Towards Trustworthy Survival Analysis with Machine Learning Models

Abdallah Alabdallah

DOCTORAL THESIS | Halmstad University Dissertations no. 128

Towards Trustworthy Survival Analysis with Machine Learning Models

Halmstad University Dissertations no. 128
ISBN 978-91-89587-73-1 (printed)
ISBN 978-91-89587-72-4 (pdf)
Publisher: Halmstad University Press, 2025 | www.hh.se/hup
Print: Media-Tryck, Lund

Printed by Media-Tryck, Lund 2025 NORDIC SWAN ECOLABEL 3041 0903

Abstract

Survival Analysis is a major sub-field of statistics that studies the time to an event, like a patient's death or a machine's failure. This makes survival analysis crucial in critical applications like medical studies and predictive maintenance. In such applications, safety is critical creating a demand for trustworthy models. Machine learning and deep learning techniques started to be used, spurred by the growing volume of collected data. While this direction holds promise for improving certain qualities, such as model performance, it also introduces new challenges in other areas, particularly model explainability. This challenge is general in machine learning due to the black-box nature of most machine learning models, especially deep neural networks (DNN). However, survival models usually output functions rather than point estimates like regression and classification models which makes their explainability even more challenging task.

Other challenges also exist due to the nature of time-to-event data, such as censoring. This phenomenon happens due to several reasons, most commonly due to the limited study time, resulting in a considerable number of studied subjects not experiencing the event during the study. Moreover, in industrial settings, recorded events do not always correspond to actual failures. This is because companies tend to replace machine parts before their failure due to safety or cost considerations resulting in noisy event labels. Censoring and noisy labels create a challenge in building and evaluating survival models.

This thesis addresses these challenges by following two tracks, one focusing on explainability and the other on improving performance. The two tracks eventually merge providing an explainable survival model while maintaining the performance of its black-box counterpart.

In the explainability track, we propose two post-hoc explanation methods based on what we define as Survival Patterns. These are patterns in the predictions of the survival model that represent distinct survival behaviors in the studied population. We propose an algorithm for discovering the survival patterns upon which the two post-hoc explanation methods rely. The first method, SurvSHAP, utilizes a proxy classification model that learns the relationship between the input space and the discovered survival patterns. The proxy model is then explained using the SHAP method resulting in per-pattern explanations. The second post-hoc method relies on finding counterfactual explanations that

would change the decision of the survival model from one source survival pattern to another. The algorithm uses Particle Swarm Optimization (PSO) with a tailored objective function to guarantee certain explanation qualities in plausibility and actionability.

On the performance track, we propose a Variational Encoder-Decoder model for estimating the survival function using a sampling-based approach. The model is trained using a regression-based objective function that accounts for censored instances assisted with a differentiable lower bound of the concordance index (C-index). In the same work, we propose a decomposition of the C-index where we found out that it can be expressed as a weighted harmonic average of two quantities; one quantifies the concordance among the observed event cases and the other quantifies the concordance between observed events and censored cases. The two quantities are weighted by a factor that balances the contribution of event and censored cases to the total C-index. Such decomposition uncovers hidden differences among survival models that seem equivalent based on the C-index. We also used genetic programming to search for a regression-based loss function for survival analysis with an improved concordance ability. The search results uncovered an interesting phenomenon, upon which we propose the use of the continuously differentiable Softplus function instead of the sharp-cut Relu function for handling censored cases. Lastly in the performance track, we propose an algorithm for correcting erroneous observed event labels that can be caused by preventive maintenance activities. The algorithm adopts an iterative expectation-maximization-like approach utilizing a genetic algorithm to search for better event labels that can maximize a surrogate survival model's performance.

Finally, the two tracks merge and we propose CoxSE a Cox-based deep neural network model that provides inherent explanations while maintaining the performance of its black-box counterpart. The model relies on the Self-Explaining Neural Networks (SENN) and the Cox Proportional Hazard formulation. We also propose CoxSENAM, an enhancement to the Neural Additive Model (NAM) by adopting the NAM structure along with the SENN loss function and type of output. The CoxSENAM model demonstrated better explanations than the NAM-based model with enhanced robustness to noise.

To my family.

Acknowledgements

I sincerely thank my supervisors, Sepideh Pashami, Mattias Ohlsson, and Thorsteinn Rögnvaldsson. Your support, guidance, and mentorship have been central to my doctoral journey. Your knowledge and dedication have shaped both this thesis and my research. I am especially grateful for your valuable feedback, constructive advice, and encouragement, which have helped me grow as a researcher.

I am also deeply thankful to my colleagues and peers. Your collaboration and friendship have made this journey so much richer. Our discussions, brainstorming sessions, and mutual support have made the work more enjoyable and sparked new ideas and directions.

A special thanks to the faculty and staff at Halmstad University for their ongoing support and for providing the resources I needed to complete this research.

Finally, I am forever grateful to my family for their unwavering encouragement throughout this challenging process.

This thesis would not have been possible without the contributions and support of all these amazing individuals. Thank you for being part of my journey.

Abdallah Alabdallah
December, 2024

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. IEEE 9th International Conference on Data Science and Advanced Analytics (DSAA) 2022.

PAPER II: **Understanding Survival Models through Counterfactual Explanations**
Abdallah Alabdallah, Jakub Jakubowski, Sepideh Pashami, Szymon Bobek, Mattias Ohlsson, M. Rögnvaldsson T., Grzegorz J. Nalepa. The 24th International Conference on Computational Science (ICCS) 2024.

PAPER III: **The Concordance Index Decomposition: A Measure for a Deeper Understanding of Survival Prediction Models**
Abdallah Alabdallah, Mattias Ohlsson, Sepideh Pashami, Thorsteinn Rögnvaldsson. Artificial Intelligence in Medicine Journal, 2024.

PAPER IV: **Improving Concordance Index in Regression-based Survival Analysis: Evolutionary Discovery of Loss Function for Neural Networks**
Mohammed Ghaith Altarabichi, Abdallah Alabdallah, Sepideh Pashami, Thorsteinn Rögnvaldsson, Sławomir Nowaczyk, Mattias Ohlsson. The Genetic and Evolutionary Computation Conference (GECCO) 2024.

PAPER V: **Discovering Premature Replacements in Predictive Maintenance Time-to-Event Data**
Abdallah Alabdallah, Thorsteinn Rögnvaldsson, Yuantao Fan, Sepideh Pashami, Mattias Ohlsson. The Asia Pacific Con-

ference of the Prognostics and Health Management Society (PHMAP) 2023.

PAPER VI: **CoxSE: Exploring the Potential of Self-Explaining Neural Networks with Cox Proportional Hazards Model for Survival Analysis**
Abdallah Alabdallah, Omar Hamed, Mattias Ohlsson, Thorsteinn Rögnvaldsson, Sepideh Pashami. *Submitted*.

Contents

List of Figures

1. Introduction

1.1 Introduction

Survival analysis originally arose in statistics as a set of techniques to study time-to-event data. The main challenge that gave rise to this field is the special nature of time-related studies where the timeline is bounded resulting in a phenomenon referred to as *censoring*. This phenomenon occurs when some studied subjects do not experience the event of interest during the study period. This causes these subjects to have recorded survival times less than their actual event times resulting in one type of censoring called *right censoring*. Other types of censoring exist, however, right censoring is the most common.

Many statistical and machine learning models have been proposed to study time-to-event data, which can handle censored cases, focusing on estimating various related quantities like the survival function or the hazard function [2–6]. Generative deep-learning models have also been introduced to estimate time-to-event distribution. More specifically, the two main paradigms, Generative Adversarial Networks (GAN) [7] and Variational Autoencoders (VAE) [8] were adapted with customized objective functions to handle censored examples in order to estimate individual survival distributions. Namely, the Deep Adversarial Time-to-event model (DATE) [9] extended the GAN model using a continuous-time regression-based objective function. The function is an extension of the Mean Squared Error (MSE) function which is used for observed event cases with an extra term to handle censored cases. On the other hand, the Variational Survival Inference (VSI) model [10] extended the VAE model predicting a discrete time-to-event distribution. Nevertheless, the use of generative models for survival modeling is still underexplored.

The improvements in survival modeling over time driven by the vast amounts of data collected and the advancements in AI systems focusing on enhancing model performance, raised concerns regarding user trust in these systems. Trustworthiness in AI systems is a broad topic that encompasses a wide range of factors and their interactions that must be addressed throughout the lifecycle of the AI system. Besides model accuracy, many aspects should be considered when building trustworthy AI systems including, explainability, robustness, transparency, fairness, reproducibility, and generalization [11; 12].

Trust is particularly crucial in survival analysis models due to their applications in medical studies and safety-critical areas like predictive maintenance. In this thesis, we narrow our focus to the explainability aspect while aiming to enhance the performance of survival models.

1.2 Challenges

Explainability is a complex subject, as no single explanation technique can fully address the diverse needs of all application areas. Various explanation techniques have been developed to interpret machine learning models, focusing on global or local explanations, feature attribution, example-based explanations like counterfactuals or prototypes, and more. Each of these methods highlights a specific aspect of the model's behavior. The improvements in survival analysis performance achieved by employing more advanced models and techniques, such as deep learning, have come at the expense of explainability. This is particularly concerning in survival analysis, where the sensitivity of application areas makes clarity and understanding essential for effective decision-making. Most survival models predict functions, like survival or hazard functions, rather than point estimates, as is common in regression or classification models. This has created a demand for explanation methods specifically tailored to survival models. Furthermore, post-hoc explanation methods require access to the training dataset to function effectively, emphasizing the need for models that can be trained to provide both predictions and explanations.

On the other hand, *Censoring* is the main challenge in survival analysis. Even though the time-to-event data for censored subjects is incomplete, it still provides valuable information about their survival beyond the censoring point. Instead of neglecting censored subjects, survival models use customized loss functions exploiting the partial information in these subjects. In regular regression analysis, where the target variable is fully observed, the Mean Squared Error (MSE) (or the Mean Absolute Error (MAE)) is usually used as a loss function. However, with the existence of censoring, a modification of the MSE function, referred to as the Mean Squared Censored Error (MSCE) [13], Equation 1.1, is used. This modification to the MSE function is intuitive and facilitates the use of censored cases up to their censoring time. It uses the MSE function for the observed event cases, similar to regular regression. Whereas, for the censored cases it uses a ReLU-shaped function which only penalizes when the prediction is less than the censoring time.

$$MSCE(t,\hat{t}) = \mathbb{E}_{\mathbf{z}\sim P_e(\mathbf{z})}[(t-\hat{t})^2] + \mathbb{E}_{\mathbf{z}\sim P_c(\mathbf{z})}[max(0,t-\hat{t})^2], \qquad (1.1)$$

where $\mathbf{z}$ represents the features of the subject, t is the time-to-event of subject z, and $\hat{t}$ is the predicted time-to-event. The notation $P_e(\mathbf{z})$ and $P_c(\mathbf{z})$, refers to the distributions of the observed events, and censored cases, respectively. The choice of the ReLU function for censored cases is motivated by the lack of information about the subject after the censoring time. However, a recent direction in research suggests going beyond the hand-engineered choices in neural networks [14–16], including functions like the activation [17] and the loss functions [18], and using search algorithms to find specialized better-performing functions.

Censoring not only complicates the estimation of time-to-event distribution but also makes the evaluation of the goodness of fit a challenging task. Many researchers refrain from using for example the MSE for evaluating survival prediction as it can only be used for observed events which is usually a small fraction of the data. Instead, other evaluation metrics that account for censored cases are used, most notably the concordance index (C-index) [19]. The C-index only considers the ranking of the predictions, however, its intuitive interpretation and holistic view of observed and censored cases make it a preferred metric in survival analysis. Nevertheless, the C-index treats observed and censored cases differently, considering only a pair of observations comparable if both are observed events or if the observed event is earlier than the censored case. This relation between these two types of comparable pairs complicates the models' comparison and creates the need for further analysis.

A new challenge arises in industrial settings where a machine component might be replaced before its failure due to safety constraints or false fault reports. Such replacements, termed premature replacements, introduce noise to the time-to-event data as the replacement time is considered a failure while the replaced component is still healthy. Discovering and correcting such cases is crucial for better understanding the survival behavior of the studied component.

1.3 Research Questions

This thesis is centered around the trustworthiness of survival models focusing on explainability and performance.

In this work, we pursued two primary directions: one aims to enhance the model's explainability, while the other seeks to improve its predictive performance, with the ultimate goal of combining these directions to develop inherently explainable models while maintaining high performance. Accordingly, the research questions guiding this thesis are framed as follows:

1. How do we explain the predictions of an existing survival model? In

particular, this question focuses on the behavior side of the mode, and how to handle the survival models' special nature.

2. How do we explain the performance differences (in terms of the evaluation metric) among survival models?

3. How to optimize Neural Networks for time-to-event distribution estimation?

4. How do we deal with noisy event labels in time-to-event data?

5. How can we approach building a high-performing survival model that is explainable by design?

1.4 Contributions

By addressing the aforementioned research questions, our contributions can be conceptualized as two overlapping areas, as depicted in Figure 1.1. One area emphasizes explainability, while the other focuses on modeling. The final work, situated in the overlapping region, integrates both aspects, thus concluding the thesis by simultaneously addressing modeling and explainability.

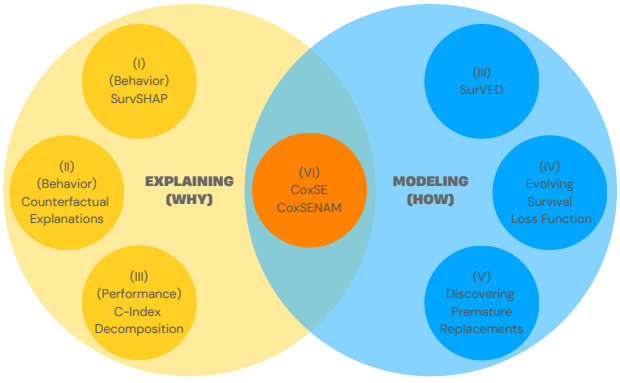

Figure 1.1: Overall picture of the thesis

The following is a detailed list of contributions of this thesis:

- We proposed an algorithm for discovering patterns in a survival model's predictions, denoted Survival Patterns, which identify significantly dif-

ferent survival behaviors in the studied populations. Based on such patterns we propose two explanation algorithms:

  - Using Shapley Values, in Paper I, we propose SurvSHAP, a proxy-based algorithm for explaining survival models with feature attribution answering for the research question 1.
  - Utilizing Particle Swarm Optimization (PSO) with a customized loss function, in Paper II, we presented an algorithm to find plausible and actionable counterfactual examples that suggest alternative survival behavior answering the research question 1.

- In Paper III, we derived a decomposition of the C-index that revealed unseen differences between seemingly similar survival models. It also helped in understanding the behavior of different survival models in response to different dataset characteristics. This work addresses answering the research question 2.

- In Paper III, we proposed a variational-inference-based deep-learning generative model for estimating time-to-event distribution utilizing a regression-based loss function combined with a ranking loss. This work seeks to address the research question 3.

- Building on the regression-based loss function employed in Paper III, we investigated the potential for improving its ranking capabilities. In Paper IV we utilized genetic algorithms to search for specialized survival analysis regression-based loss functions for neural networks. This work contributes to answering the research question 3.

- In Paper V, we proposed an algorithm for discovering premature replacements in time-to-event industrial data. This work addresses answering the research question 4.

- In Paper VI, relaying on the Self-Explaining Neural Networks (SENN) and the Cox Proportional Hazards model, we proposed CoxSE, an explainable model with a comparable performance to its black-box counterpart. We also proposed CoxSENAM as an enhancement to the Neural Additive Model (NAM), CoxNAM, by combining the NAM structure with the SENN loss function and type of output. This work explores the pros and cons of these models in terms of faithfulness, stability, and robustness to noise, aiming to answer the research question 5.

1.5 Summary of the Papers

- **Paper I: SurvSHAP: A Proxy-Based Algorithm for Explaining Survival Models with SHAP.** [20]
  Survival models usually predict functions that represent the probability of survival over time. This prohibits the use of regular machine learning explanation methods that assume single prediction like regression or classification models. To address this challenge, we propose an algorithm that relies on finding patterns in the predictions of the survival model. Such patterns, denoted Survival Patterns, represent subpopulations that are significantly different from the survival perspective, while subjects within each pattern share similar survival behavior from the survival model's perspective. Figure 1.2 shows an overall workflow of the SurvSHAP algorithm. The number of Survival Patterns is auto-

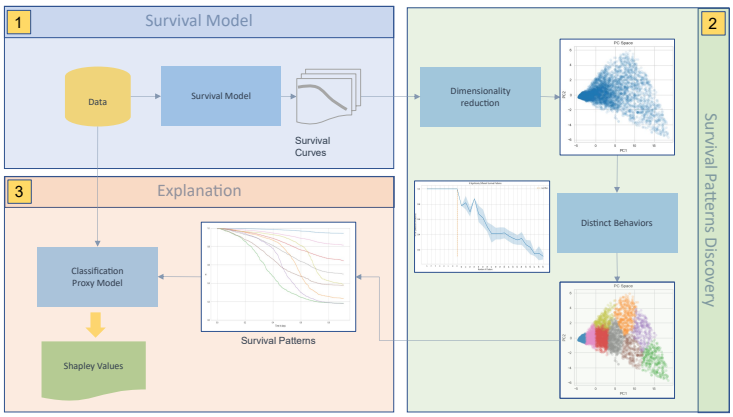

Figure 1.2: The SurvSHAP algorithm workflow [1].

matically decided by iteratively clustering the survival curves with an increasing number of clusters. Consequently, the algorithm selects the maximum number of patterns maintaining that they are significantly different from each other using the log-rank test. The resulting patterns are then used to train a proxy classification model that learns the relation between the descriptive features and the Survival Patterns. The resulting proxy model is a coarse approximation of the survival model. Finally, we use the SHAP method to explain the proxy model resulting in Shapley values quantifying the effect of a subject's features on the probability of following a certain Survival Pattern.

- **Paper II: Understanding Survival Models through Counterfactual Explanations.** [21]
  This work addresses the explanation from a different perspective relying on counterfactual explanations. Using the same algorithm for discovering Survival Patterns proposed in Paper I, the algorithm uses Particle Swarm Optimization (PSO) to find a counterfactual example for a given subject that changes the prediction of the survival model from one source survival pattern to a target one.

  Utilizing the discovered Survival Patterns, we transform the problem into a classification task. Two functions are required for the optimization process. The decision function $f(x)$, Equation 1.2, (the mapping from the features space to survival patterns) and the embedding function $z(x)$, Equation 1.3, that predicts the survival function and transform it to lower dimensional space.

$$f(\mathbf{x}) = (g_c \circ h_z \circ m_s)(\mathbf{x}) \tag{1.2}$$

$$z(\mathbf{x}) = (h_z \circ m_s)(\mathbf{x}) \tag{1.3}$$

  The two functions as part of the workflow are shown in Figure 1.3.

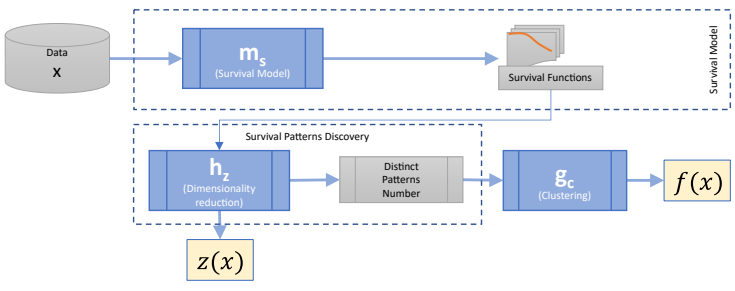

Figure 1.3: Survival Patterns decision and embedding functions [1].

  The algorithm employs the decision and embedding functions in a customized loss function with multiple parts to ensure certain qualities of the generated counterfactual examples as depicted in Figure 1.4.

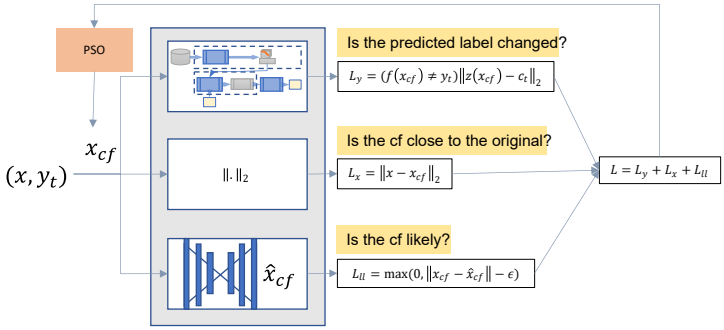

Figure 1.4: The Counterfactual Explanations workflow [1].

The first part ensures a minimal change in the input features while the second part ensures that the target survival pattern is met. Plausibility, or likelihood, is an important quality of counterfactual examples desired to ensure realistic generated examples. In this work, we utilized an autoencoder anomaly detection model to guarantee the likelihood of the generated counterfactual examples. The last quality addressed in this work is actionability. In many practical scenarios, some features cannot be changed like the age of a patient. This requires expert knowledge to decide, however, the algorithm provides a masking functionality that specifies the features that can be changed.

- **Paper III: The Concordance Index Decomposition: A measure for a deeper understanding of survival prediction models.** [22]
  The concordance index (C-index) is arguably the most commonly used metric in survival analysis. It quantifies the ranking agreement between the survival times and the predicted scores of a survival model. The C-index is a generalization of the area under the ROC curve (AUC) that accounts for censored cases. It is computed based on pair-wise comparisons between the observations. However, due to the lack of information, some pairs are not comparable. This is when the two subjects are censored cases or when the censoring time of one subject happens before the event time of the other. As such given sorted pairs, only two types of pairs are comparable: event vs. event (ee) and event vs. censored (ec). In this work, we analyzed the C-index and concluded that it can be expressed as a weighted harmonic average of two concordance quantities related to the two types of pairs CI_{ee} and CI_{ec}. The two C-indices are weighted by a third quantity, denoted α, that represents the fraction of correctly ordered *ee* pairs out of the total correctly ordered pairs. Based

on this analysis we propose the C-index decomposition as:

$$\frac{1}{CI} = \alpha \frac{1}{CI_{ee}} + (1-\alpha)\frac{1}{CI_{ec}} \tag{1.4}$$

The performance of a survival model depends on how well it handles these two types of pairs and the balance between them. Our analysis showed that two survival models can have different performances with respect to the terms of the decomposition while having similar total C-index values. Moreover, this decomposition helped to explain the superior performance of deep-learning-based survival models in one case where a fairly large dataset with a higher fraction of observed event cases was used.

We also propose a generative model based on a variational encoder-decoder neural network for estimating the time-to-event distribution. The model employed a regression-based loss function that was adapted to handle censored cases. Observed event cases can be handled using a regular regression function like the Mean Squared Error (MSE) or Mean Absolute Error (MAE). However, for censored cases, the ReLU function is used where it only penalizes if the predicted time is before the censoring time. To boost concordance, a ranking loss based on a differentiable lower bound of the C-index was used.

- **Paper IV: Improving Concordance Index in Regression-based Survival Analysis: Evolutionary Discovery of Loss Function for Neural Networks** [23]
  Inspired by a recent research direction for optimizing various aspects of neural networks like activation functions or loss functions, this work aims at exploring the potential of improving the concordance performance of the regression-based loss function used in Paper III, denoted Mean Squared Censored Error (MSCE), shown in Figure 1.5. We used genetic programming to search for specialized loss functions for various bench-mark time-to-event datasets optimizing the C-index. The loss function $\mathcal{L}(.)$ of the neural network is represented as an expression tree consisting of two main branches, one for handling observed events and the other for censored cases:

$$\mathcal{L}(t-\hat{t}) = \mathbb{E}_{\mathbf{z}\sim P_e(\mathbf{z})}[f(t-\hat{t})] + \mathbb{E}_{\mathbf{z}\sim P_c(\mathbf{z})}[g(t-\hat{t})] \tag{1.5}$$

  where $(\mathbf{z},t)$ is the subject feature vector and its corresponding survival time, and P_e and P_c represent the distributions of the observed and censored cases respectively. Upon analyzing the resulting loss functions

discovered on multiple datasets, we proposed an off-the-shelf loss function, denoted SoftPlus Mean Squared Censored Error (MSCE_{SP}) shown in Figure 1.6, utilizing MSE for the observed events while using SoftPlus for the censored cases instead of the ReLU function used in the MSCE.

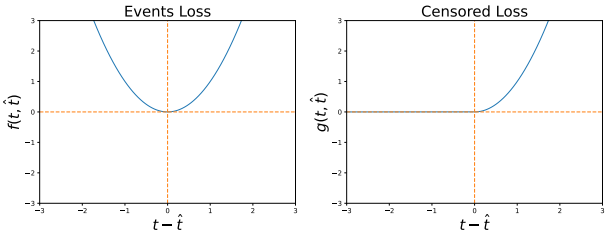

Figure 1.5: Mean Squared Censored Error (MSCE)

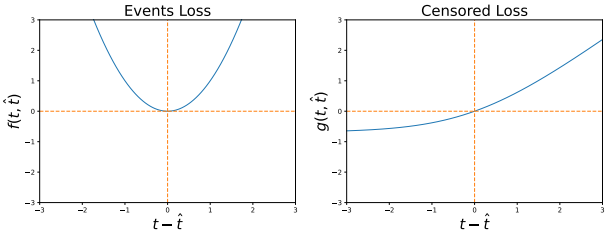

Figure 1.6: SoftPlus Mean Squared Censored Error (MSCE_{SP})

Unlike the ReLU function, the SoftPlus is a smooth function at zero with a slight gradient in the negative range. Upon analysis, such a gradient was shown to be effective in boosting the concordance in the predictions of the learned models.

- **Paper V: Discovering Premature Replacements in Predictive Maintenance Time-to-Event Data.** [24]
  In industrial environments, many component replacements are carried out proactively to mitigate the risk of failure or based on incorrect fault reports. These approaches often introduce noise in time-to-event data due to inaccurately labeled events, potentially affecting the accuracy of survival estimates for the studied component. In this work, we propose an iterative expectation-maximization-like algorithm designed to identify premature replacements utilizing genetic algorithms. The proposed

algorithm operates in two alternating phases, depicted in Figure 1.7: the Expectation phase and the Maximization phase. It divides the data into two parts, which are used interchangeably across two phases. In the Expectation phase, a surrogate survival model is fitted using one subset of the data. In the Maximization phase, the other subset is employed to search for label assignments that optimize model performance. Starting with random labels, the algorithm iteratively alternates between these two phases accumulating and aggregating the previous solutions to seed the next iteration. With each cycle, the algorithm increases confidence in the label assignments, ultimately identifying a significant portion of misclassified events.

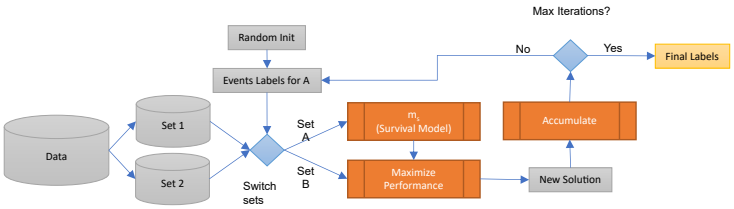

Figure 1.7: Discovering premature replacements algorithm [1]

- **Paper VI: CoxSE: Exploring the Potential of Self-Explaining Neural Networks with Cox Proportional Hazards Model for Survival Analysis** [25]
  In this work, we propose CoxSE, a self-explaining survival model adopting the Self-Explaning Neural Network (SENN) and the Cox Proportional Hazards assumption. The model learns the relevances of the features at each point by approximating the log-risk as a locally-linear function based on the Cox model formulation:

$$h(t,\boldsymbol{x}) = h_0(t)e^{f(\boldsymbol{x})} \tag{1.6}$$

$$f(\boldsymbol{x}) = \sum_{x_i \in \boldsymbol{x}} w_i(\boldsymbol{x})x_i \tag{1.7}$$

where $h_0(t)$ is an unspecified baseline hazard function, $f(\boldsymbol{x})$ is the log-risk function, and $w_i(\boldsymbol{x})$ is the relvence of the feature x_i. The model utilizes the partial likelihood function used in Cox-based models with extra regularization of the learned relevance vector $\boldsymbol{w}$ to encourage stability and robustness of the explanations. The structure of the CoxSE model is shown in Figure 1.8.

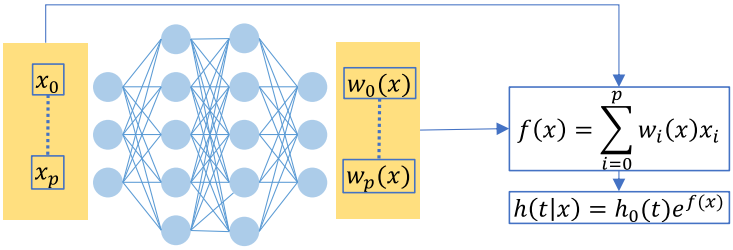

Figure 1.8: The CoxSE structure

We also propose another model, CoxSENAM, adopting the structure of Neural Additive Models (NAM) with the output and loss function of SEMM as shown in Figure 1.9. In the CoxSENAM, the log-hazard function is formulated as:

$$f(\boldsymbol{x}) = \sum_{x_i \in \boldsymbol{x}} w_i(x_i)x_i \tag{1.8}$$

where the relevance of each feature w_i only depend of the feature x_i.

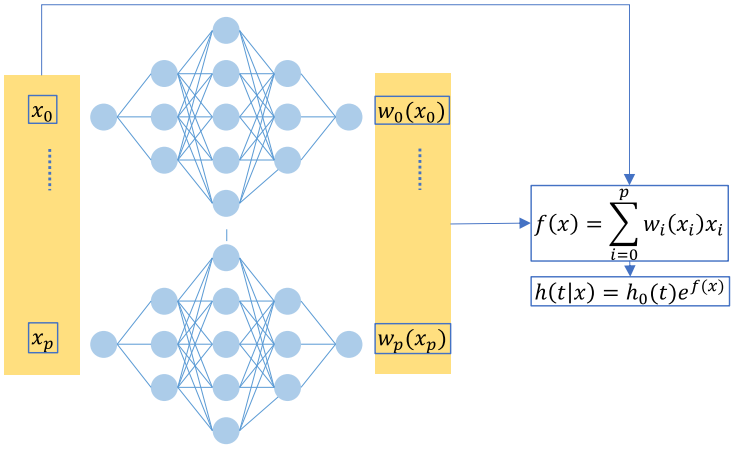

Figure 1.9: The CoxSENAM structure

The CoxSE model was shown to achieve performance comparable to its black-box counterpart while providing stable explanations. On the other hand, CoxSENAM demonstrated better robustness to noise; however, its formulation limited its ability to account for feature interactions, leading to poorer performance in such cases.

2. Background

2.1 Survival Analysis

Survival analysis models focus on studying the time of an event of interest. Such an event can be the death of a patient or the failure of a machine. Time-to-event studies are usually conducted by following a group of subjects for a certain period. During the study period, some subjects experience the event, causing their time-to-event values to be known. However, some subjects will not experience the event, resulting in censored time-to-event values, as depicted in Figure 2.1. Censoring is the main challenge that faces time-to-event studies and can occur for various reasons, such as the study ending, patients dropping out, or early replacement of a component.

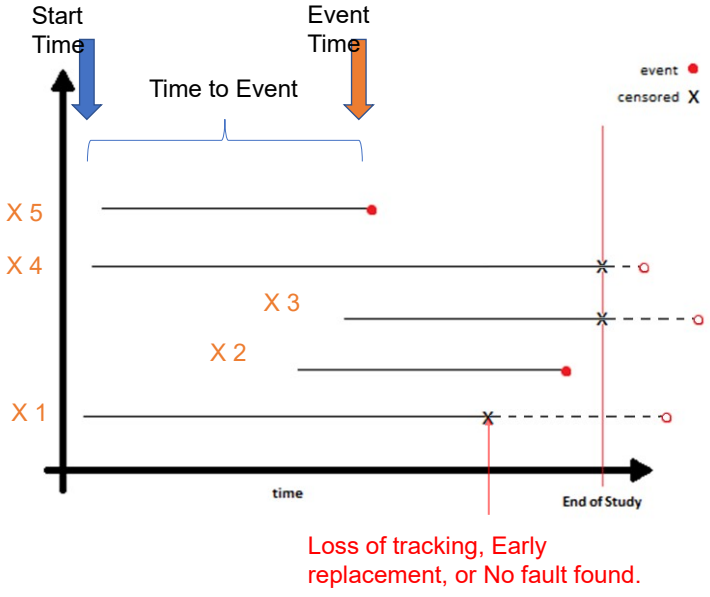

Figure 2.1: Time-To-Event Data. [1].

Given a random variable representing the event time, denoted T, the probability of the event occurring before a specific time t is $F(t) = P(T \leq t)$ referred

to as the cumulative distribution function (CDF). This can be written in terms of the event density function $f(t)$:

$$F(t) = \int_{\tau=0}^{t} f(\tau) d\tau \tag{2.1}$$

The survival function $S(t) = P(T > t)$, defined as the probability of surviving beyond time t, is the complementary CDF:

$$S(t) = 1 - F(t) \tag{2.2}$$

The hazard function $h(t)$ and the cumulative hazard function $H(t)$ are other quantities some survival models aim to estimate. The hazard function represents the instantaneous event's rate in an infinitesimal time interval δt given that the subject survived up to that time.

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t} \tag{2.3}$$

and the cumulative hazard function (CHF) $H(t)$

$$H(t) = \int_{0}^{t} h(\tau) d\tau \tag{2.4}$$

All these functions are related where:

$$S(t) = e^{-H(t)} \tag{2.5}$$

The earliest survival function estimator in the presence of censored cases is the Kaplan-Meier estimator [2]. It is a non-parametric estimator also known as the product limit estimator.

$$\hat{S}(t) = \prod_{i:t_i \leq t} \left(1 - \frac{d_i}{n_i}\right), \tag{2.6}$$

here d_i represents the number of events occurring at time t_i, and n_i denotes the number of subjects at risk at time t_i. The Kaplan-Meier estimator is a population-level estimator that does not account for the features' vector $\mathbf{x}$ that describes individual subjects.

the Cox Proportional Hazards model (CPH) [3] is the earliest model to introduce an explicit dependence on $\mathbf{x}$. The CPH model assumes a population-level baseline hazard function $h_0(t)$ independent from the features, and a time-independent log-risk that is linear in $\mathbf{x}$:

$$h(t|\mathbf{x}) = h_0(t) e^{\mathbf{w}^\intercal \mathbf{x}} \tag{2.7}$$

where **w** are the weights (parameters) that reflect the effect of the features on the hazard function. This formulation of the hazard function imposes a strong assumption that the effect of the features remains constant over time, implying that the hazard ratio between two subjects is constant and depends only on the difference in their features. In other words, the hazards are proportional, giving the model its name.

Over time, several machine learning and deep learning models have been proposed to estimate various survival-analysis-related functions. For example, Random Survival Forests (RSF) [4] extended Random Forests [26] to estimate the CHF. [27] extended the Support Vector Machine model to survival analysis combining ranking and regression of survival time. [28] proposed a gradient-boosting algorithm to estimate the hazard function in a non-parametric way optimizing a soft version of the concordance index.

More recently however, deep learning models have been introduced for survival time modeling [5; 6; 9; 10; 29–34]. Most notably, DeepSurv [5] extended the CPH model utilizing a non-linear log-risk function modeled as a neural network. Although DeepSurv introduced an improved performance over the linear CPH, it maintained the Proportional Hazards assumption. Unlike DeepSurv however, some deep learning models discretize the survival timeline. For example, DeepHit [6] estimates the probability mass function (PMF) based on a discrete output maximizing the likelihood function assessed with a ranking loss function.

Other models leverage deep generative approaches to estimate the distribution of the event times in both parametric and non-parametric ways [29; 30]. The two most notable approaches in deep generative models, Generative Adversarial Networks (GANs) [7] and Variational Autoencoders (VAEs) [8], have been extended to the domain of survival analysis. The Deep Adversarial Time-to-Event model (DATE) [9] is a survival model based on GAN that estimates the event distribution in a non-parametric manner using adversarial training. It is trained to generate $p(t|\mathbf{x})$ while penalizing fake samples $(\mathbf{x},t)$. The model utilizes a regression-based function with two terms to handle the observed events and censored cases separately. On the other hand, the Variational Survival Inference (VSI) model [10] was introduced, adopting variational inference to approximate $p(t|\mathbf{x})$. VSI is a discrete-time model that employs two encoders, $p(z|\mathbf{x})$ and $q(z|\mathbf{x},t)$, and encourages these two distributions to be similar by using Kullback-Leibler divergence which helps the model to better account for interactions between covariates and event times.

Evaluating the goodness of fit of survival models is a challenge due to the censoring problem. Various evaluation metrics have been developed to assess different aspects of a model's performance [35]. Among these, the Concordance Index (C-index) is the most widely used metric, accounting for events

and censored cases. The C-index is a measure of the probability that the predicted event times ($\hat{t}_i$, $\hat{t}_j$) of two randomly selected subjects maintain the same relative order as their true event times (t_i, t_j), i.e., $P(\hat{t}_i > \hat{t}_j | t_i > t_j)$. It is essential to recognize that not all pairs can be compared in the presence of censoring. A pair $(\mathbf{x}_i, \mathbf{x}_j)$ is considered comparable (usable) if the earliest time corresponds to an event, or if both times are events. Conversely, a pair is deemed non-comparable if the earliest time is censored or both are censored [36].

Several estimators of the C-index have been proposed, including Harrell's C-index [19] and Uno's C-index [37], which is a modified, weighted version of Harrell's C-index. Additionally, Gonen and Heller's measure [38] offers an alternative estimator based on a reversed definition of concordance. Furthermore, a time-dependent version of the C-index was introduced in [39], which considers the entire survival function over time.

2.2 Explainability in Survival Analysis

Explainability is crucial in machine learning models, particularly in high-risk domains such as healthcare and predictive maintenance. In these fields, understanding how a model arrives at its predictions is important and serves diverse purposes. For example, explainable models can assist experts in interpreting the results to identify potential model biases helping them to build more reliable models and ensuring ethical considerations. Another example of explainable models' utility is to gain insights into risk factors associated with a specific disease. This understanding enables healthcare professionals to ensure that predictions align with clinical knowledge and make more informed decisions. By uncovering key insights, explainable models contribute to accumulating valuable knowledge in the application area.

To serve the diverse proposes of explainability, several methods have been developed which can be categorized according to different criteria. The most common distinctions are global vs. local explanations, intrinsic vs. post-hoc methods, and model-specific vs. model-agnostic approaches.

Model-specific approaches rely on the internal structure and functioning of the model to generate explanations. For example, gradient-based approaches, which are widely used to explain deep learning models, focus on understanding how small changes in input data affect the model's output by analyzing the gradients [40–42]. These techniques offer insights into which features influence the model's predictions, but their applicability is generally limited to models like neural networks where gradients are meaningful. In contrast, model-agnostic methods—most notably LIME [43] and SHAP [44]—have gained significant attention for their applicability across different types of machine learning models. LIME works by locally approximating the decision

boundaries around a point of interest using a linear model, providing local explanations. SHAP, on the other hand, takes a game-theoretic approach, calculating each feature's contribution to the difference between the model's prediction and the average prediction using Shapley values [45]. Additionally, SHAP offers global explanations by aggregating Shapley values over many instances [46]. Both methods, LIME and SHAP, rely on feature attribution to explain the model's decisions by assessing the importance of each feature to the model's output.

SurvLIME [47] is an extension of the LIME framework specifically designed for survival analysis. Instead of employing a linear model to approximate the decision boundaries around a given instance, SurvLIME utilizes the Cox Proportional Hazards model which is more appropriate for handling the characteristics of survival data. The weights of the surrogate model highlight the importance of the features of the explained subject.

The SHAP method was also recently extended in SurvSHAP(t) [48] to accommodate functional output models. The method quantifies the contributions of the features to the predicted survival function at each time point providing time-dependent explanations.

Another direction focuses on example-based explanations, with one of the most intriguing being the Counterfactual-Examples-based approach, which addresses "What if" scenarios by providing alternative outcomes in hypothetical situations. These counterfactual examples offer insights into possible alternative paths that lead to different model predictions. According to [49], counterfactual examples can be generated by identifying the closest point to the original instance that changes the model's output to a predefined target. However, this method may produce unrealistic examples. This brings it closer to the concept of Adversarial Examples [50] where the purpose is to expose weaknesses and vulnerabilities in the model and improve its resilience. However, counterfactual examples are intended for end-users, such as domain experts or decision-makers, to provide explanations of the model's predictions that can be utilized to draw actions or changes to improve outcomes. To mitigate the issue of unrealistic examples, one approach is to minimize the distance between the generated counterfactual example and real-world data, as suggested by [51]. A more effective approach involves using anomaly detection models, such as Autoencoder-based models [52; 53], which ensure the realism of counterfactual examples by minimizing the reconstruction error, thus enhancing their likelihood and plausibility.

For survival models, [54] introduced a method for generating counterfactual examples based on the mean survival time, which is represented as the area under the curve (AUC) of the survival function. The approach uses Particle Swarm Optimization (PSO) to identify the smallest modification to the

input features that would adjust the AUC of the predicted survival function to a specified target value. However, the likelihood or realism of the generated counterfactual examples was not addressed in their work.

Other models rely on providing intrinsic explanations where interpretability is built directly into the model design. In these models, the structure or mechanisms inherently provide insights into how predictions are made, without the need for additional post-hoc explanation methods. A linear regression model is one of the simplest examples of an intrinsically explainable model, as it represents the target variable as a weighted sum of the input features. In this model, the coefficients or weights assigned to each feature directly indicate their relative importance and contribution to the predicted outcome, making the model's decision process straightforward and interpretable. However, the assumption of linearity often does not hold in real-world scenarios, limiting the applicability of linear models to more complex problems. To address this, Generalized Additive Models (GAM) [55] was introduced, where the target is modeled as the sum of non-linear functions of the input features, Equation 2.8. This allows GAMs to capture more complex relationships while maintaining interpretability, as each feature still contributes independently through its own function to the model's outcome.

$$g(\mathbb{E}(y|\boldsymbol{x})) = \beta + \sum_{i=0}^{p} g_i(x_i) \tag{2.8}$$

Based on the GAM model, more recently Neural Additive Models (NAM) [56] was introduced where the g_i functions are modeled as neural networks.

In Survival Analysis, the Cox Proportional Hazards (CPH) model is regarded as the counterpart to linear regression. It has been a long-standing favorite due to its interpretability, as it models the log-risk function as a linear combination of the input features, allowing for straightforward explanations. However, due to the linearity limitation, SurvNAM [57] proposed using the NAM structure combined with the Cox Proportional Hazards (CPH) model formulation as a surrogate model to explain black-box survival models. More recently CoxNAM [34] was proposed as a standalone survival model, utilizing the NAM structure to model the log-risk function within the framework of the CPH model as depicted in Figure 2.2.

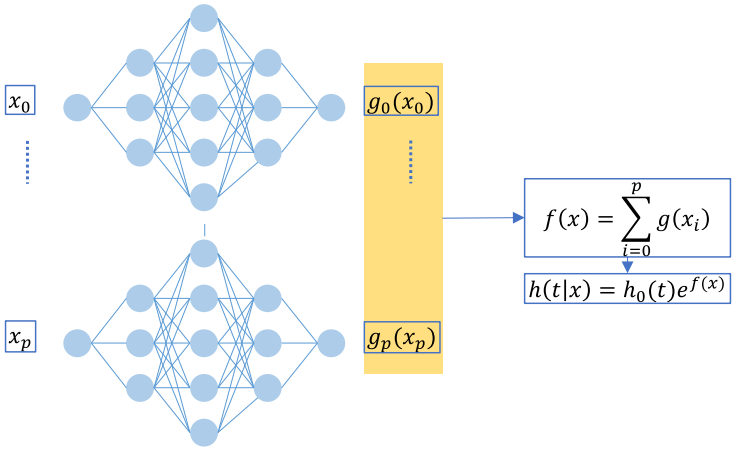

Figure 2.2: The CoxNAM Model

Although NAM-based models enable non-linear modeling, they assume that the effects of features are additive which limits their effectiveness in situations where feature interactions exist.

The Self-Explaining Neural Networks (SENN) [58], is another model that provides intrinsic explanations. This architecture features two branches: a concept encoder that generates interpretable features $\boldsymbol{h}(\boldsymbol{x})$ from the input and an input-dependent parameterizer that learns the relevance scores (weights) $\boldsymbol{w}(\boldsymbol{x})$ for these features, as illustrated in Figure 2.3.

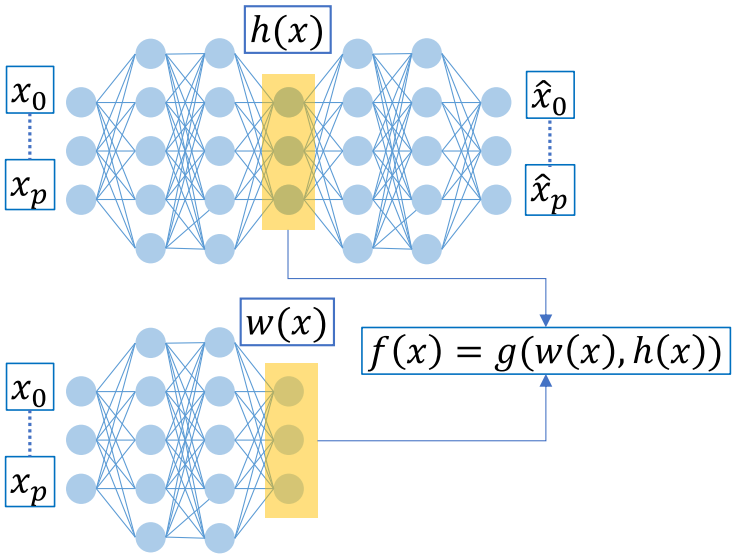

Figure 2.3: Self-Explaining Neural Networks (SENN)

The function can be written as:

$$f(x) = g(w_1(\boldsymbol{x})h_1(\boldsymbol{x}), .., w_p(\boldsymbol{x})h_p(\boldsymbol{x})) \tag{2.9}$$

where p represents the number of features and $g(.)$ denotes the aggregation function, which can be any affine function with positive weights. In the case the aggregation function is an addition, the model approximates the underlying decision boundaries with a locally-linear function, expressed as:

$$f(x) = \sum_{x_i \in \boldsymbol{x}} w_i(\boldsymbol{x})h_i(\boldsymbol{x}) \tag{2.10}$$

SENNs regularize their learned explanations to ensure that similar subjects have similar explanations enhancing the explanations' stability. This is achieved by encouraging $\boldsymbol{w}(\boldsymbol{x}_0)$ to approximate the derivative of f with respect to the concept vector $\boldsymbol{h}(\boldsymbol{x})$ in the vicinity of $\boldsymbol{x}_0$.

For tabular data, the concept vector represents the original features, i.e. $\boldsymbol{h}(\boldsymbol{x}) = \boldsymbol{x}$. This simplifies the model by eliminating the need for the autoencoder branch in the SENN structure, as well as the reconstruction component of the loss function. In the SENN model, each weight depends on all the features, enabling it to capture and model interactions between different features. Additionally, the introduced regularization provides control over the quality of the explanations, ensuring that the explanations remain interpretable and stable.

3. CONCLUDING REMARKS

3.1 Conclusions

In this thesis, we aimed to enhance trust in survival prediction models by pursuing two main tracks. The first focused on improving the explainability of these models, proposing and enhancing techniques to ensure their predictions are transparent and interpretable. The second concentrated on the modeling side, introducing various optimization techniques to achieve higher performance.

We first proposed using Survival Patterns as a basis to explain survival models. Survival Patterns are distinct survival behaviors that can be identified in the predictions of the survival model. Subjects within a certain pattern share similar survival characteristics, which can be used to describe the respective subjects. This work proposed an algorithm to discover Survival Patterns. Based on the identified patterns, two post-hoc model-agnostic explanation methods were proposed. The first relies on feature attribution through Shapley values producing faithful explanations of the underlying ground truth. The second method uses counterfactual explanations that provide alternative scenarios in which the subject follows a different Survival Pattern. This helps in studying factors affecting increases or decreases in survivability.

The main purpose of explainability methods in machine learning is to understand how models work, specifically by analyzing the relationship between the input and the output. However, we argue that understanding the model's performance is just as important, as it can lead to better and more accurate models. In this thesis, we derived a decomposition of the C-index which helped us understand why deep learning models significantly outperformed classical machine learning models, particularly when there was a larger number of observed events in the dataset. Such decomposition also revealed unseen differences between seemingly similar models in terms of C-index performance.

In the second track of this thesis, we focused on the modeling side. We proposed a variational-inference-based model for estimating time-to-event distributions. The model relies on a regression-based loss function supported by a lower bound of the C-index. The model showed discriminative performance comparable to the state-of-the-art, with a better ability to handle datasets with longer timelines than deep learning discrete-time models. We also investigated

improving the discriminative performance of the survival regression function, the Mean Squared Censored Error (MSCE), by employing an evolutionary search algorithm. The results showed that the algorithm almost always finds specialized functions that outperform the hand-crafted MSCE function. The results also showed that it is enough to optimize the part of the loss function that handles the censored cases. Upon analyzing the resulting functions, an unexpected phenomenon was revealed signifying the importance of the gradient of the censored cases part of the loss function. Based on such observation we proposed a novel loss function based on the SoftPlus function which in most cases outperformed or matched the performance of MSCE.

Premature replacements present a challenge in industrial settings, leading to noisy event labels in time-to-event data. In this regard, we proposed an iterative algorithm for discovering and correcting such labels based on genetic algorithms. The method was shown to be effective in discovering a considerable number of premature replacements providing a confidence score for each label. Such confidence score can be useful to vary the decision threshold based on the application requirements.

Finally, we concluded the thesis by joining the two tracks proposing a survival model with improved performance while providing intrinsic explanations. The model, CoxSE, is based on Self-Explaining Neural Networks (SENN) with Cox Proportional Hazards (CPH) formulation providing the explanations as locally-linear weights. The model is equipped with a loss function that optimizes the Cox partial likelihood function with two additional regularization terms, encouraging explanations' stability and robustness to noise. CoxSE matches the performance of its black-box counterpart while providing faithful and stable explanations. We also proposed CoxSENAM, a hybrid model adopting the Neural Additive Models (NAM) structure with the SENN loss function and type of output which showed an enhanced faithfulness and stability. Similar to NAM-based models, CoxSENAM is unable to model interactions between features. However, it demonstrated better robustness to noise than the CoxSE model.

3.2 Future Work Directions

3.2.1 Explainability

The Self-Explaining Neural Network we explored in **Paper VI** is a promising direction for explainable survival analysis. In this work, we only utilized a CPH model as a base survival model. Such a model makes the adaptation straightforward from linear to locally linear risk estimation, as the time component is isolated from the effect of the features. As a future direction of re-

search, this can be extended to direct time regression models where the output can be formulated as a single locally weighted average of the feature values. Moreover, it is interesting to explore its applicability to various survival models with discrete output such as DeepHit model [6]. It is also important to investigate how the SENN structure can be utilized with models that support time-dependent features. Finally, the SENN-based model was shown to be less robust to noisy features. Further investigation of different regularization techniques is needed to improve the robustness of the model which would lead to better explanations.

3.2.2 Performance

Deep neural network survival models usually utilize loss functions consisting of multiple terms optimizing multiple aspects of the estimated distribution like the likelihood, mean-squared error, and concordance. For example, many models incorporate a ranking term alongside the likelihood term in their loss function to encourage concordance in the predictions. This usually involves utilizing hand-crafted loss functions. In **Paper IV**, we utilized genetic programming to search for a loss function to improve the ranking ability of a regression-based loss function without including a separate ranking term. As a future direction of research, it is interesting to search for survival loss functions using a multi-objective optimization algorithm to simultaneously optimize different metrics like the C-index, MAE, or Brier score.

References

[1] ABDALLAH ALABDALLAH. **Machine Learning Survival Models : Performance and Explainability**, 2023. xi, 6, 7, 8, 11, 13

[2] E. L. KAPLAN AND PAUL MEIER. **Nonparametric Estimation from Incomplete Observations**. *Journal of the American Statistical Association*, **53**(282):457–481, 1958. 1, 14

[3] D. R. COX. **Regression Models and Life-Tables**. *Journal of the Royal Statistical Society. Series B (Methodological)*, **34**(2):187–220, 1972. 14

[4] HEMANT ISHWARAN, UDAYA B. KOGALUR, EUGENE H. BLACKSTONE, AND MICHAEL S. LAUER. **Random survival forests**. *Ann. Appl. Stat.*, **2**(3):841–860, 09 2008. 15

[5] 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. 15

[6] 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, 15, 23

[7] 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, 15

[8] 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, 15

[9] 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, 15

[10] 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, 15

[11] BO LI, PENG QI, BO LIU, SHUAI DI, JINGEN LIU, JIQUAN PEI, JINFENG YI, AND BOWEN ZHOU. **Trustworthy AI: From Principles to Practices**. *ACM Comput. Surv.*, **55**(9), January 2023. 1

[12] SAJID ALI, TAMER ABUHMED, SHAKER EL-SAPPAGH, KHAN MUHAMMAD, JOSE M. ALONSO-MORAL, ROBERTO CONFALONIERI, RICCARDO GUIDOTTI, JAVIER DEL SER, NATALIA DÍAZ-RODRÍGUEZ, AND FRANCISCO HERRERA. **Explainable Artificial Intelligence (XAI): What**

we know and what is left to attain Trustworthy Artificial Intelligence. *Information Fusion*, **99**:101805, 2023. 1

[13] J. KALDERSTAM. *Neural Network Approaches to Survival Analysis*. PhD thesis, Lund University, Lund, Sweden, 2015. 2

[14] RISTO MIIKKULAINEN, JASON LIANG, ELLIOT MEYERSON, ADITYA RAWAL, DANIEL FINK, OLIVIER FRANCON, BALA RAJU, HORMOZ SHAHRZAD, ARSHAK NAVRUZYAN, NIGEL DUFFY, ET AL. **Evolving deep neural networks**. In *Artificial intelligence in the age of neural networks and brain computing*, pages 293–312. Elsevier, 2019. 3

[15] MOHAMMED GHAITH ALTARABICHI. *Evolving intelligence: Overcoming challenges for Evolutionary Deep Learning*. PhD thesis, Halmstad University Press, 2024.

[16] MOHAMMED GHAITH ALTARABICHI, SŁAWOMIR NOWACZYK, SEPIDEH PASHAMI, PEYMAN SHEIKHOLHARAM MASHHADI, AND JULIA HANDL. **Rolling the dice for better deep learning performance: A study of randomness techniques in deep neural networks**. *Information Sciences*, **667**:120500, 2024. 3

[17] 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. 3

[18] 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. 3

[19] 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. 3, 16

[20] 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. 6

[21] ABDALLAH ALABDALLAH, JAKUB JAKUBOWSKI, SEPIDEH PASHAMI, SZYMON BOBEK, MATTIAS OHLSSON, THORSTEINN RÖGNVALDSSON, AND GRZEGORZ J. NALEPA. **Understanding Survival Models Through Counterfactual Explanations**. In *Computational Science – ICCS 2024*, pages 310–324, Cham, 2024. Springer Nature Switzerland. 7

[22] ABDALLAH ALABDALLAH, MATTIAS OHLSSON, SEPIDEH PASHAMI, AND THORSTEINN RÖGNVALDSSON. **The Concordance Index decomposition: A measure for a deeper understanding of survival prediction models**. *Artificial Intelligence in Medicine*, **148**:102781, 2024. 8

[23] MOHAMMED GHAITH ALTARABICHI, ABDALLAH ALABDALLAH, SEPIDEH PASHAMI, MATTIAS OHLSSON, THORSTEINN RÖGNVALDSSON, AND SŁAWOMIR NOWACZYK. **Improving Concordance Index in Regression-based Survival Analysis: Discovery of Loss Function for Neural Networks**. In *Proceedings of the Genetic and Evolutionary Computation Conference (GECCO)*, 2024. 9

[24] ABDALLAH ALABDALLAH, THORSTEINN RÖGNVALDSSON, YUANTAO FAN, SEPIDEH PASHAMI, AND MATTIAS OHLSSON. **Discovering Premature Replacements in Predictive Maintenance Time-to-Event Data**. *Proceedings of the Asia Pacific Conference of the PHM Society 2023*, **4**(1), 2023. 10

[25] ABDALLAH ALABDALLAH, OMAR HAMED, MATTIAS OHLSSON, THORSTEINN RÖGNVALDSSON, AND SEPIDEH PASHAMI. **CoxSE: Exploring the Potential of Self-Explaining Neural Networks with Cox Proportional Hazards Model for Survival Analysis**, 2024. 11

[26] LEO BREIMAN. **Random Forests**. *Machine Learning*, **45**(1):5–32, 2001. 15

[27] V. VAN BELLE, K. PELCKMANS, J.A.K. SUYKENS, AND S. VAN HUFFEL. **Additive survival least-squares support vector machines.** *Statistics in Medicine*, **29**(2):296–308, 2010. 15

[28] YIFEI CHEN, ZHENYU JIA, DAN MERCOLA, AND XIAOHUI XIE. **A gradient boosting algorithm for survival analysis via direct optimization of concordance index**. *Comput Math Methods Med*, **2013**:873595, November 2013. 15

[29] RAJESH RANGANATH, ADLER PEROTTE, NOÉMIE ELHADAD, AND DAVID BLEI. **Deep Survival Analysis**. In FINALE DOSHI-VELEZ, JIM FACKLER, DAVID KALE, BYRON WALLACE, AND JENNA WIENS, editors, *Proceedings of the 1st Machine Learning for Healthcare Conference*, **56** of *Proceedings of Machine Learning Research*, pages 101–114, Northeastern University, Boston, MA, USA, 18–19 Aug 2016. PMLR. 15

[30] XENIA MISCOURIDOU, ADLER PEROTTE, NOEMIE ELHADAD, AND RAJESH RANGANATH. **Deep Survival Analysis: Nonparametrics and Missingness**. In FINALE DOSHI-VELEZ, JIM FACKLER, KEN JUNG, DAVID KALE, RAJESH RANGANATH, BYRON WALLACE, AND JENNA WIENS, editors, *Proceedings of the 3rd Machine Learning for Healthcare Conference*, **85** of *Proceedings of Machine Learning Research*, pages 244–256, Palo Alto, California, 17–18 Aug 2018. PMLR. 15

[31] BINGZHONG JING, TAO ZHANG, ZIXIAN WANG, YING JIN, KUIYUAN LIU, WENZE QIU, LIANGRU KE, YING SUN, CAISHENG HE, DAN HOU, LINQUAN TANG, XING LV, AND CHAOFENG LI. **A deep survival analysis method based on ranking**. *Artificial Intelligence in Medicine*, **98**:1–9, 2019.

[32] CHIRAG NAGPAL, XINYU LI, AND ARTUR DUBRAWSKI. **Deep Survival Machines: Fully Parametric Survival Regression and Representation Learning for Censored Data With Competing Risks**. *IEEE Journal of Biomedical and Health Informatics*, **25**(8):3163–3175, 2021.

[33] SHI HU, EGILL FRIDGEIRSSON, GUIDO VAN WINGEN, AND MAX WELLING. **Transformer-Based Deep Survival Analysis**. In RUSSELL GREINER, NEERAJ KUMAR, THOMAS ALEXANDER GERDS, AND MIHAELA VAN DER SCHAAR, editors, *Proceedings of AAAI Spring Symposium on Survival Prediction - Algorithms, Challenges, and Applications 2021*, **146** of *Proceedings of Machine Learning Research*, pages 132–148. PMLR, 22–24 Mar 2021.

[34] LIANGCHEN XU AND CHONGHUI GUO. **CoxNAM: An interpretable deep survival analysis model**. *Expert Systems with Applications*, **227**:120218, 2023. 15, 18

[35] 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. 15

[36] FRANK E. HARREL JR., KERRY L. LEE, AND DANIEL B. MARK. **MULTIVARIABLE PROGNOSTIC MODELS: ISSUES IN DEVELOPING MODELS, EVALUATING ASSUMPTIONS AND ADEQUACY, AND MEASURING AND REDUCING ERRORS**. *Statistics in Medicine*, **15**(4):361–387, 1996. 16

[37] 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. 16

[38] M. GÖNEN AND G. HELLER. **Concordance probability and discriminatory power in proportional hazards regression**. *Biometrika*, **92**(4):965–970, 2005. 16

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

[40] KAREN SIMONYAN, ANDREA VEDALDI, AND ANDREW ZISSERMAN. **Deep inside convolutional networks: Visualising image classification models and saliency maps**. *arXiv preprint arXiv:1312.6034*, 2013. 16

[41] 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.

[42] 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. 16

[43] 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. 16

[44] 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. 16

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

[46] 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. 17

[47] 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. 17

[48] 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. 17

[49] 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. 17

[50] CHRISTIAN SZEGEDY, WOJCIECH ZAREMBA, ILYA SUTSKEVER, JOAN BRUNA, DUMITRU ERHAN, IAN GOODFELLOW, AND ROB FERGUS. **Intriguing properties of neural networks**, 2014. 17

[51] 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. 17

[52] 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. 17

[53] 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. 17

[54] MAXIM KOVALEV, LEV UTKIN, FRANK COOLEN, AND ANDREI KONSTANTINOV. **Counterfactual Explanation of Machine Learning Survival Models**. *Informatica*, **32**(4):817–847, jan 2021. 17

[55] TREVOR HASTIE AND ROBERT TIBSHIRANI. **Generalized Additive Models**. *Statistical Science*, **1**(3):297–310, 1986. 18

[56] RISHABH AGARWAL, LEVI MELNICK, NICHOLAS FROSST, XUEZHOU ZHANG, BEN LENGERICH, RICH CARUANA, AND GEOFFREY E HINTON. **Neural additive models: Interpretable machine learning with neural nets**. *Advances in neural information processing systems*, **34**:4699–4711, 2021. 18

[57] LEV V. UTKIN, EGOR D. SATYUKOV, AND ANDREI V. KONSTANTINOV. **SurvNAM: The machine learning survival model explanation**. *Neural Networks*, **147**:81–102, 2022. 18

[58] 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. 19