



## Machine learning survival models for relapse prediction in a early stage lung cancer patient

Title	Machine learning survival models for relapse prediction in a early stage lung cancer patient
Author(s)	Timilsina, Mohan;Buosi, Samuele;Janik, Adrianna;Minervini, Pasquale;Costabello, Luca;Torrente, Maria;Provencio, Mariano;Calvo, Virginia;Camps, Carlos;Ortega, Ana L.;Garcia Campelo, M.Rosario;del Barco, Edel;Bosch-Barrera, Joaquim;Nováček, Vít
Publication Date	2023-08-02
Publisher	IEEE
Repository DOI	<a href="https://doi.org/10.1109/IJCNN54540.2023.10191078">10.1109/IJCNN54540.2023.10191078</a>

# Machine Learning Survival Models for Relapse Prediction in a Early Stage Lung Cancer Patient

Mohan Timilsina  
Data Science Institute  
University of Galway  
Galway, Ireland

mohan.timilsina@insight-centre.org

Samuele Buosi  
Data Science Institute  
University of Galway  
Galway, Ireland

samuele.buosi@insight-centre.org

Adrianna Janik  
Accenture Lab  
Dublin, Ireland

adrianna.janik@accenture.com

Pasquale Minervini  
University College London  
London, United Kingdom

p.minervini@ucl.ac.uk

Luca Costabello  
Accenture Lab  
Dublin, Ireland

luca.costabello@accenture.com

Maria Torrente  
Department of Medical Oncology  
Puerta de Hierro-Majadahonda University Hospital  
Madrid, Spain

maria.torrente@salud.madrid.org

Mariano Provencio  
Department of Medical Oncology  
Puerta de Hierro-Majadahonda University Hospital  
Madrid, Spain

mariano.provencio@salud.madrid.org

Virginia Calvo  
Department of Medical Oncology  
Puerta de Hierro-Majadahonda University Hospital  
Madrid, Spain

vircalvo@hotmail.com

Carlos Camps  
Hospital General de Valencia  
Valencia, Spain

camps\_car@gva.es

Ana L. Ortega  
Hospital Universitario de Jaén  
Jaen, Spain

analauraortega@gmail.com

Bartomeu Massutí  
Hospital General Universitario de Alicante  
Alicante, Spain

bmassutis@seom.org

M.Rosario Garcia Campelo  
Complejo Hospitalario Universitario A Coruña  
A Coruna, Spain

MA.Rosario.Garcia.Campelo@sergas.es

Edel del Barco  
Hospital Universitario de Salamanca  
Salamanca, Spain

delbarco@usal.es

Joaquim Bosch-Barrera  
Institut Català d'Oncologia, Hospital Universitari Dr. Josep Trueta  
Girona, Spain

jbosch@iconcologia.net

Vit Novacek  
Faculty of Informatics  
Masaryk University Brno  
Brno, Czech Republic

novacek@fi.muni.cz

**Abstract**—Lung cancer is one of the leading health complications causing high mortality worldwide. The relapsing behavior of medically treated early-stage lung cancer makes this disease even more complicated. Thus predicting such relapse using a data-centric approach provides a complementary perspective for clinicians to understand the disease. In this preliminary work, we explored off-the-shelf survival models to predict the relapse of early-stage lung cancer patients. We analyzed the survival models on a cohort of 1348 early-stage non-small cell lung cancer (NSCLC) patients in different timestamps. Using the prediction explanation model SHAP (SHapley Additive exPlanations), we further explained the best-performing survival model's predictions. Our explainable predictive model is a potential tool for oncologists that address an unmet clinical need for post-treatment patient stratification based on the relapse hazard.

**Index Terms**—survival, time, event, prediction, cancer, explanation

## I. INTRODUCTION

When diagnosed with lung cancer, most patients may ask about their prognosis: “how long will I be alive” or “what is the success rate of possible treatment alternatives.” Many clinicians provide patients with statistics on cancer survival

based only on the site and stage of the tumor. Commonly used statistics include the 5-year survival rate and median survival time, e.g., a doctor can tell a specific patient with early-stage lung cancer that s/he has a 50% 5-year survival rate [1].

In the United States, an estimated 158,800 Americans died from lung cancer-related complications in 2016, approximately 27% of all cancer-related deaths [2]. Although there has been progress in cancer treatment over the last decade, the 5-year survival rate is still around 50% for surgically resected non-small cell lung cancer (NSCLC). However, even for stage I patients, 20% showed relapse within five years [1]. Early-stage NSCLC (stages I-II) patients are typically treated with complete surgical tumor resection. However, even after the entire resection of the tumor, 30–55% of patients will develop disease recurrence<sup>1</sup> To predict tumor relapse, clinicians and machine learning researchers have come up with two different use cases as shown in Figure 1.

The first one is “Does relapse happen?” and second one is “When will relapse happen?”. The models that are applied to

<sup>1</sup>Note that “recurrence” or “relapse” are used interchangeably in the paper.

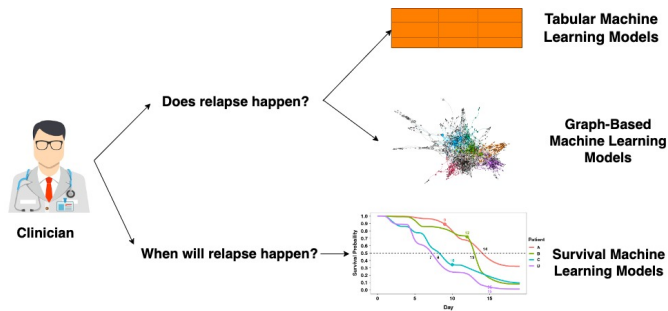


Fig. 1. Data driven approaches to identify the tumor relapse.

the first question are typically classification models trained on tabular data where there is a list of features, and the outcome variable is a binary relapse event [3], [4]. Similarly, the same problem can also be solved using a graph-based approach where data is modeled as a graph and the task cast as “link prediction” [5]. Models that are applied to the second question are called survival models, where data is also in the tabular format, but the outcome variable is time and events [6]. Here, we focus on the second question because predicting the time to the occurrence of clinical events is critical for disease prognosis and therapeutic decision [7]. Another important reason is that the classification models are not able to include censored<sup>2</sup> observations that could not be fully specified before the specified time because the outcome of these observations was unknown [8]. There are three types of censoring (i) right-censoring, (ii) left-censoring, and (iii) interval-censoring. We focus on right-censoring because it is a prevalent type of censoring and frequently occurs in cancer medical research for incomplete observation [9].

Survival models can take censoring into account and incorporate this uncertainty: instead of predicting the event, they predict the probability that an event happens at a particular time. In the context of survival analysis, statistical predictive models are typically used. One such classical method is called Nomogram. Nomogram is the graphical representation of the predictive model that can be used for the prognostication of clinical events such as relapse, disease-specific survival, or overall survival for a given patient [10]. By using a concise chart of an outcome-risk predictive model, a nomogram derives the risk probability of a specific event, such as lung cancer-specific survival (LCSS) [11]. For improving the accuracy of lung cancer survival estimations, Cox proportional hazard models have gained popularity as a way of predicting outcomes. The Cox proportional hazards (Cox) model is traditionally used to predict the clinical outcomes or hazard functions corresponding to specific time units. Cox is a popular survival model because of its strong theoretical and statistical foundation. Moreover, these models work under the linearity assumptions. However, real-world clinical characteristics might have non-linear patterns in the data. For example, both

<sup>2</sup>Censored subjects are those who have not experienced the event of interest within the observation window.

high and low blood sugar levels increase the mortality of the patients [12]. Therefore, there is a need for better solutions that focus on non-linear variables [13].

For this task, we have used a cohort of 1348 early-stage NSCLC patients from the CLARIFY project — a European project focused on monitoring health status and quality of life after the cancer treatment<sup>3</sup>. Using these real-world clinical data, we explore various survival models ranging from simple classical linear models or flexible ensemble-based models to advanced deep survival models on predicting relapse at different timestamps using standard evaluation metrics.

## II. RELATED WORK

In this section, we briefly survey the classic survival modeling method, ensemble-based survival models, and deep neural network-based survival analysis method.

**Survival Models:** In survival data, most of the data analyses have been conducted using parametric and semiparametric regression models, especially the Cox Proportional Hazards (CPH) model [14]. Survival analysis methods such as the Kaplan–Meier, Nelson–Aalen estimators, and the proportional hazards models are the most common approach for predicting prognosis, relapse, and other clinical outcomes of lung cancer patients [3]. However, the problem with CPH is that it assumes the hazard of each predictor does not change over time. The violations of such assumption lead to erroneous scientific findings [15]. It is because if the hazard of variable increases or decreases over time, the usual CPH model will ignore these time-dependent changes [16].

**Non-Linear Survival Models:** Survival models based on support vector machines (SVM) can automatically incorporate non-linearities using non-additive kernels, and interactions are automatically incorporated [17]. SVM-based models do not assume a true underlying function for which the parameters need to be estimated. In the same vein, different variants of survival models exist, also known as ensemble survival models. Ensemble models combine the predictions of multiple models to produce more accurate, robust, and reliable predictions than single learner models at the cost of interpretability [18]. The trendy ensemble survival model variant is random survival forest [19], and gradient boosting [20]. A gradient boosting model is identical to a Random Survival Forest because it depends on multiple base learners to produce an overall prediction but differs in how those are combined. While a Random Survival Forest fits a set of Survival Trees separately and averages their predictions, a gradient-boosted model is established serially in a greedy stage-wise fashion. Random Survival Forest model is considered adaptive to data and is regarded as better than traditional survival analysis methods [21]. Compared to other tree ensembles such as bagging, random subspace, and rotation forest, random forest is the most efficient algorithm in model training [22].

From a neural network perspective, researchers have applied three main types of neural networks to the problem of non-

<sup>3</sup><https://www.clarify2020.eu/>

TABLE I  
QUALITATIVE COMPARISON BETWEEN DIFFERENT SURVIVAL  
MODELS. ×: ABSENCE; ✓: PRESENCE

	CPH	SSVM	Ensemble-based Survival Model	Deep Learning-based Survival Model
Support non-linearity	×	✓	✓	✓
Interpretability by design	✓	×	×	×
No parameters to tune	✓	×	×	×
Extensive feature engineering	✓	×	×	×

linear survival analysis. These include variants of: (i) classification methods [23], (ii) time-encoded methods [24], (iii) and risk-predicting methods [25]. This third type is a feed-forward neural network (NN) that estimates an individual’s risk of failure. Faraggi-Simon’s network is seen as a non-linear extension of the Cox proportional hazards model. However, these models have improved on the CPH model by relaxing the specific functional relationship between predictors and the hazard function in the standard CPH model while maintaining the other central assumption– that the hazard rate is constant over time. Therefore, it limits the potential capacity of deep neural networks to learn complex representations of risk and, in particular, to capture the time-dependent influence of predictors on survival [26]. To address this problem recently, deep survival models [27] have been applied to CPH models to solve the problem of non-linear survival analysis. Often it is laborious to choose the perfect survival model because each model possesses its advantages and disadvantages, which requires comprehensive expertise of each model. Recently several deep neural network-based survival models have been developed. However, deep learning survival models [27] are computationally expensive to training and validating [11]. Another caveat of these models is the process of predictions. Deep learning networks function like black boxes, making it difficult to determine how the network makes decisions and interprets the results.

Table I shows the qualitative comparison between the state-of-the-art survival models. CPH models are interpretable by design, and do not require parameters to tune. However, CPH cannot support non-linear relationships. While non-linear survival methods, such as neural networks and survival forests, can inherently model these high-level interaction terms but needs to do hyperparameter tuning.

### III. METHODS

**Data Format:** The survival analysis data-points has 3 major elements:  $(X_i, E_i, T_i) \forall_i$ ,

- $X_i$  is a p-dimensional feature vector.
- $E_i$  is the event such that  $E_i = 1$ , for the event that has happened else  $E_i = 0$ .  $E$  is the relapse in our case.
- $T_i = \min(t_i, c_i)$  is the observed time, with  $t$  is the actual time of the event and  $c_i$  is the time of censoring.
- $i$  is the individual or patient or event of interest in the datasets.

**Survival Function:** The survival function  $S(t)$  indicates probability of the event which has not occurred by time  $t$ :

$$S(t) = P(T \geq t) = 1 - F(t) = \int_t^{\infty} f(x) dx \quad (1)$$

$T$  denote a continuous non-negative random variable corresponding to a patient’s survival time with the corresponding probability density function  $f(t)$  and the cumulative distribution function being  $F(t)$  at a given time  $t$ .

#### Models:

- **Cox proportional hazard (CoxPH):** It is a semi-parametric model that focuses on modeling the hazard function by assuming that its time component and the feature component are proportional. This model is considered a baseline and often appealing because its coefficients can be interpreted in terms of the hazard ratio, which often provides valuable insight.
- **Survival Support Vector Machine (SSVM):** It can account for complex, non-linear relationships between features and survival using kernel trick. A kernel function implicitly maps the input features into high-dimensional feature spaces where a hyperplane can state survival. It makes SSVM extremely versatile and applicable to a wide range of data.
- **Deep Survival Model (DeepSurv):** It is a multilayer perceptron model which predicts the risk of an event. The output of the model is the single node, which estimates the risk function parameterized by the network. The core technical details of the DeepSurv are available in [27].
- **Gradient Boosting Survival Model (GBM):** Gradient Boosting does not attribute to one unique model but an adaptable scheme to optimize many loss functions. It follows convention by combining the predictions of multiple base learners to obtain a robust model. The base learners are often very straightforward models that are marginally superior to random guessing, which is why they are also referred to as weak learners. The predictions are aggregated in an additive manner, where the summation of each base model enhances (or “boosts”) the overall model.
- **Random Forest Survival Model (RFSM):** A Random Survival Forest assures that individual trees are de-correlated by i) constructing each tree on a distinctive bootstrap sample of the initial training data and ii) at each node, only assess the split criterion for a randomly selected subset of features and thresholds. Predictions are formed by adding predictions of individual trees in the ensemble.

**Evaluations:** The model is evaluated using two different performance metric.

- **Concordance Index or C-index:** It is a generalization of the area under the ROC curve (AUC) that can account for censored data. It represents the global assessment of the model discrimination power: this is the model’s ability to

TABLE II  
HYPERPARAMETERS SEARCH SPACE USED FOR TUNING THE SURVIVAL MODELS.

Model	Hyperparameters and Search Space	scoring function
CPH	no parameter required	-
SSVM	C: [1e-3, 1e-2, 1e-1, 1, 10] gamma: [1e-5, 1e-3, 1, 10]	Concordance Index
DeepSurv	num_nodes: [32],[32, 32],[64, 64],[128, 128] dropout: [0.1, 0.2, 0.3, 0.4] batch_size: [32, 64, 128, 256] epoch: [10, 20, 40, 60, 120] learning_rate: LogSpace[1e-3, 1e-2, 1] optimizer: [Adam, SGD] activation: [RELU, SELU]	Concordance Index
RFSM	Estimators: [10, 50, 100, 200] max Depth: [3, 4, 5, 6, 7, 8, 9]	Concordance Index
GBM	Estimators: [10, 50, 100, 200] max Depth: [3, 4, 5, 6, 7, 8, 9]	Concordance Index

correctly provide a reliable ranking of the survival times based on the individual risk scores. It is computed as:

$$C\text{-index} = \frac{\# \text{concordant pairs}}{\# \text{concordant pairs} + \# \text{discordant pairs}} \quad (2)$$

C-index = 1 means the best model prediction, and C-index = 0.5 means a random prediction.

- **Integrated Brier Score (IBS):** It is the aggregation of the Brier score (BS), which is used to evaluate the accuracy of a predicted survival function at a given time  $t$ . IBS provides an overall calculation of the model performance at all available times. It is computed as:

$$IBS(t_{max}) = \frac{1}{t_{max}} \int_0^{t_{max}} BS(t) dt \quad (3)$$

The BS lies between 0 and 1, 0 being the best possible value.

The survival models are trained using ten-fold cross-validation. The model parameters are identified using grid search in the training sets. For each survival model, we execute a set of 10 trials to find the optimal hyperparameter configuration, where each trial corresponds to a single hyperparameter configuration chosen from a predefined search space. We then choose hyperparameters corresponding to the best-performing configuration as the model's best hyperparameters for each of the inspected models. The hyperparameters and scoring function used by different survival model is shown in Table II.

#### IV. EXPERIMENTS

Here we answer the following questions:

- *Q1- Censoring Windows:* How do the censoring windows affect the performance of survival models for predicting tumor relapse?
- *Q2- Explainability:* How does the best-performing survival model yield its prediction?

**Datasets:** This study uses electronic health records (EHRs) of lung cancer patients collected and stored by the Spanish Lung Cancer Group (SLCG). The cohort contains 1348 early-stage (stage I or II) NSCLC patients, where 491 (36.4%) of

TABLE III

A COHORT OF THE EARLY STAGE PATIENTS WHICH WERE UTILIZED TO CREATE THE DATASETS. COMPARE WITH LUNG CANCER IN SPAIN: INFORMATION FROM THE THORACIC TUMORS REGISTRY (TTR STUDY) FOR A SET OF NSCLC PATIENTS CHARACTERISTICS (PROVENCIO ET AL. 2019 [28]) AND [3], [5] FOR A COHORT OF ALL NSCLC PATIENTS IN THE DATASET.

Features		Relapse	Survival	Total	
		491 (36.4%)	857 (63.6%)	1348 (100%)	
Age	Mean (Range)	65.9 (33-88)	65.7 (31 - 118)	65.7 (31-118)	
Gender	Male	384 (38.0%)	626(62.0%)	1010(74.9%)	
	Female	107 (31.7%)	231 (68.3%)	338 (25.1%)	
Smoking	Current/Previous	436 (37.1%)	739 (62.9%)	1175 (87.2%)	
	Non Smoker	55 (31.8%)	118 (68.2%)	173 (12.8%)	
Cancer Stage	I	1 (100.0%)	0 (0.0%)	1 (0.0742%)	
	IA	73 (28.7%)	181 (71.3%)	254 (18.8%)	
	IA1	8 (32.0%)	17 (68.0%)	25 (1.85%)	
	IA3	8 (17.4%)	38 (82.6%)	46 (3.41%)	
	IA2	9 (13.6%)	57 (86.4%)	66 (4.9%)	
	IIA	107 (45.9%)	126 (54.1%)	233 (17.3%)	
	IB	154 (39.0%)	241 (61.0%)	395 (29.3%)	
	IIB	131 (39.9%)	197 (60.1%)	328 (24.3%)	
	T stage	T2a	202 (40.6%)	296 (59.4%)	498 (36.9%)
		T1a	54 (32.3%)	113 (67.7%)	167 (12.4%)
T2b		73 (42.0%)	101 (58.0%)	174 (12.9%)	
T1b		55 (25.7%)	159 (74.3%)	214 (15.9%)	
T3		81 (39.5%)	124 (60.5%)	205 (15.2%)	
T1c		14 (20.0%)	56 (80.0%)	70 (5.19%)	
Tx		12 (60.0%)	8 (40.0%)	20 (1.48%)	
N stage		N0	387 (34.4%)	738 (65.6%)	1125 (83.5%)
	N1	104 (46.6%)	119 (53.4%)	223 (16.5%)	
M stage	M0	491 (36.4%)	857 (63.6%)	1348 (100.0%)	
Tumor size	Mean (Range)	36.0 (0.8-110.0)	33.2 (1.5-110.0)	34.6 (0.8-110.0)	
	0	243 (29.2%)	588 (70.8%)	831 (61.6%)	
	1	216 (46.7%)	247 (53.3%)	463 (34.3%)	
	2	26 (60.5%)	17 (39.5%)	43 (3.19%)	
	3	5 (71.4%)	2 (28.6%)	7 (0.519%)	
Ecog status	4	1(100%)	0(0.0%)	1 (0.0742%)	
	Non specified	137 (39.3%)	212 (60.7%)	349 (25.9%)	
	Moderately	55 (26.8%)	150 (73.2%)	205 (15.2%)	
	Poorly	44 (42.7%)	59 (57.3%)	103 (7.64%)	
	Undifferentiated	3 (50.0%)	3 (50.0%)	6 (0.445%)	
Surgery	Well	55 (43.0%)	73 (57.0%)	128 (9.5%)	
	434 (32.2%)	835 (61.9%)	1269 (94.1%)		
	Chemotherapy	353 (26.2%)	305 (22.6%)	658 (48.8%)	
Radiotherapy	38 (2.82%)	21 (1.56%)	59 (4.38%)		

these patients had tumor relapse after successful treatment. The patient's average age in the dataset is 65.9 for those who relapsed. For those with disease-free survival, it is 65.7. The patient data used in this study was collected under the provisions of the Law 14/2007 on Biomedical Research at all times along with the confidentiality of the data of patients according to the requirements of the law by EU General Data Protection Regulation 2016/679 (GDPR).

In Table III, we provide the summary of the patient cohort. When training various machine learning models, we distinguish between the following types of features extracted from the EHRs:

- **General feature** contains the generic information about the patient such as age, sex, race, smoking habit, and family cancer history of the patient.
- **Diagnosis feature** details the information on the tumor classification, histology at the time of diagnosis, along with the symptoms and the patient history.
- **Treatment feature** includes the details of any chemotherapy, radiotherapy, or surgical procedures the patient underwent during their treatment.
- **All feature** includes the union of **General**, **Diagnosis** and **Treatment** features.

**Implementation:** All the algorithms used in the experiments are implemented using dedicated Python packages. For the

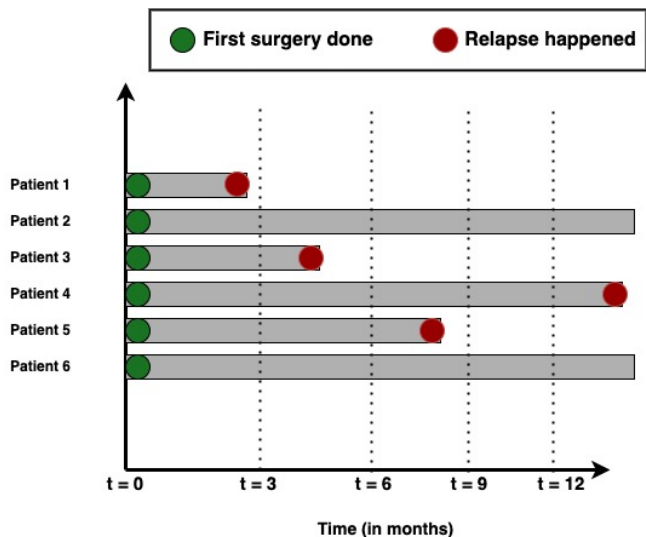


Fig. 2. Relapse data with different censoring windows. Surgeries are done at  $t=0$  for the patients. For e.g. at  $t=3$  months, outcome for Patient 5 is censored but relapse will be observed for  $t=9$  months.

linear and ensemble-based models, scikit-survival<sup>4</sup> is used and for the deep learning based survival model PySurvival<sup>5</sup> is used.

### Q1 Censoring Windows

Figure 2 is a pictorial representation of our patient relapse data. Observations are synchronized at  $t = 0$ , which is the time at which the patients receive the first surgery. If the event of the tumor relapse is not within a chosen time interval, e.g.,  $t = 3$  months, this would be a censored data point.

We consider four different durations for the censoring window, namely 3, 6, 9, and 12 months after the patient’s first surgery. Table IV provides the result of our experiment. The IBS score for SSVM is NA because the idea behind formulating the survival problem using SVM in clinical application is based on a ranking problem. Therefore, the algorithm is only able to define risk groups and not the prediction of the survival time. However, other models can compute survival functions; therefore, the IBS score is reported.

For C-index, the higher values and for IBS, the lower value is considered the best performing model. The last column presents the relative improvement of the best-performing model (GBM) over the baseline model (CPH). We found that GBM achieved the best result in C-index and IBS in all the censoring windows. The highest C-index of 0.66 is observed for GBM in  $t = 12$  months, and for the lowest IBS, there is a tie between CPH, RFSM, and GBM for  $t = 3$  months. Whereas in all the censoring windows, GBM has the lowest IBS. The highest relative improvement of 16.25% in the C-index is observed for  $t = 6$  months windows. Therefore, in this data GBM is superior to other survival models adapted for relapse prediction for different censoring windows. One

<sup>4</sup>[https://scikit-survival.readthedocs.io/en/stable/user\\_guide/00-introduction.html](https://scikit-survival.readthedocs.io/en/stable/user_guide/00-introduction.html)

<sup>5</sup><https://square.github.io/pysurvival/intro.html>

crucial observation is that DeepSurv, a deep learning-based survival model, did not outperform the ensemble-based model in our dataset. One reason for that might be that deep learning is perfect for unstructured data, and its strength lies in the ability to learn the latent feature extraction (essentially, that is the opposite of tabular data). However, in our dataset, we already have manually extracted clinical features therefore we did not see any significant improvement using the deep survival model, even though it still performed better than the linear CPH model.

We applied the Wilcoxon signed rank test, a non-parametric test that does not rely on assumptions that data belongs to any particular distribution and is used for comparison between different machine learning models [29]. Thus we compute the p-value for the Wilcoxon signed-rank test between GBM and the other compared survival models to check the significance of GBM against the competitors for the C-index obtained from 10-fold cross-validation.

In Table V, we observe a significant difference in prediction performed by GBM with respect to CPH, SSVM, and DeepSurv in month 12. For month 9, GBM has a significant difference in prediction with CPH and SSVM. Similarly, in month 6, GBM significantly differs in prediction with CPH. However, in month 3 there is no significant difference with respect to predictions performed by CPH, SSVM, DeepSurv, and RFSM. That means we have no evidence that GBM is superior to CPH, SSVM, DeepSurv, and RFSM for the month 3 prediction.

### Q2 Explainability

Given the low variability of the evaluation metrics for C-index and IBS when across 10 folds of the data, we chose a random test data partition (75% training and 25% testing) to compute SHAP (SHapley Additive exPlanations(SHAP))<sup>6</sup> values for the best-performing survival model. SHAP values provide us the flexibility to visualize the global and local feature importance for the survival models. SHAP is based on a game theoretic approach to explaining the output of any machine learning model [9]. In our case, for the censoring window  $t = 12$ , GBM performs the best, as shown in Table IV. Therefore we are going to use this model for the global and local explanation.

The bar plot in Figure 3 shows those features in order of importance by summarising all the individual predictions to provide an overall picture of their impact on a global scale for 12 months censoring window. It is based on the mean absolute value of the SHAP values of each feature. The barplot shows that comorbidity (diagnostic feature), and age ( general feature), are ranked first and second, which may indicate a high risk of relapse. We then search oncological literature to see if we can find supporting evidence for this high feature ranking. We found in the study by Ludbrook et al. [30] that age and comorbidity are associated with high tumor recurrences and are less likely to be treated with

<sup>6</sup><https://github.com/slundberg/shap>

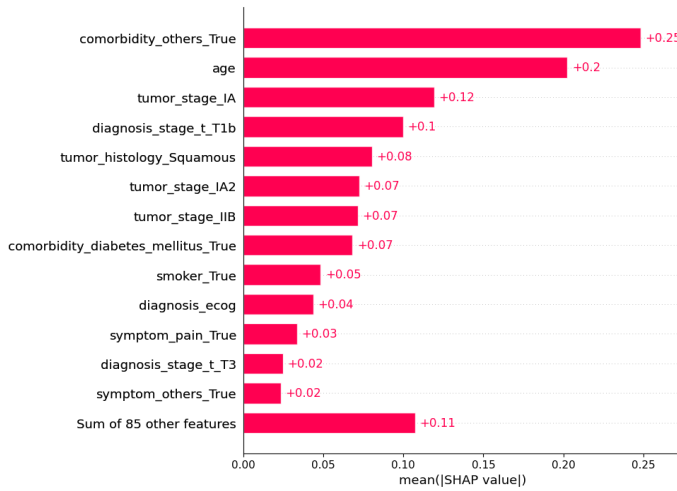


Fig. 3. Bar Plot of the important features for  $t=12$  months.

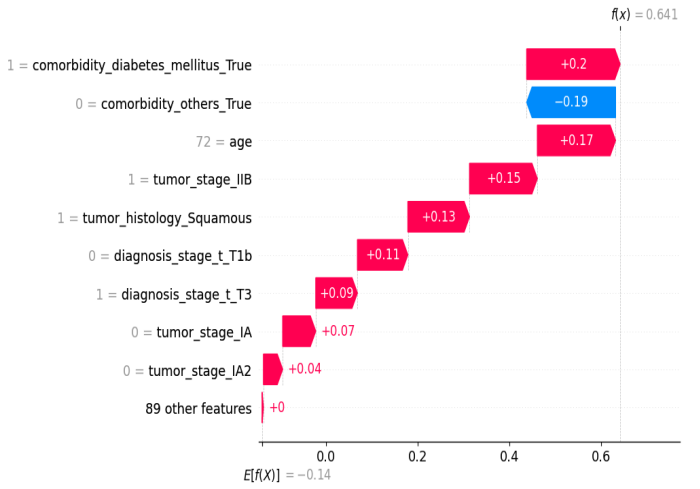


Fig. 4. SHAP explanation with a waterfall plot of features contributing to the prediction, in red increasing the prediction score, in blue decreasing for the GBM model for  $t = 12$  months.

TABLE IV

PREDICTIVE PERFORMANCE OF DIFFERENT SURVIVAL MODELS FROM CENSORING WINDOWS OF 3,6,9 AND 12 MONTHS FOR PREDICTING RELAPSE. THE FIGURE AFTER  $\pm$  IS THE STANDARD DEVIATION OBTAINED FROM THE 10 FOLD CROSS VALIDATION OF C-INDEX AND IBS SCORE.  $\uparrow$  IS FOR IMPROVEMENT IN C-INDEX AND  $\downarrow$  IS FOR DECREMENT IN IBS.

Censoring Windows	Evaluation Metric	CPH	SSVM	DeepSurv	RFSM	GBM	Relative Improvement
$t = 3$	C-index	0.632 $\pm$ 0.132	0.607 $\pm$ 0.090	0.612 $\pm$ 0.073	0.598 $\pm$ 0.159	<b>0.642<math>\pm</math> 0.165</b> $\uparrow$	1.58%
	IBS	<b>0.015<math>\pm</math>0.006</b> $\downarrow$	NA	0.018 $\pm$ 0.007	<b>0.015<math>\pm</math> 0.006</b> $\downarrow$	<b>0.015<math>\pm</math>0.006</b> $\downarrow$	0%
$t = 6$	C-index	0.566 $\pm$ 0.093	0.615 $\pm$ 0.055	0.596 $\pm$ 0.073	0.631 $\pm$ 0.134	<b>0.658<math>\pm</math> 0.104</b> $\uparrow$	16.25%
	IBS	0.032 $\pm$ 0.008	NA	0.035 $\pm$ 0.008	0.032 $\pm$ 0.008	<b>0.031<math>\pm</math> 0.008</b> $\downarrow$	-3.12%
$t = 9$	C-index	0.554 $\pm$ 0.068	0.612 $\pm$ 0.061	0.595 $\pm$ 0.04	0.624 $\pm$ 0.088	<b>0.644<math>\pm</math>0.057</b> $\uparrow$	16.24%
	IBS	0.048 $\pm$ 0.01	NA	0.05 $\pm$ 0.01	0.047 $\pm$ 0.01	<b>0.046<math>\pm</math>0.01</b> $\downarrow$	-4.16%
$t = 12$	C-index	0.581 $\pm$ 0.047	0.633 $\pm$ 0.067	0.617 $\pm$ 0.047	0.653 $\pm$ 0.066	<b>0.667<math>\pm</math>0.054</b> $\uparrow$	14.97%
	IBS	0.061 $\pm$ 0.011	NA	0.063 $\pm$ 0.011	0.06 $\pm$ 0.01	<b>0.059<math>\pm</math> 0.01</b> $\downarrow$	-3.27%

TABLE V

WILCOXON SIGNED RANK TEST BETWEEN BEST PERFORMING SURVIVAL MODEL WITH OTHER SURVIVAL MODELS AT SIGNIFICANCE LEVEL  $\alpha = 0.05$ . THE BOLD FIGURES INDICATE SIGNIFICANT VALUE.

	Months = 3			
	CPH	SSVM	DeepSurv	RFSM
GBM	0.769	0.845	0.998	0.556
	Months = 6			
	GBM	<b>0.003</b>	0.375	0.275
	Months = 9			
	GBM	<b>0.003</b>	<b>0.001</b>	0.131
	Months = 12			
	GBM	<b>0.001</b>	<b>0.037</b>	<b>0.003</b>

chemotherapy and surgery. Similarly, other factors, such as smoking habits (general features) and tumor stages (diagnostic features), also contributed to the high risk of relapse [31]. Similarly, on a local scale in Figure 4, we picked one patient in a test set for demonstration purposes. The waterfall chart demonstrates the patient's features contributing to the prediction (positively/negatively). Again, there are two colors, red and blue. Red means positive, and blue means a negative

contribution to the model. For this patient, the GBM survival model predicts 0.641 chances of relapsing. Comorbidity with diabetes mellitus, age of 72, and tumor stage IIB are the highest contributing factors.

## V. CONCLUSION AND FUTURE WORK

Cox's proportional hazards model is the most popular survival model because it is easy to interpret once trained. However, if prediction performance is the main objective, more sophisticated, non-linear, or ensemble models might produce better results. In our study, the popular Cox proportional hazard model did not perform very well compared to other models. It is because the Cox models are not designed to handle larger numbers of patient features (that may, however, be critical for good predictive performance), which might have caused the model to overfit the data, resulting in weak performances compared to other models for unseen data. The other reason is that the Cox model is a linear model that cannot capture the non-linearity relationship in the data. The RFSM, GBM, DeepSurv and SSVM models can, on the other hand, detect complex, non-linear relationships amongst the features, giving them an advantage over linear Cox models. We also did not observe a statistically significant difference in predictive performance between RFSM, GBM, and SSVM models for

the censoring windows of months=3. This indicates that there are several well-performing options to choose from when developing new survival models of clinical relevance in the area of lung cancer relapse prediction. Last but not least, we have implemented a prototype explanation subsystem based on SHAP values to interpret the relapse prediction results for the best-performing survival model for a specific censoring window period. This is considered a must for modern clinical decision support systems, as clinicians cannot be expected to follow black-box predictions [32].

The major limitation of this study is that we have not done any clinical validation and verification. However, it is a part of our future work where we would train survival models in retrospective data and test on prospective patients to assess the models' quality to predict the likelihood of patients who did indeed relapse in the follow-up period. This will allow for a much more realistic assessment of the survival models and overcome the main limitation of the presented in the study.

#### ACKNOWLEDGEMENT

We would like to acknowledge Science Foundation Ireland (SFI/12/RC/2289\_P2) and the CLARIFY project (European Commission under the grant number 875160) for funding this research.

#### REFERENCES

- [1] B. Lee, S. H. Chun, J. H. Hong, I. S. Woo, S. Kim, J. W. Jeong, J. J. Kim, H. W. Lee, S. J. Na, K. S. Beck *et al.*, "Deepbts: prediction of recurrence-free survival of non-small cell lung cancer using a time-binned deep neural network," *Scientific reports*, vol. 10, no. 1, pp. 1–10, 2020.
- [2] L. Cui, H. Li, W. Hui, S. Chen, L. Yang, Y. Kang, Q. Bo, and J. Feng, "A deep learning-based framework for lung cancer survival analysis with biomarker interpretation," *BMC bioinformatics*, vol. 21, no. 1, pp. 1–14, 2020.
- [3] S. K. Mohamed, B. Walsh, M. Timilsina, M. Torrente, F. Franco, M. Provencio, A. Janik, L. Costabello, P. Minervini, P. Stenertorp *et al.*, "On predicting recurrence in early stage non-small cell lung cancer," in *AMIA Annual Symposium Proceedings*, vol. 2021. American Medical Informatics Association, 2021, p. 853.
- [4] M. Timilsina, S. Bousi, D. Fey, A. Janik, M. Torrente, M. Provencio, A. Bermúdez, E. Carcereny, L. Costabello, D. Abreu, M. Cobo, R. Castro, R. Bernabé, M. Guirado, P. Minervini, and V. Nováček, "Integration of clinical information and imputed aneuploidy scores to enhance relapse prediction in early stage lung cancer patients," in *AMIA 2022, American Medical Informatics Association Annual Symposium, Washington, DC, USA, November 5-9, 2022*. AMIA, 2022.
- [5] A. Janik, M. Torrente, L. Costabello, V. Calvo, B. Walsh, C. Camps, S. K. Mohamed, A. L. Ortega, V. Nováček, B. Massutí *et al.*, "Machine learning-assisted recurrence prediction for early-stage non-small-cell lung cancer patients," *arXiv preprint arXiv:2211.09856*, 2022.
- [6] D. W. Kim, S. Lee, S. Kwon, W. Nam, I.-H. Cha, and H. J. Kim, "Deep learning-based survival prediction of oral cancer patients," *Scientific reports*, vol. 9, no. 1, pp. 1–10, 2019.
- [7] Y. Gao and Y. Cui, "Clinical time-to-event prediction enhanced by incorporating compatible related outcomes," *PLOS digital health*, vol. 1, no. 5, p. e0000038, 2022.
- [8] J. Xiao, M. Mo, Z. Wang, C. Zhou, J. Shen, J. Yuan, Y. He, Y. Zheng *et al.*, "The application and comparison of machine learning models for the prediction of breast cancer prognosis: Retrospective cohort study," *JMIR medical informatics*, vol. 10, no. 2, p. e33440, 2022.
- [9] Y. Li, L. Sun, D. S. Burstein, and K. D. Getz, "Considerations of competing risks analysis in cardio-oncology studies: Jacc: Cardiooncology state-of-the-art review," *JACC: CardioOncology*, vol. 4, no. 3, pp. 287–301, 2022.
- [10] V. P. Balachandran, M. Gonen, J. J. Smith, and R. P. DeMatteo, "Nomograms in oncology: more than meets the eye," *The lancet oncology*, vol. 16, no. 4, pp. e173–e180, 2015.
- [11] Y. She, Z. Jin, J. Wu, J. Deng, L. Zhang, H. Su, G. Jiang, H. Liu, D. Xie, N. Cao *et al.*, "Development and validation of a deep learning model for non-small cell lung cancer survival," *JAMA network open*, vol. 3, no. 6, pp. e205 842–e205 842, 2020.
- [12] L. W. Arnold and Z. Wang, "The hba1c and all-cause mortality relationship in patients with type 2 diabetes is j-shaped: a meta-analysis of observational studies," *The Review of Diabetic Studies*, vol. 11, no. 2, 2014.
- [13] K. E. Kopecky, D. Urbach, and M. L. Schwarze, "Risk calculators and decision aids are not enough for shared decision making," *JAMA surgery*, vol. 154, no. 1, pp. 3–4, 2019.
- [14] J. E. Schottinger, C. D. Jensen, N. R. Ghai, J. Chubak, J. K. Lee, A. Kamineni, E. A. Halm, C. Sugg-Skinner, N. Udaltsova, W. K. Zhao *et al.*, "Association of physician adenoma detection rates with postcolonoscopy colorectal cancer," *JAMA*, vol. 327, no. 21, pp. 2114–2122, 2022.
- [15] X. Xue, X. Xie, M. Gunter, T. E. Rohan, S. Wassertheil-Smoller, G. Y. Ho, D. Cirillo, H. Yu, and H. D. Strickler, "Testing the proportional hazards assumption in case-cohort analysis," *BMC medical research methodology*, vol. 13, no. 1, pp. 1–10, 2013.
- [16] Z. Zeng, Y. Gao, J. Li, G. Zhang, S. Sun, Q. Wu, Y. Gong, and C. Xie, "Violations of proportional hazard assumption in cox regression model of transcriptomic data in tcga pan-cancer cohorts," *Computational and structural biotechnology journal*, vol. 20, pp. 496–507, 2022.
- [17] V. Van Belle, K. Pelckmans, J. Suykens, and S. Van Huffel, "Support vector machines for survival analysis," in *Proceedings of the Third International Conference on Computational Intelligence in Medicine and Healthcare (CIMED2007)*, 2007, pp. 1–8.
- [18] M. C. Rendleman, B. J. Smith, G. Canahuate, T. A. Braun, J. M. Buatti, and T. L. Casavant, "Representative random sampling: an empirical evaluation of a novel bin stratification method for model performance estimation," *Statistics and computing*, vol. 32, no. 6, p. 101, 2022.
- [19] H. Ishwaran, U. B. Kogalur, E. H. Blackstone, and M. S. Lauer, "Random survival forests," *The annals of applied statistics*, vol. 2, no. 3, pp. 841–860, 2008.
- [20] G. Ridgeway, "The state of boosting," *Computing science and statistics*, pp. 172–181, 1999.
- [21] Z. Zhao, Y. Tian, Z. Yuan, P. Zhao, F. Xia, and S. Yu, "A machine learning method for improving liver cancer staging," *Journal of Biomedical Informatics*, vol. 137, p. 104266, 2023.
- [22] M. F. Amasyali and O. K. Ersoy, "Classifier ensembles with the extended space forest," *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 3, pp. 549–562, 2013.
- [23] W. N. Street, "A neural network model for prognostic prediction." in *ICML*. Citeseer, 1998, pp. 540–546.
- [24] E. Biganzoli, P. Boracchi, L. Mariani, and E. Marubini, "Feed forward neural networks for the analysis of censored survival data: a partial logistic regression approach," *Statistics in medicine*, vol. 17, no. 10, pp. 1169–1186, 1998.
- [25] D. Faraggi and R. Simon, "A neural network model for survival data," *Statistics in medicine*, vol. 14, no. 1, pp. 73–82, 1995.
- [26] C. Lee, W. Zame, J. Yoon, and M. Van Der Schaar, "Deephit: A deep learning approach to survival analysis with competing risks," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, no. 1, 2018.
- [27] J. L. Katzman, U. Shaham, A. Cloninger, J. Bates, T. Jiang, and Y. Kluger, "DeepSurv: personalized treatment recommender system using a cox proportional hazards deep neural network," *BMC medical research methodology*, vol. 18, no. 1, pp. 1–12, 2018.
- [28] M. Provencio, E. Carcereny, D. Rodríguez-Abreu, R. López-Castro, M. Guirado, C. Camps, J. Bosch-Barrera, R. García-Campelo, A. L. Ortega-Granados, J. L. González-Larriba *et al.*, "Lung cancer in Spain: information from the thoracic tumors registry (tr study)," *Translational lung cancer research*, vol. 8, no. 4, p. 461, 2019.
- [29] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *The Journal of Machine learning research*, vol. 7, pp. 1–30, 2006.
- [30] J. J. Ludbrook, P. T. Truong, M. V. MacNeil, M. Lesperance, A. Webber, H. Joe, H. Martins, and J. Lim, "Do age and comorbidity impact treatment allocation and outcomes in limited stage small-cell lung cancer? a community-based population analysis," *International Journal*

*of Radiation Oncology\* Biology\* Physics*, vol. 55, no. 5, pp. 1321–1330, 2003.

- [31] C. M. Tammemagi, C. Neslund-Dudas, M. Simoff, and P. Kvale, “Smoking and lung cancer survival: the role of comorbidity and treatment,” *Chest*, vol. 125, no. 1, pp. 27–37, 2004.
- [32] J. M. Durán and K. R. Jongsma, “Who is afraid of black box algorithms? on the epistemological and ethical basis of trust in medical ai,” *Journal of Medical Ethics*, vol. 47, no. 5, pp. 329–335, 2021.