LICENTIATE THESIS

# Machine Learning Survival Models: Performance and Explainability

Abdallah Alabdallah

Halmstad University | School of Information Technology
LicentiateThesis | Halmstad University Dissertations no. 108


# Machine Learning Survival Models: Performance and Explainability

Abdallah Alabdallah

LICENTIATE THESIS | Halmstad University Dissertations no. 108

Machine Learning Survival Models: Performance and Explainability

Halmstad University Dissertations no. 108
ISBN 978-91-89587-30-4 (printed)
ISBN 978-91-89587-29-8 (pdf)
Publisher: Halmstad University Press, 2023 | www.hh.se/hup
Printer: Media-Tryck, Lund

# Machine Learning Survival Models: Performance and Explainability –

Abdallah Alabdallah

# Abstract

Abstract:

Survival analysis is an essential statistics and machine learning field in various critical applications like medical research and predictive maintenance. In these domains understanding models' predictions is paramount. While machine learning techniques are increasingly applied to enhance the predictive performance of survival models, they simultaneously sacrifice transparency and explainability.

Survival models, in contrast to regular machine learning models, predict functions rather than point estimates like regression and classification models. This creates a challenge regarding explaining such models using the known off-the-shelf machine learning explanation techniques, like Shapley Values, Counterfactual examples, and others.

Censoring is also a major issue in survival analysis where the target time variable is not fully observed for all subjects. Moreover, in predictive maintenance settings, recorded events do not always map to actual failures, where some components could be replaced because it is considered faulty or about to fail in the future based on an expert's opinion. Censoring and noisy labels create problems in terms of modeling and evaluation that require to be addressed during the development and evaluation of the survival models.

Considering the challenges in survival modeling and the differences from regular machine learning models, this thesis aims to bridge this gap by facilitating the use of machine learning explanation methods to produce plausible and actionable explanations for survival models. It also aims to enhance survival modeling and evaluation revealing a better insight into the differences among the compared survival models.

In this thesis, we propose two methods for explaining survival models which rely on discovering survival patterns in the model's predictions that group the studied subjects into significantly different survival groups. Each pattern reflects a specific survival behavior common to all the subjects in their respective group. We utilize these patterns to explain the predictions of the studied model in two ways. In the first, we employ a classification proxy model that can capture the relationship between the descriptive features of subjects and the learned survival patterns. Explaining such a proxy model using Shapley Values provides insights into the feature attribution of belonging to a

specific survival pattern. In the second method, we addressed the "what if?" question by generating plausible and actionable counterfactual examples that would change the predicted pattern of the studied subject. Such counterfactual examples provide insights into actionable changes required to enhance the survivability of subjects.

We also propose a variational-inference-based generative model for estimating the time-to-event distribution. The model relies on a regression-based loss function with the ability to handle censored cases. It also relies on sampling for estimating the conditional probability of event times. Moreover, we propose a decomposition of the C-index into a weighted harmonic average of two quantities, the concordance among the observed events and the concordance between observed and censored cases. These two quantities, weighted by a factor representing the balance between the two, can reveal differences between survival models previously unseen using only the total Concordance index. This can give insight into the performances of different models and their relation to the characteristics of the studied data.

Finally, as part of enhancing survival modeling, we propose an algorithm that can correct erroneous event labels in predictive maintenance time-to-event data. we adopt an expectation-maximization-like approach utilizing a genetic algorithm to find better labels that would maximize the survival model's performance. Over iteration, the algorithm builds confidence about events' assignments which improves the search in the following iterations until convergence.

We performed experiments on real and synthetic data showing that our proposed methods enhance the performance in survival modeling and can reveal the underlying factors contributing to the explainability of survival models' behavior and performance.

To my family.

# Acknowledgements

I would like to express my heartfelt gratitude to my supervisors, Mattias Ohlsson, Thorsteinn Rögnvaldsson, and Sepideh Pashami, for their unwavering support, guidance, and invaluable mentorship throughout my doctoral journey. Your expertise, patience, and commitment to my academic growth have been instrumental in shaping this thesis and my overall research experience. Your feedback, constructive criticism, and tireless dedication have pushed me to strive for excellence and have enriched my understanding of my field.

I am also deeply thankful to my colleagues and fellow researchers, whose camaraderie and collaborative spirit have made this academic journey more enjoyable and intellectually stimulating. Your discussions, insights, and shared experiences have broadened my perspective and have inspired me to explore new ideas and approaches.

I would like to extend my gratitude to the staff and faculty at Halmstad University, whose support and resources have played a crucial role in the successful completion of this research.

I am indebted to my family for their unwavering love and encouragement throughout this demanding endeavor.

This thesis would not have been possible without the support and contributions of all these individuals. Thank you all.

Abdallah Alabdallah
October, 2023

# List of Papers

The following papers, referred to in the text by their Roman numerals, are included in this thesis.

PAPER I: **SurvSHAP: A Proxy-Based Algorithm for Explaining Survival Models with SHAP**
Abdallah Alabdallah, Sepideh Pashami, Thorsteinn Rögnvaldsson, Mattias Ohlsson. (2022). IEEE 9th International Conference on Data Science and Advanced Analytics (DSAA)

PAPER II: **Understanding Survival Models through Counterfactual Explanations**
Abdallah Alabdallah, Jakub Jakubowski, Sepideh Pashami, Szymon Bobek, Mattias Ohlsson, M. Rögnvaldsson T., Grzegorz J. Nalepa. *submitted.*

PAPER III: **The Concordance Index Decomposition: A Measure for a Deeper Understanding of Survival Prediction Models**
Abdallah Alabdallah, Mattias Ohlsson, Sepideh Pashami, Thorsteinn Rögnvaldsson. *submitted.*

PAPER IV: **Discovering Premature Replacements in Predictive Maintenance Time-to-Event Data**
Abdallah Alabdallah, Thorsteinn Rögnvaldsson, Yuantao Fan, Sepideh Pashami, Mattias Ohlsson. In PHM Society Asia Pacific Conference 2023 Sep 11.

# Contents

# List of Figures

# 1. Introduction

## 1.1 Introduction

Survival analysis is a prominent branch of Statistics, used to analyze time-to-event data in various fields, like medical studies and predictive maintenance. The main challenge that gave rise to this field is the time-boundedness of time-to-event studies. This caused the target variable (time-to-event) to be not fully observed, a phenomenon referred to as censoring. This challenge complicates both the estimation of the time-to-event distribution, the evaluation, and the explanation of the estimation models.

However, although the target variable of the censored cases is not fully observed, it contains the information that the subject has survived up to a certain time, and including such information can correct part of the bias resulting from censoring and improve the performance of the prediction models. Many traditional and deep learning methods were proposed that can handle both, observed and censored cases, like [1–5] to name a few.

Survival analysis is mainly interested in estimating the event's distribution through estimating one of its related functions like the hazard function estimated by the Cox Proportional Hazards Model (CPH) [2], or the Cumulative Hazard function estimated by the Random Survival Forests (RSF) [3]. More recently, deep generative models adapted to handle censored examples were developed for events time distribution estimation and were shown to be powerful in capturing intricate patterns and relationships in large and complex datasets. The two most popular generative models' paradigms in machine learning, the Generative Adversarial Networks (GAN) [6] and Variational Autoencoders (VAE) [7] were extended to handle survival data. Namely, the Deep Adversarial Time-to-event model (DATE) [8] extends GAN utilizing a regression-based function consisting of two terms to handle the events and the censored cases separately. The Variational Survival Inference (VSI) model [9], is another method based on the variational inference to estimate a discrete survival time.

## 1.2 Challenges

One major difference between Survival models and regular machine learning models is the type of output. While machine learning models usually output a regression value or a class label prediction in regression and classification cases, respectively, Survival models usually output functions. This creates a challenge regarding explaining such models using the known machine learning explanation techniques, like Shapley Values, Counterfactual examples, and others.

Another major difference is the existence of censoring in the time-to-event data. Censoring is the phenomenon where the target variable (time-to-event) is not fully observable for all the subjects under study. Censoring gives survival models their special nature in terms of modeling and evaluation. More specifically, having partial information about the target variable prohibits the use of regular evaluation metrics like the mean squared error (MSE) leading to the use of less informative metrics like the Concordance Index (C-index) relying on ranking instead. The C-index as a summary statistic is the most used metric in survival analysis for its intuitive interpretation, and that it considers both observed and censored event cases. However, by relying on ranking, the C-index discards information about the actual times of occurrence of events. Also, the fact that it is computed based on the comparable pairs of observed and censored events, makes it hide information that can be insightful comparing seemingly similar models' performances.

From the modeling perspective, machine learning generative modeling is a promising technique for estimating the time-to-event distribution. However, this requires special handling for the censoring problem to allow the model to make use of the partial information present in the censored cases.

Moreover, in predictive maintenance settings, recorded events do not always map to actual failures, where some components could be replaced because it is considered faulty or about to fail in the near future based on an expert's opinion. Such premature replacements which are recorded as failures in the time-to-event data create some kind of noisy labels that can compromise the performance of the trained survival model.

## 1.3 Research questions

This thesis explores the different aspects of the performance and explainability of machine learning survival models. This research aims to address the following two main key research questions:

- How can survival models be explained? This can break into more con-

crete research questions: How different survival behaviors can be described with feature attributions? How to suggest alternative behavior for improving survival behavior? How can the differences in performance between survival models be understood?

- How to improve the performance of machine learning survival models? In more detail, how can advanced Machine Learning techniques like Generative models be incorporated to estimate time-to-event distribution? How to deal with noisy event labels in survival modeling e.g. in predictive maintenance settings.

By investigating these questions, this thesis seeks to contribute to the growing body of knowledge surrounding machine learning survival models and their application in real-world scenarios, with a focus on both performance enhancement and interpretability, ultimately advancing the field of survival analysis.

## 1.4 Contributions

- We have presented an algorithm to find Survival Patterns that can be used to identify risk groups that have significantly different survival behaviors (Paper I).

- Based on Survivl Patterns and Shapley Values, we have presented an algorithm that can explain the behavior of survival models which works for Proportional and Non-proportional hazards' models (Paper I).

- Based on Survivl Patterns, we presented an algorithm that can find plausible and actionable counterfactual explanations (Paper II).

- We derived a decomposition of the concordance index which showed that it is a harmonic weighted average of two quantities that can give a better understanding of survival models' performances (Paper III).

- We presented a new variational-inference-based generative survival model that enhances survival modeling and achieves performance comparable to the state of the art (Paper III).

- We presented an iterative algorithm based on survival analysis and genetic algorithms that can discover incorrectly labeled events which can enhance survival modeling (Paper IV).

## 1.5 Summary of the papers

- **Paper I: SurvSHAP: A Proxy-Based Algorithm for Explaining Survival Models with SHAP.**
  Survival models usually predict functions, like the survival or the hazard functions. In the case of non-proportional hazards, survival functions can intersect in which case the area under the survival curve (AUC) is not representative of the curve. This prohibits the use of the AUC as a single output value when applying regular machine learning explainability methods like Shapely Values. In this work, we propose an algorithm that discovers survival patterns in the predictions of a survival model. Such patterns represent subgroups of the population that are significantly different from the survival perspective where subjects that follow a certain pattern share similar survival characteristics. Based on the discovered patterns we employ a classification proxy model that learns the mapping between descriptive features and the survival patterns leading to a coarse approximation of the survival model. In the final step, we explain the proxy model with Shaple Values that produce feature attributions for each survival pattern which we consider as an explanation of the survival model.

- **Paper II: Understanding Survival Models through Counterfactual Explanations.**
  Based on the same algorithm proposed in Paper I, in this work, we extend the use of survival patterns with the help of Particle Swarm Optimization (PSO) to find counterfactual examples seeking the minimum change to the covariates, that changes the predicted survival function from one pattern to a predefined target pattern. One of the important aspects of a counterfactual example is the plausibility (or the likelihood) of the generated example. For that sake, we employed an autoencoder-based anomaly detection model to ensure that the generated counterfactual example is a plausible subject. Actionability is another important aspect that we considered in this work. In various scenarios, some features of the subject under study can not be changed like the age of the patient. In this regard, we introduced a mask to the optimization algorithm to block the change of certain features. In the end, the algorithm will find counterfactual examples that are plausible and actionable with minimum change to subjects' features.

- **Paper III: The Concordance Index Decomposition: A measure for a deeper understanding of survival prediction models.**
  The concordance Index is the most commonly used metric in survival

analysis that relies on comparing the concordance of pairs of subjects. However, when analyzing it, we found out that it is a harmonic weighted average of two quantities resampling the concordance in two subsets of pairs; event vs. event ($CI_{ee}$) and event vs. censored ($CI_{ec}$). The weight is the fraction of correctly ordered ($ee$) pairs out of the total correctly ordered pairs which we call $\alpha$. Two survival models can have different performances with respect to the decomposition terms while having similar C-index values. In this work, we propose the C-index decomposition as it gives a deeper understanding of models' differences previously unseen due to the averaging in the total C-index.

We also propose a variational-inference-based generative model with a regression-based loss function that can handle continuous survival time. We also utilized a ranking term in the loss function to encourage concordance.

- **Paper IV: Discovering Premature Replacements in Predictive Maintenance Time-to-Event Data.**
  In industrial settings, a significant fraction of component replacements are performed as a proactive response to the risk of failure. This results in noisy event labels in time-to-event data that can affect the use of such data to estimate the survival of the studied component. In this work, we propose an evolutionary-based iterative algorithm to discover premature replacements. The algorithm splits the data into two parts that will be used interchangeably in two phases. In the first phase (Expectation phase) one part of the data is used for fitting a survival model, and in the second phase (Maximization phase) the other part of the data is used to search for label assignments that maximize the fitted model performance. Starting with random labels, the algorithm iterates over these two phases and accumulates the found labels to seed the next iteration. Over iteration, the algorithm builds confidence in label assignments finding a significant fraction of the wrongly labeled events.

# 2. Background

## 2.1 Survival Analysis

Survival Analysis is the branch of Statistics focusing on the examination of time-to-event data. A significant challenge with time-to-event data is that the target variable (time) is partially unobserved for a considerable number of subjects under study. This phenomenon is commonly known as censoring. Censoring arises due to multiple reasons, primarily because of the finite duration of the study, during which some subjects (whether they are patients or machines) survive beyond the study's endpoint, as depicted in Figure 2.1. These subjects are referred to as censored cases.

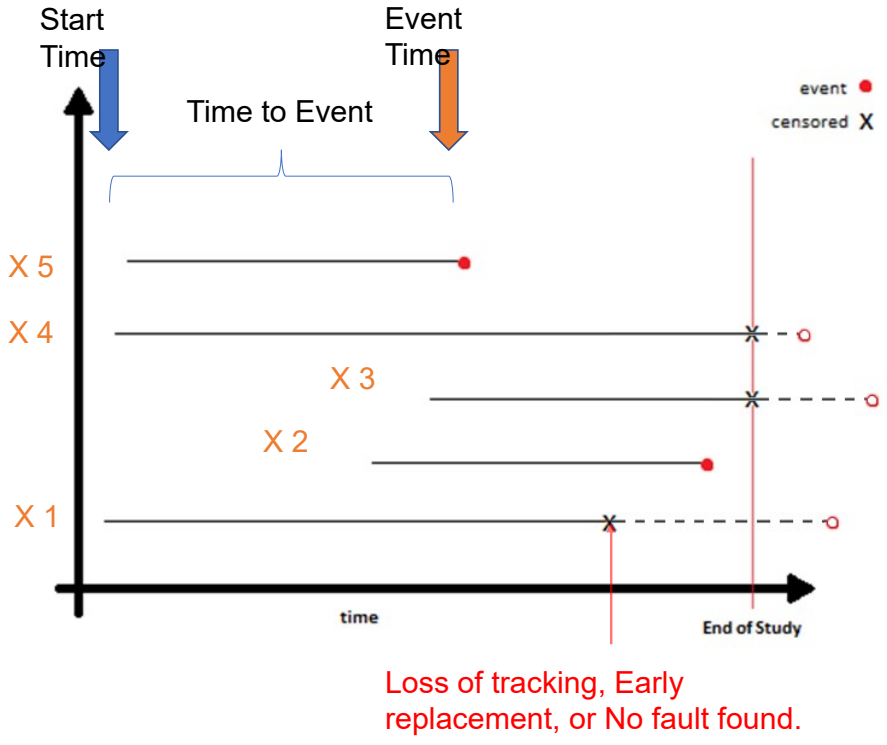



**Figure 2.1:** Time-To-Event Data.

The main outcome of survival analysis studies is the Survival Function $S(t)$, which represents the probability of surviving beyond time $t$:

$$S(t) = P(T > t) \tag{2.1}$$

where $T$ is the event time (e.g., time of death of a patient or time of failure of a machine).

The Kaplan-Meier estimator [1], also known as the product limit estimator, is the earliest method to estimate the survival function $S(t)$. It estimates the survival function in a non-parametric way:

$$\hat{S}(t) = \prod_{i:t_i \leq t} \left(1 - \frac{d_i}{n_i}\right), \tag{2.2}$$

where $d_i$ is the number of events occured at time $t_i$, and $n_i$ is the number of the subjects at risk at time $t_i$. The Kaplan-Meier estimator is a population-level estimator and does not consider the covariates $\mathbf{x}$ that describe subjects.

The earliest model to introduce an explicit dependence on $\mathbf{x}$ was the Cox Proportional Hazard model (CPH), as described by Cox in 1972 [2]. The CPH model assumes the existence of a baseline hazard function at the population level and that $\mathbf{x}$ has a linear and time-independent influence on the logarithm of the hazard function: og of the hazard function:

$$h(t|\mathbf{x}) = h_0(t)e^{\mathbf{w}^\intercal \mathbf{x}} \tag{2.3}$$

where $h_0(t)$ is an unknown baseline function, and $\mathbf{w}$ are the weights (parameters) that reflect the effect of the features on the hazard function.

Random Survival Forests (RSF), introduced by Ishwaran et al. in 2008 [3], is a machine learning technique that extends the Random Forests method proposed by Breiman in 2001 [10] to the domain of survival analysis. An RSF is composed of multiple survival trees, and its node-splitting criterion aims at maximizing the survival difference of the resulting nodes using the log-rank statistical test [11]. Each survival tree of the ensemble computes the Cumulative Hazard Function (CHF) for its leaf nodes in a non-parametric manner, based on the instances falling within those nodes during training. As a final outcome, RSF predicts the CHF for a subject by averaging the predictions generated by all the trees in the ensemble.

More recently, with the advancements achieved by deep learning techniques, many deep learning models were introduced for modeling survival time. One such model is DeepSurv, presented by Katzman et al. in 2018 [4]. DeepSurv is a direct extension of the Cox Proportional Hazard (CPH) model, where it replaces the CPH linear predictor with a deep neural network. Importantly, DeepSurv, like the CPH model, adheres to the proportional hazards assumption.

Other deep learning models adopt a discretized approach to the survival timeline. Notably, DeepHit, introduced by Lee et al. in 2018 [5], estimates the probability mass function based on discrete outputs.

An important aspect of survival analysis is handling *censored* cases, e.g., hospitalized patients who do not experience a relapse before the end of a study, equipment that is replaced before a breakdown, or equipment that has not experienced a breakdown yet. Censoring is very common in clinical studies and may arise for different reasons. It is possible for a patient not to experience the event (death or relapse, for example) during the time of the study. Also, a patient might experience a different event, making it impossible to follow up on the event of interest.

Censoring also creates the problem of evaluating the goodness of fit while the target variable is not fully observed. Several evaluation metrics have been proposed to measure different aspects of a model's performance [12]. However, the Concordance Index (C-index) is one of the most commonly used metrics that consider both events and censored cases. It quantifies the rank correlation between actual survival times and a model's predictions. Multiple estimators of the C-index have been proposed, like Harrel's C-index [13], Uno's C-index [14] that is a modified weighted version of Harrel's C-index, and Gonen and Heller's measure [15], which is an alternative estimator based on the reversed definition of concordance. A time-dependent version of the C-index was proposed in [16], which takes the whole survival function into consideration.

## 2.2 Explainability in Survival Analysis

Explainability is essential in machine learning models especially when the application domain involves high risk like healthcare and predictive maintenance. The increasing interest in explainability led to the development of many explanation methods that try to address different aspects of machine learning models' behavior. Some of these methods are model-specific and depend on the model's mechanisms to generate explanations like gradient-based methods which explain deep learning models [17–19]. However, model-agnostic methods, most notably LIME [20] and SHAP [21], gained more attention for their applicability to various types of machine-learning models. LIME depends on approximating the decision boundaries locally around the point of interest with a linear model and provides local explanations. On the other hand, SHAP method adopts a game-theoretic approach by computing the contributions of features to the difference between the model's prediction and the average prediction using Shapley Values [22]. SHAP also provides global explanations based on the aggregation of Shapley values of many instances [23]. The afore-

mentioned methods, in general, rely on feature attribution trying to explain the model's decision based on features' contribution to the output values.

SurvLIME [24] is an extension to LIME which uses the Cox Proportional Hazard model instead of the linear model in the vicinity of the exampled example. The SHAP method was also very recently extended in SurvSHAP(t) [25] to handle functional output models and provide time-dependent explanation.

Other directions rely on example-based explanations. One of the most interesting among them is the Counterfactual-Examples-based approach which tries to answer the "What if" question based on providing parallel-universe scenarios. Such examples provide insights into alternative paths on which different outcome is observed. Counterfactual examples can be generated, as proposed by [26], by finding the closest point to the original subject which satisfies the condition of changing the output of the model to a predefined target. However, such an approach can lead to unrealistic examples. To handle unrealistic examples, one can minimize the distance between the generated counterfactual example and the observed data [27]. A more efficient solution can be using an anomaly detection model, like Autoencoder-based models [28; 29], which relies on minimizing the reconstruction error of the generated counterfactual examples to ensure their likelihood.

For survival models, [30] proposed a method for finding counterfactual examples based on the mean survival time; i.e. the area under the curve (AUC) of the survival function. The method utilizes Particle Swarm Optimization (PSO) to find the minimum change to the input example that would change AUC of the predicted survival function to a predefined target value. However, they did not discuss the likelihood of the generated counterfactual examples.

# 3. Method

## 3.1 Paper I

In this work, we present SurvSHAP [31], a model-agnostic algorithm for explaining survival models. Given a trained survival model and its predicted survival functions (box 1 in Figure 3.1). The algorithm comprises three steps as shown in Figure 3.1(boxes 2 and 3):

- Distinct Survival Patterns Discovery.
- Proxy Model.
- Explanations.

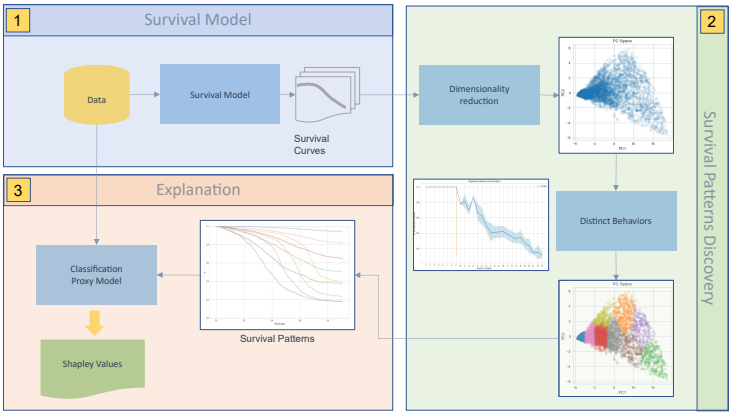



**Figure 3.1:** SurvSHAP workflow.

**Distinct Survival Patterns Discovery**: This step aims at subgrouping the survival model's predictions into the maximum number of distinct survival behaviors within the population. To achieve this, the algorithm employs a clustering approach with pairwise comparisons using the log-rank test [11]. Survival curves are represented as multidimensional vectors, and Principal Components Analysis (PCA) is used to create a lower-dimensional representation ($Z$) of these curves.

Subsequently, the algorithm searches for the maximum number of survival patterns by iteratively clustering the curves in the reduced-dimensional space (Z). At each iteration, it conducts log-rank pairwise comparisons between resulting clusters and calculates the percentage of significantly different groups relative to the total comparisons made. Finally, the algorithm determines $k*$, which represents the largest number of patterns that yield the highest percentage of unique survival patterns. This choice of 'k*' ensures that further division of sub-populations would not yield survival patterns distinguishable from one another.

**Proxy Model**: In this step we employ a classifier that learns the mapping between the inputs (X) and the discovered survival patterns (C). This makes the proxy model capture the coarse behavior of the survival model which transforms the survival model into a classification model.

**Explanations**: At this step, SHAP (SHapley Additive exPlanations) method is employed to explain the proxy model. These explanations offer descriptions of the discovered survival patterns in the model predictions and serve as explanations of the behavior of the survival model.

## 3.2 Paper II

Based on the same algorithm proposed in Paper I for discovering survival patterns, this work extends the use of survival patterns to explain survival models based on counterfactual explanations. For each studied subject, we utilize the Particle Swarm Optimization (PSO) algorithm to search for a counterfactual example optimizing an objective function that satisfies four criteria:

- Achieving Target Output Change: The generated counterfactual examples achieve the desired change in the survival model prediction.
- Minimal Input Change: It aims to make minimal changes to the input features.
- Plausible Counterfactuals: Counterfactual examples generated are plausible i.e., realistic.
- Actionable Counterfactuals: The generated counterfactual examples adhere to domain-specific constraints.

The workflow of this method is slightly different from the one mentioned in (Paper I). Utilizing the discovered Survival Patterns, we transform the problem into a classification task where the decision function $f(x)$, Equation 3.1, (the mapping from the features space to survival patterns) is the composition of

three functions as shown in Figure 3.2. $m_s$ is the survival model that predicts the survival curve, $h_z$ lowers the dimensionality of the predicted curve, and $g_c$ is the clustering model that predicts the survival patterns based on the nearest centroid.

$$f(\mathbf{x}) = (g_c \circ h_z \circ m_s)(\mathbf{x}) \tag{3.1}$$

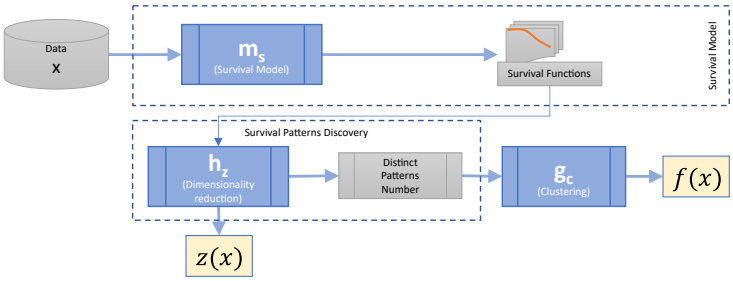



**Figure 3.2:** Survival Counterfactuals decision function parts $m_s$, $h_z$, and $g_c$.

Given a subject $\mathbf{x}$ and a target survival pattern $t$, the algorithm uses PSO to search for a counterfactual example $\mathbf{x}_{cf}$ that changes the predicted survival curve to the predefined target Survival Pattern.

The objective function consists of three weighted terms:

**Change in Target Output** ($\mathcal{L}_y$): Realizes the desired change to the target Survival Pattern.

$$\mathcal{L}_y = \mathbb{1}((f(\mathbf{x}_{cf}) \neq t)\|z(\mathbf{x}_{cf}) - c_t\|_2 \tag{3.2}$$

where $c_t$ is the centroid of the target survival pattern in the lower dimensional space and $z(\mathbf{x}_{cf}) = (h_z \circ m_s)(\mathbf{x}_{cf})$ is the lower dimensional representation of counterfactual example predicted survival curve.

**Minimal Input Change** ($\mathcal{L}_x$): Encourages minimal changes to input features.

$$\mathcal{L}_x(\mathbf{x}_{cf}) = \|\mathbf{x} - \mathbf{x}_{cf}\|_p \tag{3.3}$$

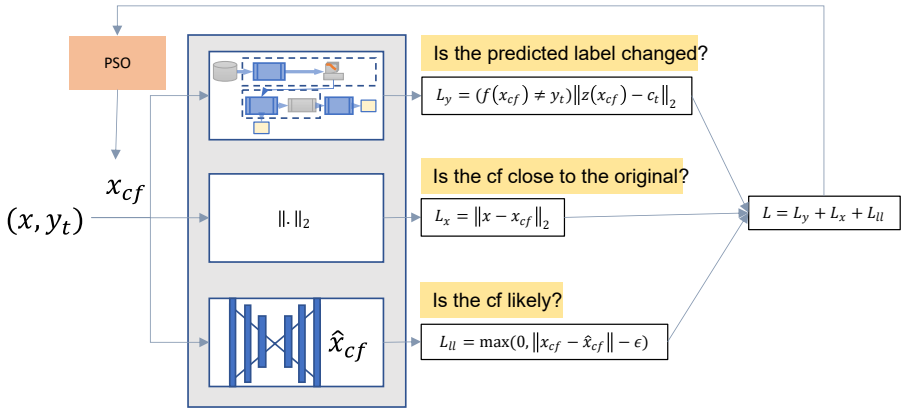
**Likelihood of Counterfactuals** ($\mathcal{L}_{LL}$): Ensures plausibility of counterfactual examples and considers the anomaly score of the Autoencoder model trained on the same dataset as the survival model.

$$\mathcal{L}_{AE} = \text{ReLU}(\|\mathbf{x}_{cf} - \mathbf{x}'_{cf}\|_p - A_t) \tag{3.4}$$

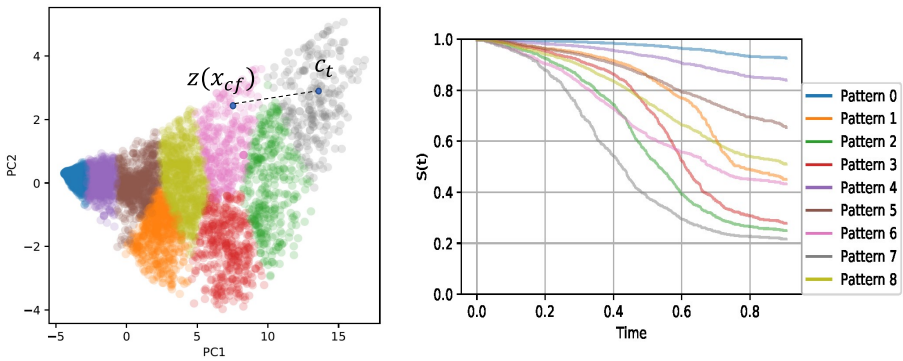
where $A_t$ is the anomaly threshold estimated based on the quantiles of the autoencoder reconstruction error and the *ReLU* function stops the effect of this

term when the anomaly score of the counterfactual example is less than the threshold.

Figure 3.3a shows the full counterfactual explanations' optimization workflow with the three parts of the objective function. It also shows in Figure 3.3b an illustration of Survival Patterns and their embedding space depicting the distance between the embedding of the survival function of the counterfactual example being optimized and the embedding of the Survival Pattern's center.




**(a)** Counterfactual Explanations workflow




**(b)** Survival Patterns (right) and their embedding space (left).

**Figure 3.3:** a) The full counterfactual explanations optimization workflow. b) The Survival Patterns with their representation in the embedding space color-coded and distance $\|z(\mathbf{x}_{cf}) - c_t\|_2$ between the counterfactual example embedding and the embedding of the target survival pattern's center.

The actionability of counterfactual examples is guaranteed by masking features that cannot be controlled in real-life applications, which are specified by domain experts. This constraint ensures that generated counterfactuals adhere to practical limitations.

Figure 3.4 shows counterfactual examples generated for two subjects with and without using the likelihood loss. It shows that the counterfactuals with the likelihood loss were closer to the distribution of the target survival pattern. This is also reflected in the anomaly scores of the counterfactuals shown in the figure.

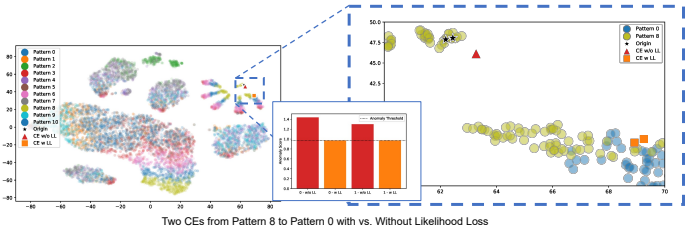



**Figure 3.4:** Counterfactual Examples with/out using the likelihood loss.

The method offers another option based on Survival Scores (area under the survival curve) which treats the survival problem as a regression problem. This option works well when survival curves do not intersect (proportional hazards).

## 3.3 Paper III

[32] The C-index measures the agreement between predicted and actual event times in survival analysis. It quantifies the probability that the predicted time for one subject exceeds that of another, given the actual event order. It is important to note that not all pairs can be compared when there is censoring. Pairs are only comparable when the earlier subject is an event, resulting in two types of comparable pairs event vs. event ($ee$) and event vs. censored ($ec$).

Mathematically, the C-index is represented as the probability of concordance ($CI = P(o)$), where $o$ denotes whether a pair is concordant or discordant.

We designate $CI_{ee}$ as the C-index for event-event instances and $CI_{ec}$ as the C-index for event-censored cases. Additionally, we introduce the symbol $\alpha$ to represent the conditional probability that a pair is an event-event pair ($ee$) given that it is correctly ordered.

$$CI_{ee} \equiv P(o|ee) \tag{3.5}$$

$$CI_{ec} \equiv P(o|ec) \tag{3.6}$$

$$\alpha \equiv P(ee|o) = 1 - P(ec|o) \tag{3.7}$$

The C-index can be written as a weighted harmonic average of the two terms $CI_{ee}$ and $CI_{ec}$ weighted by $\alpha$.

$$\frac{1}{CI} = \alpha \frac{1}{CI_{ee}} + (1-\alpha)\frac{1}{CI_{ec}} \tag{3.8}$$

The weight $\alpha$ denotes the fraction of correctly ordered *ee* pairs out of the total concordant pairs. It quantifies how much of the C-index relates to correctly ordering *ee* pairs compared to *ec* pairs and is influenced by the model's performance and dataset characteristics.

$\alpha^*$ represents the optimal value of $\alpha$ when all pairs are correctly ordered. A "balanced" predictor can achieve $\alpha = \alpha^*$ by equally scoring event-event and event-censored pairs.

We define the $\alpha$-Deviation as the difference between $\alpha$ and $\alpha^*$. A positive $\alpha$-Deviation indicates a predictor excelling in ordering *ee* pairs, while a negative $\alpha$-Deviation suggests better performance with *ec* pairs.

Figure 3.5 is a schematic illustration of the C-index decomposition terms in the space of comparable pairs. Interestingly, it shows the scores of three different models that have the same C-index, however, different scores with respect to the decomposed terms.

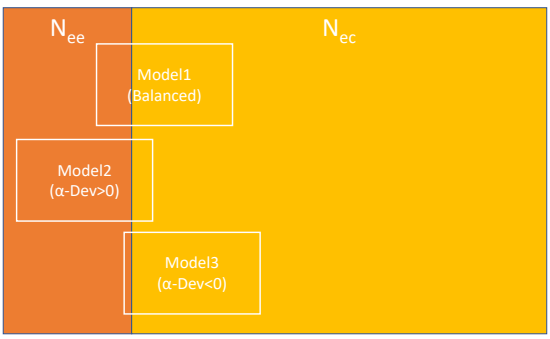



**Figure 3.5:** C-index Decomposition illustrated in the space of pairs.

On the modeling side, we propose a new generative survival model for estimating time-to-event distribution based on a variational encoder-decoder model. The model, $G_\theta(\mathbf{x})$, learns the conditional probability distribution of the event's accruing time $P(t|\mathbf{x})$ using a regression loss function consisting of four terms.

The first term $L_e$, Equation 3.9, is the mean absolute error estimated based on observed event cases. The second term $L_c$, Equation 3.10, is a truncated version of the mean absolute error, which only penalizes when the predicted event's time is less than the censoring time. The third term $L_{KL}$, Equation 3.11, is the regularization term active on the latent layer. Finally, $C_{lb}$, Equation 3.12

is the negative of a lower bound of the concordance index which, by minimizing, helps to maximize the concordance index.

$$L_e = \mathbb{E}_{\mathbf{x}\sim P_e(\mathbf{x})}\left[|t - G_\theta(\mathbf{x})|\right] \tag{3.9}$$
$$L_c = \mathbb{E}_{\mathbf{x}\sim P_c(\mathbf{x})}\left[\max(0, t - G_\theta(\mathbf{x}))\right] \tag{3.10}$$
$$L_{KL} = KL(P(\mathbf{z}|\mathbf{x}), N(0,1)) \tag{3.11}$$

$$C_{lb}(\theta, \varepsilon) = -\frac{1}{|\varepsilon|} \sum_{(\mathbf{x}_i, \mathbf{x}_j)\in\varepsilon} \left(1 + \frac{\log \sigma(G_\theta(\mathbf{x}_i) - G_\theta(\mathbf{x}_j))}{\log 2}\right) \tag{3.12}$$

The final loss function is a weighted sum of the four terms $L_e$, $L_c$, $L_{KL}$, and $C_{lb}$.

## 3.4 Paper IV

In predictive maintenance settings, a considerable number of components are replaced before they fail while recorded as failures in the maintenance log. From the perspective of Survival analysis, such premature replacements should be treated as a censored case. In order to discover such cases, in this work, we use the mentioned assumption and survival analysis to search for better labels of events that maximize the performance of the survival model. We propose an evolutionary-based iterative algorithm to discover premature replacements [33] in the log of time-to-event data assuming that we know a rough estimate of their fraction. The algorithm splits the data into two parts that will be used interchangeably in two phases, Expectation and Maximization.

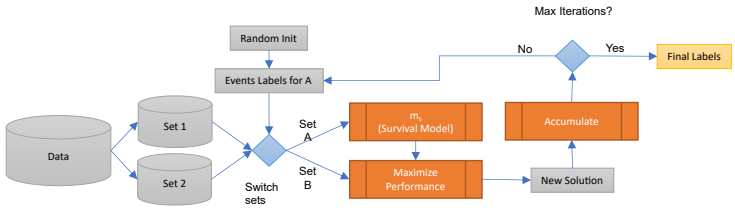



**Figure 3.6:** SurvPRD workflow

- Expectation Phase: A Cox Proportional Hazards Model (CPH) is fitted to the first part of the data (training data) with random events' labels at the first iteration. However, in the following iterations, we use the

aggregated solutions found in the previous iterations as a result of the Maximization phase.

- Maximization Phase: We use the fitted model in the Expectation phase, and use a genetic algorithm to find event assignments that maximize the performance of the model of the second part of the data (validation data). The found solution will be aggregated with the previous solutions found on this part of the data and used to seed the next iteration of the algorithm.

The two parts of the dataset are switched and the algorithm repeats for a certain number of iterations until convergence as shown in Figure 3.6.

Over iterations, the algorithm gains confidence in the labels where the subjects that are likely to be actual failures will be more frequently selected as events by the algorithm.

# 4. CONCLUDING REMARKS

## 4.1 Conclusion and future work directions

This thesis explores two tracks in machine learning survival analysis. The first focuses on the explainability of survival models in which we proposed two model-agnostic methods for post-hoc survival model explanations. The two methods are based on the same framework which finds survival patterns that distinctly categorize subjects into different survival behaviors that are significantly different from each other. In the first method, we utilized survival patterns to build a proxy classification model which we then explained with Shapley Values. Whereas in the second method, we search for counterfactual examples with a minimal change to the subjects of interest which changes the survival model's prediction from one survival pattern to a predefined target pattern.

Understanding machine learning models' behavior is the main goal of explainability methods where the focus is to understand the relation between the input and the output of the model. However, we believe that understanding the performance is as important which can lead to better modeling. In this regard, we proposed a decomposition of the concordance index which revealed unseen differences between models' performances. This decomposition showed that with smaller datasets with high censoring percentages, there were no big differences between classical and deep learning survival models. However, as the number of observed events increases, deep-learning models make better use of the events improving the ranking between event pairs and converging to a higher total C-index.

The second track of this thesis aims to improve survival models' performance. In particular, we proposed a continuous-time variational-inference-based generative model that learns the survival time distribution conditioned on the subject features using an encoder-decoder neural network structure. Moreover, in this track, we also explored the case of noisy event labels, a problem that is observed in industrial time-to-event data. We proposed an iterative algorithm based on genetic algorithms to discover such cases based on maximizing a surrogate survival model C-index performance.

As a future work, we are planning to bridge the gap between the aforementioned tracks, i.e., modeling and explanation. Self-explainable neural networks (SENN) [34] is a neural network structure that learns locally-linear approximation of the decision boundaries making the explanation an intrinsic property of the model. With added regularizations, such explanations can be controlled to meet certain desiderata of robustness. Our plan is to use SENNs for survival analysis, exploring their potential to produce faithful and stable explanations while maintaining good survival modeling.

On the other hand, there is an interesting growing direction in using evo-

lutionary algorithms for optimizing different aspects of neural networks where it is used to find new activation and loss functions [35; 36] that can outperform the out-of-the-box functions. Such direction is even more interesting for survival analysis due to the special nature of time-to-event data where loss functions are designed to achieve different goals in survival modeling. Such loss functions usually consist of multiple terms to handle observed and censored cases or encourage a better ranking in the predictions. Our plan is to explore the use of evolutionary algorithms in finding loss functions for neural-networks-based survival models which can enhance the performance and draw insights into the choices of different loss functions in different scenarios.

# References


[1] E. L. KAPLAN AND PAUL MEIER. **Nonparametric Estimation from Incomplete Observations**. *Journal of the American Statistical Association*, **53**(282):457–481, 1958. 1, 8

[2] D. R. COX. **Regression Models and Life-Tables**. *Journal of the Royal Statistical Society. Series B (Methodological)*, **34**(2):187–220, 1972. 1, 8

[3] HEMANT ISHWARAN, UDAYA B. KOGALUR, EUGENE H. BLACKSTONE, AND MICHAEL S. LAUER. **Random survival forests**. *Ann. Appl. Stat.*, **2**(3):841–860, 09 2008. 1, 8

[4] 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):24, 2018. 8

[5] CHANGHEE LEE, WILLIAM ZAME, JINSUNG YOON, AND MIHAELA VAN DER SCHAAR. **DeepHit: A Deep Learning Approach to Survival Analysis With Competing Risks**. *Proceedings of the AAAI Conference on Artificial Intelligence*, **32**(1), Apr. 2018. 1, 9

[6] IAN GOODFELLOW, JEAN POUGET-ABADIE, MEHDI MIRZA, BING XU, DAVID WARDE-FARLEY, SHERJIL OZAIR, AARON COURVILLE, AND YOSHUA BENGIO. **Generative Adversarial Nets**. In Z. GHAHRAMANI, M. WELLING, C. CORTES, N. LAWRENCE, AND K. Q. WEINBERGER, editors, *Advances in Neural Information Processing Systems*, **27**, pages 2672–2680. Curran Associates, Inc., 2014. 1

[7] DIEDERIK P. KINGMA AND MAX WELLING. **Auto-Encoding Variational Bayes**. *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014. 1

[8] PAIDAMOYO CHAPFUWA, CHENYANG TAO, CHUNYUAN LI, COURTNEY PAGE, BENJAMIN GOLDSTEIN, LAWRENCE CARIN DUKE, AND RICARDO HENAO. **Adversarial Time-to-Event Modeling**. In JENNIFER DY AND ANDREAS KRAUSE, editors, *Proceedings of the 35th International Conference on Machine Learning*, **80** of *Proceedings of Machine Learning Research*, pages 735–744, Stockholmsmä ssan, Stockholm Sweden, 10–15 July 2018. PMLR. 1

[9] ZIDI XIU, CHENYANG TAO, AND RICARDO HENAO. **Variational Learning of Individual Survival Distributions**. In *CHIL '20: Proceedings of the ACM Conference on Health, Inference, and Learning*, pages 10–18. ACM, 2020. 1

[10] LEO BREIMAN. **Random Forests**. *Machine Learning*, **45**(1):5–32, 2001. 8

[11] RICHARD PETO AND JULIAN PETO. **Asymptotically Efficient Rank Invariant Test Procedures**. *Journal of the Royal Statistical Society. Series A (General)*, **135**(2):185–207, 1972. 8, 11

[12] M. SHAFIQUR RAHMAN, GARETH AMBLER, BABAK CHOODARI-OSKOOEI, AND RUMANA Z. OMAR. **Review and evaluation of performance measures for survival prediction models in external validation settings**. *BMC Medical Research Methodology*, **17**(60), 2017. 9

[13] FRANK E. HARRELL JR., ROBERT M. CALIFF, DAVID B. PRYOR, KERRY L. LEE, AND ROBERT A. ROSATI. **Evaluating the Yield of Medical Tests**. *JAMA*, **247**(18):2543–2546, 05 1982. 9

[14] H. UNO, T. CAI, M.J. PENCINA, R.B. D'AGOSTINO, AND L.J. WEI. **On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data.** *Statistics in Medicine*, **30**(10):1105–1117, 2011. 9

[15] M. GÖNEN AND G. HELLER. **Concordance probability and discriminatory power in proportional hazards regression**. *Biometrika*, **92**(4):965–970, 2005. 9

[16] L. ANTOLINI, P. BORACCHI, AND E. BIGANZOLI. **A time-dependent discrimination index for survival data**. *Statistics in Medicine*, **24**(24):3927–3944, 2005. 9

[17] KAREN SIMONYAN, ANDREA VEDALDI, AND ANDREW ZISSERMAN. **Deep inside convolutional networks: Visualising image classification models and saliency maps**. *arXiv preprint arXiv:1312.6034*, 2013. 9

[18] ANH NGUYEN, ALEXEY DOSOVITSKIY, JASON YOSINSKI, THOMAS BROX, AND JEFF CLUNE. **Synthesizing the preferred inputs for neurons in neural networks via deep generator networks**. In D. LEE, M. SUGIYAMA, U. LUXBURG, I. GUYON, AND R. GARNETT, editors, *Advances in Neural Information Processing Systems*, **29**. Curran Associates, Inc., 2016.

[19] MUKUND SUNDARARAJAN, ANKUR TALY, AND QIQI YAN. **Axiomatic Attribution for Deep Networks**. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 3319–3328. JMLR.org, 2017. 9

[20] MARCO TULIO RIBEIRO, SAMEER SINGH, AND CARLOS GUESTRIN. **"Why Should I Trust You?" Explaining the Predictions of Any Classifier**. *In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)*, 2016. 9

[21] SCOTT M LUNDBERG AND SU-IN LEE. **A Unified Approach to Interpreting Model Predictions**. In I. GUYON, U. VON LUXBURG, S. BENGIO, H. WALLACH, R. FERGUS, S. VISHWANATHAN, AND R. GARNETT, editors, *Advances in Neural Information Processing Systems*, **30**. Curran Associates, Inc., 2017. 9

[22] L. S. SHAPLEY. *17. A Value for n-Person Games*, pages 307–318. Princeton University Press, Princeton, 1953 [cited 2023-08-29]. 9

[23] 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**:56–67, 2020. 9

[24] 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. 10

[25] MATEUSZ KRZYZIŃSKI, MIKOŁAJ SPYTEK, HUBERT BANIECKI, AND PRZEMYSŁAW BIECEK. **SurvSHAP(t): Time-dependent explanations of machine learning survival models**. *Knowledge-Based Systems*, **262**:110234, 2023. 10

[26] SANDRA WACHTER, BRENT MITTELSTADT, AND CHRIS RUSSELL. **Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR**. *Harvard Journal of Law & Technology*, 2017. 10

[27] SUSANNE DANDL, CHRISTOPH MOLNAR, MARTIN BINDER, AND BERND BISCHL. **Multi-Objective Counterfactual Explanations**. In *Parallel Problem Solving from Nature – PPSN XVI*, pages 448–469. Springer International Publishing, 2020. 10

[28] AMIT DHURANDHAR, PIN-YU CHEN, RONNY LUSS, CHUN-CHEN TU, PAISHUN TING, KARTHIKEYAN SHANMUGAM, AND PAYEL DAS. **Explanations based on the Missing: Towards Contrastive Explanations with Pertinent Negatives**. In S. BENGIO, H. WALLACH, H. LAROCHELLE, K. GRAUMAN, N. CESA-BIANCHI, AND R. GARNETT, editors, *Advances in Neural Information Processing Systems*, **31**. Curran Associates, Inc., 2018. 10

[29] ARNAUD VAN LOOVEREN AND JANIS KLAISE. **Interpretable Counterfactual Explanations Guided by Prototypes**. In *Machine Learning and Knowledge Discovery in Databases. Research Track: European Conference, ECML PKDD 2021, Bilbao, Spain, September 13–17, 2021, Proceedings, Part II*, page 650–665, Berlin, Heidelberg, 2021. Springer-Verlag. 10

[30] MAXIM KOVALEV, LEV UTKIN, FRANK COOLEN, AND ANDREI KONSTANTINOV. **Counterfactual Explanation of Machine Learning Survival Models**. *Informatica*, **32**(4):817–847, jan 2021. 10

[31] ABDALLAH ALABDALLAH, SEPIDEH PASHAMI, THORSTEINN RÖGNVALDSSON, AND MATTIAS OHLSSON. **SurvSHAP: A Proxy-Based Algorithm for Explaining Survival Models with SHAP**. In *2022 IEEE 9th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 1–10, 2022. 11

[32] ABDALLAH ALABDALLAH, MATTIAS OHLSSON, SEPIDEH PASHAMI, AND THORSTEINN RÖGNVALDSSON. **The Concordance Index decomposition: a measure for a deeper understanding of survival prediction models**, 2022. 15

[33] ABDALLAH ALABDALLAH, THORSTEINN RÖGNVALDSSON, YUANTAO FAN, SEPIDEH PASHAMI, AND MATTIAS OHLSSON. **Discovering Premature Replacements in Predictive Maintenance Time-to-Event Data**. In *Proceedings of the Asia Pacific Conference of the PHM Society 2023*, **4**, 2023. 17

[34] DAVID ALVAREZ MELIS AND TOMMI JAAKKOLA. **Towards Robust Interpretability with Self-Explaining Neural Networks**. In S. BENGIO, H. WALLACH, H. LAROCHELLE, K. GRAUMAN, N. CESA-BIANCHI, AND R. GARNETT, editors, *Advances in Neural Information Processing Systems*, **31**. Curran Associates, Inc., 2018. 20

[35] G. BINGHAM, W. MACKE, AND R. MIIKKULAINEN. **Evolutionary optimization of deep learning activation functions**. In *Proceedings of the 2020 Genetic and Evolutionary Computation Conference*, pages 289–296, 2020. 21

[36] S. GONZALEZ AND R. MIIKKULAINEN. **Improved training speed, accuracy, and data utilization through loss function optimization**. In *IEEE Congress on Evolutionary Computation (CEC)*, pages 1–8, 2020. 21

School of Information Technology

ISBN: 978-91-89587-30-4 (printed)
Halmstad University Dissertations, 2023

HALMSTAD
UNIVERSITY