepSurv performed the worst in terms of IBS (0.226 ± 0.009).
Although neural models offer many benefits and flexibility, in this case Deepsurv was not the best fit, probably because the dataset is not large enough to leverage its power without overfitting. Moreover, Deep Learning models generally require a larger number of features.

The feature importance plots also revealed interesting insights. It is possible to see that Cox and Weibull are very aligned and agree that the three most important predictors of survival are blastoma morphology, small cell carcinoma morphology, and people aged over 80.

ML-models RSF and GBSA also showed consistent permutation importance values, implying that variables like blastoma, small cell carcinoma and adenocarcinoma morphologies, together with older age groups and sex have the greatest impact on the model's prediction. Tobacco also appears as an irrelevant variable for the ML-models. Overall, it has been seen that tree-based models effectively highlight the key predictors of survival, therefore emphasizing the clinical significance of tumor-related variables in a simple way.

The lack of feature or permutation importance for DeepSurv highlights one of the key trade-offs of Deep Learning in survival analysis: while the model may capture complex non-linear patterns and interactions, there is a cost of interpretability. This is a relevant consideration in clinical scenarios, where understanding the reasoning behind a prediction might be as important as the prediction itself.

In general, no major differences were found between traditional and ML-based methods, suggesting that, at least for this dataset, there was no significant improvement in using ML models compared to statistical ones.

As a rule of thumb, for high-dimensional data and cases where the assumptions do not excessively hold, ML-models might outperform traditional ones and provide additional information (for example, non-linear interactions). Nevertheless, for a given survival dataset with a small number of predictors, traditional models such as Cox should be considered first, as they are more interpretable, clinically accepted and computationally faster.

5.2. Assumptions, limitations and future work

A key limitation of this thesis is the presence of right-censored data, where the event of interest is not observed for all patients during the follow-up period, reducing the effective sample size and statistical power.
Administrative censoring is also observed, which occurs when all patients are followed only up to a fixed study date. This makes that patients diagnosed later have shorter follow-up times by default, creating time-dependent bias.
Moreover, most survival models assume that the reason for censoring is not related to the risk of the event. However, in the real world, this assumption might not always hold (for example, patients may leave the hospital due to deteriorating health, which is informative).

Other strong assumptions that have been made are the ones concerning the traditional survival models. Those are the PH assumption in the case of the Cox model and the Weibull data distribution in the case of the parametric model.

While the failed PH assumption in the first Cox model propelled the research and development of the Weibull model, other options might have been as valid. For instance, it would have been interesting to modify the Cox model to account for non-proportional hazards by including an interaction between a predictor and time or by stratifying by a predictor.

Moreover, an undeniable limitation was faced due to the incompleteness of the data. The tumor staging, for example, would have been a very interesting variable to study and include in the models. Unfortunately, the hospital started tracking this variable recently, so most of the data for this column was incomplete.
Smoking data also came from an unreliable source, which lead to confusing results.

The small size of the dataset and the few predictors have also made the comparison with ML-models a bit unfair. A bigger dataset would have been useful to see if ML-based models can offer a better alternative and overcome the limitations traditional models have. With this dataset, no bigger differences have been observed.
Regarding the DL model (DeepSurv), feature work may include the use of a validation set for early stopping as well as regularization techniques such as dropout or L2 weight decay. Furthermore, adding automated hyperparameter optimization and exploring deeper or other neural architectures could enhance the predictive performance and generalizability of the DL approach.

In addition, this study is based on retrospective hospital data, which inherently carries the risk of selection bias. It is possible that certain confounding factors may not be adequately controlled and could influence the observed survival outcomes.

Future research may focus on incorporating prospective cohorts and enhance the dataset with additional clinical and molecular biomarkers that could improve the performance of the models and reveal new patterns.

Reproducibility and validation of AI models across diverse populations, regions and healthcare systems is also a burden that needs to be addressed for robustness and fairness [69].

5.3. Conclusions

This thesis highlights the strengths and limitations of both traditional and ML-based survival analysis methods when applied to lung cancer patient data.

Traditional models have been proven to be valuable tools for survival analysis due to their simplicity, interpretability and strong clinical acceptance. Their transparent assumptions allow healthcare professionals and engineers to understand and trust the model outputs.

Despite the potential and the slightly better accuracy, ML models can be less interpretable due to their complexity and intricate relationships between numerous parameters. This makes it hard to understand and explain how predictions are derived, which is a critical consideration in clinical applications [19].

The results of this work emphasize the vital importance of selecting a survival modeling approach that aligns with the characteristics of the dataset (size and dimensionality) and the goals of the analysis (for example, whether to prioritize interpretability, robustness, and reproducibility).

In an era where AI and complex ML models are often overhyped and sometimes unnecessarily applied, this thesis demonstrates that traditional statistical methods can still perform remarkably well in certain clinical datasets. While ML-based models have a lot of potential, their benefits must be weighted against their complexity, especially in healthcare domains where clarity and trust are crucial.

Comparisons aside, this thesis has also identified key predictors of survival in lung cancer patients, quantified them through HRs and visualized them with survival curves. Variables such as sex, age, and tumor morphology have been consistently proven to be determinant for patient outcomes.

Therefore, these findings contribute to a deeper understanding of the disease and reinforce the potential of data-driven approaches in oncology.

6. Ethical-social impact, sustainability, and diversity

6.1. Ethical-social impact

The use of survival analysis models in clinical scenarios implicitly comes with important ethical and social implications, especially if these models are designed to support diagnosis, prognosis, or treatment decisions. In such contexts, aspects like fairness and reliability are crucial.

Although model fairness was not directly optimized in this work, recognizing the limitations of biased datasets is critical, since it could result in unequal treatment of patients based on attributes such as gender, age or socio-economic status, among others. Basically, if certain groups are underrepresented in the data, the model may not perform as well for them.

Additionally, the type of model chosen can also influence its ethical implications. Traditional statistical models like Cox are generally more interpretable and allow healthcare professionals to understand how each variable contributes to the prediction obtained. In contrast, ML models (especially DL approaches) tend to act as “black boxes,” making it harder to explain predictions. This lack of transparency can raise concerns in healthcare, where the decisions need to be justified and trusted.

Moreover, uncontrolled external factors might have impacted the observed outcomes. For instance, people living in rural areas tend to be diagnosed later, which can produce a worse prognosis.

Future work should focus on assessing and improving model fairness, but also on understanding how healthcare system disparities can influence the data and outcomes.

On the other hand, the use of local hospital data creates models for a specific population with the potential to reflect regional health patterns. This could help to reduce health inequalities in areas that have less resources, although it is important to highlight that such models may not generalize well to other regions.

To conclude, the tools explored in this thesis can help achieve better data-driven clinical decisions based on patients risk and survival predictions. This can ultimately save hospital resources and time, leading to more available quality time for patients and less economic burden for all. Identifying risk factors and vulnerable patients can be crucial for improving patient's outcomes and move towards personalized medicine.

6.2. Sustainability

This research was conducted entirely remotely and only with digital methods, therefore minimizing the need for travel, physical resources, and laboratory-based experimentation. This reduces the overall environmental footprint during the months when the thesis was developed. Also, by using retrospective hospital data, the study eliminates the need for patient recruitment and reduces the environmental and financial costs associated with clinical trials.

Additionally, from a sustainability perspective, the use of tabular data and classical models significantly reduces computational demands. Traditional survival models such as Cox or Weibull regression require little energy to train and evaluate, which makes them an environmentally friendly choice.
ML models, while more computationally intensive, can still be efficient if carefully designed.

Finally, by improving patient's outcomes and detecting vulnerabilities early, the energy associated with hospital expenses and certain treatments can decrease.

6.3. Diversity

The present study is based on a cohort of lung cancer patients from two hospitals in Lleida, which limits the ethnical and geographical diversity of the dataset. Although the sample contains certain variations in terms of age, sex, and clinical characteristics, the results may not generalize well when it comes to broader or more heterogeneous populations.

Moreover, the small sample size of the dataset also limits the diversity inside the specific region of Lleida.

A more diverse dataset that includes data from multiple regions and healthcare systems would improve the generalizability and equity of the developed models. In addition, this would ensure that the results can be extrapolated to other regions.

7. References

1. World Health Organization. Lung cancer [Internet]. Geneva: World Health Organization; 2023 [cited 2025 Apr 6]. Available from: https://www.who.int/news-room/fact-sheets/detail/lung-cancer

2. Nicolson M. ES05.01 Lung Cancer Survival: Progress and Challenges. Journal of Thoracic Oncology. 2019 Oct; Volume 14, Issue 10, S24. Available from: https://www.jto.org/article/S1556-0864(19)30770-1/fulltext

3. Lococo F, Ghaly G, Chiappetta M, Flamini S, Evangelista J, Bria E, Stefani A, Vita E, Martino A, Boldrini L, et al. Implementation of artificial intelligence in personalized prognostic assessment of lung cancer: A narrative review. Cancers (Basel). 2024 May 10;16(10):1832. doi: 10.3390/cancers16101832. Available from: https://www.mdpi.com/2072-6694/16/10/1832

4. Herbst RS, Heymach JV, Lippman SM. Lung cancer. N Engl J Med. 2008;359(13):1367–80. doi:10.1056/NEJMra0802714. Available from: https://www.nejm.org/doi/full/10.1056/NEJMra0802714

5. Zappa C, Mousa SA. Non-small cell lung cancer: current treatment and future advances. Transl Lung Cancer Res. 2016 Jun;5(3):288–300. doi:10.21037/tlcr.2016.06.07. PMID: 27413711; PMCID: PMC4931124. Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC4931124/

6. Nooreldeen R, Bach H. Current and future development in lung cancer diagnosis. Int J Mol Sci. 2021;22(16):8661. doi:10.3390/ijms22168661. Available from: https://www.mdpi.com/1422-0067/22/16/8661

7. National Health Service (NHS). Lung cancer - Diagnosis [Internet]. London: NHS; 2022 [cited 2025 Apr 9]. Available from: https://www.nhs.uk/conditions/lung-cancer/diagnosis/

8. Mayo Clinic Staff. Lung cancer: Diagnosis and treatment. Mayo Clinic. Published May 2023. Available from: https://www.mayoclinic.org/diseases-conditions/lung-cancer/diagnosis-treatment/drc-20374627

9. American Cancer Society. How to detect lung cancer: Tests for lung cancer [Internet]. Atlanta (GA): American Cancer Society; [cited 2025 Apr 9]. Available from: https://www.cancer.org/cancer/types/lung-cancer/detection-diagnosis-staging/how-diagnosed.html

10. Cui L, Li H, Hui W, Zhang X, Mo G, Liu Z, et al. A deep learning-based framework for lung cancer survival analysis with biomarker interpretation. BMC Bioinformatics. 2020;21:112. doi:10.1186/s12859-020-3431-z.Available from: https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-020-3431-z

11. Rebasa P. Conceptos básicos del análisis de supervivencia. Cir Esp. 2005;78(4):222-230. doi:10.1016/S0009-739X(05)70923-4. Available from: https://www.elsevier.es/es-revista-cirugia-espanola-36-articulo-conceptosbasicos-del-analisis-supervivencia-13079636

12. Clark TG, Bradburn MJ, Love SB, Altman DG. Survival analysis part I: basic concepts and first analyses. Br J Cancer. 2003 Jul 21;89(2):232–8. doi:10.1038/sj.bjc.6601118. PMID: 12865907; PMCID: PMC2394262. Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC2394262/

13. Schober P, Vetter TR. Survival analysis and interpretation of time-to-event data: the tortoise and the hare. Anesth Analg. 2018 Sep;127(3):792–8. doi:10.1213/ANE.0000000000003653. PMID: 30015653; PMCID: PMC6110618. Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC6110618/

14. Spooner A, Chen E, Sowmya A, Sachdev P, Kochan NA, Trollor J, Brodaty H. A comparison of machine learning methods for survival analysis of high-dimensional clinical data for dementia prediction. Sci Rep. 2020 Nov 23;10(1):20410. doi: 10.1038/s41598-020-77220-w. Available from: https://www.nature.com/articles/s41598-020-77220-w#citeas

15. Lococo F, Ghaly G, Chiappetta M, Flamini S, Evangelista J, Bria E, Stefani A, Vita E, Martino A, Boldrini L, et al. Implementation of artificial intelligence in personalized prognostic assessment of lung cancer: a narrative review. Cancers (Basel). 2024;16(10):1832. doi:10.3390/cancers16101832. Available from: https://www.mdpi.com/2072-6694/16/10/1832

16. You HS, Gao CX, Wang HB, Luo SS, Chen SY, Dong YL, Lyu J, Tian T. Concordance of treatment recommendations for metastatic non-small-cell lung cancer between Watson for Oncology system and medical team. Cancer Manag Res. 2020 Mar 16;12:1947-1958. doi: 10.2147/CMAR.S244932. PMID: 32214852; PMCID: PMC7083631. Available from: https://pubmed.ncbi.nlm.nih.gov/32214852/

17. Huang Y, Li J, Li M, Aparasu RR. Application of machine learning in predicting survival outcomes involving real-world data: a scoping review. BMC Med Res Methodol. 2023;23(1):268. doi:10.1186/s12874-023-02078-1. Available from: https://bmcmedresmethodol.biomedcentral.com/articles/10.1186/s12874-023-02078-1#citeas

18. Wang Y, Flowers CR, Li Z, Huang X. CondiS: A conditional survival distribution-based method for censored data imputation overcoming the hurdle in machine learning-based survival analysis. J Biomed Inform. 2022 Jul;131:104117. doi:10.1016/j.jbi.2022.104117. PMID:

35690348; PMCID: PMC10099458. Available from: https://pubmed.ncbi.nlm.nih.gov/35690348/

19. Kolasseri AE, Venkataramana B. Comparative study of machine learning and statistical survival models for enhancing cervical cancer prognosis and risk factor assessment using SEER data. Sci Rep. 2024;14(1):22203. doi:10.1038/s41598-024-72790-5. Available from: https://www.nature.com/articles/s41598-024-72790-5

20. Meltzer R. What is Random Forest? [Internet]. CareerFoundry Blog. 2023 Aug 31 [cited 2025 Apr 11]. Available from: https://careerfoundry.com/en/blog/data-analytics/what-is-random-forest/

21. Ishwaran H, Kogalur UB, Blackstone EH, Lauer MS. Random survival forests. Ann Appl Stat. 2008 Sep;2(3):841–860. doi: 10.1214/08-AOAS169. Available from: https://projecteuclid.org/journals/annals-of-applied-statistics/volume-2/issue-3/Random-survival-forests/10.1214/08-AOAS169.full

22. Paksoy T. Machine learning methods for survival data [master's thesis]. İzmir: Dokuz Eylül University, Graduate School of Natural and Applied Sciences; 2021.

23. Pölsterl S. Using Random Survival Forests. In: scikit-survival 0.24.1 [Internet]. scikit-survival developers; c2015–2025 [cited 2025 Apr 11]. Available from: https://scikit-survival.readthedocs.io/en/stable/user_guide/random-survival-forest.html

24. Pölsterl S. Gradient Boosted Models. In: scikit-survival 0.24.1 [Internet]. scikit-survival developers; c2015–2025 [cited 2025 Apr 11]. Available from: https://scikit-survival.readthedocs.io/en/stable/user_guide/boosting.html

25. Katzman JL, Shaham U, Cloninger A, Bates J, Jiang T, Kluger Y. DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. BMC Med Res Methodol. 2018;18(1):24. doi:10.1186/s12874-018-0482-1. Available from: https://bmcmedresmethodol.biomedcentral.com/articles/10.1186/s12874-018-0482-1#citeas

26. Löschmann L, Smorodina D. Deep Learning for Survival Analysis [Internet]. Seminar Information Systems (WS19/20); 2020 Feb 6 [cited 2025 Apr 11]. Available from: https://humboldt-wi.github.io/blog/research/information_systems_1920/group2_survivalanalysis/#deeplearning_sa

27. Pandas: McKinney, W. (2010). *Data Structures for Statistical Computing in Python*. In Proceedings of the 9th Python in Science Conference, 51–56.

28. NumPy: Harris, C. R., et al. (2020). *Array programming with NumPy*. Nature, 585(7825), 357–362.

29. Matplotlib: Hunter, J. D. (2007). *Matplotlib: A 2D Graphics Environment*. Computing in Science & Engineering, 9(3), 90–95.

30. Seaborn: Waskom, M. L. (2021). *Seaborn: statistical data visualization*. Journal of Open Source Software, 6(60), 3021.

31. Scikit-learn: Pedregosa, F., et al. (2011). *Scikit-learn: Machine Learning in Python*. Journal of Machine Learning Research, 12, 2825–2830.

32. XGBoost: Chen, T., & Guestrin, C. (2016). *XGBoost: A Scalable Tree Boosting System*. In Proceedings of the 22nd ACM SIGKDD, 785–794.

33. scikit-survival: Pölsterl, S. (2020). *scikit-survival: A Library for Time-to-Event Analysis Built on Top of scikit-learn*. Journal of Machine Learning Research, 21(212), 1–6.

34. PyTorch: Paszke, A., et al. (2019). *PyTorch: An Imperative Style, High-Performance Deep Learning Library*. Advances in Neural Information Processing Systems, 32, 8026–8037.

35. PyCox: Kvamme, H., Borgan, Ø., & Scheel, I. (2019). *Time-to-Event Prediction with Neural Networks and Cox Regression*. Journal of Machine Learning Research, 20(129), 1–30.

36. Lifelines: Davidson-Pilon, C. (2019). *Lifelines: survival analysis in Python*. https://github.com/CamDavidsonPilon/lifelines

37. Therneau TM. A package for survival analysis in R. [Internet]. 2023 [cited 2025 Apr 18]. Available from: https://cran.r-project.org/package=survival

38. Kassambara A, Kosinski M, Biecek P. survminer: Drawing Survival Curves using 'ggplot2'. [Internet]. 2023 [cited 2025 Apr 18]. Available from: https://cran.r-project.org/package=survminer

39. Wickham H. ggplot2: Elegant Graphics for Data Analysis. 2nd ed. Springer-Verlag New York; 2016.

40. Sjoberg DD, Whiting K, Curry M, Lavery J, Larmarange J. gtsummary: Presentation-Ready Data Summary and Analytic Result Tables. [Internet]. 2023 [cited 2025 Apr 18]. Available from: https://cran.r-project.org/package=gtsummary

41. Iannone R, Cheng J, Schloerke B. gt: Easily Create Presentation-Ready Display Tables. [Internet]. 2023 [cited 2025 Apr 18]. Available from: https://cran.r-project.org/package=gt

42. Bache SM, Wickham H. magrittr: A Forward-Pipe Operator for R. [Internet]. 2022 [cited 2025 Apr 18]. Available from: https://cran.r-project.org/package=magrittr

43. Jackson C. flexsurv: Flexible Parametric Survival Models. [Internet]. 2023 [cited 2025 Apr 18]. Available from: https://cran.r-project.org/package=flexsurv

44. Fox J, Weisberg S. An R Companion to Applied Regression. 3rd ed. Thousand Oaks, CA: Sage; 2019.

45. Schroeder MS, Culhane AC, Quackenbush J, Haibe-Kains B. survcomp: Performance assessment and comparison for survival analysis. *Bioinformatics*. 2011;27(3):320–1.

46. Kuhn M. caret: Classification and Regression Training. *J Stat Softw*. 2008;28(5):1–26.

47. Carbonetto P. survex: Explainable AI for Survival Models. [Internet]. 2023 [cited 2025 Apr 18]. Available from: https://cran.r-project.org/package=survex

48. Mogensen UB, Ishwaran H, Gerds TA. Evaluating random forests for survival analysis using prediction error curves. *Stat Med*. 2012;31(11-12):968–83.

49. Gerds TA, Kattan MW, Schumacher M. Estimating the error prediction of survival models in the presence of censoring. *Biostatistics*. 2008;9(3):524–39.

50. Ministerio de Sanidad. *Códigos de causas de defunción según la CIE* [Internet]. Madrid: Ministerio de Sanidad; [cited 2025 Apr 18]. Available from: https://www.sanidad.gob.es/estadEstudios/estadisticas/estadisticas/estMinisterio/mortalidad/docs/codigos_causas_cie.pdf

51. Fritz A, Percy C, Jack A, Shanmugaratnam K, Sobin L, Parkin DM, Whelan S, editors. International classification of diseases for oncology: ICD-O. 3rd ed, 1st rev. Geneva: World Health Organization; 2013.

52. May L, Shows K, Nana-Sinkam P, Li H, Landry JW. Sex differences in lung cancer. Cancers (Basel). 2023 Jun 8;15(12):3111. doi: 10.3390/cancers15123111. PMID: 37370722; PMCID: PMC10296433. Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC10296433/

53. Brock BA, Mir H, Flenaugh EL, Oprea-Ilies G, Singh R, Singh S. Social and biological determinants in lung cancer disparity. Cancers (Basel). 2024 Jan 31;16(3):612. doi: 10.3390/cancers16030612. PMID: 38339362; PMCID: PMC10854636. Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC10854636/

54. DATAtab Team. DATAtab: Online Statistics Calculator [Internet]. Graz, Austria: DATAtab e.U.; 2025 [cited 2025 Apr 11]. Available from: https://datatab.net/tutorial/log-rank-test

55. DATAtab Team. Cox Proportional Hazards Survival Regression [Internet]. Graz, Austria: DATAtab e.U.; 2025 [cited 2025 Apr 11]. Available from: https://datatab.net/tutorial/cox-regression

56. Haile SR. Weibull AFT Regression Functions in R [Internet]. Zurich: Epidemiology, Biostatistics and Prevention Institute, University of Zurich; 2023 Sep 27 [cited 2025 Apr 11]. Available from: https://cran.r-project.org/web/packages/SurvRegCensCov/vignettes/weibull.pdf

57. Shamsudduha M. Groundwater dynamics and arsenic mobilisation in Bangladesh: a national-scale characterisation [PhD thesis]. London: University College London; 2011. Available from: https://www.researchgate.net/figure/Weibull-model-assumption-is-checked-with-a-cloglog-plot-of-log-log1-Ft-and-logt_fig33_295718674

58. Lee S. A guide to the Akaike Information Criterion for model selection. Number Analytics https://www.numberanalytics.com/blog/guide-to-akaike-information-criterion-model-selection

59. kjytay. What is Harrell's C-index? Statistical Odds & Ends [Internet]. 2019 Oct 26 [cited 2025 Apr 11]. Available from: https://statisticaloddsandends.wordpress.com/2019/10/26/what-is-harrells-c-index/

60. Park SY, Park JE, Kim H, Park SH. Review of statistical methods for evaluating the performance of survival or other time-to-event prediction models (from conventional to deep learning approaches). Korean J Radiol. 2021 Oct;22(10):1697-1707. doi: 10.3348/kjr.2021.0223. PMID: 34269532; PMCID: PMC8484151. Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC8484151/

61. Zhou H, Cheng X, Wang S, Zou Y, Wang H. *SurvMetrics: Predictive Evaluation Metrics in Survival Analysis* [R package]. Version 0.5.1. Published 2025 Jan 16. Available from: https://cran.r-project.org/web/packages/SurvMetrics/SurvMetrics.pdf

62. Cavalcante T, Ospina R, Leiva V, Cabezas X, Martin-Barreiro C. Weibull Regression and Machine Learning Survival Models: Methodology, Comparison, and Application to Biomedical Data Related to Cardiac Surgery. *Biology (Basel)*. 2023 Mar 13;12(3):442. doi:10.3390/biology12030442. PMID: 36979135; PMCID: PMC10045304. Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC10045304/#sec3-biology-12-00442

63. Pölsterl S. Evaluating survival models. In: *scikit-survival 0.24.1 User Guide* [Internet]. 2025 [cited 2025 Apr 11]. Available from: https://scikit-survival.readthedocs.io/en/stable/user_guide/evaluating-survival-models.html

64. Nahhas RW. *Introduction to regression methods for public health using R* [Internet]. 2025 [cited 2025 Apr 11]. Available from: https://www.bookdown.org/rwnahhas/RMPH/

65. van Meerbeeck, J. P., Fennell, D. A., & De Ruysscher, D. K. (2011). Small-cell lung cancer. *The Lancet, 378*(9804), 1741–1755. https://doi.org/10.1016/S0140-6736(11)60165-7

66. Mok TS, Wu YL, Thongprasert S, Yang CH, Chu DT, Saijo N, et al. Gefitinib or carboplatin-paclitaxel in pulmonary adenocarcinoma. *N Engl J Med*. 2009 Sep 3;361(10):947–57. doi:10.1056/NEJMoa0810699. Available from: https://pubmed.ncbi.nlm.nih.gov/19692680/

67. Tao L, Wang R, Gao YT, Yuan JM. Impact of postdiagnosis smoking on long-term survival of cancer patients: the Shanghai cohort study. Cancer Epidemiol Biomarkers Prev. 2013 Dec;22(12):2404-11. doi: 10.1158/1055-9965.EPI-13-0805-T. PMID: 24319070; PMCID: PMC3919701. Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC3919701/

68. Yu J, Zhang Y, Liu Z, He Y, Pei Y, Zhang R, Peng X, Fang F. Association of smoking with the survival of patients with brain metastasis of lung cancer. Front Neurol. 2023 Mar 13;14:1036387. doi: 10.3389/fneur.2023.1036387. Available from: https://www.frontiersin.org/journals/neurology/articles/10.3389/fneur.2023.1036387/full

69. Lococo F, Ghaly G, Chiappetta M, Flamini S, Evangelista J, Bria E, et al. Implementation of artificial intelligence in personalized prognostic assessment of lung cancer: A narrative review. *Cancers (Basel)*. 2024 May 10;16(10):1832. doi:10.3390/cancers16101832. Available from: https://pubmed.ncbi.nlm.nih.gov/38791910/

Annex

A1. Code availability

The code used in this Master Thesis can be found in the following GitHub repository:

https://github.com/IhonaMaria/Predicting-Lung-Cancer-Survival-MT

Note that there are three available files:

1. Jupyter notebook with the preprocessing steps followed before the Survival Analysis.
2. Rmarkdown file containing the code for the traditional Survival Analysis.
3. Jupyter notebook with the Python code for the ML-based Survival Analysis.

A2. Proportional Hazard assumption test

	Chi-squared	Degrees of freedom	p-value
sex	0.211	1	0.64629
age_group	2.409	4	0.66108
tobacco	3.369	1	0.06645
morphology	24.305	6	0.00046
GLOBAL	30.466	12	0.00238

Table 15.- Schoenfeld residuals test results for the general Cox model. The Chi-squared and Degrees of freedom are used to compute the p-values.

	Chi-squared	Degrees of freedom	p-value
age_group	2.438	4	0.65580
morphology	24.220	6	0.00048
sex	0.215	1	0.64309
tobacco	3.256	1	0.07117

sex:tobacco	2.867	1	0.09042
GLOBAL	31.367	13	0.00298

Table 16.- Schoenfeld residuals test results for the Cox model with interaction term sex*tobacco. The Chi-squared and Degrees of freedom are used to compute the p-values.

	Cox model Men			Cox model Women		
	Chi-squared	Degrees of freedom	p-value	Chi-squared	Degrees of freedom	p-value
age_group	3.583	4	0.47	1.88	4	0.76
tobacco	0.934	1	0.33	2.66	1	0.10
morphology	9.340	6	0.16	9.85	6	0.13
GLOBAL	13.849	11	0.24	12.27	11	0.34

Table 17.- Schoenfeld residuals test results for the sex stratified Cox model (Men and Women). The Chi-squared and Degrees of freedom are used to compute the p-values.

A3. Hyperparameter ranges of the ML models

Hyperparameter	Description	Range Explored	Selected Value
n_estimators	Number of trees in the forest	[50, 100, 150, 200]	50
min_samples_split	Minimum number of samples to split a node	[2, 5, 10, 15]	10
min_samples_leaf	Minimum samples required at a leaf node	[1, 5, 10, 15, 20]	15
max_features	Features considered at each split	["sqrt", "log2", 0.5, 0.75, 1.0]	sqrt

Table 18.- Hyperparameter ranges and selected value for the RSF model.

Hyperparameter	Description	Range Explored	Selected Value
n_estimators	Number of boosting stages (trees)	[50, 100, 150, 200]	150
learning_rate	Step size used in updating predictions	[0.01, 0.05, 0.1, 0.2]	0.1
max_depth	Maximum depth of individual trees	[2, 3, 4, 5]	2
min_samples_split	Minimum number of samples to split an internal node	[2, 5, 10]	10
min_samples_leaf	Minimum number of samples required at a leaf node	[1, 5, 10]	5
subsample	Fraction of samples used for fitting each base learner	[0.6, 0.8, 1.0]	0.8

Table 19.- Hyperparameter ranges and selected value for the GBSA model.

A4. Project organization

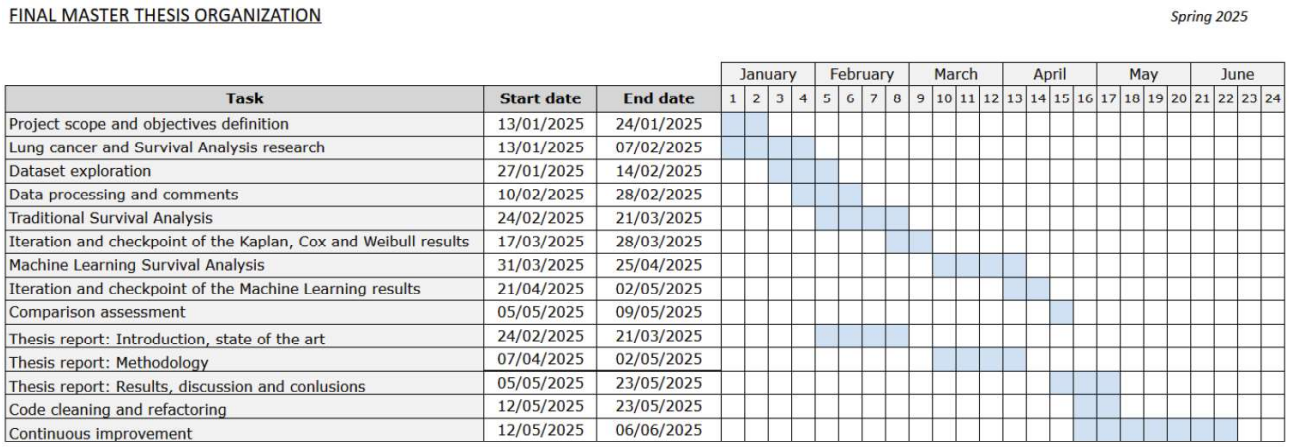


Figure 17.- Gantt diagram chart for project organization planning.