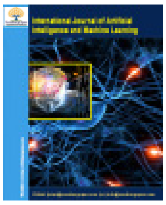
SvedbergOpen
DISSEMINATION OF KNOWLEDGE

International Journal of Artificial Intelligence and Machine Learning
Publisher's Home Page: https://www.svedbergopen.com/



Research Paper Open Access

Machine Learning Algorithms for Survival Analysis: Advantages, Disadvantages, and Examples

Diego Vallarino[1*]

[1]Independent Researcher, Madrid, Spain. E-mail: diego.vallarino@gmail.com

Article Info

Volume 4, Issue 1, January 2024
Received : 12 August 2023
Accepted : 09 December 2023
Published : 05 January 2024
doi: 10.51483/IJAIML.4.1.2024.10-21

Abstract

This paper studies the application of survival machine learning models in management for outcome prediction based on the medical literature. Twenty survival models and over ten survival machine learning algorithms were analyzed to find their key advantages and disadvantages. In the first half of this study, we examine and evaluate the most prevalent models in terms of their similarities and differences, as well as their data types and evaluation strategies. We also highlight the concepts that all machine learning algorithms for survival analysis must adhere to. Four machine learning algorithms from each family (trees, multi-task, kernel, and deep network) were used to analyze a breast cancer dataset and two additional simulated datasets using the R coxed package. The results indicate how machine learning algorithms might be used to recommend medicines and improve population health by analyzing survival. Moreover, we establish the ideal approaches to use based on more than twelve limitations, such as suppressed data.

Keywords: *Survival analysis, Statistical inference, Survival machine learning, Time to event analysis*

1. Introduction

Survival analysis is essential for predicting patients' time-to-event outcomes and aiding healthcare practitioners in making the best treatment decisions (Wang *et al.*, 2019), not only in disease analysis or monitoring procedures but also in assisting with the quantitative and qualitative improvement of preventive medicine (lifestyle interventions, vaccine efficacy, screening programs, among others). In addition to its use in healthcare, survival analysis plays a key role in decision-making across a variety of disciplines of management.

Survival analysis approaches are frequently used in marketing, finance, risk assessment, and bankruptcy analysis to make educated decisions. By applying survival analysis, firms are able to evaluate the time-to-event consequences of numerous elements and adjust their strategy appropriately (Zelenkov, 2020). In recent

* *Corresponding author: Diego Vallarino, Independent Researcher, Madrid, Spain. E-mail: diego.vallarino@gmail.com*

years, machine learning algorithms have evolved into remarkable tools for survival analysis, providing precise and trustworthy estimates (Wang *et al.*, 2019; Hu *et al.*, 2021; Yuan *et al.*, 2022).

In light of the significance of machine learning models, the primary goal of this study is to identify the most frequent machine learning approaches for predictive survival analysis. This article provides a detailed examination of machine learning models for survival analysis, which may aid healthcare professionals and researchers in selecting the best suitable model for their datasets.

The structure of the paper is as follows. Based on the availability of the experiment's associated data, the first part provides an explanation of the models. In turn, objective data and qualitative indicators will be used to show the different methods for evaluating machine learning models. In addition to the strengths and flaws of the core models, the faults shared by all machine learning models for survival analysis will be analyzed.

The second portion analyzes in depth the use of four machine learning approaches to survival analysis. Presented is a comparison between a dataset including the R survival package and a dataset on breast cancer. The conclusion of the publication contains the study's findings and recommendations for future research paths.

2. Machine Learning Survival Models

Depending on the availability of data in this research, we undertake a comprehensive examination of numerous methodologies for survival analysis (Wang *et al.*, 2019). In addition, we highlight the main weaknesses of these algorithms and give a comprehensive analysis of the several measures used to evaluate their effectiveness. Finally, we explore the advantages and disadvantages of these machine learning models for survival analysis, which may be used to choose the optimal approach for various datasets.

2.1. Discussion of Algorithms Based on the Availability of Some Type of Data

Survival algorithms may predict a patient's survival after a diagnosis, but their usefulness is highly dependent on the availability and quality of patient data. There are several types of data that may affect the use of these methods. For instance, basic patient data consisting of demographic and clinical information such as sickness stage, comorbidities, and treatment (Hair and Fávero, 2019; Maharana *et al.*, 2022).

Using datasets containing censored data, competitive risk data, or even data that displays longitudinal patient information might potentially be problematic (Barrett *et al.*, 2011). Survival algorithms are applicable despite some restrictions (Jin *et al.*, 2021; Cuperlovic-Culf, 2018).

In the next paragraphs, we will discuss the implications of this kind of data for survival analysis.

2.1.1. Baseline Agent Data

Essential to the development of a survival strategy for healthcare practitioners is basic patient information. Along with clinical data such as illness stage, comorbidities, and treatment history, demographic information such as age, gender, race, and ethnicity may have a substantial influence on a patient's survival rate. Although we have evidenced the characteristics of the baseline data with the most representative example in the literature on survival analysis, the most characteristics apply to the analysis of customers in the purchasing process, employees in the work process in the company, or the company in its life process over the years.

Developing a survival strategy requires in-depth understanding and analysis of several variables that might influence a patient's prognosis (Jin *et al.*, 2021; Cuperlovic-Culf, 2018). Various algorithms, including trees, forests, neural networks, deep learning, multitasking, boosting, and "others," may be used by healthcare providers to create survival forecasts (Thenmozhi *et al.*, 2019; Zhao *et al.*, 2022). To minimize mistakes and biases, it is vital to consider the limits and restrictions of these algorithms while generating predictions (Azodi *et al.*, 2020). Therefore, fundamental patient data is crucial for constructing accurate survival algorithms for successful patient care, but it is often insufficient for achieving a satisfactory performance in machine learning models.

2.1.2. Censored Data

The idea of suppressing data is characteristic of survival data. If the event of interest is death or bankruptcy of a company, the event time is censored for participants who are still alive at the conclusion of the research. This implies that the statistical analysis must continue without knowledge of the subject's date of death (Jiang, 2022; Vinzamuri *et al.*, 2017; . Basak *et al.*, 2022).

The only information available on his death is that it occurred after the conclusion of the research. In general, people who drop out of follow-up research are censored since they are often lost to follow-up and the timing of their occurrence is unclear (Raghunathan, 2004). The date of the occurrence is unobserved, but it is not a missing data point either, since these two categories of unobserved data have distinct properties and empirical interpretations (Yuan *et al.*, 2022).

For right-censored topics, the sole known fact is that their incident happened after the censorship period. If the research had been continued (or if the volunteers had not dropped out), ultimately the result of interest would have been seen for all participants. Conventional statistical approaches for analyzing survival data assume censoring to be independent or non-informative (Khan and Zubek, 2008).

This implies that, at a given point in time, the subjects who remain in follow-up have the same future risk for the occurrence of the event as those who are no longer followed (either due to censorship or abandonment of the study), as if the losses to follow-up were random and therefore not informative (Basak *et al.*, 2022).

Current research clearly demonstrates that the handling of censored data is essential in order to have an accurate view of the survival analysis experiment to be conducted (Jiang, 2022). Therefore, the focus of this study will be to determine the optimal ways for integrating censored data, both from the right (the most prevalent in analytic models) and from the left. The latter have not been widely analyzed in the literature, although time-to-event statistical analysis may give a lot of hints (Yuan *et al.*, 2022; Cui *et al.*, 2020).

When dealing with survival data, it is typical to encounter censored data, which happens when the precise timing of an event is unknown, but it is known that the event did not occur before or after a certain period. There are three forms of censored data: right-censored data, interval censored data, and left-censored data. There are a number of excellent algorithms available for managing massive quantities of filtered data (Yuan *et al.*, 2022; Cui *et al.*, 2020).

Survival Random Forest is one method that can effectively manage restricted data. It is an approach for machine learning that builds numerous decision trees and combines their predictions (Jin *et al.*, 2021; Zhao *et al.*, 2022; Jin Ziwei and Shang, 2020). Multi-Tasking Linear Regression (MTLR) is an additional approach that can effectively manage censored data. It employs a Bayesian technique to estimate the survival time distribution and is beneficial when dealing with many outcomes (Wang *et al.*, 2017). XGboost is another a well-known algorithm that can deal with enormous amounts of censored data with both continuous and categorical variables (Barnwal *et al.*, 2022).

In the following portion of this study, we shall give evidence supporting assertions.

2.1.3. Competitive Events/Risk Data

A competing risk is an occurrence that impedes the occurrence of the primary event of interest (Yuan *et al.*, 2022; . Barrett *et al.*, 2011; Nevo and Gorfine, 2020; Nevo *et al.*, 2022). For instance, in a study where the primary outcome was time to default, receive family help was included as a competing event. Therefore, a person who receives family help to pay the debt no longer runs the risk of defaulting on the credit we are analyzing. A subject will not be seen to default after accessing family support to pay off that specific debt, regardless of the duration of the person's follow-up.

In healthcare research including various categories of occurrences with nonfatal outcomes, conflicting risks may exist when deciding which kind of event occurred first. Thus, a study with three kinds of occurrences is possible: the diagnosis of heart illness, the diagnosis of cancer, and death. Each kind of occurrence is a competing danger, as a cancer diagnosis before a heart disease or death prevents the latter two from occurring first (Nevo and Gorfine, 2020).

Conventional techniques to survival data analysis assume the lack of conflicting dangers. The hazards are considered to be independent if information describing a subject's risk of experiencing one kind of event does not transmit information regarding the subject's risk of experiencing the other type of event (Nevo and Gorfine, 2020; Gorfine and Zucker, 2022). The tactics mentioned in this article are relevant in both circumstances in which competing risks are independent of one another and situations in which competing risks are not independent of one another. In biological applications, biology often exposes at least some association between contradictory hazards, which in many instances may be exceedingly severe.

Consequently, distinct competing dangers may be rare in biological applications. When analyzing survival data with conflicting risks, analysts usually exclude individuals when a competing event occurs. Consequently, if the outcome is time to death from cardiovascular causes, an analyst may consider a subject censored if that subject dies from reasons other than cardiovascular. Censoring persons at the point of death from causes other than cardiovascular disease may be challenging.

As with conventional survival analysis, the purpose of competitive event data analysis is to assess the probability of an event among numerous possible occurrences across time, allowing people to fail competitive events.

Therefore, recognizing that competing events will always develop when longitudinally analyzing several patients and calculating the time to event is essential for the research proposed in this proposal (Nevo *et al.*, 2022). The primary challenge is to develop a statistical method capable of analyzing the relevance of competing events and appreciating the value of the information they provide for survival analysis (Peng and Xiang, 2019). Other competing events that are not considered important a priori may become so due to their effects on subsequent temporal sequences, within or outside the period of research (Gorfine and Zucker, 2022).

The following are techniques for addressing the problem of event risk. Competitive Survival Analysis (CSA) is a statistical method for examining survival data in situations where several events have the ability to influence the outcome of interest. Using competition data and the Cox proportional hazards model, CSA simulates the hazard rates of both the event of interest and the competing event. This allows for a more precise prediction of the intended outcome while accounting for the effect of competing events (Hong *et al.*, 2022).

The Cumulative Incidence Function (CIF) is a statistical technique used to estimate the probability that an event of interest will occur in the setting of competing risks. The CIF computes the marginal probability of meeting the event of interest at a particular time point, taking competing events into consideration. This provides for a more accurate prediction of the probability of the event of interest when competing occurrences are taken into account (Lambert,, 2017).

Lunn-McNeil (LM) is a statistical technique used to forecast the probability of seeing the event of interest vs a competing event in the presence of competing risks. Developing a dummy variable for the competing event, which is subsequently included into the Cox proportional hazards model, is required. This allows for a more accurate estimation of the event's hazard rate while correcting for the impact of competing events (Huszti *et al.*, 2011).

2.1.4. Longitudinal Agent Data

Longitudinal patient data consists of information obtained over time on a patient's health status or other factors that may affect their health. This data may be rather diverse, ranging from changes in the individual's income to work status. This kind of information may provide valuable insight into the patient's overall health and its progression over time. Covariates are variables that may be associated with the outcome of interest but are not of primary concern; their effects may be accounted for by including them in the analysis (Thenmozhi *et al.*, 2019).

Survival analysis may leverage longitudinal patient data to improve forecast accuracy and provide a more comprehensive exploration of the factors affecting the result of interest. Over time, information on a patient's medication use, diet, or exercise habits, for example, might be collected and incorporated as research factors. These factors provide additional information that may have an indirect effect on the patient's survival; including them into the research may improve the accuracy of prediction (Thenmozhi *et al.*, 2019; Nevo *et al.*, 2022).

In addition, longitudinal data is often collected on a regular basis and offers information on the evolution of a patient's health over time. This information may be used to replicate the time-varying effects of variables and provide a more comprehensive analysis of how these factors affect the intended outcome.

As we have identified, the more data in our collection, the more information we can extract from them. This suggests that as we build survival analysis experiments with a larger number and variety of data points, the performance of machine learning algorithms will improve and their prediction potential will increase.

Obviously, it is essential to recognize that it is pertinent to comprehend the most effective strategies for managing data with specific features and peculiarities, such as NAs, missing data, censored data, and competitive events, among other forms of data.

2.2. Common Weaknesses for Survival Machine Learning Algorithms and Some Solutions

Survival analysis is performed with the use of machine learning algorithms, despite the fact that these techniques have comparable limits and dangers (Jin *et al.*, 2021; Cuperlovic-Culf, 2018; Libbrecht and Noble, 2015; Tarca *et al.*, 2007). In this part, we will examine these restrictions and some possible remedies.

The most fundamental disadvantage of utilizing machine learning models in survival analysis is the lack of interpretability. Comprehending "black box" machine learning models is difficult (Azodi *et al.*, 2020; Guidotti *et al.*, 2018). However, they do not show the underlying correlations between the factors employed to make such projections (Miller, 2018). This is a crucial criterion to bear in mind when using machine learning models to survival analysis, since understanding the underlying correlations between the variables might alter therapy and other intervention efforts (Chai *et al.*, 2021; . Zhou *et al.*, 2022).

In addition to interpretability, overfitting is an issue when using machine learning models to survival analysis (Libbrecht and Noble, 2015). Overfitting occurs when a model captures too much of the noise in the data and does not generalize well to new data. This may lead to erroneous predictions since the model is unable to accurately capture the underlying data linkages. In survival analysis, overfitting may result in inaccurate forecasts of the time until an event occurs (Tarca *et al.*, 2007).

Although this is a common flaw in many machine learning investigations, it becomes glaringly apparent in survival analysis algorithms due to the addition of a new layer of complexity, the time variable. This increases the complexity of the overfitting issue (Yin *et al.*, 2022).

Thirdly, the use of machine learning models in survival analysis for prediction is constrained by the need for an enormous amount of data in each dataset (Azodi *et al.*, 2020;). Training machine learning algorithms often requires large amounts of data, which may be difficult to collect in the medical and social sciences.

Moreover, datasets generally include hidden information, which may lead to biased results. Censored data refers to observations in which the outcome of interest, such as death or disease recurrence, is absent (Jiang, 2022; Vinzamuri *et al.*, 2017; Basak *et al.*, 2022). Although it is not a unique flaw of machine learning algorithms, the likelihood of discovering repressed information increases when processing vast amounts of data; hence, it must be seen as a relative weakness in these instances.

Using visualization to examine machine learning model predictions may also reveal a variable's underlying relationships. The use of data augmentation techniques, such as synthetic data (Haradal *et al.*, 2018; Pérez *et al.*, 2023), may help alleviate the issue of inadequate data.

Using data augmentation techniques, it is feasible to generate more data points for use in training the model. Stratification, which includes separating the study into multiple time periods, is a frequent strategy for reducing the time-effect-in-covariances shortcomings. Thus, the effect of time on the result may be examined with more precision. Using time-dependent covariates allows us to analyze how the link between a certain variable and the result varies over time (Jin *et al.*, 2021; Haradal *et al.*, 2018; Pérez *et al.*, 2023; Mumuni and Mumuni, 2022).

2.3. Specific Pros and Cons for Survival Machine Learning Algorithms

There are potential answers to these problems despite the limitations of machine learning techniques in survival analysis. Creating predictions using ensembles of machine learning models is one approach. Ensembles combine the predictions of several models to get a more precise forecast (Jin *et al.*, 2021; Azodi *et al.*, 2021; Wang *et al.*, 2017). This may reduce the risk of overfitting and improve the forecast's accuracy.

The majority of machine learning models can adapt to fresh data. This is a distinguishing feature between machine learning models and classical survival models. Algorithms may "learn" by accessing fresh data, which has a similar distribution to the training data. If the behavior of the new data is comparable to that of the training data, there is no need to retrain the algorithm (Alyass *et al.*, 2015).

The following matrix compares the performance of various machine learning models for survival analysis based on several characteristics, including censored data, missing data, small and large number of observations, number of variables, overfitting, interpretability, covariate independence, computational time, hyperparameter robustness competitive events, and non-linear relationships (Table 1).

Each column in the matrix represents a feature that a single model can handle. The values range from 1 (poor performance) to 5 (excellent performance) (good performance).

According to the matrix, there is no one model that excels across all criteria. Rather, each model has benefits and weaknesses, and the model selected should be based on the particular characteristics of the data and the subject of the study.

Table 1: Essential Features of Machine Learning Survival Models

	Surv. Tree	**Random Forest**	**MTLR**	**SVM**	**XG Boosting**	**Deep Surv**	**Deep Hit**
Censored Data	2	5	4	3	4	3	3
Missing Data	2	3	2	2	2	3	3
Small number Observations	4	4	5	4	4	4	4
Big number of Observations	2	5	3	3	4	5	5
Number Variables	3	4	4	3	4	5	5
Over fitting	5	4	4	4	4	2	2
Interpretability	5	3	3	2	2	2	2
Independency	4	3	2	2	4	5	5
Computational time	4	4	4	2	3	3	3
Hyper-param Sensitive	4	4	4	4	3	2	2
Competitive Events	2	5	4	3	4	3	3
Non-linear relationships	3	4	4	3	3	4	4
Processing time (+10 var x +40.000 obs.)	4	4	4	2	4	3	3
Punctuation	**44**	**52**	**47**	**37**	**43**	**45**	**45**

Note: A valuation of 1 = low ML performance and 5 = high ML performance.

Source: Own Elaboration

If the dataset contains missing or censored data, decision tree-based models, such as Survival Tree and Survival Random Forest, may be suitable (Basak *et al.*, 2022; Jin Ziwei and Shang, 2020; Bertsimas *et al.*, 2022). If the dataset comprises a large number of observations and variables, deep learning-based models such as DeepSurv and DeepHit may be preferable due to their potential to detect complicated patterns in the data (Lee *et al.*, 2018; Lee *et al.*, 2020; Miscouridou *et al.*, 2018; Hao *et al.*, 2021). For datasets with a small number of observations, MTLR may be preferred since it can control the small sample size and reduce overfitting (Wang *et al.*, 2017).

The decision tree-based model (Survival Tree) is the most interpretable of the matrix models when it comes to interpretability (Basak *et al.*, 2022). It is important to note, however, that the majority of models in the matrix have poor interpretability, a common flaw of machine learning models.

In terms of computation time, decision tree-based models (Survival Tree and Random Forest) and MTLR are the fastest, but deep learning-based models (DeepSurv and DeepHit) are slower due to their complex (Basak *et al.*, 2022; Wang *et al.*, 2017; Chai *et al.*, 2021; Cartocci *et al.*, 2021). Consequently, the machine learning model used should rely on the characteristics of the dataset and the research problem being addressed. It is essential to evaluate the pros and drawbacks of each model and choose the one that best matches the analysis's goals.

3. Case Studies

We demonstrate the efficacy of four machine learning models on three distinct datasets: SimulatedA, SimulatedB, and NKI Breast Cancer Data (Lum *et al*, 2013). The SimulatedA is a data frame containing 2000 observations that were created using the coxed R package (Kropko and Harden, 2019). It has 10 variables, of which 30% are censored (Figure 1). For SimulatedB, the same library with the same properties has been utilized. The sole modification implemented was the consideration of 80% of censored data (Figure 2).

The NKI Breast Cancer Data (a variant of this dataset is available in the survival R package) contains survival information for 272 breast cancer patients, making it the largest dataset we evaluated (Figure 3).

Our empirical investigation sheds light on the efficacy of several machine learning models for survival analysis and their applicability to a variety of data sources. Following this, we shall display the accomplished outcomes.

For the assessment and comparison of models, we have used the Cindex, which permits a comparative evaluation of the predictive ability of the models and the codes described below.

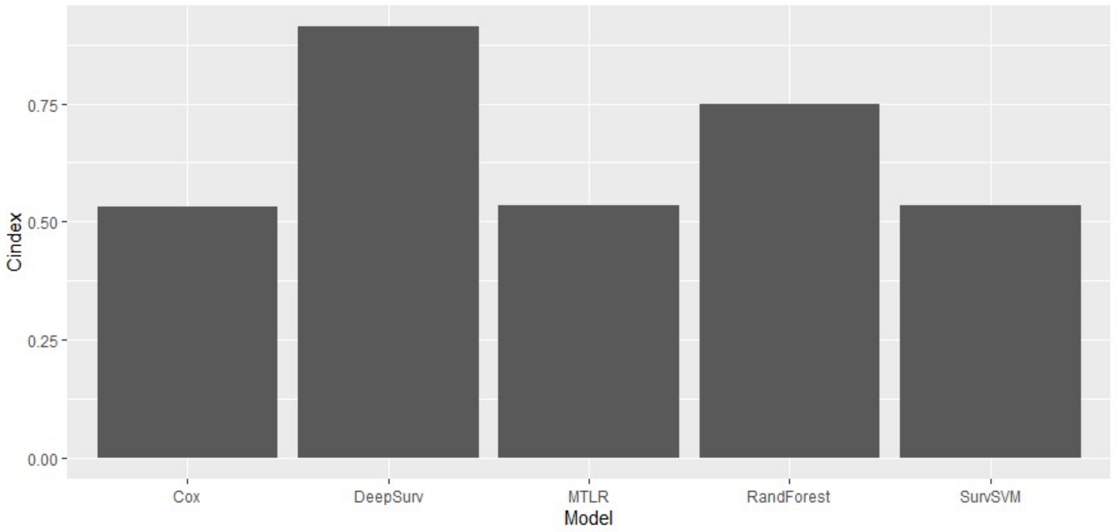


Figure 1: Results from Different Machine Learning Models on SimulatedA Dataset

Source: Own Elaboration

In two instances, the data suggests that the DeepSurv model fared the best, with a Cindex of 0.84512 in the first test. This implies that the model's estimated probability of the event happening are well calibrated and that the model can distinguish between patients who will experience the event and those who will not, as well as the time at which the event occurs.

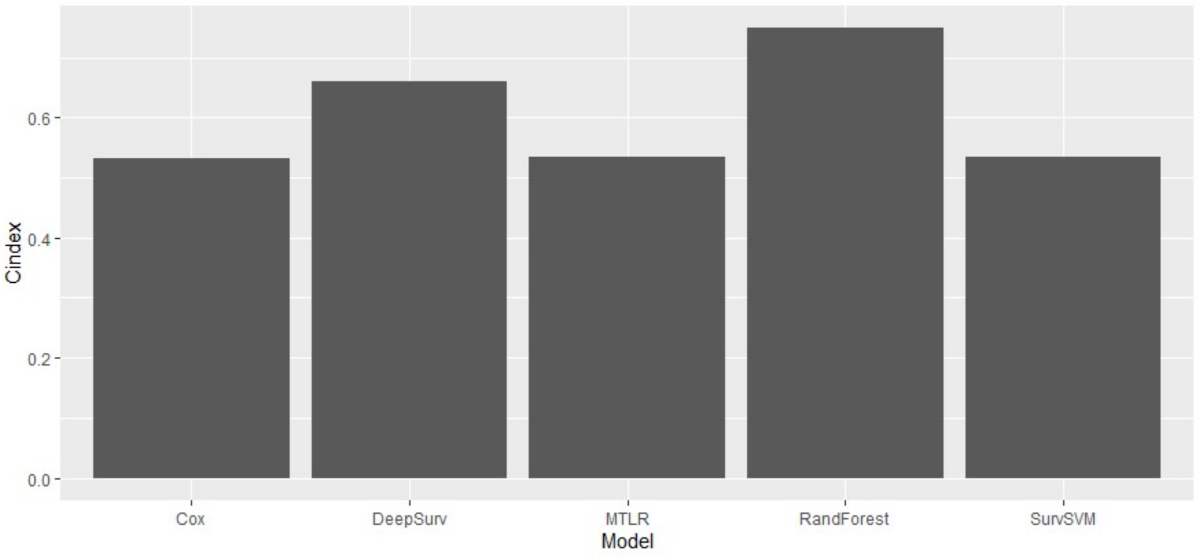


Figure 2: Results from Different Machine Learning Models on SimulatedB Dataset

Source: Own Elaboration

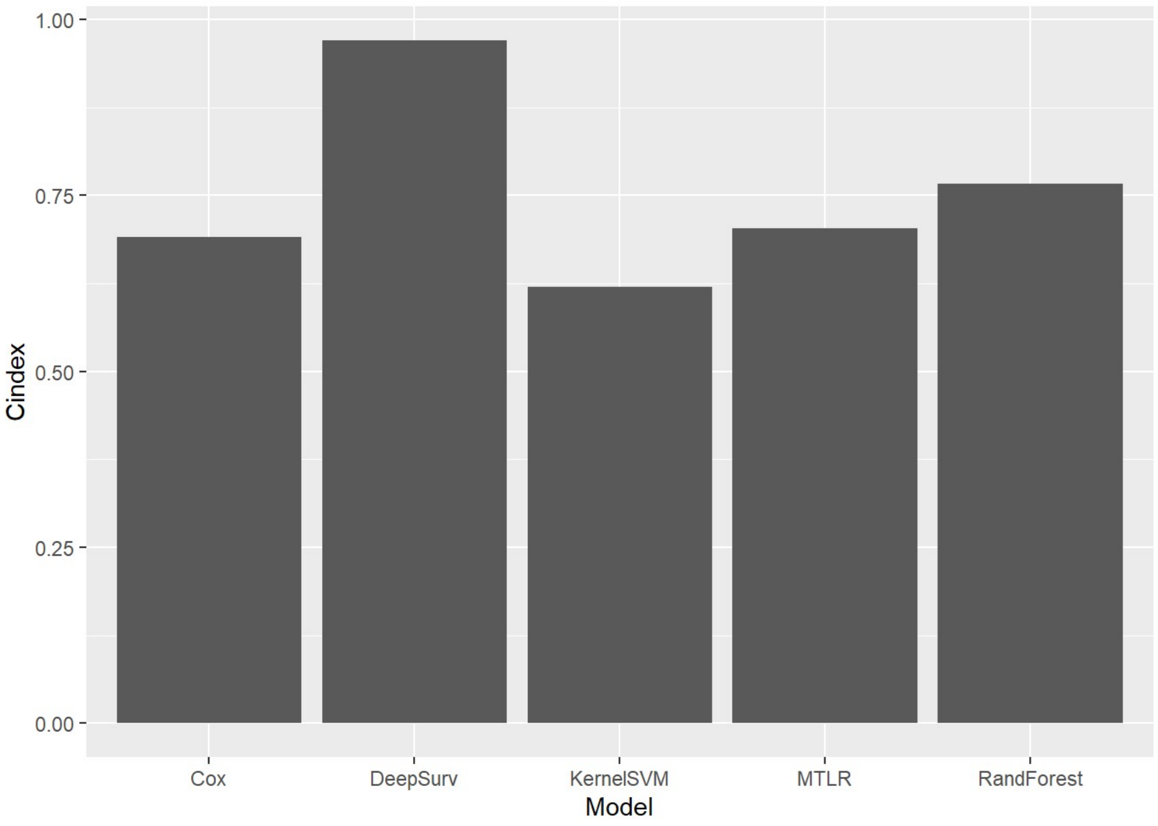


Figure 3: Results from Different Machine Learning Models on NKI Breast Cancer Data Dataset

Source: Own Elaboration

Table 2: Code for the Main Survival Machine Learning Models

Algorithm	R code
Survival Tree	LTRCART.obj <- LTRCART(Surv(time, End, status) ~ karno + age + trt, data=data.train_tree)LTRCIT.obj <- LTRCIT(Surv(time, End, status) ~ karno + age + trt, data=data.train_tree)
MTLR Model	mtlr(Surv(time, status)~., data = data.train, nintervals = 9)
XGBoost	xgboost(data, label = output_vector, max.depth = 4, eta = 1, nthread = 2, nrounds = 10,objective = "binary:logistic")
Random Forest	rfsrc(Surv(time, status) ~ ., data.train)
Survival Kernel SVM	survivalsvm(Surv(time, status) ~ ., data = data.train, type = "regression", gamma.mu = 1, opt.meth = "quadprog", kernel = "lin_kernel")
DeepSurv Model	deepsurv(data = data.train, frac = 0.3, activation = "relu", num_nodes = c(4L, 8L, 4L, 2L), dropout = 0.5, early_stopping = TRUE, batch_size = 32L, epochs = 100L)
DeepHit Model	deephit(data = data.train, frac = 0.3, activation = "relu", num_nodes = c(4L, 8L, 4L, 2L), dropout = 0.1, early_stopping = TRUE, epochs = 100L, batch_size = 32L)

Source: Own Elaboration

All of the other models, including Cox, MTLR, Random Forest, and Kernel SVM, performed well, but not as well as DeepSurv. The discrepancies in performance may be attributable to a number of variables, such as changes in the model's architecture, the kind of loss function used, or the hyperparameters selected for training.

When we adjust the proportion of suppressed data, something occurs. We have observed that the absolute and relative performance of models with identical attributes (with the same hyperparameters) is distinct. It is obvious what occurs when two datasets have identical characteristics, with the sole difference being the proportion of suppressed data.

In the cases of SimulatedA and SimulatedB, the relative performance of the DeepSurv algorithm has reduced significantly, as previously mentioned. Consistent with what was provided in the preceding part, we may infer that the amount of censored data had an effect on this method.

It is important to mention at this time that the standard parameters listed in Table 2 were utilized to create each of the models. In future research, it would be useful to go more into the analyses that specific model hyperparameter adjustments may yield.

4. Conclusion

In conclusion, despite the fact that machine learning algorithms have made significant strides in addressing some of the inadequacies of traditional survival analysis methods, there is still potential for improvement. All machine learning models have weaknesses like overfitting, interpretability, data quantity and quality, and temporal effects in variables. Although ensemble machine learning solutions have been developed to overcome these shortcomings, hyperparameter sensitivity persists in the most complex models.

In addition, since access to longitudinal or supplementary data is limited, many hypotheses cannot be tested retroactively, which poses a challenge to the validity of the results. Nevertheless, by modifying parameters as they gain knowledge from new data, machine learning algorithms provide enormous productivity gains.

To solve these limitations, future research should focus on developing more robust machine learning algorithms that can handle massive volumes of data while retaining their accuracy over time, as well as improving access to complementary and longitudinal data. Additionally, the issue of interpretability must be addressed to ensure that domain experts can comprehend and validate models.

Overall, machine learning algorithms are a crucial tool for survival analysis because they give excellent prediction abilities and enable researchers to uncover patterns and insights in large, complex datasets. With continuing investment in research and development, the capacity of machine learning for survival analysis will increase, boosting our understanding of disease and enhancing patient outcomes.

In order to advance the field of machine learning in survival analysis, there are a number of critical future steps that academics must investigate. A crucial step is to test different models with simulated data in order to assess their resistance to data quality, including repeatability of prediction results and hyperparameter sensitivity. By comparing the expected performance of multiple models, researchers may determine which algorithms are most effective for certain use situations.

An additional important next step is to research approaches for addressing the inadequacies of the current machine learning algorithms in survival analysis, such as overfitting, interpretability, and the temporal effect of variables. By addressing these problems, researchers will be able to develop more reliable and precise models that may be used to make treatment decisions.

Finally, it is essential to enhance the interpretability of certain algorithms so that domain experts can comprehend and assess them. This will allow clinicians and physicians to make better decisions and improve patient outcomes.

References

Alyass, A., Turcotte, M. and Meyre, D. (2015). From Big Data Analysis to Personalized MedicinefFor All: Challenges and Opportunities. *BMC Med Genomics*, 8(1), 33. doi: 10.1186/s12920-015-0108-y.

Azodi, C.B., Tang, J. and Shiu, S.H. (2020). Opening the Black Box: Interpretable Machine Learning for Geneticists. *Trends in Genetics*, 36(6), 442-455. Elsevier Ltd. doi: 10.1016/j.tig.2020.03.005.

Barnwal, A., Cho, H. and Hocking, T. (2022). Survival Regression with Accelerated Failure Time Model in XGBoost. *Journal of Computational and Graphical Statistics*, 31(4), 1292-1302. doi: 10.1080/10618600.2022.2067548.

Barrett, J.K., Siannis, F. and Farewell, V.T. (2011). A Semi-competing Risks Model for Data With Interval-censoring and Informative Observation: An Application to the MRC Cognitive Function And Ageing Study. *Stat Med*, 30(1), 1-10. doi: 10.1002/sim.4071.

Basak, P., Linero, A., Sinha, D. and Lipsitz, S. (2022). Semiparametric Analysis of Clustered Interval-censored Survival Data Using Soft Bayesian Additive Regression Trees (SBART). *Biometrics*, 78(3), 880-893. doi: 10.1111/biom.13478.

Bertsimas, D., Dunn, J., Gibson, E. and Orfanoudaki, A. (2022). Optimal Survival Trees. *Mach Learn*, 111(8), 2951-3023. doi: 10.1007/s10994-021-06117-0.

Cartocci, A., Cevenini, G. and Barbini, P. (2021). A Compartment Modeling Approach to Reconstruct and Analyze Gender and Age-grouped Covid-19 Italian Data for Decision-making Strategies. *J Biomed Inform*, 118, 103793. doi: https://doi.org/10.1016/j.jbi.2021.103793.

Chai, H., X. Zhou, Z. Zhang, J. Rao, H. Zhao, and Y. Yang, (2021). Integrating Multi-omics Data Through Deep Learning For Accurate Cancer Prognosis Prediction. *Comput Biol Med*, 134, 104481. doi: https://doi.org/10.1016/j.compbiomed.2021.104481.

Cui P. *et al.* (2020). Causal Inference Meets Machine Learning. in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Association for Computing Machinery, August, 3527-3528. doi: 10.1145/3394486.3406460.

Cuperlovic-Culf, M. (2018). Machine Learning Methods for Analysis of Metabolic Data and Metabolic Pathway Modeling," *Metabolites*, 8(1). doi: 10.3390/metabo8010004.

Gorfine M. and D. M. Zucker, (2022). Shared Frailty Methods for Complex Survival Data: A Review of Recent Advances. May, [Online]. Available: http://arxiv.org/abs/2205.05322

Guidotti, R., A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, (2018). A Survey of Methods for Explaining Black Box Models. *ACM Comput Surv*, 51(5). doi: 10.1145/3236009.

Hair J.F. and Fávero, L.P. (2019). Multilevel Modeling For Longitudinal Data: Concepts and Applications. *RAUSP Management Journal*, 54(4), 459-489, Dec. doi: 10.1108/RAUSP-04-2019-0059.

Hao, L., Kim, J. Kwon, S. and Do Ha, I. (2021). Deep Learning-based Survival Analysis for High-dimensional Survival Data. *Mathematics*, 9(11). doi: 10.3390/math9111244.

Haradal, S., Hayashi, H. and Uchida, S. (2018). Biosignal Data Augmentation Based on Generative Adversarial Networks. in *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 368-371. doi: 10.1109/EMBC.2018.8512396.

Hong, C., Yi, F. and Huang, Z. (2022). Deep-CSA: Deep Contrastive Learning for Dynamic Survival Analysis With Competing Risks," *IEEE J Biomed Health Inform*, 26(8), 4248-4257, doi: 10.1109/JBHI.2022.3161145.

Hu, L., Ji, J. and Li, F. (2021). Estimating Heterogeneous Survival Treatment Effect In Observational Data Using Machine Learning. *Stat Med*, 40(21), 4691-4713, Sep. doi: 10.1002/sim.9090.

Huszti, E., Abrahamowicz, M., Alioum, A.and Quantin, C. (2011). Comparison of Selected Methods for Modeling of Multi-State Disease Progression Processes: A Simulation Study. *Commun Stat Simul Comput*, 40(9), 1402-1421. doi: 10.1080/03610918.2011.575505.

Jin Ziwei and Shang. (2020). RFRSF: Employee Turnover Prediction Based on Random Forests and Survival Analysis. in W. and W.H. and Z.R. and Z.Y. Huang Zhisheng and Beek (Ed.), *Web Information Systems Engineering – WISE 2020*, 503-515. Cham: Springer International Publishing.

Jiang, R. (2022). A Novel Parameter Estimation Method for the Weibull Distribution on Heavily Censored Data. *Proc Inst Mech Eng O J Risk Reliab*, 236(2), 307-316. doi: 10.1177/1748006X19887648.

Jin, P., Haider, H., Greiner, R., Wei, S. and Häubl, G. (2021). Using Survival Prediction Techniques to Learn Consumer-specific Reservation Price Distributions. *PLoS One*, 16(4). doi: 10.1371/journal.pone.0249182.

Khan, F.M. and Zubek, V.B. (2008). Support Vector Regression for Censored Data (SVRc): A Novel Tool for Survival Analysis. in *2008 Eighth IEEE International Conference on Data Mining*, 863-868. doi: 10.1109/ICDM.2008.50.

Kropko J. and Harden, J.J. (2019). Coxed: An R Package for Computing Duration-Based Quantities from the Cox Proportional Hazards Model. *The R Journal*. 11, 38. 10.32614/RJ-2019-042.

Lambert, P.C. (2017). The Estimation and Modeling of Cause-specific Cumulative Incidence Functions Using Time-dependent Weights. *Stata J*, 17(1), 181-207. doi: 10.1177/1536867X1701700110.

Lee, C., Yoon, J. and van der Schaar, M. (2020). Dynamic-DeepHit: A Deep Learning Approach for Dynamic Survival Analysis With Competing Risks Based on Longitudinal Data. *IEEE Trans Biomed Eng*, 67(1), 122-133. doi: 10.1109/TBME.2019.2909027.

Lee, C., Zame, W.R., Yoon, J. and Van Der Schaar, M. (2018). DeepHit: A Deep Learning Approach to Survival Analysis with Competing Risks. [Online]. Available: www.aaai.org

Libbrecht, M.W. and Noble, W.S. (2015). Machine Learning Applications in Genetics and Genomics. *Nat Rev Genet*, 16(6), 321-332. doi: 10.1038/nrg3920.

Lum P.Y. *et al.* (2013). Extracting Insights from the Shape of Complex Data Using Topology. *Sci Rep*, 3(1), 1236. doi: 10.1038/srep01236.

Maharana, K., Mondal, S. and Nemade, B. (2022). A Review: Data Pre-processing and Data Augmentation Techniques. *Global Transitions Proceedings*, 3(1), 91-99. doi: https://doi.org/10.1016/j.gltp.2022.04.020.

Miller, T. (2019). Explanation in Artificial Intelligence: Insights From the Social Sciences. *Artif Intell*, 267, 1-38. doi: https://doi.org/10.1016/j.artint.2018.07.007.

Miscouridou, X., Perotte, A., Noémie, E. and Ranganath, R. (2018). Deep Survival Analysis: Nonparametrics and Missingness. *Proceedings of Machine Learning Research*, 85, 1-12.

Mumuni, A. and Mumuni, F. (2022). Data Augmentation: A Comprehensive Survey of Modern Approaches. *Array*, 16, 100258. doi: https://doi.org/10.1016/j.array.2022.100258.

Nevo, D. and Gorfine, M. (2020). Causal Inference for Semi-competing Risks Data. Oct. [Online]. Available: http://arxiv.org/abs/2010.04485

Nevo, D., Blacker, D., Larson, E.B. and Haneuse, S. (2022). Modeling Semi-competing Risks Data As a Longitudinal Bivariate Process. *Biometrics*, 78(3), 922-936, Sep. doi: 10.1111/biom.13480.

Peng, M., and Xiang, L. (2019). Joint Regression Analysis for Survival Data in the Presence of Two Sets of Semi-competing Risks. *Biometrical Journal*, 61(6), 1402-1416, Nov. doi: 10.1002/bimj.201800137.

Pérez, J., Arroba, P. and Moya, J.M. (2023). Data Augmentation Through Multivariate Scenario Forecasting in Data Centers using Generative Adversarial Networks. *Applied Intelligence*, 53(2), 1469-1486. doi: 10.1007/s10489-022-03557-6.

Raghunathan, T.E. (2004). What Do We Do with Missing Data? Some Options for Analysis of Incomplete Data. *Annu Rev Public Health*, 25(1), 99-117. doi: 10.1146/annurev.publhealth.25.102802.124410.

Tarca, A.L., Carey, V.J., wen Chen, X., Romero, R. and DrÎghici, S. (2007). Machine Learning and Its Applications to Biology. *PLoS Computational Biology*, 3(6). doi: 10.1371/journal.pcbi.0030116.

Thenmozhi, M., Jeyaseelan, V., Jeyaseelan, L., Isaac, R. and Vedantam, R. (2019). Survival Analysis in Longitudinal Studies for Recurrent Events: Applications and Challenges. *Clin Epidemiol Glob Health*, 7(2), 253-260. doi: https://doi.org/10.1016/j.cegh.2019.01.013.

Vinzamuri, B., Li, Y. and Reddy, C.K. (2017). Pre-Processing Censored Survival Data using Inverse Covariance Matrix based Calibration, *IEEE Transactions on Knowledge and Data Engineering*, 29(10), 2111-2124.

Wang, L., Li, Y., Zhou, J., Zhu, D. and Ye, J. (2017). Multi-task Survival Analysis. in *2017 IEEE International Conference on Data Mining (ICDM)*, 485-494. doi: 10.1109/ICDM.2017.58.

Wang, P., Li, Y. and Reddy, C.K. (2019). Machine Learning for Survival Analysis: A Survey. *ACM Comput Surv*, 51(6). doi: 10.1145/3214306.

Wang, P., Li, Y. and Reddy, C.K. (2017). Machine Learning for Survival Analysis: A Survey. Aug. [Online]. Available: http://arxiv.org/abs/1708.04649

Yin, Q., Chen, W., Wu, R. and Wei, Z. (2022). Cox-ResNet: A Survival Analysis Model Based on Residual Neural Networks for Gene Expression Data in *2022 IEEE International Conference on Networking, Sensing and Control (ICNSC)*, 1-6. doi: 10.1109/ICNSC55942.2022.10004157.

Yuan, H. *et al.* (2022). AutoScore-Imbalance: An Interpretable Machine Learning Tool for Development of Clinical Scores With Rare Events Data. *J Biomed Inform*, 129, 104072. doi: https://doi.org/10.1016/j.jbi.2022.104072.

Zelenkov, Y. (2020). Bankruptcy Prediction Using Survival Analysis Technique. in *Proceedings - 2020 IEEE 22nd Conference on Business Informatics, CBI 2020*, Institute of Electrical and Electronics Engineers Inc., Jun. pp. 141-149. doi: 10.1109/CBI49978.2020.10071.

Zhao, Z.L., Yu, H.J. and Cheng, F. (2022). An Analysis of Factors Affecting Agricultural Tractors' Reliability Using Random Survival Forests Based on Warranty Data. *IEEE Access*, 10, 50183-50194. doi: 10.1109/ACCESS.2022.3172348.

Zhou, F., Fu, L., Li, Z. and Xu, J. (2022). The Recurrence of Financial Distress: A Survival Analysis. *Int J Forecast*, 38(3), 1100-1115, Jul. doi: 10.1016/j.ijforecast.2021.12.005.

Cite this article as: Diego Vallarino (2024). Machine Learning Algorithms for Survival Analysis: Advantages, Disadvantages, and Examples. *International Journal of Artificial Intelligence and Machine Learning*, 4(1), 10-21. doi: 10.51483/IJAIML.4.1.2024.10-21.