Tree-Based Approaches for Interpretable Modeling in Healthcare

Juliette Murris

HeKA, Inria Paris, Inserm, Université Paris Cité, Pierre Fabre R&D
Paris, France
juliette.murris@inria.com

Introduction

In oncology, cancer relapse, tumour progression or death are often used to measure treatment effect (Figure 1). The principles of survival analysis aim to model the time to the event of interest (Jenkins 2005). AI has advanced survival analysis by adapting traditional machine learning (ML) algorithms (Wang, Li, and Reddy 2019), which are now key in personalized medicine and improving patient care (Quazi 2022). Typically, ensemble-based approaches have extended survival analysis, overcoming traditional model assumptions with methods like random survival forests (RSF) (Ishwaran et al. 2008), gradient boosted survival trees (Hothorn et al. 2006), or optimal survival trees (Bertsimas et al. 2022). While these tree-based approaches have demonstrated great performance in various settings (Grinsztajn, Oyallon, and Varoquaux 2022; Yabaci and Sigirli 2022; Penny-Dimri et al. 2023), they are considered as black-boxes, i.e. we do not have direct access to models' internal reasoning.

Interpretability methods for common classification and regression tasks are broadly adopted (Molnar 2020), but they lag for survival models (Langbein et al. 2024). Improving interpretability in a survival framework is essential to meet healthcare regulatory requirements, troubleshoot survival models and maintain the integrity of medical decision-making. To answer this need, this work presents several contributions around tree-based methods and interpretability in survival analysis:

- We examine the assessment of AI algorithms by health regulators and identify key technical requirements for their interpretability and explainability (Farah et al. 2023);
- We develop an extension of the random survival forests algorithm to handle occurrences of multiple events with a model-agnostic interpretability method (Murris et al. 2024);
- We underline the need for model-specific interpretability methods for survival ML algorithm.

Background

Regulatory need for interpretability in healthcare. Clinicians are increasingly sensitive to the integration of AI tools into their daily routines, and they need to trust the decisions made by algorithms (LaRosa and Danks 2018). If the reasoning behind these decisions is opaque, it can lead to skepticism and reluctance to adopt the technology. Besides, regulatory oversight is provided by health technology assessment (HTA) bodies, which carry a broader responsibility towards all healthcare stakeholders. [1] Developing HTA standards in interpretability for AI-based MDs could also streamline market and patient access across countries (Singh 2022). For instance, the FDA emphasizes the importance of *"understanding of a model's intended integration into clinical workflow (interpretability and explicability)"* (US. FDA, Health Canada and MHRA 2021). However, despite the recognized importance of these criteria, a standardized methodology for their measurement is still missing.

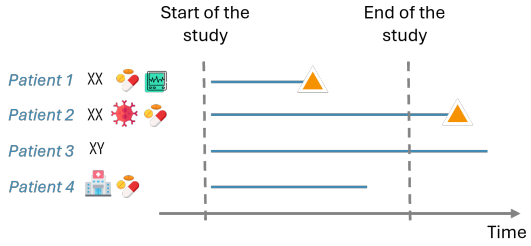

Figure 1: Patient data collected for survival analysis includes the event of interest, the time, and relevant features (e.g., demographics and comorbidities).

Survival analysis and patient outcomes. Survival analysis refers to the analysis of time-to-event data and is particularly appropriate when the time between exposure (e.g. diagnosis or start of treatment) and the event is clinically relevant. Censoring is defined in cases when the event of interest has not occurred for some patients by the end of the study or before they are lost to follow-up (Figure 1). For this reason, survival analysis requires specific tools to handle censored data for time-to-event endpoints for the robust estimations of the associated survival probabilities. The core

[1] HTA is a multidisciplinary process which refers to the systematic evaluation of properties, effects, and/or impacts of health technology (Organization et al. 2011).

task is to estimate for each patient two functions over time. First, the survival function is the probability of not having experienced the event by time t: $S(t) = \mathbf{P}(T > t)$, with $S(0) = 1$ and $\lim_{t\to\infty} S(t) = 0$. Then, the hazard function is the probability of presenting the event of interest in a small time interval around t: $\lambda(t) = \lim_{\Delta t\to 0} \frac{P(t\leq T<t+\Delta t|T\geq t)}{\Delta t}$. Various ML algorithms now have their survival counterparts and are effectively employed to answer medical questions (Wang, Li, and Reddy 2019). The RSF algorithm from (Ishwaran et al. 2008) embodies a powerful ensemble learning technique and has been extended to model several phenomena, such as competing risks, or longitudinal data (Ishwaran et al. 2014; Devaux et al. 2023).

Accomplished work

Interpretability for survival analysis. First, we needed to better understand the concepts of interpretability and explainability of AI/ML algorithms in healthcare. To this end, we systematically reviewed all evaluation criteria used by health authorities worldwide for the assessment of AI-based MDs in (Farah et al. 2023). This study resulted in providing existing tools and methods to elucidate how and why algorithms work to hold stakeholders more accountable for decisions made. We have also drawn recommendations based on the level of risk of the algorithm under assessment. We believe we contributed to raising awareness of these concepts for their widespread adoption in response to ethical issues.

Second, we reviewed the current state of interpretability methods in survival framework. We carried out a comprehensive illustration of three common survival ML models, with up-to-date interpretability methods such as SurvSHAP (Krzyziński et al. 2023) and SurvLIME (Kovalev, Utkin, and Kasimov 2020). We are now finalising this project using different open source datasets in order to draw up guidance and assessment of how each interpretability method works, and provide recommendations.

A tree-based approach to capture multiple clinical events. In medical research, patients may face recurrent disease relapses, frequent hospitalizations, or repeated surgeries. However, existing extensions of RSF focused solely on the first occurrence of a given event (Murris et al. 2023). As part of the PhD work, we developed a new RSF algorithm to handle multiple clinical events introduced in (Murris et al. 2024). To better suit the context of oncology studies, we have also incorporated the ability to handle the presence of a fatal terminal event. The flexibility of our approach has been demonstrated in various simulated scenarios, including the processing of missing and high-dimensional data, attesting to the robustness and adaptability of the algorithm. A practical demonstration on open-source data was carried out to show both the impact on real data and the optimisation of the hyperparameters. We are currently finalising a clinical study using our extended RSF to better understand multiple hospital readmissions after cancer surgery in all French healthcare facilities. By taking all events occurences into account, our approach is tailored as closely as possible to patient follow-up, enabling more precise clinical predictions.

We also introduced a new performance metric based on the concordance index (Harrell et al. 1984) to account for all event occurences for each individual. The C-index from (Murris et al. 2024) is given by:

$$\hat{\mathbb{C}} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} \mathbb{1}_{r_i>r_j} \times \mathbb{1}_{\hat{r}_i>\hat{r}_j}}{\sum_{i=1}^{n}\sum_{j=1}^{n} \mathbb{1}_{r_i>r_j}} \tag{1}$$

where r_i and $\hat{r}_i$ are the observed and predicted event rates, respectively. This metric is unbiased (Uno et al. 2011) and is one of the firsts to account multiple events, alongside (Kim, Schaubel, and McCullough 2018; Bouaziz 2024). For interpretability purpose, we have included permutation feature importances as a model-agnostic method in our extension of RSF based on the above C-index.

While this RSF extension enables strong performance gain when dealing with multiple events survival data, this approach can be seen as a black-box in the same way as other common ensemble methods. This paves the way towards our planned next steps described in the next section.

Next steps

Tree-based ML algorithms for classification or regression benefit from model-specific interpretability methods. TreeSHAP (Lundberg, Erion, and Lee 2018) is an extended algorithm to compute SHAP (SHapley Additive exPlanation) specifically designed for tree-based models and reduces the computational cost of explanations (Lundberg and Lee 2017). SHAP computation associated with an instance i is based on a path-dependent feature perturbation algorithm (consistently detailed in (Lundberg, Erion, and Lee 2018)). Basically, for each leaf and each feature i on the path to this leaf, the following are calculated:

- The proportion of subsets S at the leaf that contain i and the proportion of subsets S that do not contain i;
- For each cardinality, the proportion of the sets of that cardinality contained at the leaf.

The SHAP contribution of feature i is then computed as

$$\phi_i = \sum_{j=1}^{L} \sum_{P\in S_j} \frac{w(|P|, j)}{M_j \binom{M_j - 1}{|P|}} (p_o^{i,j} - p_z^{i,j}) v_j \tag{2}$$

where S_j is the set of present feature subsets at leaf j, M_j is the length of the path and $w(|P|, j)$ is the proportion of all subsets of cardinality P at leaf j, $p_o^{i,j}$ and $p_z^{i,j}$ represent the fractions of subsets that contain or do not contain feature i respectively, and v_j is the value of the leaf with index j, i.e. the model output.

The time dependent nature of TreeSHAP can be accommodated by incorporating explanations that are computed and evaluated on the survival function (which is the probability of each individual experiencing a clinical event). The SurvSHAP algorithm by (Krzyziński et al. 2023) extended SHAP for any functional output of a (machine learning) survival model and generates explanations for all time points.

This way, combining both TreeSHAP and SurvSHAP would lead to a model-specific interpretability method for tree-based survival models (like our RSF previously introduced) and open broader possibilities for interpretability in survival machine learning.

Acknowledgments

Author Murris would like to express her sincere gratitude to her supervisors Prof. S. Katsahian and Dr. A. Lavenu. This research was supported by a CIFRE grant from the Association Nationale de la Recherche et de la Technologie with Pierre Fabre (number 2020/1701).

References

Bertsimas, D.; Dunn, J.; Gibson, E.; and Orfanoudaki, A. 2022. Optimal survival trees. *Machine learning*, 111(8): 2951–3023.

Bouaziz, O. 2024. Assessing model prediction performance for the expected cumulative number of recurrent events. *Lifetime Data Analysis*, 30(1): 262–289.

Devaux, A.; Helmer, C.; Genuer, R.; and Proust-Lima, C. 2023. Random survival forests with multivariate longitudinal endogenous covariates. *Statistical Methods in Medical Research*, 32(12): 2331–2346.

Farah, L.; Murris, J. M.; Borget, I.; Guilloux, A.; Martelli, N. M.; and Katsahian, S. I. 2023. Assessment of performance, interpretability, and explainability in artificial intelligence–based health technologies: what healthcare stakeholders need to know. *Mayo Clinic Proceedings: Digital Health*, 1(2): 120–138.

Grinsztajn, L.; Oyallon, E.; and Varoquaux, G. 2022. Why do tree-based models still outperform deep learning on tabular data?

Harrell, F. E.; Lee, K. L.; Califf, R. M.; Pryor, D. B.; and Rosati, R. A. 1984. Regression modelling strategies for improved prognostic prediction. *Statistics in medicine*, 3(2): 143–152.

Hothorn, T.; Bühlmann, P.; Dudoit, S.; Molinaro, A.; and Van Der Laan, M. J. 2006. Survival ensembles. *Biostatistics*, 7(3): 355–373.

Ishwaran, H.; Gerds, T. A.; Kogalur, U. B.; Moore, R. D.; Gange, S. J.; and Lau, B. M. 2014. Random survival forests for competing risks. *Biostatistics*, 15(4): 757–773.

Ishwaran, H.; Kogalur, U. B.; Blackstone, E. H.; and Lauer, M. S. 2008. Random survival forests.

Jenkins, S. P. 2005. Survival analysis. *Unpublished manuscript, Institute for Social and Economic Research, University of Essex, Colchester, UK*, 42: 54–56.

Kim, S.; Schaubel, D. E.; and McCullough, K. P. 2018. A C-index for recurrent event data: Application to hospitalizations among dialysis patients. *Biometrics*, 74(2): 734–743.

Kovalev, M. S.; Utkin, L. V.; and Kasimov, E. M. 2020. SurvLIME: A method for explaining machine learning survival models. *Knowledge-Based Systems*, 203: 106164.

Krzyziński, M.; Spytek, M.; Baniecki, H.; and Biecek, P. 2023. SurvSHAP (t): time-dependent explanations of machine learning survival models. *Knowledge-Based Systems*, 262: 110234.

Langbein, S. H.; Krzyziński, M.; Spytek, M.; Baniecki, H.; Biecek, P.; and Wright, M. N. 2024. Interpretable Machine Learning for Survival Analysis. arXiv:2403.10250.

LaRosa, E.; and Danks, D. 2018. Impacts on trust of healthcare AI. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 210–215.

Lundberg, S. M.; Erion, G. G.; and Lee, S.-I. 2018. Consistent individualized feature attribution for tree ensembles. *arXiv preprint arXiv:1802.03888*.

Lundberg, S. M.; and Lee, S.-I. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.

Molnar, C. 2020. *Interpretable machine learning*. Lulu. com.

Murris, J.; Bouaziz, O.; Jakubczak, M.; Katsahian, S.; and Lavenu, A. 2024. Random survival forests for the analysis of recurrent events for right-censored data, with or without a terminal event.

Murris, J.; Charles-Nelson, A.; Tadmouri Sellier, A.; Lavenu, A.; and Katsahian, S. 2023. Towards filling the gaps around recurrent events in high dimensional framework: a systematic literature review and application. *Biostatistics & Epidemiology*, 7(1): e2283650.

Organization, W. H.; et al. 2011. Health technology assessment of medical devices.

Penny-Dimri, J. C.; Bergmeir, C.; Reid, C. M.; Williams-Spence, J.; Perry, L. A.; and Smith, J. A. 2023. Tree-based survival analysis improves mortality prediction in cardiac surgery. *Frontiers in Cardiovascular Medicine*, 10: 1211600.

Quazi, S. 2022. Artificial intelligence and machine learning in precision and genomic medicine. *Medical Oncology*, 39(8): 120.

Singh, H. 2022. Fair, Robust, and Data-Efficient Machine Learning in Healthcare. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, 914–914.

Uno, H.; Cai, T.; Pencina, M. J.; D'Agostino, R. B.; and Wei, L.-J. 2011. On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Statistics in medicine*, 30(10): 1105–1117.

US. FDA, Health Canada and MHRA. 2021. Good Machine Learning Practice for Medical Device Development: Guiding Principles. https://www.fda.gov/medical-devices/software-medicaldevice-samd/good-machine-learning-practice-medical-device-development-guiding-principles.

Wang, P.; Li, Y.; and Reddy, C. K. 2019. Machine learning for survival analysis: A survey. *ACM Computing Surveys (CSUR)*, 51(6): 1–36.

Yabaci, A.; and Sigirli, D. 2022. Comparison of tree-based methods used in survival data. *Statistics in Transition new series*, 23(1): 21–38.