SurvSHAP(t): Time-dependent explanations of machine learning survival models

Mateusz Krzyziński
Faculty of Mathematics and Information Science
Warsaw University of Technology
mateusz.krzyzinski.stud@pw.edu.pl

Mikołaj Spytek
Faculty of Mathematics and Information Science
Warsaw University of Technology

Hubert Baniecki
Faculty of Mathematics and Information Science
Warsaw University of Technology

Przemysław Biecek
Faculty of Mathematics and Information Science
Warsaw University of Technology

Abstract

Machine and deep learning survival models demonstrate similar or even improved time-to-event prediction capabilities compared to classical statistical learning methods yet are too complex to be interpreted by humans. Several model-agnostic explanations are available to overcome this issue; however, none directly explain the survival function prediction. In this paper, we introduce SurvSHAP(t), the first time-dependent explanation that allows for interpreting survival black-box models. It is based on SHapley Additive exPlanations with solid theoretical foundations and a broad adoption among machine learning practitioners. The proposed methods aim to enhance precision diagnostics and support domain experts in making decisions. Experiments on synthetic and medical data confirm that SurvSHAP(t) can detect variables with a time-dependent effect, and its aggregation is a better determinant of the importance of variables for a prediction than SurvLIME. SurvSHAP(t) is model-agnostic and can be applied to all models with functional output. We provide an accessible implementation of time-dependent explanations in Python at `https://github.com/MI2DataLab/survshap`.

***K*eywords** survival analysis · censored data · Cox proportional hazards model · random survival forest · interpretability · explainable AI

1 Introduction

Machine learning has been gaining popularity in practical applications to solve various problems. Especially in the medical domain, its capabilities prove highly advantageous. Topol [58] states that every medical professional will work with AI technology in the future, particularly deep learning. Unfortunately, many machine learning models, especially complex ones such as deep neural networks [31, 37, 66], are considered black-box models, i.e., it is not possible to know directly what influences their prediction internally [9]. Such knowledge proves helpful for explaining and examining the model of interest.

It enables to verify that the predictions are made on the same basis that human domain experts would make them and increases trust in model predictions. The fact that black-box models do not come with readily available explanations has been a hindrance in their widespread adoption, as in many areas, including medicine, it is crucial to know what affects the model output. Therefore, the need for research on explanation methods of such complex models has been postulated [3, 21]. However, often worse-performing but more common and well-established models are still preferred in many settings [5].

This phenomenon is evident in the field of survival analysis. In this area, the most common are parametric or semi-parametric statistical models with the semi-interpretable Cox Proportional Hazards model [13] (CPH or Cox model in short) at the forefront [38, 42]. However, due to the limitations of CPH in modeling complex dependencies and its

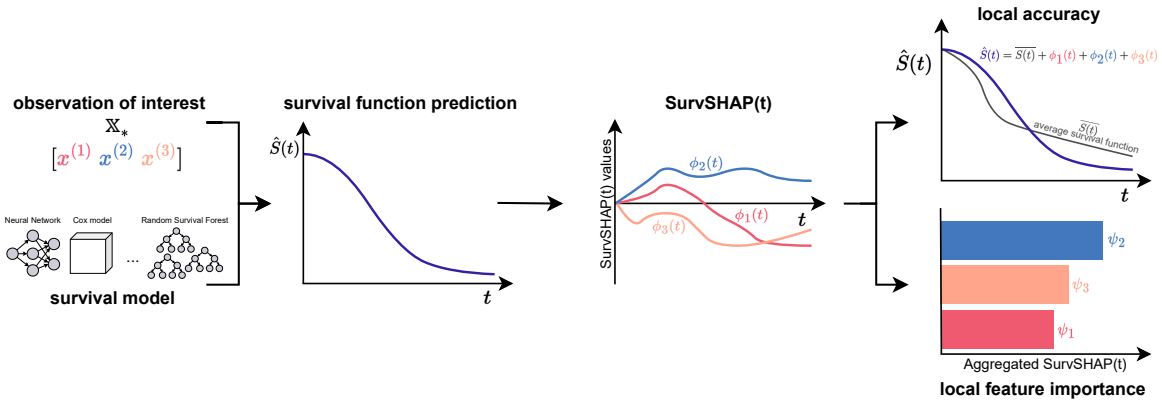

Figure 1: SurvSHAP(t) allows for time-dependent explainability of any survival model predictions. SurvSHAP(t) values add up to the survival function predicted by the model and aggregated over time can be treated as a local importance ranking of features.

strong assumptions that are often not met [50], machine learning models, which are more flexible, have become used in practice, both in healthcare [16, 25, 57] and in other fields [10, 54]. The growing popularity of this class of models motivates the need to develop methods for explaining their predictions.

To the best of our knowledge, there are few existing post-hoc explanation methods specific to complex survival models. These are counterfactual explanations [33], a technique based on the use of neural additive models called SurvNAM [59], and different versions of SurvLIME [34, 35] – an approach inspired by LIME [48]. However, in none of these methods, the time dimension, crucial for predicting the survival conditional probability distribution, is included in the final explanation.

Thus, we aim to extend the available techniques by presenting the new method called SurvSHAP(t) that fills this gap (see diagram in Figure 1). The proposed solution generalizes SHapley Additive exPlanations (SHAP) [40] to survival models. However, the method can be applied to any model with functional output. In the survival analysis setting, the time dimension is used in the proposed explanations – they provide an insight into how each variable influences the model's response (survival function) at each time point. We refer to this property as *time-dependent* explainability.

The contributions of this paper can be summarized as follows:

1. We introduce SurvSHAP(t) – the first time-dependent explanation that allows for interpreting any survival model with functional output. It is based on SHAP with solid theoretical foundations and a broad adoption among machine learning practitioners.
2. We prove that SurvSHAP(t) meets the local accuracy property and accurately explains the model predictions in the form of a survival function, describing variable contributions across the entire analyzed time range. The conducted experiments confirm that SurvSHAP(t) is able to detect variables with a time-dependent effect, and its aggregation is a better determinant of the importance of variables for a prediction than SurvLIME.
3. We provide an accessible implementation of both SurvSHAP(t) and SurvLIME in Python. Source code for both methods and the data generation is available on GitHub at `https://github.com/MI2DataLab/survshap`.

2 Related work

Machine learning survival models. The most fundamental and frequently used approach to survival tasks is applying the Cox Proportional Hazards model, which is a semi-parametric model. It has limitations such as a degradation of performance when working with high dimensional data or correlated variables. A strong proportional hazards assumption has highlighted the need for developing methods based on classical machine learning techniques appropriately adapted to the censored data [65].

One popular survival machine learning model, which is used in our experiments, is Random Survival Forest (RSF) [27]. The most common splitting rule for forming individual decision trees is the log-rank splitting rule based on the log-rank test [51]. The single survival tree prediction for an individual is a function computed for all individuals in the same tree terminal node; most often a cumulative hazard function estimated using the Nelson-Aalen estimator [1, 46]. The entire prediction of RSF is the function averaged over all trees, which makes them able to predict complicated survival functions. Another class of models applied to survival analysis is Gradient Boosting Machines (GBM) [49]. A GBM performs a greedy stage-wise process with the objective of optimizing a selected function. In the basic case, it maximizes the log-partial likelihood known from the Cox model. The likelihood can also be substituted with a differentiable approximation of the concordance index – a metric used to evaluate the survival models' performance [11]. Implementations of these models are accessible through open-source software like `scikit-survival` [47] Python package. In R, many popular packages can be adapted for survival predictions, including `randomForestSRC` [26] and `gbm` [23].

Early adaptations of neural networks to survival analysis were proposed in the 1990s. Faraggi and Simon [17] proposed a non-linear proportional hazards model trained using the partial likelihood function. A similar idea is also used in more recent approaches to survival prediction, such as Cox-nnet [12] – the main difference is the modification of the loss function. Modern deep learning models, such as DeepSurv [31] and Cox-Time [36], also often model a prognostic index corresponding to the linear predictor from the Cox model. However, there are also architectures that return more flexible predictions without relying on the Cox model [20, 37, 67]. Another notable deep learning approach is DeepOmix, which has been proposed as an interpretable model for multi-omics data [66].

Unfortunately, relatively few survival deep learning models have open-source implementations. Some of the solutions are available in `pycox` [36] and `auton-survival` [45] Python packages, or `survivalmodels` [56] R package.

Explanations of machine learning models. *Interpretability* has many definitions, and one widely-used formulation is *the degree to which a human can consistently understand model predictions* [32]. Many complex machine and deep learning models are not directly interpretable, which leads to the development of eXplainable AI (XAI) [24]. Explanation of machine learning models can be divided into two categories: local methods, which are used to explain the model's predictions for a particular observation, and global ones, which provide information about the overall behavior of the model. The most widespread model-agnostic local explanations are additive feature attributions: Local Interpretable Model-agnostic Explanations (LIME) [48] and SHapley Adaptive exPlanations (SHAP) [40].

LIME [48] uses a local interpretable surrogate model fitted to the vicinity (a new dataset generated artificially) of an individual observation. For tasks of regression and classification, linear and logistic regression models are used, whose coefficients have clear interpretations.

SHAP [40] is based on the Shapley value framework from game theory [53], which was introduced into interpreting machine learning predictions in [62, 63]. Shapley values characterize how much each feature influences the model prediction in relation to the baseline average. Attributions are calculated as a mean change to the prediction after adding the examined feature to each possible subset of the model's features. KernelSHAP [40] is proposed as an exact estimation of Shapley values inspired by LIME. Exact KernelSHAP has two main limitations: high computation cost and producing misleading explanations when features are dependent. The first can be improved when explaining tree-ensemble models like random forests with the efficient TreeSHAP algorithm [41]. The second can be overcome by a more thoughtful generation of neighbourhood samples under the curse of dimensionality [2]. Many more explanations based on the SHAP framework like TimeSHAP [7] and FastSHAP [29] advance our understanding of complex learning algorithms.

Alike implementations of survival machine learning models, explanations for classification and regression models are available through open-source software both in Python (`dalex` [4], `SHAP` [40]) and R (`DALEX` [8], `shapr` [52]).

Explanations of machine learning survival models. Some explanations of predictive models for the more standard regression and classification tasks can be adapted for survival models with the use of single-point risk predictions or aggregations of survival functions as done in the study conducted by Moncada-Torres et al. [44]. However, this leads to the loss of information contained in the survival distribution, especially in the case of complex models algorithms

modeling flexible survival functions. In contrast, SurvSHAP(t) provides explanations of the whole distribution in the form of the survival function.

To overcome the mentioned shortcomings, Kovalev et al. [35] propose to adapt LIME into the SurvLIME method by using another object describing survival distribution – the cumulative hazard function (CHF) – as the basis for calculations. The Cox Proportional Hazards model is used as a surrogate model, whose coefficients are fitted by optimizing a loss function based on the distances between CHFs predicted by the local surrogate model and the black-box model. Like in LIME, the optimization problem is based on a sample of weighted observations from the local area around the point of interest. SurvLIME uses the L^2 metric to calculate the distance between functions. SurvLIME-KS [34] is an extension to the method that uses Kolmogorov-Smirnov bounds for constructing sets of predicted CHFs, which helps to robustify the explanation. However, these methods take into account the distribution only in the computation phase, and the obtained results are the coefficients of the Cox Proportional Hazards model (single values). SurvSHAP(t) extends the dependency on the distribution by providing time-dependent explainability, i.e., it returns an explanation for the entire support of the considered distribution (time range) while being able to aggregate it into meaningful single values.

Moreover, counterfactual explanations for survival analysis models have been proposed, in which the survival function is used to find the counterfactual [33]. Specifically, the difference between the two survival functions, for the original point of interest and the counterfactual, is based on the mean time-to-event distance for the set of observations. This method falls under a different explanation methods category than SurvSHAP(t) as it does not return variable attributions.

Despite the existence of several explanation methods, further development is needed in the field of XAI for survival analysis, and certain problems are apparent. Firstly, none of the known approaches supply time-dependent explanations. Another major deficiency is the lack of publicly available implementations of the methods described in the literature. Therefore, we present SurvSHAP(t) with the expected properties of the survival model explanation along with its implementation.

3 Preliminaries

3.1 Mathematical background of survival analysis

Survival analysis deals with tasks based on censored data. That means we have incomplete information about an individual's survival time for a part of the population from the dataset.

Usually, the case of right censoring is considered – during the study, the event of interest, e.g., a patient's death, is not observed for some part of the population. That means the observed event time is less than or equal to the actual survival time.

Mathematically, a given instance i is represented as a triplet $(\mathbf{x}_i, y_i, \delta_i)$, where $\mathbf{x}_i = [x_i^1, x_i^2, \ldots, x_i^p] \in \mathbb{R}^p$ indicates the covariates vector; δ_i is the indicator of event of interest's occurrence; and y_i stands for the observed time (either survival time T_i when $\delta_i = 1$ or censoring time C_i when $\delta_i = 0$). Thus, one should acknowledge that T_i is a latent value for censored observations. The primary objective of survival analysis is the estimation of T_j for an instance j with covariates vector $\mathbf{x}_j$. Most often, instead of predicting a single time moment, a certain function of time is the output.

The first key object is the survival function (1) which describes the probability of an individual surviving until time t without experiencing the event, i.e.,

$$S(t) = \mathbb{P}(T > t) = 1 - \mathbb{P}(T \leq t). \tag{1}$$

Another fundamental concept is the hazard function (2) that can be interpreted as the conditional failure rate in a short (infinitesimal) time interval, provided that the event has not occurred by time t, which is defined as

$$h(t) = \lim_{t \to 0} \frac{\mathbb{P}(t \leq T < t + \Delta t \,|\, T \geq t)}{\Delta t} = \frac{f(t)}{S(t)}, \tag{2}$$

where $f(t) = -\frac{\mathrm{d}S(t)}{\mathrm{d}t}$ is the event of interest's density function.

Therefore, the survival function is connected with the hazard function and can be rewritten as

$$S(t) = \exp(-H(t)), \tag{3}$$

where $H(t) = \int_0^t h(s)\,\mathrm{d}s$ is called the cumulative hazard function.

Moreover, based on the survival function, selecting the appropriate aggregation makes it possible to obtain other predictions of the type of the single values, e.g., time until an event of interest, risk score [55].

3.2 Variable importance ranking in CPH and SurvLIME

The coefficients of a Cox Proportional Hazards model can be used to rank the relative importance of variables for a given prediction. We achieve this by comparing the absolute values of the coefficients multiplied by the values of its covariates $|x^{(d)} \cdot b^{(d)}|$ for each variable d – higher values indicate higher importance. Note that simply comparing the values of **b** coefficients does not inform the end-user about the importance of variables as they can be of vastly different scales. The interpretation of a coefficient $b^{(d)}$ on its own is that an increase of the covariate $x^{(d)}$ by one unit indicates that the hazard rate for the given observation will be $\exp(b^{(d)})$ times higher than the actual value.

We later use this method of ranking variables when comparing SurvLIME to SurvSHAP(t) in Section 6.2, as the SurvLIME explanation takes the form of Cox model coefficients.

4 Time-dependent explanations of machine learning survival models

Let $\mathbb{D} = \{(\mathbf{x}_i, y_i, \delta_i) : i = 1, 2, \ldots, n\}$ be the survival dataset used for training the black-box model. Moreover, assume that $t_1 < t_2 < \ldots < t_m$ are distinct times to event of interest from the set $\{y_i : \delta_i = 1; i = 1, 2, \ldots, n\}$. For each individual described by a covariates vector $\mathbf{x}$, the model returns the individual's survival distribution $\hat{S}(t, \mathbf{x})$ (i.e., the distribution of the event of interest occurring over $\mathbb{R}_{\geq 0}$). The returned object is the survival function as it uniquely determines the distribution. Note that most often, the value of $\hat{S}(t, \mathbf{x})$ is known for all $t \in \{t_1, \ldots, t_m\}$ and for the remaining points, interpolations or step functions are used.

The idea behind the proposed SurvSHAP(t) method is to use the survival function not only to compute an explanation but also to present its results. For this purpose, for the observation of interest $\mathbf{x}_*$ at any selected time point t the algorithm assigns an attribution (importance value) $\phi_t(\mathbf{x}_*, d)$ to the value of each feature $x^{(d)}$, $d \in \{1, 2, \ldots, p\}$, included in the model. In this way, the SurvSHAP(t) functions $[\phi_{t_1}(\mathbf{x}_*, d), \phi_{t_2}(\mathbf{x}_*, d), \ldots, \phi_{t_m}(\mathbf{x}_*, d)]$ are generated for every predictor d. These functions describe the time-dependent influence of covariates on the prediction of the model.

We implement SurvSHAP(t) estimation in two ways, which are based on regression and classification approaches. The first is the Shapley sampling values algorithm. Let $e^D_{t,\mathbf{x}_*} = \mathbb{E}[\hat{S}(t, \mathbf{x})|\mathbf{x}^D = \mathbf{x}^D_*]$ be the expected value for a conditional distribution where conditioning applies to all covariates from the set D.

The contribution of predictor d in time point t is calculated as

$$\phi_t(\mathbf{x}_*, d) = \frac{1}{|\Pi|} \sum_{\pi \in \Pi} e^{\mathrm{before}(\pi, d) \cup \{d\}}_{t,\mathbf{x}_*} - e^{\mathrm{before}(\pi, d)}_{t,\mathbf{x}_*}, \tag{4}$$

where Π is a set of all permutations of p variables and $\mathrm{before}(\pi, d)$ denotes a subset of predictors that are before d in the ordering $\pi \in \Pi$. For easier comparison between different models and time points, this value can be normalized to obtain values on a common scale (from -1 to 1) according to the formula

$$\phi^*_t(\mathbf{x}_*, d) = \frac{\phi_t(\mathbf{x}_*, d)}{\sum_{j=1}^{p} |\phi_t(\mathbf{x}_*, j)|}. \tag{5}$$

It should be noted that thanks to the property that the expected value of the random vector is the vector of the expected values, this operation can be vectorized and performed simultaneously for all selected time points. However, due to the computational cost, permutation sampling is used in the case of high-dimensional models.

Another way to calculate SurvSHAP(t) faster is to use the Shapley kernel [40] and weighted linear regression with functional responses and scalar covariates [18]. Here, we need to define sample coalitions $z_j \in \{0, 1\}^p$, $j \in \{1, 2, \ldots, J\}$ where the value of 1 indicated the presence of corresponding feature in coalition, and the mapping function $h_x : \{0, 1\}^p \to \mathbb{R}^p$ that converts binary vectors into the original input space (1 represents the original value). In this setting, the Shapley kernel remains the same, and the weight given to each binary vector z is

$$w(z) = \frac{p - 1}{\binom{p}{s} s(p - s)}, \tag{6}$$

where s is the number of ones in z. Let Z be the matrix of all binary vectors. Then SurvSHAP(t) is estimated as

$$\Phi = (Z^T W Z)^{-1} Z^T W Y, \tag{7}$$

where W is the diagonal matrix consisting of Shapley kernel weights, and Y is the matrix whose rows contain the survival function values predicted by the model F for the mapping of each row of the Z matrix by the h_x function. Each row of the resulting $p \times r$ matrix Φ contains an explanation for a single variable included in the model.

Thanks to such algorithm structure, SurvSHAP(t) preserves the desired SHAP properties, stated in [40] extended to consider the time-dependent nature of the explanation: local accuracy, missingness, and consistency.

In the context of this study, the property of local accuracy can be defined as:

$$\forall_t \, \hat{S}(t, \mathbf{x}) = e_t^{\emptyset} + \sum_{d=1}^{p} \phi_t(\mathbf{x}, d). \tag{8}$$

We use SurvSHAP(t) to calculate feature importance by aggregating the time-dependent function as

$$\psi(\mathbf{x}, d) = \int_0^{t_m} |\phi_t(\mathbf{x}, d)| \, \mathrm{d}t. \tag{9}$$

5 Evaluation metrics

We introduce metrics used in Section 6 for assessing the quality of explanations.

Local accuracy (10) is a time-dependent adaptation of a local accuracy metric proposed by Lundberg et al. [41]. It is calculated as the normalized standard deviation of the difference between the black-box model's output and the explanation as follows:

$$\sigma(t) = \sqrt{\frac{\mathbb{E}(\hat{S}(t, \mathbf{x}) - \sum_i \phi_t(\mathbf{x}, i))^2}{\mathbb{E}\hat{S}(t, \mathbf{x})^2}}. \tag{10}$$

Lower values of this metric indicate that for that specific time point, the sum of contributions of variables is closer to the actual output of the model. For methods that meet the local accuracy property, the values of this metric are zero.

Another metric we propose to evaluate the ability of an explanation method to show the variables whose effect changes in time is Changing Sign Proportion (CSP). Its purpose is to numerically measure for what fraction of explained observations the value of SurvSHAP(t) was positive and negative for at least α of the considered time period. Specifically, it is defined as

$$CSP_{\alpha, t_\mathrm{s}, t_\mathrm{e}} = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1} \left(\left| \bigcup_{t \in [t_\mathrm{s}, t_\mathrm{e}]} \{t : \phi_t(\mathbf{x}_i, d) \geq 0\} \right| > \alpha \cdot |[t_\mathrm{s}, t_\mathrm{e}]| \right) \mathbb{1} \left(\left| \bigcup_{t \in [t_\mathrm{s}, t_\mathrm{e}]} \{t : \phi_t(\mathbf{x}_i, d) \leq 0\} \right| > \alpha \cdot |[t_\mathrm{s}, t_\mathrm{e}]| \right), \tag{11}$$

where n is the number of samples in the dataset, t_s, t_e are start and end time points of selected time range, $\phi_t(\mathbf{x}_i, d)$ represents the Shapley value for observation i, feature d and time t and $|\cdot|$ is the length of the line segments. For variables with a time-dependent nature of an effect, the metric values should be greater than for variables whose type of effect is constant. In the case of models that do not take into account time-dependent effects, e.g., CPH, the values should be 0.

For evaluating the correlation between rankings of variables by their importance for prediction obtained from the explanation method and known ground-truth orderings (see Section 3.2), we use the additive hyperbolic Kendall's τ_h rank correlation coefficient [60]. We give the average of the coefficients obtained for all analyzed observations as a final measure.

In order to assess the SurvSHAP(t) quality against the ground-truth Shapley values, we use the GT-Shapley metric. It is adapted from [41] by applying it to each considered time point as follows:

$$\forall_t \, \rho(t) = \frac{1}{n} \sum_{i=1}^{n} \texttt{Pearson}([\phi_t(\mathbf{x}_i, d)]_{1 \leq d \leq p}, [\phi_t^{true}(\mathbf{x}_i, d)]_{1 \leq d \leq p}), \tag{12}$$

where the values $\phi_t^{true}(\mathbf{x}_i, d)$ are acquired using SurvSHAP(t) on a background sample with a much larger number of observations ($N = 10000$) generated from the same distribution.

Additionally, we define a metric based directly on residuals between obtained explanations and ground-truth values to measure how difficult it is for SurvSHAP(t) to explain a given variable d. It is given as follows:

$$\text{normalized RMSE}(t, d) = \sqrt{\frac{\mathbb{E}(\phi_t(\mathbf{x}, d) - \phi_t^{true}(\mathbf{x}, d))^2}{\mathbb{E}\phi_t^{true}(\mathbf{x}, d)^2}}. \tag{13}$$

6 Experiments

Evaluation of machine learning explanations presents many challenges [3, 61] as (1) in general, no ground truth of explanation is available [39], and (2) one explains an imprecise black-box model provided imperfect data [30]. Thus, for a comprehensive evaluation scheme, we divide our experiments into three steps:

1. Measuring local accuracy and time-dependence on synthetic data
2. Comparison with SurvLIME to show that SurvSHAP(t) is able to detect variables with a time-dependent effect, and its aggregation is a better determinant of the importance of variables for a prediction than SurvLIME.
3. Showing in a real-world use case that SurvSHAP(t) properly explains machine learning survival models predicting cancer.

6.1 Evaluating explanations on synthetic data

Setup. For the first experiment, data is generated synthetically in order to demonstrate that SurvSHAP(t) explanations work correctly for variables that have a time-dependent effect, provided that the used model can make use of such dependencies. The dataset `EXP1` consisting of $N = 1000$ observations is generated using the method suggested for generating time-dependent effects by Crowther and Lambert [14]. The base hazard function is defined as

$$h_0(t) = \exp(-17.8 + 6.5t - 11\sqrt{t} \cdot \ln t + 9.5\sqrt{t}), \tag{14}$$

and for a chosen observation from the dataset, the hazard function is of the form

$$h(t) = h_0(t) \cdot \exp[(-0.9 + 0.1t + 0.9\ln(t))x^{(1)} + 0.5x^{(2)} - 0.2x^{(3)} + 0.1x^{(4)} + 10^{-6}x^{(5)}]. \tag{15}$$

The coefficients were chosen such that the covariate $x^{(1)}$ has a time-dependent effect, $x^{(2)}, x^{(3)}$, and $x^{(4)}$ are covariates with constant effect and $x^{(5)}$ is insignificant – represents random noise. Variables $x^{(1)}$ and $x^{(2)}$ are binary, sampled from the binomial distribution, such that $\mathbb{P}(x^{(1)} = 0) = \mathbb{P}(x^{(1)} = 1) = \mathbb{P}(x^{(2)} = 1) = \mathbb{P}(x^{(2)} = 0) = 0.5$, whereas $x^{(3)} \sim \mathcal{N}(10, 2)$, $x^{(4)} \sim \mathcal{N}(20, 4)$, and $x^{(5)} \sim \mathcal{N}(0, 1)$.

To generate the survival times T_i, the process described by Crowther and Lambert [14] is followed. In the first step, the hazard function (15) is integrated numerically to obtain the cumulative hazard function, which is then transformed into the survival function using formula (3). Then Brent's iterative root finding method is applied to function $g(t) = S(t, \mathbf{x}) - U$, where $U \sim U[0, 1]$. The found root is used as the true (latent) survival time T_i for a vector of covariates $\mathbf{x}_i$.

In order to determine observed times y_i based on generated survival times T_i, a method proposed by Wan [64] is used. For each observation, two values are generated from the uniform distribution: $C_{l,i} \sim U[11, 16]$, which can be interpreted as the time to administrative censoring event, and $C_{r,i} \sim U[0, 24]$ which denotes the time to the occurrence of a right censoring event. If both those values are higher than the generated time T_i then $\delta_i = 1$, otherwise $\delta_i = 0$. The observed time y_i is defined as $y_i = \min\{T_i, C_{ri}, C_{li}\}$, which in this case translates into a censoring rate of 0.331.

Results. The first experiment intends to show how the SurvSHAP(t) works. Thus, we fit Cox Proportional Hazards and Random Survival Forest models to the generated dataset and calculate the prediction explanations for each observation. The models' performance expressed in the Brier score measure [22] is presented in Figure 2. RSF has an integrated Brier score equal to 0.097 and, because of its ability to model complex dependencies, outperforms CPH, for which the integrated Brier score is 0.167.

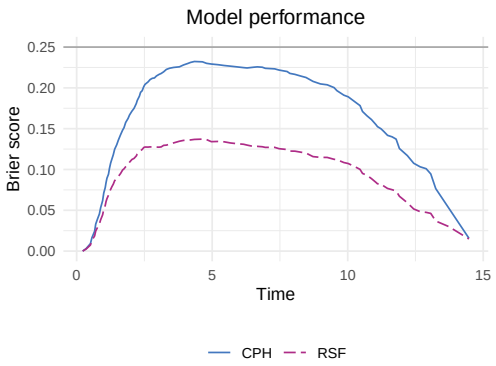

Figure 2: Time-dependent performance of the RSF and CPH models measured by Brier score (**lower** is better; Brier score of 0.25 indicates random predictions).

In Figure 3 the SurvSHAP(t) functions for a selected prediction are presented. In the first row, SurvSHAP(t) functions of each variable are shown for each of the two models, whereas in the second row, they are normalized according to formula (5). Positive SurvSHAP(t) values indicate that a given variable has increased the survival function by that much, while negative values indicate a decrease. It can be seen that the variable $x^{(1)}$, which has a time-dependent effect (positive at the beginning, negative later), is correctly modeled in RSF but not in CPH. This shows that SurvSHAP(t) is capable of finding such differences between models (i.e., it explains the model, not data) and therefore is useful for validating if models consider time-dependent variables. Indeed, by looking at the normalized SurvSHAP(t) values, we can see that the variable effects for the CPH are constant over time (narrow boxplots). Moreover, RSF assigned part of the changing impact over time to other variables – it is expected as a non-parametric model without knowledge of a specific form of the time-dependent effect has difficulties with its precise separation (i.e., determination of the source of this effect).

Another benchmark we perform is checking if the additivity property of SHAP is retained. For this purpose, we computed the time-dependent version of the local accuracy metric defined in (10).

Low values of the local accuracy metric (shown in Figure 4) on the order of 10^{-7} indicate that the property is preserved. The fact that they seem to rise with time is also expected as it is difficult for models to make good predictions near the end of the examined time, where many observations are censored.

As the data used for this experiment is synthetically generated, we know that the variable $x^{(1)}$ has a time-dependent effect, positive at the beginning and negative later. Therefore, if a model uses this fact in its prediction, it should be noticeable in the explanations, such that in some proportion of the considered time period, the effect is negative and in some – positive. We use the Changing Signs Proportion metric (11) defined earlier with start and end time points fixed at 0.1 and 0.9 quantiles of the times included in the data, respectively.

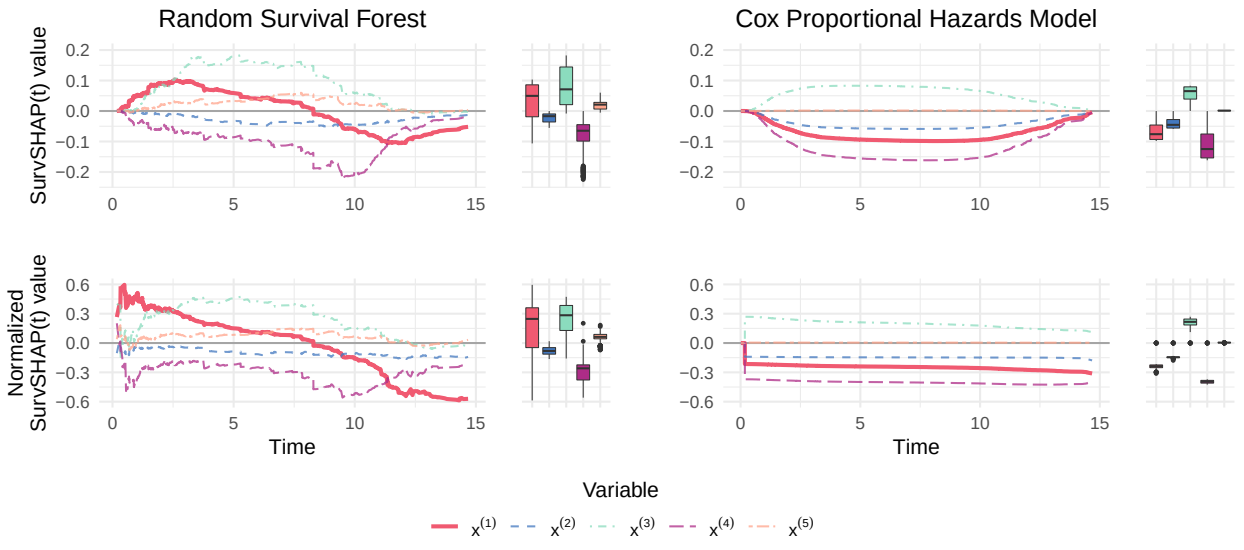

Figure 3: SurvSHAP(t) for the selected observation and two models trained on the dataset `EXP1`.

The value should be low for variables with a constant effect on the prediction and noticeably higher for the variables whose effect changes in time. It should be very low for models we know cannot take time-dependent variables into account, e.g., the Cox Proportional Hazards model. Table 1 presents the values of $CSP_{0.05}$ for each variable and for both models. We see that the metric is high for the $x^{(1)}$ variable for RSF and low for other variables. For CPH, the metric is close to 0, as variables can only have either positive or negative influence *independent of time*.

Table 1: Comparison of the $CSP_{0.05}$ metric values for each variable between Random Survival Forest and Cox Model.

Variable	RSF	CPH
$x^{(1)}$	0.954	0.000
$x^{(2)}$	0.127	0.000
$x^{(3)}$	0.154	0.072
$x^{(4)}$	0.274	0.066
$x^{(5)}$	0.481	0.015

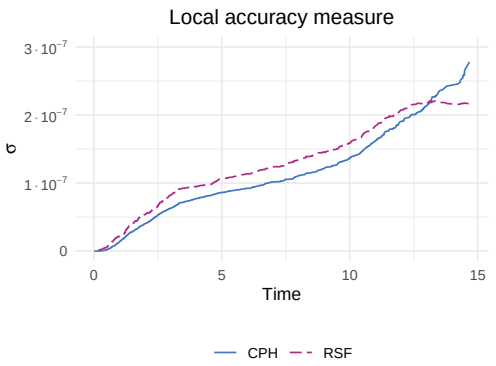

Figure 4: Sanity check of local accuracy (additivity) property for two models trained on the dataset `EXP1`.

Another dependency we measured is the value of the GT-Shapley metric (12) for explanations of CPH and RSF models. For each time point, we calculate the Pearson's correlation between SurvSHAP(t) generated on the basis of the `EXP1` dataset and a larger, artificially generated sample of observations (as we know the underlying distribution of data). The high scores shown in Figure 5 confirm that the explanations are stable: a greater number of points to estimate SurvSHAP(t) does not meaningfully change the explanations.

We also evaluated the explanations of CPH and RSF models using the normalized RMSE metric (13). The comparison is visualized in Figure 6. It is clearly visible that for both models, the attribution of the time-dependent variable $x^{(1)}$ is difficult to explain. We suspect that time-dependent variables need a bigger background of observations to be explained correctly. Another fact worth noting is that explanations for the variable $x^{(5)}$ perform much worse in the CPH model – the plot needs to be cropped to show useful information. For this feature, the attributed SurvSHAP(t) values are close to 0, as we see in Figure 3, so a reason for the high value of the metric might be the numerical errors occurring when normalizing such low values.

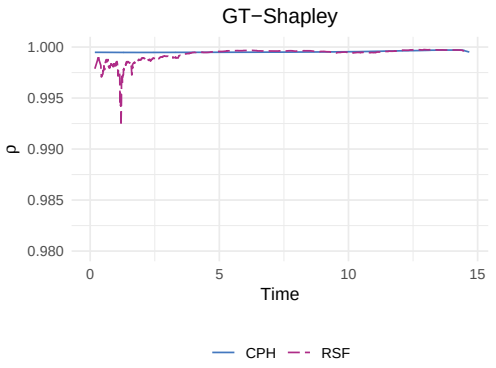

Figure 5: Comparison of the GT-Shapley metric for explanations of CPH and RSF.

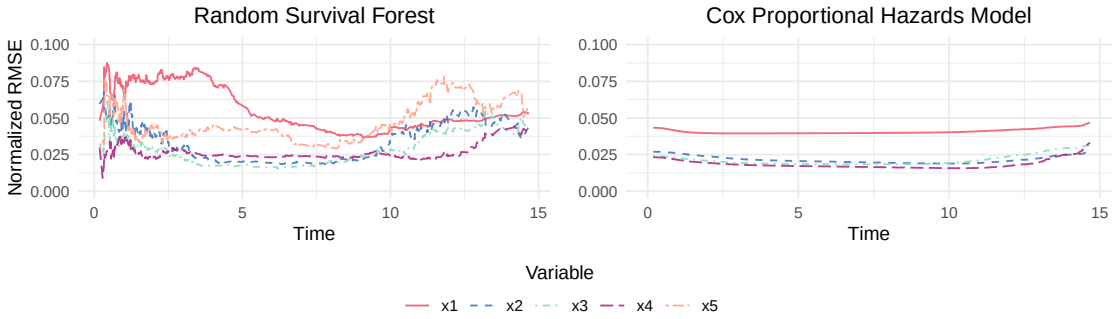

Figure 6: Comparison of normalized RMSE of SurvSHAP(t) for each variable between RSF and CPH (**lower** is better). **Note:** variable $x^{(5)}$ is off the scale for CPH.

6.2 Comparison to SurvLIME

Setup. We aim to compare the explanations of SurvSHAP(t) with those provided by SurvLIME and therefore follow the same experimental setup as in [35]. We use two datasets where the vector of covariates $\mathbf{x}_i \in \mathbb{R}^5$ is generated from the uniform distribution on a 5-dimensional sphere with a predefined radius. The center of the sphere in `dataset0` is $(0, 0, 0, 0, 0)$, whereas `dataset1` is sampled from a sphere centered around $(4, -8, 2, 4, 2)$. Both spheres have the radius $R = 8$. The survival times for these data are generated according to the method proposed by Bender et al. [6] with the following formula:

$$y_i = \left(\frac{-\ln U}{\lambda \exp(\mathbf{b}^T \mathbf{x}_i)} \right)^{1/v}, \tag{16}$$

where $U \sim U[0, 1]$ and

- in `dataset0`: $\lambda = 10^{-5}, v = 2, \mathbf{b}^T = (10^{-6}, 0.1, -0.15, 10^{-6}, 10^{-6})$,
- in `dataset1`: $\lambda = 10^{-5}, v = 2, \mathbf{b}^T = (10^{-6}, -0.15, 10^{-6}, 10^{-6}, -0.1)$.

Event indicators δ_i are generated from the binomial distribution, such that $\mathbb{P}(\delta_i = 1) = 0.9$ and $\mathbb{P}(\delta_i = 0) = 0.1$. Each dataset consists of $N = 1000$ observations divided into train and test sets in the 9:1 proportion. We use the test sets to train the model and then evaluate it and explain it in the test setting.

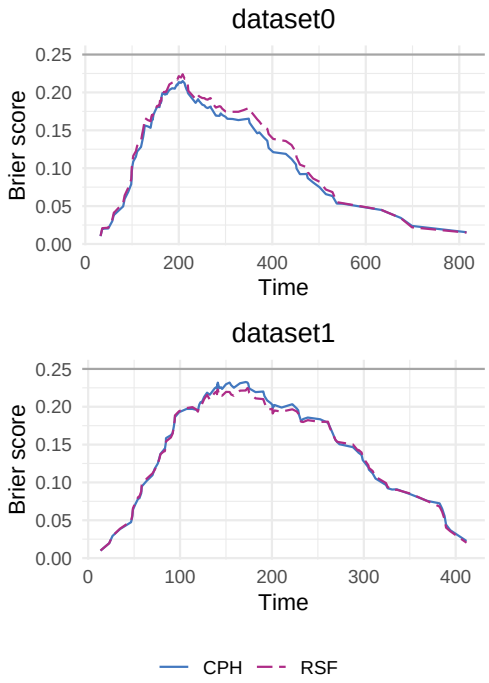

Figure 7: Time-dependent performance of the RSF and CPH models measured by Brier score (**lower** is better; Brier score of 0.25 indicates random predictions).

Results. For this experiment, four models were fitted, i.e., the Cox Proportional Hazards and Random Survival Forest models for each of the two datasets. For `dataset0` the values of integrated Brier score indicating the performance of the models were 0.103 for RSF and 0.097 for CPH, and for `dataset1` RSF achieved a score of 0.137, whereas CPH 0.139. The values of the Brier score for the entire considered time range are presented in Figure 7. The first performed test confirms that SurvSHAP(t) preserves the local accuracy property, whereas SurvLIME does not. The values of normalized standard deviations of local accuracy (10) for RSF model trained on `dataset0` are presented in Figure 8. We see that the value for SurvSHAP(t) is close to 0 across the whole time range, while for SurvLIME, it is significantly larger. This is expected as SurvSHAP(t) can explain all functions, whereas the explanation of SurvLIME always takes the form of a Cox model's survival function.

Both SurvLIME and SurvSHAP(t) can be used to assess the relative importance of variables on the local level (i.e., for selected prediction). First, we fit Cox Proportional Hazards model to obtain the ground-truth variable importance ranking via the method described in Section 3.2 to both datasets. We calculate the rankings of variable attributions using SurvLIME and SurvSHAP(t) for each observation from the test set. Then we use the additive hyperbolic Kendall's τ_h rank correlation coefficient to determine the similarity between the true and explanation rankings. Finally, we average this value across all observations. Table 2 presents the results proving better feature importance ranking estimation of SurvSHAP(t) in the glass-box evaluation scheme.

Table 2: Average τ_h correlations of the variable importance rankings according to explanations against the ground-truth ranking in the Cox model (**higher** is better).

	SurvLIME	**SurvSHAP(t)**
`dataset0`	0.763	0.917
`dataset1`	0.454	0.745

Consecutively, we fit a black-box Random Survival Forest to both datasets and explain its predictions using both methods. The ground-truth importance ranking of the RSF black-box model is not available.

Therefore, we imitate the ranking using permutational variable importance [19] with the integrated Brier score as a loss function. Figure 9 visualizes the aggregation of variable rankings over 100 observations in the test set. Each horizontal bar represents the fraction of observations for which the variable represented as a given color was ranked 1st, 2nd, etc. The correct ordering, obtained by permutational variable importance, is presented in the legend of the Figure. We observe that the global aggregation of local SurvLIME rankings is close to random – for `dataset0` almost every place has a uniform distribution of variables, whereas SurvSHAP(t) decidedly attributes one feature to the first and second place in the ranking.

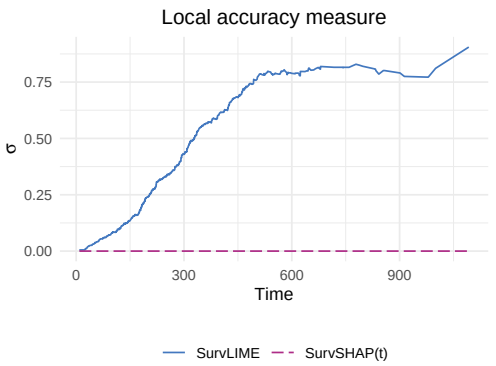

Figure 8: Normalized standard deviation of the difference between black-box model output and the explanation (**lower** is better). **Note:** the curve for SurvSHAP(t) coincides with the x-axis.

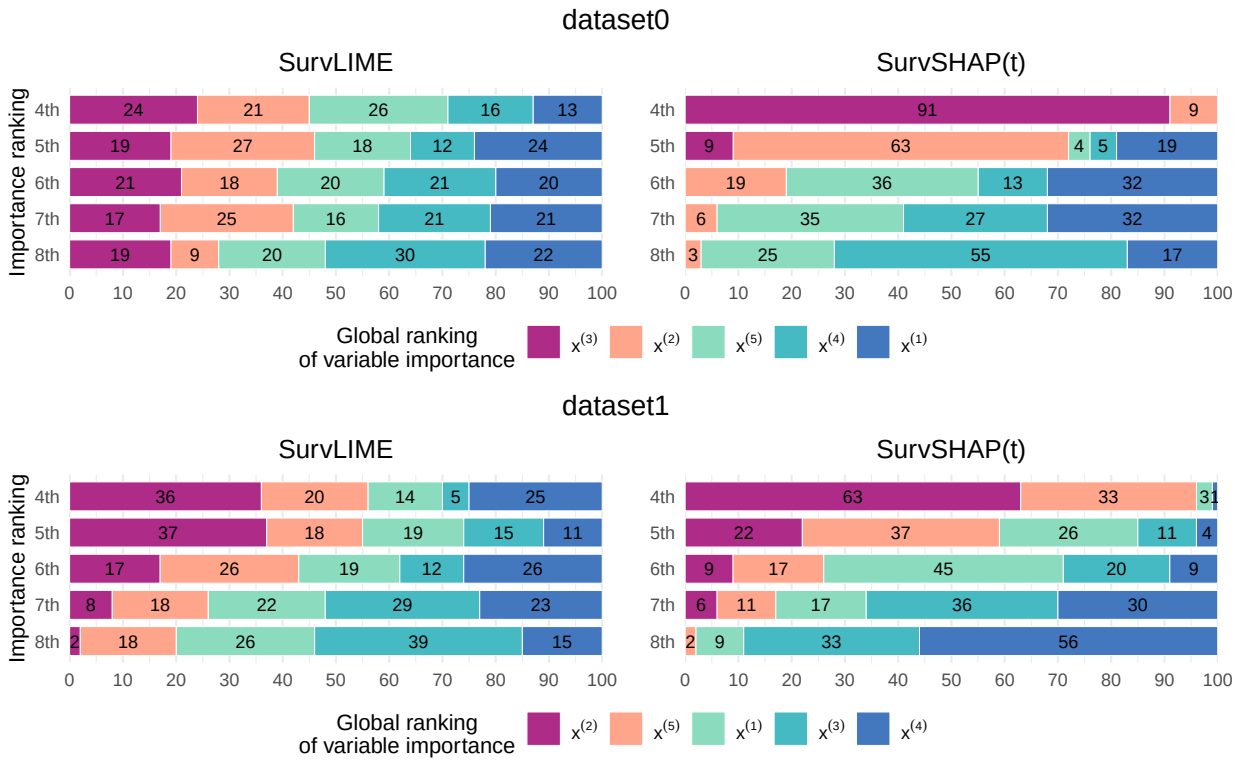

Figure 9: Juxtaposition of local and global importance rankings for 100 predictions of the RSF model fitted to `dataset0` (**top**) and `dataset1` (**bottom**). In each case, there are two globally important variables and three less important ones– the colors are specifically sorted to show the global ranking of features. We observe that SurvSHAP(t) maintains the majority observations per each consecutive variable (specifically in `dataset1` the majorities are represented by 63 for $x^{(2)}$ in 1st, 37 for $x^{(2)}$ in 2nd, 45 for $x^{(1)}$ in 3rd, 36 for $x^{(3)}$ in 4th, and 56 for $x^{(4)}$ in 5th) outperforming SurvLIME, which provides more uniformly distributed rankings (especially in `dataset0`).

6.3 Real-world use case: predicting cancer survival

The main motivation for the development of SurvSHAP(t) is the potential of such a method in practical applications. In analyses of medical time-to-event data, time-dependent effects, i.e., non-proportional hazards, often occur [15, 28, 43].

In order to show the use case of the SurvSHAP(t), we apply the method to two models trained on real-world dataset `UM dataset`. The analyzed cohort includes 164 patients diagnosed with uveal melanoma treated by primary enucleation without any prior therapies [16]. We use eight selected variables and limit the sample to 155 observations with complete data. In this case, we also use two algorithms: Cox model and Random Survival Forest. The performance of the models measured with the Brier score is presented in Figure 10. The integrated Brier score for the RSF model is 0.085, while the CPH model has a score of 0.119.

An exemplary single prediction explanation for the RSF model is presented in Figure 11. It shows that the model assigned a time-dependent effect to one of the variables (ciliary body infiltration). We also see which variables have the greatest impact on the obtained prediction. With domain knowledge, it allows for the assessment of whether a given model output is reliable.

Figure 10: Time-dependent performance of the RSF and CPH models measured by Brier score (**lower** is better; Brier score of 0.25 indicates random predictions).

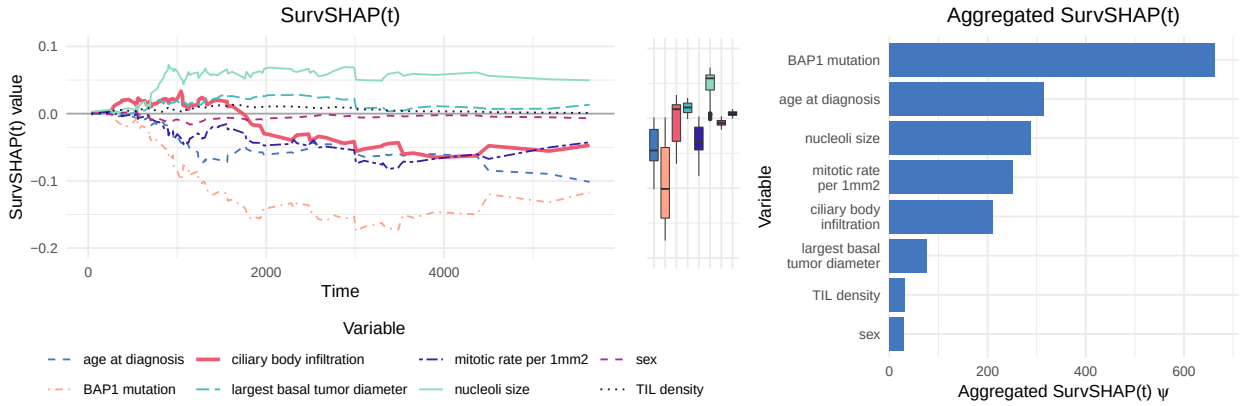

Figure 11: SurvSHAP(t) for the selected observation and RSF model trained on the `UM dataset`.

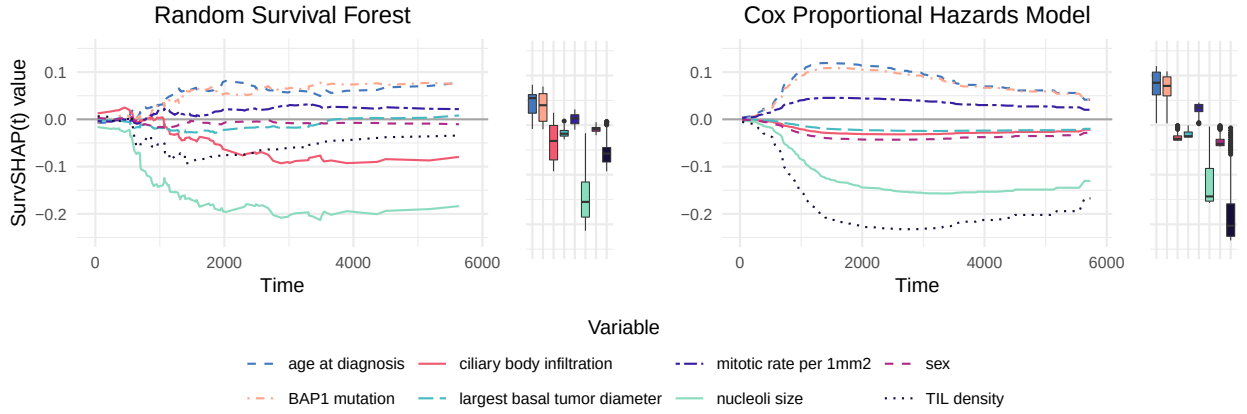

Figure 12: SurvSHAP(t) for the selected observation and two models trained on the `UM dataset`.

In Figure 12, explanations of the decisions of both models for the same patient are presented. They show some inconsistencies between the models. In-depth analysis enables the physician to decide which prediction is more adequate, helping to develop personalized medicine.

For such a small sample and real complex data, it is difficult to obtain ground-truth Shapley values. Moreover, due to the presence of many binary or categorical variables with level 0 in the data, it becomes impossible to use the importance of variables in the Cox model as defined in Section 3.2. Therefore, we imitate the global variable importance ranking for both models by calculating the permutational variable importance once again. Further, we compare the importance rankings of the variables obtained by aggregating SurvSHAP(t) with those from the Cox model found by SurvLIME.

The results are presented in Figure 13, where the variables within one bar are sorted by the global importance. Again, one can observe the superiority of SurvSHAP(t) aggregation over the coefficients derived from the SurvLIME method. For the Cox model, SurvLIME most often indicates the largest basal tumor diameter as the most important variable, which is the third least important variable globally. SurvSHAP(t) indicates age as the most important variable for the biggest percentage of the CPH predictions, which is consistent with the global ranking. Our method coped even better in the context of assessing the importance of variables in the RSF model. It identified nucleoli size and the BAP1 mutation indicator as the most important variables 124 times in total. These two variables have similar importance globally (the mean increase of the integrated Brier score after a permutation is 0.0476 and 0.0451, respectively).

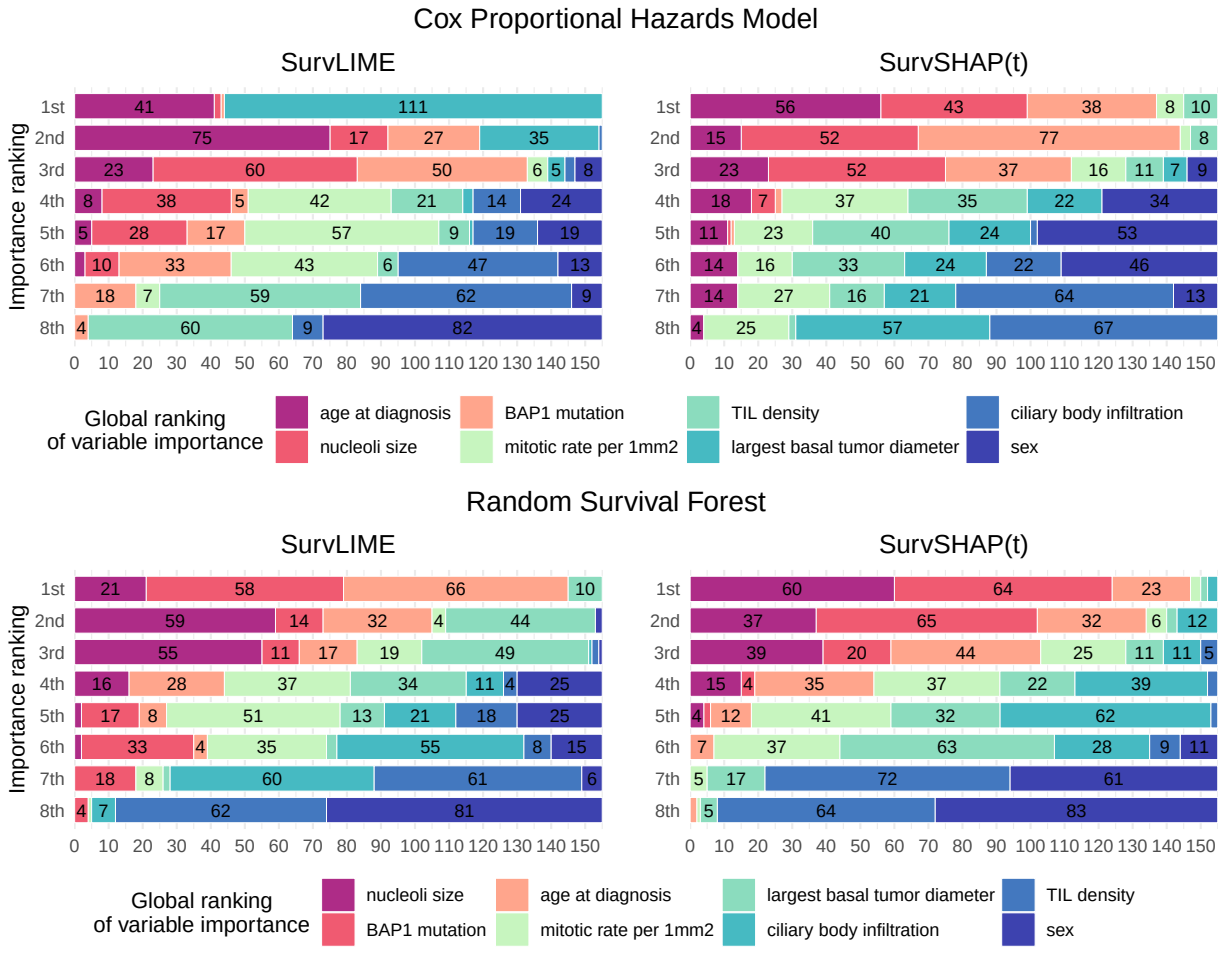

Figure 13: Juxtaposition of local and global importance rankings for predictions of the CPH model (**top**) and RSF (**bottom**). The colors are specifically sorted from purple to blue to show the global ranking of features.

7 Discussion

SurvSHAP(t) extends the idea of SHAP to a broad class of models working on survival data. It is a subsequent method designed to explain survival functions, but the first one based on the concept of Shapley values and the first that provides time-dependent explanations.

On the differences to SurvLIME. SHAP has been chosen as the basis of this method because it is one of the most widely used approaches for explaining black-box models [24]. It was also suggested as an area of possible further research by Kovalev et al. [35]. It is important to notice that the approach to explanation is different from the one proposed in SurvLIME. For the calculation, the entire output function is considered in both methods, but SurvLIME finds the closest possible CHF among the outputs of a Cox model, whereas SurvSHAP(t) does not have this limitation. The authors of SurvLIME state that the independence of the Cox model's covariate effect on time is an advantage, as it makes the optimization problem significantly simpler. However, it has one drawback – SurvLIME cannot properly explain variables whose effect changes in time. SurvSHAP(t) does not make use of the Cox model, so it is able to explain variables that have a time-dependent effect. The change is also visible in the context of the final explanation form. SurvLIME outputs the found coefficients of a Cox model (each feature's attribution described by a single value), while SurvSHAP(t) produces functions of time-dependent importance for each variable. Moreover, coefficients of the SurvLIME explanation do not directly indicate the importance of features, i.e., the magnitude of the feature is also a contributing factor. The proposed method presents importance at all time points explicitly.

Software implementation. The inclusion of code implementing both SurvSHAP(t) and SurvLIME in Python is another key point of the conducted research, as it allows for the application of explanations to existing models. The code is tailored to work with the models implemented in the `scikit-survival` Python package [47]. The produced plots give the user an intuitive visualization of the SurvSHAP(t) explanations – one can see what the influence of a particular variable is at any chosen time, even if they do not have previous experience with XAI.

Limitations. Another thing worth noting is the fact that SurvSHAP(t), as an extension of SHAP, inherits many of its drawbacks. One of them is that reported values might be misleading if the model is not additive [9]. Another practical limitation is the fact that the computation of Shapley values is time-consuming, which is amplified even more by the fact that the calculation needs to be done for many time points.

8 Conclusion

This paper introduces a new local feature attribution method for survival models with sound theoretical guarantees like local accuracy. It has been illustrated and validated using synthetically generated data and compared with the SurvLIME method. Moreover, an example of the developed method in a real-life use case is presented. The method's source code, the experiments carried out in this study, and a source code of the so far *not* implemented SurvLIME method are contributed.

SurvSHAP(t) is the first explanation that presents its final results as time-dependent functions. We believe this sets a new direction for research at the intersection of survival analysis and XAI. It is worth pointing out that the approach could be generalized for any model producing functional output. Future works should consider the possibility of aggregating the SurvSHAP(t) function across data distribution to introduce global explanations of machine learning survival models.

We anticipate that an accessible visualization of SurvSHAP(t) can popularize explainability methods in domains where survival analysis is applied. Our contribution benefits various stakeholders, e.g., physicians and bioinformaticians, in extracting knowledge from data and model analysis. We recommend applying SurvSHAP(t) to explain RSF and deep learning models in scenarios where only the CPH model was previously considered in practice.

Acknowledgments

We would like to thank Mai P. Hoang, MD, from Harvard Medical School, Boston, MA, USA, and Piotr Donizy, MD, from Department of Clinical and Experimental Pathology, Wroclaw Medical University, Wroclaw, Poland, for valuable discussions on the presented method and providing data on survival in uveal melanoma. We also thank Anna Kozak and Katarzyna Woźnica for their valuable comments about the study. This work was financially supported by the NCBiR grant INFOSTRATEG-I/0022/2021-00 and NCN Sonata Bis-9 grant 2019/34/E/ST6/00052.

References

[1] Odd Aalen. Nonparametric Inference for a Family of Counting Processes. *The Annals of Statistics*, 6(4):701–726, 1978. doi:10.1214/aos/1176344247.

[2] Kjersti Aas, Martin Jullum, and Anders Løland. Explaining individual predictions when features are dependent: More accurate approximations to Shapley values. *Artificial Intelligence*, 298:103502, 2021. doi:10.1016/j.artint.2021.103502.

[3] Anna Markella Antoniadi, Yuhan Du, Yasmine Guendouz, Lan Wei, Claudia Mazo, Brett A. Becker, and Catherine Mooney. Current Challenges and Future Opportunities for XAI in Machine Learning-Based Clinical Decision Support Systems: A Systematic Review. *Applied Sciences*, 11(11), 2021. ISSN 2076–3417. doi:10.3390/app11115088.

[4] Hubert Baniecki, Wojciech Kretowicz, Piotr Piatyszek, Jakub Wisniewski, and Przemyslaw Biecek. dalex: Responsible Machine Learning with Interactive Explainability and Fairness in Python. *Journal of Machine Learning Research*, 22(214):1–7, 2021.

[5] Falco J Bargagli Stoffi, Gustavo Cevolani, and Giorgio Gnecco. Simple models in complex worlds: Occam's razor and statistical learning theory. *Minds and Machines*, 32(1):13–42, 2022.

[6] Ralf Bender, Thomas Augustin, and Maria Blettner. Generating survival times to simulate Cox proportional hazards models. *Statistics in Medicine*, 24(11):1713–1723, 2005. doi:10.1002/sim.2059.

[7] João Bento, Pedro Saleiro, André F. Cruz, Mário A.T. Figueiredo, and Pedro Bizarro. TimeSHAP: Explaining Recurrent Models through Sequence Perturbations. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, pages 2565–2573, 2021. doi:10.1145/3447548.3467166.

[8] Przemyslaw Biecek. DALEX: Explainers for Complex Predictive Models in R. *Journal of Machine Learning Research*, 19(84):1–5, 2018.

[9] Przemyslaw Biecek and Tomasz Burzykowski. *Explanatory Model Analysis*. Chapman and Hall/CRC, New York, 2021. ISBN 9780367135591. doi:10.1201/9780429027192.

[10] Pavle Boškoski, Matija Perne, Martina Rameša, and Biljana Mileva Boshkoska. Variational Bayes survival analysis for unemployment modelling. *Knowledge-Based Systems*, 229:107335, 2021. doi:10.1016/j.knosys.2021.107335.

[11] Yifei Chen, Zhenyu Jia, Dan Mercola, and Xiaohui Xie. A Gradient Boosting Algorithm for Survival Analysis via Direct Optimization of Concordance Index. *Computational and Mathematical Methods in Medicine*, 2013. doi:10.1155/2013/873595.

[12] Travers Ching, Xun Zhu, and Lana X. Garmire. Cox-nnet: An artificial neural network method for prognosis prediction of high-throughput omics data. *PLOS Computational Biology*, 14(4):1–18, 2018. doi:10.1371/journal.pcbi.1006076.

[13] David Cox. Regression Models and Life-Tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2):187–220, 1972.

[14] Michael J. Crowther and Paul C. Lambert. Simulating biologically plausible complex survival data. *Statistics in Medicine*, 32(23):4118–4134, 2013. doi:10.1002/sim.5823.

[15] Piotr Donizy, Grazyna Pietrzyk, Agnieszka Halon, Cyprian Kozyra, Tserenchunt Gansukh, Hermann Lage, Pawel Surowiak, and Rafal Matkowski. Nuclear-cytoplasmic PARP-1 expression as an unfavorable prognostic marker in lymph node-negative early breast cancer: 15-year follow-up. *Oncology Reports*, 31(4):1777–1787, 2014.

[16] Piotr Donizy, Mateusz Krzyziński, Anna Markiewicz, Paweł Karpiński, Krzysztof Kotowski, Artur Kowalik, Jolanta Orłowska-Heitzman, Bożena Romanowska-Dixon, Przemysław Biecek, and Mai P Hoang. Machine Learning Models Demonstrate that Clinicopathologic Variables are Comparable to Gene Expression Prognostic Signature in Predicting Survival in Uveal Melanoma. *European Journal of Cancer*, In Press.

[17] David Faraggi and Richard Simon. A neural network model for survival data. *Statistics in Medicine*, 14(1):73–82, 1995. doi:https://doi.org/10.1002/sim.4780140108.

[18] Julian J Faraway. Regression analysis for a functional response. *Technometrics*, 39(3):254–261, 1997. doi:10.2307/1271130.

[19] Aaron Fisher, Cynthia Rudin, and Francesca Dominici. All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *Journal of Machine Learning Research*, 20(177):1–81, 2019.

[20] Eleonora Giunchiglia, Anton Nemchenko, and Mihaela van der Schaar. RNN-SURV: A Deep Recurrent Model for Survival Analysis. In *International Conference on Artificial Neural Networks (ICANN)*, 2018. doi:10.1007/978-3-030-01424-7_3.

[21] Randy Goebel, Ajay Chander, Katharina Holzinger, Freddy Lecue, Zeynep Akata, Simone Stumpf, Peter Kieseberg, and Andreas Holzinger. Explainable AI: the new 42? In *International Cross-Domain Conference for Machine Learning and Knowledge Extraction (CD-MAKE)*, 2018. doi:10.1007/978-3-319-99740-7_21.

[22] Erika Graf, Claudia Schmoor, Willi Sauerbrei, and Martin Schumacher. Assessment and comparison of prognostic classification schemes for survival data. *Statistics in Medicine*, 18(17-18):2529–2545, 1999.

[23] Brandon Greenwell, Bradley Boehmke, Jay Cunningham, and GBM Developers. *gbm: Generalized Boosted Regression Models*, 2020. URL `https://CRAN.R-project.org/package=gbm`. R package version 2.1.8.

[24] Andreas Holzinger, Anna Saranti, Christoph Molnar, Przemyslaw Biecek, and Wojciech Samek. Explainable AI Methods – A Brief Overview. In *Workshop on Extending Explainable AI Beyond Deep Models and Classifiers (ICML XXAI)*, 2022. doi:10.1007/978-3-031-04083-2_2.

[25] Farhad Imani, Ruimin Chen, Conrad Tucker, and Hui Yang. Random forest modeling for survival analysis of cancer recurrences. In *IEEE International Conference on Automation Science and Engineering (CASE)*, pages 399–404, 2019.

[26] Hemant Ishwaran and Udaya B. Kogalur. *randomForestSRC*, 2022. URL `https://cran.r-project.org/package=randomForestSRC`. R package version 3.1.1.

[27] Hemant Ishwaran, Udaya B. Kogalur, Eugene H. Blackstone, and Michael S. Lauer. Random survival forests. *The Annals of Applied Statistics*, 2(3):841–860, 2008. doi:10.1214/08-AOAS169.

[28] Ismail Jatoi, William F Anderson, Jong-Hyeon Jeong, and Carol K Redmond. Breast cancer adjuvant therapy: time to consider its time-dependent effects. *Journal of clinical oncology*, 29(17):2301, 2011.

[29] Neil Jethani, Mukund Sudarshan, Ian Connick Covert, Su-In Lee, and Rajesh Ranganath. FastSHAP: Real-Time Shapley Value Estimation. In *International Conference on Learning Representations (ICLR)*, 2022.

[30] Yunzhe Jia, Eibe Frank, Bernhard Pfahringer, Albert Bifet, and Nick Lim. Studying and Exploiting the Relationship Between Model Accuracy and Explanation Quality. In *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD)*, pages 699–714, 2021.

[31] Jared L Katzman, Uri Shaham, Alexander Cloninger, Jonathan Bates, Tingting Jiang, and Yuval Kluger. DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. *BMC Medical Research Methodology*, 18(1):1–12, 2018. doi:10.1186/s12874-018-0482-1.

[32] Been Kim, Rajiv Khanna, and Oluwasanmi O Koyejo. Examples are not enough, learn to criticize! criticism for interpretability. In *Neural Information Processing Systems (NeurIPS)*, 2016.

[33] Maxim Kovalev, Lev Utkin, Frank Coolen, and Andrei Konstantinov. Counterfactual Explanation of Machine Learning Survival Models. *Informatica*, 32(4):817–847, 2021. doi:10.15388/21-INFOR468.

[34] Maxim S. Kovalev and Lev V. Utkin. A robust algorithm for explaining unreliable machine learning survival models using the Kolmogorov–Smirnov bounds. *Neural Networks*, 132:1–18, 2020. doi:10.1016/j.neunet.2020.08.007.

[35] Maxim S. Kovalev, Lev V. Utkin, and Ernest M. Kasimov. SurvLIME: A method for explaining machine learning survival models. *Knowledge-Based Systems*, 203:106164, 2020. doi:10.1016/j.knosys.2020.106164.

[36] Håvard Kvamme, Ørnulf Borgan, and Ida Scheel. Time-to-Event Prediction with Neural Networks and Cox Regression. *Journal of Machine Learning Research*, 20(129):1–30, 2019.

[37] Changhee Lee, William Zame, Jinsung Yoon, and Mihaela van der Schaar. DeepHit: A Deep Learning Approach to Survival Analysis With Competing Risks. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2018. doi:10.1609/aaai.v32i1.11842.

[38] Seungyeoun Lee and Heeju Lim. Review of statistical methods for survival analysis using genomic data. *Genomics & Informatics*, 17(4), 2019.

[39] Yang Liu, Sujay Khandagale, Colin White, and Willie Neiswanger. Synthetic Benchmarks for Scientific Research in Explainable Machine Learning. In *Neural Information Processing Systems (NeurIPS Datasets and Benchmarks Track)*, 2021.

[40] Scott M Lundberg and Su-In Lee. A Unified Approach to Interpreting Model Predictions. In *Neural Information Processing Systems (NeurIPS)*, 2017.

[41] Scott M Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*, 2(1):56–67, 2020. doi:10.1038/s42256-019-0138-9.

[42] Susan Mallett, Patrick Royston, Rachel Waters, Susan Dutton, and Douglas G Altman. Reporting performance of prognostic models in cancer: a review. *BMC medicine*, 8(1):1–11, 2010.

[43] Tony S Mok, Yi-Long Wu, Sumitra Thongprasert, Chih-Hsin Yang, Da-Tong Chu, Nagahiro Saijo, Patrapim Sunpaweravong, Baohui Han, Benjamin Margono, Yukito Ichinose, et al. Gefitinib or carboplatin–paclitaxel in pulmonary adenocarcinoma. *New England Journal of Medicine*, 361(10):947–957, 2009.

[44] Arturo Moncada-Torres, Marissa C van Maaren, Mathijs P Hendriks, Sabine Siesling, and Gijs Geleijnse. Explainable machine learning can outperform Cox regression predictions and provide insights in breast cancer survival. *Scientific Reports*, 11(1):1–13, 2021.

[45] Chirag Nagpal, Willa Potosnak, and Artur Dubrawski. auton-survival: an Open-Source Package for Regression, Counterfactual Estimation, Evaluation and Phenotyping with Censored Time-to-Event Data. *arXiv preprint arXiv:2204.07276*, 2022.

[46] Wayne Nelson. Theory and Applications of Hazard Plotting for Censored Failure Data. *Technometrics*, 14(4): 945–966, 1972. doi:10.2307/1267144.

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

[48] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *International Conference on Knowledge Discovery and Data Mining (KDD)*, 2016. doi:10.1145/2939672.2939778.

[49] Greg Ridgeway. The state of boosting. *Computing Science and Statistics*, pages 172–181, 1999.

[50] Eliana Rulli, Francesca Ghilotti, Elena Biagioli, Luca Porcu, Mirko Marabese, Maurizio D'Incalci, Rino Bellocco, and Valter Torri. Assessment of proportional hazard assumption in aggregate data: a systematic review on statistical methodology in clinical trials using time-to-event endpoint. *British Journal of Cancer*, 119(12):1456–1463, 2018.

[51] Mark Robert Segal. Regression Trees for Censored Data. *Biometrics*, 44(1):35–47, 1988. doi:10.2307/2531894.

[52] Nikolai Sellereite, Martin Jullum, and Annabelle Redelmeier. *shapr: Prediction Explanation with Dependence-Aware Shapley Values*, 2022. https://norskregnesentral.github.io/shapr/, https://github.com/NorskRegnesentral/shapr.

[53] Lloyd Stowell Shapley. *A Value for n-Person Games*, pages 307–318. Princeton University Press, 2016. doi:10.1515/9781400881970-018.

[54] Brett Snider and Edward A McBean. Improving urban water security through pipe-break prediction models: Machine learning or survival analysis. *Journal of Environmental Engineering*, 146(3):04019129, 2020.

[55] Raphael Sonabend. *A Theoretical and Methodological Framework for Machine Learning in Survival Analysis. Enabling Transparent and Accessible Predictive Modelling on Right-Censored Time-to-Event Data*. Ph. D. Thesis, University College London, 2021. URL `https://discovery.ucl.ac.uk/id/eprint/10072700`.

[56] Raphael Sonabend. *survivalmodels: Models for Survival Analysis*, 2022. URL `https://CRAN.R-project.org/package=survivalmodels`. R package version 0.1.13.

[57] Annette Spooner, Emily Chen, Arcot Sowmya, Perminder Sachdev, Nicole A Kochan, Julian Trollor, and Henry Brodaty. A comparison of machine learning methods for survival analysis of high-dimensional clinical data for dementia prediction. *Scientific reports*, 10(1):1–10, 2020.

[58] Eric J Topol. High-performance medicine: the convergence of human and artificial intelligence. *Nature Medicine*, 25(1):44–56, 2019. doi:10.1038/s41591-018-0300-7.

[59] Lev V. Utkin, Egor D. Satyukov, and Andrei V. Konstantinov. SurvNAM: The machine learning survival model explanation. *Neural Networks*, 147:81–102, 2022. doi:10.1016/j.neunet.2021.12.015.

[60] Sebastiano Vigna. A Weighted Correlation Index for Rankings with Ties. In *International Conference on World Wide Web (WWW)*, pages 1166–1176, 2015. ISBN 9781450334693. doi:10.1145/2736277.2741088.

[61] Giulia Vilone and Luca Longo. Notions of explainability and evaluation approaches for explainable artificial intelligence. *Information Fusion*, 76:89–106, 2021. doi:10.1016/j.inffus.2021.05.009.

[62] Erik Štrumbelj and Igor Kononenko. An Efficient Explanation of Individual Classifications using Game Theory. *Journal of Machine Learning Research*, 11(1):1–18, 2010.

[63] Erik Štrumbelj and Igor Kononenko. Explaining Prediction Models and Individual Predictions with Feature Contributions. *Knowledge and Information Systems*, 41(3):647–665, 2014. doi:10.1007/s10115-013-0679-x.

[64] Fei Wan. Simulating survival data with predefined censoring rates under a mixture of non-informative right censoring schemes. *Communications in Statistics - Simulation and Computation*, 51(7):3851–3867, 2022. doi:10.1080/03610918.2020.1722838.

[65] Ping Wang, Yan Li, and Chandan K Reddy. Machine learning for survival analysis: A survey. *ACM Computing Surveys*, 51(6):1–36, 2019. doi:10.1145/3214306.

[66] Lianhe Zhao, Qiongye Dong, Chunlong Luo, Yang Wu, Dechao Bu, Xiaoning Qi, Yufan Luo, and Yi Zhao. DeepOmix: A scalable and interpretable multi-omics deep learning framework and application in cancer survival analysis. *Computational and Structural Biotechnology Journal*, 19:2719–2725, 2021. doi:10.1016/j.csbj.2021.04.067.

[67] Lili Zhao and Dai Feng. Deep Neural Networks for Survival Analysis Using Pseudo Values. *IEEE Journal of Biomedical and Health Informatics*, 24(11):3308–3314, 2020. doi:10.1109/JBHI.2020.2980204.