Proceedings of the 20th Conference on Computer Science and Intelligence Systems (FedCSIS) pp. 647–652
DOI: 10.15439/2025F6352
ISSN 2300-5963 ACSIS, Vol. 43

Machine learning for survival analysis: a comparative study on intensive care unit (ICU) patient data and simulations

Lukáš Boček[1, 2]
ORCiD: 0009-0006-9576-715X
[1]Department of Statistics and Probability
Faculty of Informatics and Statistics
Prague University of Economics and Business
W. Churchill's square 4, 13067 Prague, Czech Republic
Email: l.bocek9825@gmail.com
&
[2]DataSentics, Inc.
Washingtonova 1599/17, 11000 Prague, Czech Republic
Email: lukas.bocek@datasentics.com

Lubomír Štěpánek[1, 2, 3]
ORCiD: 0000-0002-8308-4304
[1]Department of Statistics and Probability
[2]Department of Mathematics
Faculty of Informatics and Statistics
Prague University of Economics and Business
W. Churchill's square 4, 13067 Prague, Czech Republic
Email: lubomir.stepanek@vse.cz
&
[3]Institute of Biophysics and Informatics
First Faculty of Medicine
Charles University
Salmovská 1, 12000 Prague, Czech Republic
Email: lubomir.stepanek@lf1.cuni.cz

***Abstract*—Survival analysis focuses on modeling the time until a specific event occurs, often in the presence of censored observations. While classical methods like the Cox model are widely used, modern machine learning (ML) approaches offer greater flexibility and predictive power. This paper compares classical and ML-based survival models on both real-world and simulated datasets. We demonstrate that techniques like CoxBoost and penalized Cox regression outperform tree-based models like Random Survival Forests in most settings. Explainable Artificial Intelligence (AI) tools are applied to improve the transparency and interpretability of model predictions.**

***Index Terms*—Survival analysis, machine learning, Cox proportional hazards model, random survival forests, CoxBoost, penalized Cox model, ICU patients, explainable artificial intelligence**

I. Introduction

SURVIVAL analysis is a branch of statistics concerned with the time until an event of interest occurs. The event of interest can be, for example, the death of a patient or the time until the failure of a machine. In this situation, a phenomenon called data censoring is often present, as typically, not all units experienced an event of interest during the monitoring period.

The most widely used model in survival analysis is the Cox model, introduced in 1972 by [1]. The Cox model does not make assumptions regarding the hazard function, which provides necessary flexibility, but makes parametric assumptions about the effect of the covariates on the hazard function. The assumptions of the model are relatively strict and may lead to biased results if violated. This is often pointed out as a potential drawback.

Machine learning is a subfield of artificial intelligence that leverages knowledge from statistics and computer science. In the last decades, it has established itself as a highly flexible and powerful tool capable of solving a wide variety of problems. In general, machine learning requires little to no assumption to be made and is able to discover complex relations in the data. This raises an important question: How can survival analysis benefit from incorporating machine learning methods. Addressing this question is the primary focus of this work.

This paper presents a comprehensive evaluation of six models: Cox proportional hazard model, Cox model with elastic net regularization, CoxBoost, Random Survival Forests (RSF), Conditional Inference Forests (CIF), and Oblique RSF. We evaluated their performance on a real-world Intensive Care Unit (ICU) dataset as well as on simulated datasets with known ground truth.

II. Related work

The use of machine learning and statistical methods for survival analysis has been attracting a lot of attention in recent years. Survival clustering analysis for time-to-event data was proposed in [2], robust nonparametric survival curves comparison in [3], [4], while support vector machines in [5], [6], together with implementation in R. All the above mentioned approaches offer rather a methodological framework on usage the ML in survival analysis.

Another approach how to utilize machine learning in survival analysis appeared in a publication [7], where the authors decomposed the dependent two-dimensional time-event variable into two components. Thereafter, the authors separated

the analysis into two parts, the occurrence of the event being a classification task and time-to-event estimation being a regression task. This approach was later improved and used for the prediction of a COVID-19 blood antibody decrease, which could help to identify individuals who should receive boosting vaccination when a new variant of COVID-19 would appear [8].

Furthermore, the use of deep learning techniques in the field of survival analysis started to seem promising and become one of the main topics for research. In an article [9], the authors utilized neural networks in order to create a personalized treatment recommender system. Authors in an article [10] employed deep learning techniques for competing risk analysis. A comprehensive review about deep learning algorithms used in survival analysis was lately given by [11], which also discusses deep learning algorithms capable of handling time-dependent covariates. Additionally, the review outlines how deep learning can be used to integrate multimodal data – including image, text, and tabular inputs – into survival modeling.

III. Methodology

This section outlines the analytical framework used in our study. We detail the specific traditional and machine learning survival models evaluated, the metrics employed for assessing their predictive performance, and the explainable AI techniques used to interpret their behaviour.

A. Survival and ML Models

We evaluated six survival analysis models from both traditional and machine learning families:

- **Cox Proportional Hazards (Cox PH)**: The most popular model for the analysis of survival data is the Cox proportional hazards model (in literature, it often appears as the Cox PH model or simply the Cox model). The Cox model is a semiparametric model, it makes no assumptions about the hazard function but makes a parametric assumption regarding the effect of the covariates on the hazard function. The fact that the hazard function is not needed provides a significant advantage, as in many situations the true form of the hazard function might be unknown or too complex [12].
  The Cox model is most often stated in the following form,
  $$h(t|\mathbf{x}) = h_0(t)\exp(\mathbf{x}\boldsymbol{\beta}), \tag{1}$$
  where
  - $h(t|\mathbf{x})$ is the hazard function at time t given covariates $\mathbf{x}$,
  - $h_0(t)$ is the baseline hazard function at time t,
  - $\mathbf{x}$ is the vector of covariate (i.e., it is eventually a matrix),
  - $\boldsymbol{\beta}$ is the vector of regression coefficients.
- **Random Survival Forests (RSF)**: Random survival forests (RSF) is a method introduced in article [13], and it presents modification for analysis of right-censored survival data of the popular random forest algorithm. The authors argue that extending the random forest algorithm for the purpose of analyzing survival data is of great value as with traditional methods, nonlinear effects of variables must be modeled by transformations, and ad hoc approaches are often needed. On the other hand, random forests handle these difficulties automatically.
  The RSF algorithm, as described in [13] is as follows:
  (i) Draw B bootstrap samples from the original data. Each bootstrap sample excludes, on average, 37% of the data, these data are called out-of-bag data (OOB data).
  (ii) Grow a survival tree for each bootstrap sample. At each node of the tree, randomly select p variables as candidates for the splitting. The node is split using the variable that maximizes the survival difference between the child nodes.
  (iii) Grow the tree under the constraint that each terminal node should contain at least one unique event of interest.
  (iv) Calculate a cumulative hazard function (CHF) for each tree. Obtain the ensemble cumulative hazard function by averaging across all trees.
  (v) Use the OOB data to compute the prediction error for the ensemble CHF.
- **CoxBoost**: CoxBoost is an extension of the Cox model that incorporates boosting algorithm introduced in [14]. Boosting is a popular, iterative ensemble method that builds a strong predictive model by sequentially combining multiple weak models. The boosting algorithm initially sets equal weights to all observations, then for each successive (m-th) iteration, where $m \in \{2, 3, \ldots, M\}$, the weights for each observation are modified. At each (m-th) step, the misclassified observations have their weights increased, whereas correctly classified observations have their weights decreased. As the algorithm proceeds, observations that are difficult to classify correctly gain greater influence [15].
- **Penalized Cox (Elastic Net)**: As the amount of collected data rapidly grows, supported by advancements in detection techniques, high-dimensional settings are becoming increasingly common across most domains. For classical cases, with significantly more observations than predictors, the Cox model tends to work well. However, in situations where the number of predictors is close to or even exceeds the number of observations, the Cox model tends to output degenerate behavior, as all parameters β_j are converging towards $\pm\infty$, [16].
  To address this challenge, Simon et al. [16] proposed an algorithm that incorporates the elastic net penalty as
  $$\lambda\left(\alpha\sum_{i=1}^{p}|\beta_i| + \frac{1}{2}(1-\alpha)\sum_{i=1}^{p}\beta_i^2\right), \quad \alpha \in [0,1], \tag{2}$$
  which is to be minimized by the algorithm and is a mixture of the L_1 (LASSO) and L_2 (ridge regression) penalties. LASSO tends to work well for sparsity problems, as it tends to choose only a few nonzero coefficients. On the other hand, in the presence of correlations between the predictors, ridge regression tends to perform better, but sets no coefficients to exactly zero. The elastic net combines the strength of both approaches and as α changes between 0 and 1, the approach is changing to be more ridge-like or more LASSO-like [16].

- **Conditional Inference Forests (CIF)**: Conditional inference trees were introduced in [17]. The authors argue that the majority of recursive partitioning algorithms, such as CART [18], have fundamental problems in overfitting and selection bias towards covariates with many possible splits [17].
  To address these concerns, conditional inference trees adopt a more statistically principled approach that takes into account the distributional properties of the data. The core of this approach lies in determining the association between covariates and the response, which allows for an unbiased selection of variables, irrespective of their scale or the number of categories they might have [17].
- **Oblique Random Survival Forests (ORSF)**: A potential modification of the random survival forests described above is extending the splits to allow a linear combination of the predictors instead of using only a single predictor. This extension is applied in oblique random survival forests, as described in [19]. Authors argue that the described approach might improve the random survival forests especially when the predictors are correlated and some are irrelevant to the survival outcome, which is often the case when working, for example, with large medical databases.

B. Model Evaluation

In order to assess model performance, there is a need for metrics that can quantify the prediction performance of the survival models. A commonly used measure is *Harrell's concordance index* [20], which measures how well the model ranks two random individuals in terms of survival time. It is defined as a ratio of concordant pairs (correctly ordered pairs) to all comparable pairs. In order for the random pair i and j to be comparable, the sample with lower observed time y had to experience an event, i.e., $y_j > y_i$ and $\delta_i = 1$, where δ_i is a binary event indicator. A pair is concordant if the estimated risk predicted by the survival model is higher for observations with lower survival time (in other words, if the survival time is in general low, the risk had to be high because it shortened the survival time). Otherwise the pair is called discordant. The concordance index is then computed as

$$C = \frac{\text{number of concordant pairs}}{\text{number of all comparable pairs}}. \tag{3}$$

C. Explainability

With the introduction of machine learning methods into survival analysis, there emerged a need to deepen our understanding of these algorithms. As stated in [21], the successful adoption of machine learning in healthcare critically depends on techniques that enhance our ability to interpret model outputs.

In response to this challenge, an entire subfield has developed, commonly known as Explainable AI (XAI). This subfield focuses on creating techniques that make machine learning and artificial intelligence solutions more transparent and provide better understanding of the process behind decision making of the model [22].

However, until recently, these techniques were not designed to handle censoring and provide explanations for survival models. To fill this gap, [23] proposed a `survex` framework implemented into R, with the goal of empowering stakeholders with model understanding and building trust in machine learning models. The `survex` framework is designed to be model agnostic and, hence, can be applied to any survival model that returns predictions in the form of survival or cumulative hazard function [23].

Furthermore, methods dedicated to explain survival models such as SurvSHAP(t) [24] or SurvLIME [25] are also incorporated in the `survex` package along with explanations tailored to incorporate the time dimension. Moreover, local explanations that refers to prediction as well as global explanations regarding the whole model are provided [23].

D. Real-world data

Real-world dataset chosen for the purpose of the practical application of survival analysis methods, was MIMIC-III [26]. MIMIC-III is a large relational database consisting of 26 tables containing information about 38,597 adult patients and 7,870 neonates that were admitted to critical care units (ICU) to the Beth Israel Deaconess Medical Center in Boston, Massachusetts, between 2001 and 2012. Information about patients' demographics, diagnoses, laboratory tests, physiological measurements, drug codes, and many more are collected. Different types of information are kept separately in different tables [26].

The dataset underwent extensive preprocessing, the details of which are beyond the scope of this paper but are reproducible and available upon request. The final dataset contained 11,435 unique patients out of which 5,036 experienced an event of interest (i.e., death resulting from a severe health condition). This equals to a censoring level of roughly 56 %.

E. Simulated data

The goal of the analysis in the simulated dataset was to further assess the qualities of the individual algorithms in simulation settings that represent different scenarios. Another objective was to examine how well the models were able to recognize variables that have a real impact on the outcome and which models are more likely to get confused by variables representing only random noise. This can be done only under the simulation study, where these variables are known ahead.

The simulated dataset was created using standard Weibull distribution, which incorporates time-dependent effects of the variables. For simplification, a formula of how such a generating model can look for one covariate is illustrated in equation (4). In this example, it is assumed that the time-dependent effect of the covariate is induced by interaction with log time [27],

$$h_i(t) = \gamma \lambda t^{\gamma-1} \exp(\beta_0 X_{i,j} + \beta_1 X_{i,j} \cdot \log(t)), \tag{4}$$

where

- $h_i(t)$ is the hazard function for the i-th individual at time t,
- γ is the shape parameter of the Weibull distribution,
- λ is the scale parameter of the Weibull distribution,
- $X_{i,j}$ is the j-th covariate value for i-th individual,
- β_0 is log hazard ratio for the j-th covariate,
- β_1 specifies the amount by which the log hazard ratio changes for the j-th covariate, for every unit increase in $\log(t)$.

To test model robustness in noisy environments, a dataset containing 32 variables was created, with only 8 of these variables having a real impact on the outcome. The values of each variable were simulated from common probability distributions such as normal, log-normal or Bernoulli.

IV. Results and Discussion

A. Real-world data

For the analysis of real-world data, 80/20 train/test split was conducted and patient's diagnosis and demographics were used as predictors.

On the test set, RSF performed slightly better initially, but Cox-based models (especially CoxBoost and Cox with elastic net regularization) showed better or equal performance over the long term. Overall, performance was very similar.

Fig. 1 display comparison of the performance, using concordance index, over all used models. We observe that performance stability, particularly at later time points (when only a smaller number of individuals who have not experienced the event of interest remain, making prediction more difficult), is achieved by the CoxBoost and Penalized Cox models, whereas the traditional Cox proportional hazards model demonstrates poorer predictive performance.

Traditional Cox model found age, gender (males having 11.5% higher risk), admission type (emergency/urgent increasing risk), and diagnosis (sepsis being most dangerous, increasing risk by 33.8 % vs. Altered Mental Status (AMS)) to be significant predictors.

Random Survival Forest identified age, diagnosis, and admission type as most important, but ranked ethnicity/insurance higher than gender, unlike Cox model.

For better understanding of the model's behaviour, we used the R package `survex` [23] and displayed the predicted survival function for a new patient for different values of the variable of our interest (diagnosis). In this case, we identified the most typical patient who experienced the event of interest (death) as a 73-year-old female who, in reality, was diagnosed with congestive heart failure and died 71 days after the first admission. The practical implications of this explainable finding are straightforward – a newly admitted patient who is a 73-year-old female should be treated with maximum care, as she is at the highest risk of death given this diagnosis.

This demonstrates how powerful the machine learning algorithms could be in understanding complex relationships, which are often present in real-world data.

As was mentioned in the methodology section, the Cox model is constrained by strong assumptions – one of which being the proportional hazard ratios. This constraint is evident in Fig. 2, where the survival curves for individual diagnoses do not intersect. In contrast, the Random Survival Forest (RSF) model imposes no such assumption, enabling more flexible modeling of time-dependent effects. This flexibility is illustrated in Fig. 3 – for example, pneumonia initially presents lower risk compared to seizure or intracranial hemorrhage, but its predicted risk eventually exceeds both over time.

B. Simulated data

The algorithms used in the simulation part were the same as the algorithms used in real-world data analysis. Model performance was evaluated by concordance index on the test set after splitting 80 % of the data for training and 20 % for testing, can be seen in Table I. The size of the dataset before splitting is denoted by n. The best performance under each setting is bolded.

Initially, a dataset with only 100 observations was created, leaving only about 80 for training and 20 observations for

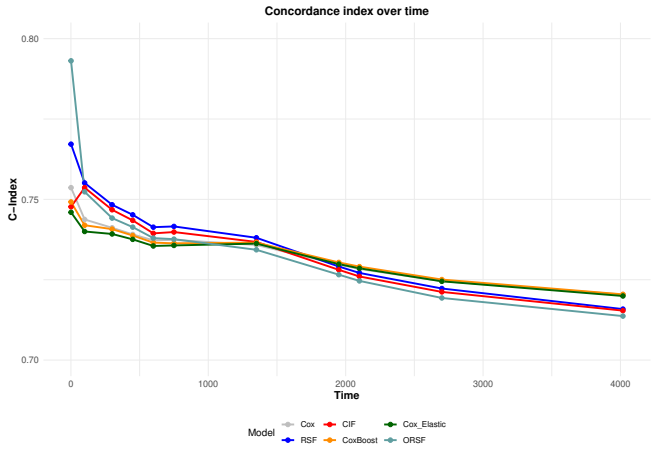

Fig. 1. Concordance index over time for all models.

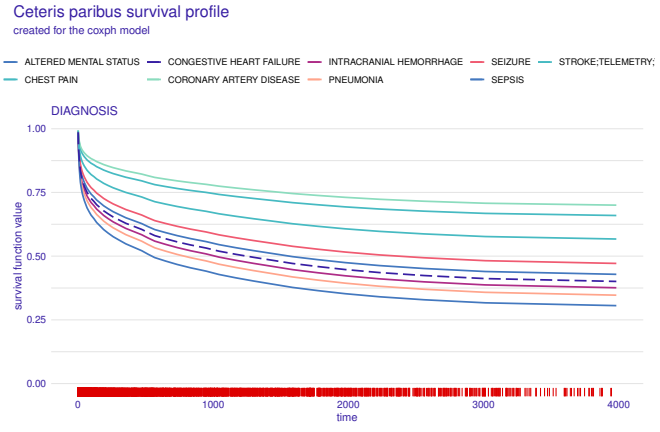

Fig. 2. Predictions for patient 16680 under different diagnoses – Cox PH model.

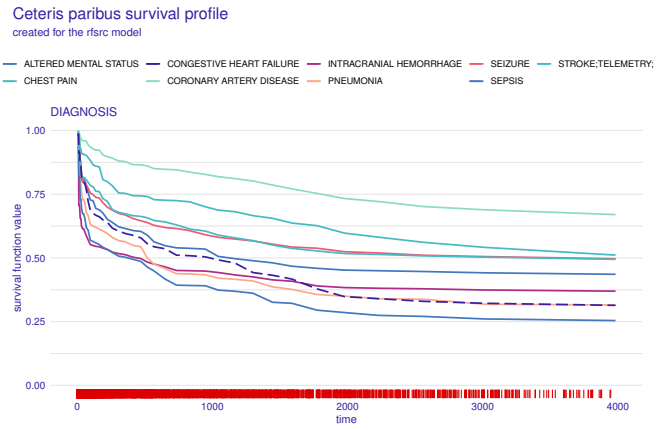

Fig. 3. Predictions for patient 16680 under different diagnoses – RSF.

testing. Under the Weibull distribution with scale 0.025 and shape 1.0, the Cox model with elastic net regularization performed the best on the test set. However, as the Weibull distribution was changed, presenting more risky environments with less censoring, the traditional Cox model surpassed the Cox model with regularization. Moreover, it is interesting to note that under the Weibull distribution with a parameter of scale 0.070 and shape 1.1, the conditional inference forests, the CoxBoost and the oblique random survival forest all had concordance index below 0.5, meaning that it was worse than random guessing (!).

Subsequently, the number of instances was increased to 2500 in order to see how a bigger dataset would influence the performances of each algorithm and also to see if the level of censoring would still be as influential as it was with the small dataset. First of all, from Table I, it is visible that once the size of the dataset is significantly increased, the influence of the censoring is decreased, and the performances of the algorithms do not tend to differ that much among different distribution settings. The best performing models, under this size of the dataset, were the CoxBoost and the Cox model with elastic net regularization.

TABLE I
CONCORDANCE INDEX FOR EACH OF THE ALGORITHM UNDER EACH SETTING

Weibull parameters	W(1.0, 0.025)	W(1.0, 0.060)	W(1.1, 0.070)
n=100			
Cox	0.788	**0.701**	**0.625**
RSF	0.765	0.591	0.554
CIF	0.671	0.468	0.391
CoxBoost	0.782	0.552	0.489
Cox_elastic	**0.882**	**0.701**	0.620
ORSF	0.647	0.565	0.467
n=2500			
Cox	0.619	0.641	0.647
RSF	0.623	0.641	0.635
CIF	0.608	0.628	0.632
CoxBoost	**0.633**	0.634	**0.654**
Cox_elastic	0.620	**0.642**	0.649
ORSF	0.621	0.638	0.640
n=5000			
Cox	0.638	0.627	0.626
RSF	0.629	0.618	0.614
CIF	0.636	0.625	0.619
CoxBoost	0.640	0.633	0.627
Cox_elastic	**0.641**	**0.634**	**0.630**
ORSF	0.628	0.627	0.622

Finally, the size of the dataset was even more increased to 5000 observations. This was done in order to see whether adding more instances would result in better model performance. In comparison with the setting, where only 2500 observations were generated, the performances improved only a little and, in some cases, got even a little bit worse. From that we can conclude that beyond a certain point, adding additional observations, which come from the same distribution, is bringing smaller and smaller value. The Penalized Cox model appears to be the best-performing model for large datasets, which aligns with our expectations, as penalization can effectively address the imbalance between the number of observations and covariates.

Additionally, we assessed each model's ability to distinguish between informative variables and random noise under different simulation settings. Under the settings of only 100 observations, algorithms like the RSF or the traditional Cox model tended to struggle in recognizing important variables and were only able to do better once the number of observations was significantly increased. On the other hand, the *Cox model with elastic net regularization was able to recognize the majority of important variables even under the setting of only 100 observations*. This likely contributed to its superior performance under that setting.

V. CONCLUSION AND FUTURE RESEARCH IDEAS

Overall, machine learning approaches demonstrated promising performance that was at least competitive with the traditional Cox model on both real-world and simulated data. Moreover, machine learning models offer valuable flexibility in capturing complex, non-linear relationships between covariates and survival outcomes. Applying techniques such as elastic net regularization combined with the Cox model under a penalized framework proved especially effective in scenarios with limited data availability or an imbalance between observations and covariates.

Future research could explore deep learning and multimodal data integration, which may enhance performance by providing richer patient information. Additionally, leveraging time-dependent covariates could reveal important patterns from measurements collected during a patient's stay. Moreover, future work should incorporate formal statistical testing (e.g., bootstrapped confidence intervals for the concordance index) to enable more rigorous model comparisons. While this study focused on a single ICU dataset, evaluating models across multiple real-world datasets would help assess generalizability across healthcare settings.

VI. Acknowledgment

This paper is supported by the grant IG410035 with no. F4/51/2025, which has been provided by the Internal Grant Agency of the Prague University of Economics and Business.

References

[1] D. R. Cox, "Regression models and life-tables," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 34, no. 2, pp. 187–220, 1972. [Online]. Available: http://www.jstor.org/stable/2985181

[2] P. Chapfuwa, C. Li, N. Mehta, L. Carin, and R. Henao, "Survival cluster analysis," Feb. 2020. doi: https://dx.doi.org/10.48550/ARXIV.2003.00355

[3] L. Štěpánek, F. Habarta, I. Malá, and L. Marek, "Analysis of asymptotic time complexity of an assumption-free alternative to the log-rank test," in *Proceedings of the 2020 Federated Conference on Computer Science and Information Systems*, vol. 21. IEEE, Sep. 2020. ISSN 2300-5963 p. 453–460. [Online]. Available: http://dx.doi.org/10.15439/2020F198

[4] ——, "Non-parametric comparison of survival functions with censored data: A computational analysis of greedy and monte carlo approaches," in *Proceedings of the 19th Conference on Computer Science and Intelligence Systems (FedCSIS)*, vol. 39. Polish Information Processing Society, Oct. 2024. ISSN 2300-5963 p. 725–730. [Online]. Available: http://dx.doi.org/10.15439/2024F223

[5] V. Van Belle, K. Pelckmans, J. Suykens, and S. Van Huffel, "Support vector machines for survival analysis," 2007.

[6] C. Fouodo, I. König, C. Weihs, A. Ziegler, and M. Wright, "Support vector machines for survival analysis with r," *The R Journal*, vol. 10, no. 1, p. 412, 2018. doi: https://dx.doi.org/10.32614/rj-2018-005

[7] L. Štěpánek, F. Habarta, I. Malá, L. Marek, and F. Pazdírek, "A machine-learning approach to survival time-event predicting: Initial analyses using stomach cancer data," in *2020 International Conference on e-Health and Bioengineering (EHB)*. IEEE, Oct. 2020. doi: https://dx.doi.org/10.1109/ehb50910.2020.9280301

[8] L. Štěpánek, F. Habarta, I. Malá, L. Štěpánek, M. Nakládalová, A. Boriková, and L. Marek, "Machine learning at the service of survival analysis: Predictions using time-to-event decomposition and classification applied to a decrease of blood antibodies against covid-19," *Mathematics*, vol. 11, no. 4, p. 819, Feb. 2023. doi: https://dx.doi.org/10.3390/math11040819

[9] J. L. Katzman, U. Shaham, A. Cloninger, J. Bates, T. Jiang, and Y. Kluger, "Deepsurv: personalized treatment recommender system using a cox proportional hazards deep neural network," *BMC Medical Research Methodology*, vol. 18, no. 1, Feb. 2018. doi: https://dx.doi.org/10.1186/s12874-018-0482-1

[10] C. Lee, W. Zame, J. Yoon, and M. Van der Schaar, "Deephit: A deep learning approach to survival analysis with competing risks," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, Apr. 2018. doi: https://dx.doi.org/10.1609/aaai.v32i1.11842

[11] S. Wiegrebe, P. Kopper, R. Sonabend, B. Bischl, and A. Bender, "Deep learning for survival analysis: a review," *Artificial Intelligence Review*, vol. 57, no. 3, Feb. 2024. doi: https://dx.doi.org/10.1007/s10462-023-10681-3

[12] F. E. J. Harell, *Regression Modeling Strategies*, 2nd ed., ser. Springer eBook Collection. Cham: Springer, 2015. ISBN 9783319194257

[13] H. Ishwaran, U. B. Kogalur, E. H. Blackstone, and M. S. Lauer, "Random survival forests," *Annals of Applied Statistics 2008, Vol. 2, No. 3, 841-860*, vol. 2, no. 3, Sep. 2008. doi: https://dx.doi.org/10.1214/08-aoas169

[14] H. Binder and M. Schumacher, "Incorporating pathway information into boosting estimation of high-dimensional risk prediction models," *BMC Bioinformatics*, vol. 10, no. 1, Jan. 2009. doi: https://dx.doi.org/10.1186/1471-2105-10-18

[15] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2009. ISBN 978-0-387-84857-0

[16] N. Simon, J. Friedman, T. Hastie, and R. Tibshirani, "Regularization paths for cox's proportional hazards model via coordinate descent," *Journal of Statistical Software*, vol. 39, no. 5, 2011. doi: https://dx.doi.org/10.18637/jss.v039.i05

[17] T. Hothorn, K. Hornik, and A. Zeileis, "Unbiased recursive partitioning: A conditional inference framework," *Journal of Computational and Graphical Statistics*, vol. 15, no. 3, pp. 651–674, Sep. 2006. doi: https://dx.doi.org/10.1198/106186006x133933

[18] L. Breiman, J. Friedman, C. J. Stone, and R. Olshen, *Classification and Regression Trees*, 1st ed. New York: Chapman and Hall/CRC, 1984. ISBN 0-534-98053-8

[19] B. C. Jaeger, S. Welden, K. Lenoir, J. L. Speiser, M. W. Segar, A. Pandey, and N. M. Pajewski, "Accelerated and interpretable oblique random survival forests," *Journal of Computational and Graphical Statistics*, vol. 33, no. 1, pp. 192–207, Aug. 2023. doi: https://dx.doi.org/10.1080/10618600.2023.2231048

[20] F. E. Harrell, "Evaluating the yield of medical tests," *JAMA: The Journal of the American Medical Association*, vol. 247, no. 18, p. 2543, May 1982. doi: https://dx.doi.org/10.1001/jama.1982.03320430047030

[21] M. A. Ahmad, A. Teredesai, and C. Eckert, "Interpretable machine learning in healthcare," in *2018 IEEE International Conference on Healthcare Informatics (ICHI)*. IEEE, Jun. 2018. doi: https://dx.doi.org/10.1109/ichi.2018.00095

[22] F. Xu, H. Uszkoreit, Y. Du, W. Fan, D. Zhao, and J. Zhu, *Explainable AI: A Brief Survey on History, Research Areas, Approaches and Challenges*. Springer International Publishing, 2019, pp. 563–574. ISBN 9783030322366

[23] M. Spytek, M. Krzyziński, S. H. Langbein, H. Baniecki, M. N. Wright, and P. Biecek, "survex: an r package for explaining machine learning survival models," 2023.

[24] M. Krzyziński, M. Spytek, H. Baniecki, and P. Biecek, "Survshap(t): Time-dependent explanations of machine learning survival models," 2022. doi: https://dx.doi.org/10.48550/ARXIV.2208.11080

[25] M. S. Kovalev, L. V. Utkin, and E. M. Kasimov, "Survlime: A method for explaining machine learning survival models," 2020.

[26] A. E. Johnson, T. J. Pollard, L. Shen, L.-w. H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. Anthony Celi, and R. G. Mark, "Mimic-iii, a freely accessible critical care database," *Scientific Data*, vol. 3, no. 1, May 2016. doi: https://dx.doi.org/10.1038/sdata.2016.35

[27] S. L. Brilleman, R. Wolfe, M. Moreno-Betancur, and M. J. Crowther, "Simulating survival data using the simsurv r package," *Journal of Statistical Software*, vol. 97, no. 3, 2021. doi: https://dx.doi.org/10.18637/jss.v097.i03