IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 17 July 2023, accepted 28 August 2023, date of publication 5 September 2023, date of current version 12 September 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3312310

RESEARCH ARTICLE

Exploiting Censored Information in Self-Training for Time-to-Event Prediction

FATEME NATEGHI HAREDASHT[1,2], KAZEEM ADESINA DAUDA[1,2,3], AND CELINE VENS[1,2]

[1]Department of Public Health and Primary Care, Campus KULAK, KU Leuven, 8500 Kortrijk, Belgium
[2]ITEC—imec, KU Leuven, 8500 Kortrijk, Belgium
[3]Department of Mathematics and Statistics, Kwara State University, Malete 241103, Nigeria

Corresponding author: Fateme Nateghi Haredasht (fateme.nateghi@kuleuven.be)

This work was supported by KU Leuven Internal Funds under Grant 3M180314.

ABSTRACT A common problem in medical applications is predicting the time until an event of interest such as the onset of a disease, time to tumor recurrence, and time to mortality. Traditionally, classical survival analysis techniques have been used to address this problem. However, these techniques are of limited usage when considering nonlinear and interaction effects among biomarkers, and high profiling survival datasets. Although supervised machine learning techniques have shown some advantages over standard statistical methods in handling high-dimensional datasets, their application to survival analysis, particularly in the context of feature-based approaches, is at best limited. A major reason behind this is the difficulty in processing censored data, which is a common component of survival analysis. In this paper, we have transformed the time-to-event prediction problem into a semi-supervised regression problem. We utilize a self-training wrapper approach, where an outer layer guides the iterative refinement of predictions. This approach enhances the performance of our model by leveraging confident predictions from censored instances. The self-training wrapper is applied in conjunction with random survival forests as the base learner. In this approach, censored observations are introduced as partially labeled observations since their predicted time (target value) should exceed the censoring time. First, the algorithm builds a base model over the observed instances and then augments them iteratively with highly confident predictions over the censored set, using a smart stopping criterion based on the censoring time. The proposed approach has been evaluated and compared on fifteen real-world survival analysis datasets, including clinical and high-dimensional data. The ability of our proposed approach to integrate partial supervision information within a semi-supervised learning strategy has enabled it to achieve competitive performance compared to baseline models, particularly in the case of a high-dimensional regime.

INDEX TERMS Random survival forest, self-training, semi-supervised learning, survival analysis.

I. INTRODUCTION

Survival analysis is a subfield of statistics concerned with the analysis of data where the outcome of interest is the time until a particular event of interest occurs. There is widespread use of survival analysis in medicine, where events of interest might include death, tumor recurrence, and hospital discharge, among others. Censoring, which can occur for various reasons such as drop-out, is one of the main challenges of survival analysis. Observations that are censored (right-censored or left-censored) cannot provide the true survival time, as, for example, in the right-censored case, we know that the observed time is an underestimate of the survival time [1].

Traditionally, methods like Cox Proportional Hazards (CPH) and Accelerated Failure Time (AFT) models have been widely used throughout literature to overcome censoring; however, these methods have been unable to cope with real-world datasets with hundreds or thousands of features. Additionally, these models are not able to incorporate the nonlinear relationship that exists between the features [2], [3].

The field of survival analysis has adopted many supervised machine learning algorithms in recent years, but the problem

The associate editor coordinating the review of this manuscript and approving it for publication was Seifedine Kadry.

of applying these techniques directly to censored data is challenging since the time to the event is only partially known. Sometimes the task is transformed into a binary classification task (does the event happen before a certain time?), in which censored data points are either eliminated [4] or their impact is diminished by a weighting procedure [2]. A number of machine learning algorithms have been successfully modified to employ censored information in survival analysis. For example, decision trees [5], artificial neural networks (ANN) [6], and support vector machines (SVM) [7]. Among the most popular ensemble-based frameworks are bagging survival trees [8] and random survival forests [9]. There has also been an extension of more advanced learning tasks such as active learning [10] and transfer learning [11] towards survival analysis.

Although in recent years, applying supervised machine learning-based techniques in the survival analysis domain has gained attention [12], semi-supervised learning (SSL) methods [13], [14] are also briefly addressed in the survival analysis literature. The study by Bair and Tibshirani [15], combines supervised and unsupervised learning to predict survival times for cancer patients. A supervised approach is used to select a subset of genes from a gene expression dataset that correlates with survival. Then, to identify cancer subtypes, unsupervised clustering is applied to these gene subsets. Having identified such subtypes, they apply supervised learning techniques again to classify future patients into the appropriate subgroups (low-risk or high-risk) or to predict their survival. The low- and high-risk groups are created by comparing the survival time to the median survival time. For the censored patients, based on the Kaplan-Meier survival curve for all the patients, they estimate the probability that a censored case survives a specified length of time and thus belongs to the low-risk and high-risk classes, respectively.

Roy et al. [16] modeled the time-to-event prediction as a multi-target regression problem, with censored observations modeled as partially labeled. More specifically, the different event times in the dataset are viewed as binary targets. For each data instance, it is specified whether it has experienced the event or not at each time stamp, using missing values when an instance has been censored after a certain period of time. Then they apply semi-supervised predictive clustering trees and ensembles thereof to the resulting data.

Furthermore, there has been some research that models a survival analysis task as a semi-supervised learning problem by employing a self-training strategy to predict event times from observed and censored data. Both [17] and [18] treat the censored data points as unlabeled, thus not taking into account the time-to-event information that these data points provide.

Although several previous studies have applied semi-supervised learning approaches to survival data analysis, not many have utilized the underlying information contained within the censored data, which is the fact that the target value for right-censored observations should be greater than the censoring time. Liang et al. [19] disregard data points for which the model predicts a value lower than the right-censored time points. They combine Cox proportional hazard (Cox) and accelerated failure time (AFT) model in a semi-supervised set-up to predict the treatment risk and the survival time of cancer patients. Regularization is used for gene selection, which is an essential task in cancer survival analysis. The authors found that many censored data points consistently violate the constraint that the predicted survival time should be higher than the censored time, restricting the full exploitation of the censored data. Therefore, in follow-up work [20], they embedded a self-paced learning mechanism called Cox-SP-AFT in their framework to gradually introduce more complex data samples in the training process, leading to a more accurate estimation for the censored samples. To estimate the coefficients of the AFT model, they introduce a loss function derived from the constraint that the survival time must not be less than the censoring time. As a result, if the estimated survival time of a sample is less than the censoring time, then this sample must be falsely labeled, and its loss value must be positive infinity. A censored sample, however, has a square loss function if it obeys the censoring condition. Then in order to select confident samples from the censored dataset, they define a threshold (age parameter) for the loss function in which the samples with losses smaller than the age parameter (α) will be kept at the training phase, otherwise will be assigned zero weight. It should be noted that a traditional parametric model (AFT) has been used for the training process. Specifically for such high-dimensional datasets that were used in the study, an advanced machine learning model that is capable of properly handling nonlinear relations between the features could be more effective and superior to the AFT.

In this paper, using a semi-supervised learning approach, we propose a new time-to-event prediction algorithm that utilizes the underlying information contained within the censored data. Specifically, this paper utilizes the widely used self-training wrapper technique [21], [22], which builds a classifier/regressor over the labeled (in our case, observed) data points and then augments the labeled set iteratively with highly confident predictions over the unlabeled (censored) set of data. Our approach uses random survival forests as the base learner [9] and compares the proposed algorithm's predictive performance with three competing methods based on fifteen real-life healthcare datasets.

II. BACKGROUND

The proposed method is explained after a brief introduction to survival analysis, followed by an exploration of the models that have been employed in our methodology, namely the random survival forest model, followed by a discussion of the self-training models.

A. SURVIVAL ANALYSIS

Survival analysis is a widely used subfield of statistics that was originally designed to predict the lifespan of patients in

a clinical setting. The primary objective of survival analysis is to predict time-to-event distributions based on features, address factors influencing the distribution, and determine their nature.

More specifically, for a given instance i, represented by a triplet (X_i, y_i, δ_i), where $X_i \in \mathbb{R}^P$ is the feature vector; δ_i is the binary event indicator (i.e., $\delta_i = 1$ for an uncensored instance and $\delta_i = 0$ for a censored instance); and y_i denotes the observed time and is equal to the survival time T_i for an uncensored instance and C_i for a censored instance; that is,

$$y_i = \begin{cases} T_i, & \text{if } \delta_i = 1. \\ C_i, & \text{if } \delta_i = 0. \end{cases} \tag{1}$$

The objective is to estimate the time to the event of interest denoted by T_j for a new instance j based on feature predictors described by X_j. For an arbitrary time point t, survival function $S(t)$ represents a probability that a specified event will not take place earlier than time t, i.e., $S(t) = P(T > t)$ [12].

The hazard function $h(t)$, is defined as $h(t) = f(t)/S(t)$, where $f(t)$ is the density function for the time to an event and $f(t) = -\frac{d}{dt}S(t)$. More specifically, $h(t)$ represents the likelihood of the event occurring at time t given that no event has occurred before time t [23]. The Cumulative Hazard Function (CHF) is defined as $H(t) = \int_0^t h(u)du$ which results in the following equation:

$$S(t) = e^{-H(t)} \tag{2}$$

where $H(t)$ and $S(t)$ denote the cumulative hazard function and the survival function, respectively.

In contrast to the above non-parametric methods, in the semi-parametric category, the Cox model [1] is the most commonly used regression analysis approach for survival data. In spite of being based on a parametric regression model, the Cox model is described as semi-parametric due to the fact that no knowledge of the underlying distribution of time to the event of interest is required [12]. Many real-world domains have accumulated high-dimensional data due to the development of data collection and detection techniques. An example of this would be datasets where the number of features (P) is much more than the number of instances (N). Therefore, a good prediction model cannot incorporate all of the information available in the feature set. In this regard, several different penalty functions including the Lasso (Lasso-Cox) [24], the Ridge (Ridge-Cox) [25], and the Elastic-Net (EN-Cox) [26] have been developed to identify the features that are most relevant to the outcome variable among what can be tens of thousands of features.

B. RANDOM SURVIVAL FOREST MODEL

Regression trees have been extended to survival data [5], and survival trees have been used in ensemble methods, such as bagging and boosting.

Random Survival Forest (RSF) [9] is a statistical algorithm that is widely used in machine learning for predicting the survival time of an individual or an event. It is an extension of the random forest [27] algorithm, which is a well-known ensemble method that combines multiple decision trees to make predictions. The primary distinguishing feature of RSF is that the trees are trained to split data based on both the predictor variables (features) and the survival time of the observations. This enables RSF to handle censored data. The RSF algorithm functions by randomly selecting a subset of features and constructing multiple decision trees. Each tree is built using a different bootstrap sample (B) of the data and a random subset of the features. In order to split the node into two child nodes, the best candidate feature and split point should be determined by the log-rank test [28]. Optimal splitting is one that maximizes survival differences between the child nodes. A stop criterion is used to decide when to stop growing the tree structure (for example, when the number of observed instances in the terminal nodes declines below a certain threshold). By the end of the analysis, the Nelson-Aalen estimator, which is a non-parametric estimator of the cumulative hazard function CHF, is used to determine the CHF associated with each node in the tree [29].

The CHF is the same for all cases within the same terminal node. The ensemble CHF is calculated as the average over the CHF of the B survival trees.

The RSF algorithm has several advantages over traditional survival analysis methods, such as Cox regression. RSF can handle high-dimensional data, nonlinear relationships between predictor variables and the outcome, and interactions between variables. It also provides measures of feature importance, which can help identify the most important features for survival. Random survival forest has been applied to a range of problems, including predicting the survival of cancer patients, the failure time of mechanical components, and the risk of loan default. The versatility and effectiveness of RSF have made it a popular algorithm in the field of survival analysis.

C. SELF-TRAINING MODEL

As a combination of supervised and unsupervised learning, semi-supervised learning (SSL) has been used in many applications [30], [31], [32]. In order to obtain a more accurate prediction model, SSL methods seek to make use of unlabeled data as well as labeled data. In some applications, it is difficult to achieve good performance with supervised techniques due to the relatively small number of labeled instances. This is due to the fact that labeling techniques are generally expensive and time-consuming. Consequently, over the years, many SSL techniques have been proposed [13], [14]. In this article, we will focus on self-training (also called self-learning) [21] which is one of the earliest approaches in semi-supervised learning. In recent years, self-training has gained popularity and has been used in different ways like deep neural networks [33], face recognition [34], and parsing [35]. By augmenting the training set with unlabeled instances, this framework overcomes the problem of insufficient labeled data. This wrapper algorithm starts by training

Algorithm 1 Self-Training

Require: Labeled data (L), Unlabeled data (U)
Ensure: Trained model
1: Train a base model using L
2: **while** $|U| \neq 0$ **do**
3: Make predictions for U using the trained model
4: Select T instances from U, where T contains unlabeled samples with high confidence predictions
5: Label all samples in T using the trained model
6: $L = L \cup T$
7: $U = U - T$
8: Retrain the model using L
9: **end while**

a model using a base learner on the labeled data set. Then the model assigns pseudo-labels to unlabeled data using its predictions, after which it augments the labeled data set with predictions for unlabeled instances that the model is most confident in by considering these confident pseudo-labeled unlabeled data as additional labeled points (see Algorithm 1). This process of pseudo-labeling and learning a new model continues until there is no more unlabeled data to pseudo-label. The process could be stopped before adding all the unlabeled data if a certain stopping criterion is defined. The stopping condition, the number of examples to be increased per iteration, and the definition of confidence is determined based on the problem being addressed. The selection of each of these three factors is crucial, particularly since the first two are often set arbitrarily or with costly parameter optimizations.

III. THE PROPOSED METHOD

In our proposed approach, we apply a semi-supervised learning approach, the widely used self-training wrapper technique, that was explained earlier [36]. The current work presented in this paper is an extension of the previous work [37]. While the previous study focused on predicting survival outcomes in the presence of unlabeled data, this current work builds upon the earlier findings by exploring the performance of the predictive model without the use of unlabeled data.

Using the self-training wrapper technique, we build the initial model using only the observed data points, then iteratively augment it with high-confidence predictions from the censored data. In other words, we treat the censored examples as unlabeled, and the observed examples as labeled, and cast the problem as a pure semi-supervised learning problem. However, in this scenario, the censored instances are not totally unlabeled, since we know that their event time is greater than the censoring time (assuming right-censored instances). As a result, we aim to exploit this information of censored instances to introduce a smarter stopping criterion in the data augmentation process. We denote this approach as STUART: Self-Trained sUrvivAl foResT which is a self-trained random survival forest corrected with censored times. Figure 1 shows the learning process in this self-training algorithm. This technique first builds an initial model using RSF over the labeled (in our case, observed) data points and then iteratively augments the labeled set with the most confident predictions of survival time for the unlabeled dataset (censored). In order to predict the survival time for each individual, we calculate the expected future lifetime (T_p) which at a given time t_0 is the time remaining until the event, given that the event did not occur until t_0 [38]:

$$T_p = \frac{1}{S(t_0)} \int_{t_0}^{\infty} S(t)dt \tag{3}$$

where $S(t)$ is the survival function predicted by RSF and t_0 is the smallest unique event time in the sample size.

Using the variance of the individual tree predictions as a confidence measure of the ensemble predictions, we determine which unlabeled (censored) instances to add to the augmentation process. Given the inherent generation of an ensemble of decision trees by RSF, the variance emerges as a natural metric to quantify the ensemble's consensus. Notably, our methodology addresses time-to-event prediction, a context often characterized by censored observations and inherent uncertainty. The variance, as an index of dispersion, aligns well with the need to identify instances for augmentation where the ensemble's predictions exhibit broader divergence. Additionally, the calculation of variance is seamlessly integrated into the RSF framework, requiring no supplementary computations. We sort the predictions in increasing order according to the variance, and then we decide when to stop adding any new instances based on the information in the censoring time. In more detail, we know that the true event time must exceed the censoring time. We, therefore, stop the augmentation process whenever a censored instance is encountered with a predicted time T_p that is lower than its censoring time T_c. If at the end of an iteration, no instances can be augmented, the entire process is terminated. In order to avoid premature termination (prediction variances can be high, in which case adding or removing some trees from the forest could result in a substantially different T_p value and therefore, a different condition outcome), we relax the condition $T_c \leq T_p$ as follows.

We calculate a 95% tolerance interval around T_p and require T_c to be smaller than or within the tolerance interval. In other words, we allow T_c to be larger than T_p, but only if it is within its 95% tolerance interval (see Figure 2). In the case of censored examples that meet the criteria for being added to the training set, we set their status to observed with a survival time equal to T_p. Algorithm 2 describes this approach in detail. Although the training sets in the self-training approach do not contain censored data points, we still chose RSF as the base learner in order to obtain a survival function as the prediction. Moreover, RSF benefits from advantages inherent in random forest techniques: high accuracy, efficient learning times, parallelizable, feature importance scores, etc. In addition, in the result section, we compare our method to the direct application of RSF (i.e., in a non-self-training set-up),

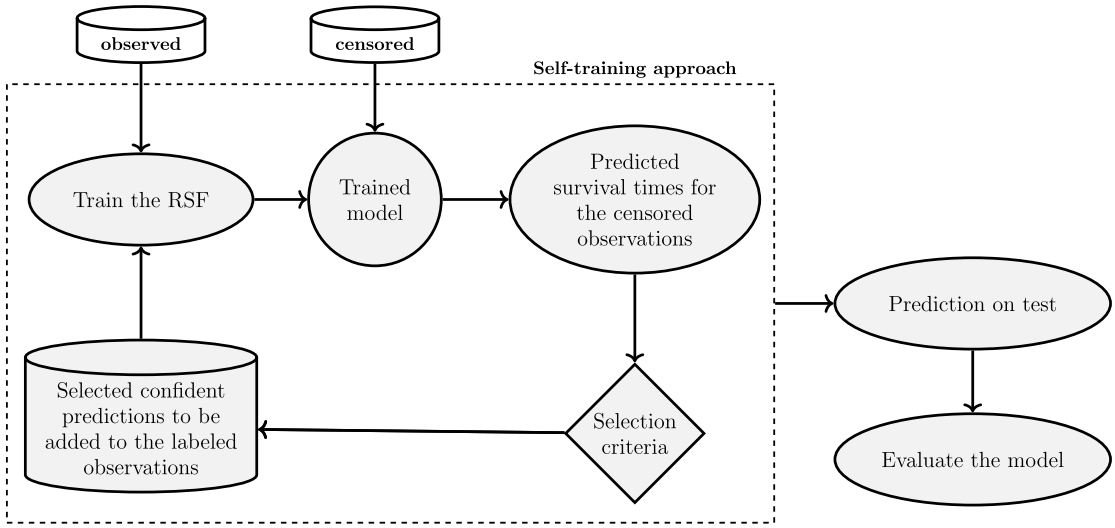


FIGURE 1. Pipeline for the proposed approach, called STUART.

Algorithm 2 STUART

Require: Observed data (observed), Censored data (censored)

Ensure: Prediction model for survival time

1: **repeat**
2: Train a base model using observed
3: Make a prediction for the survival time (T_p) of each instance in censored
4: Calculate the variance for each prediction
5: Sort the predictions based on minimum variance
6: Calculate a 95% tolerance interval for the predictions
7: Find the first instance i from the sorted predictions whose censoring time (T_c) is greater than $T_p+2\sigma$ (does not meet the criterion)
8: Remove all instances sorted before i (confident predictions) from censored and add them to the training set (observed)
9: **until** no confident predictions have been added to the training set

which is currently one of the state-of-the-art methods in the survival analysis domain, and as a result, a different base learner would complicate the interpretation of the results.

IV. EXPERIMENTAL SET-UP

In this section, we first describe the datasets in detail in Section IV-A, then we discuss the evaluation metrics in Section IV-B, and we continue with an explanation of the comparison methods and parameter instantiation in Section IV-C.

A. DATASET DESCRIPTION

During our evaluation of our proposed approach, real-life datasets with various characteristics, including those from /textit[survival] [39] in R as well as high-dimensional datasets with large numbers of observations from [40] and some from R/Bioconductor, were used. Moreover, 10 high-dimensional gene expression datasets were used ($p \gg n$) [41]. In these datasets, thousands of genes are typically expressed across a few samples ($<$ 300), contributing information about demographic characteristics, disease type, survival time, etc. As it was computationally expensive to run all the competitor methods on datasets with more than 10,000 gene expression features, we reduced the number of features to the top ten thousand features with the largest variance across all samples. Table 1 provides a description and characteristics of the datasets used in this study. The predicted outcome for all datasets is survival time (time until death).

B. PERFORMANCE EVALUATION

Harrell's concordance index (C-index) [43] is the most commonly used metric for evaluating survival models and represents the generalization of the ROC curve over all data in the survival analysis [44]. The C-index can be interpreted as the proportion of all pairs of subjects whose predicted survival times are correctly ordered among all subjects whose survival times can be predicted. Another way of putting it is that it is the probability that the predicted and observed survival times will coincide. It is feasible to rank two subjects' survival times if (1) both subjects are observed as well as (2) either one's observed survival time is smaller than the other's censored survival time [45]. Consider a set of observation and prediction values for two different instances, $(y_1, \hat{y}_1)$ and $(y_2, \hat{y}_2)$, where y_i and $\hat{y}_i$ represent the actual survival time and the predicted value, respectively. The concordance probability between these two instances can be computed as $c = Pr(\hat{y}_1 > \hat{y}_2 | y_1 > y_2)$.

C. COMPARISON METHODS AND PARAMETER INSTANTIATION

STUART was compared with representative time-to-event models: RSF [9], Lasso-Cox, and Cox-SP-AFT which was

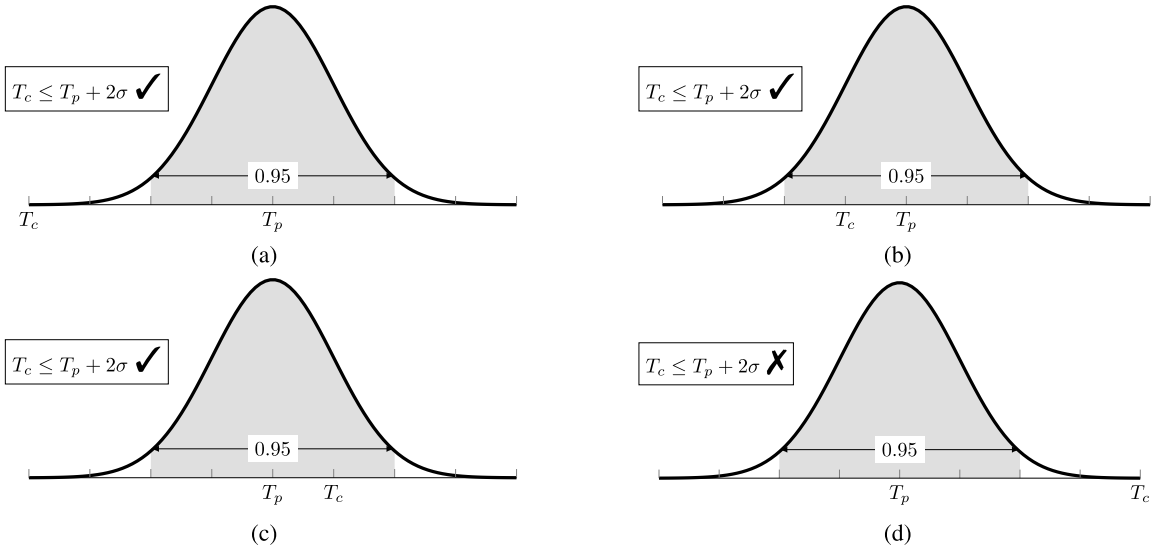


FIGURE 2. Tolerance interval corresponding to two times the standard deviation. Figures a, b, and c represent situations where the condition $T_c \leq T_p + 2\sigma$ is fulfilled, where σ is the standard deviation of the individual tree predictions, and hence, these situations are accepted by our method. In Figure d, the condition is violated.

TABLE 1. Characteristics of the used clinical and high-dimensional datasets.

Name	#Observations	#Features	Censoring rate
Veteran [39]	137	6	6%
Lung [39]	228	8	27%
PBC [39]	312	17	60%
DrAsGiven [41]	119	22122	42%
EMTAB386 [41]	129	10364	44%
GSE14764 [41]	80	13112	74%
GSE32062 [41]	260	20112	54%
Norway/Stanford Breast Cancer Data (NSBCD) [41]	115	549	67%
Sporadic lymph-node-negative patients (Veer) [41]	78	4751	56%
Dutch Breast Cancer Data (DBCD)[41]	295	4919	73%
Diffuse Large-B-Cell Lymphoma data (DLBCL) [41]	240	7399	42%
Lung adenocarcinomas (LungBeer) [41]	86	7129	72%
Acute myeloid leukemia (AML) [41]	79	54675	40%
Breast invasive carcinoma (BRCA) [42]	1080	117	86%
First National Health and Nutrition Examination Survey (NHANES I) [40]	9549	21	64%

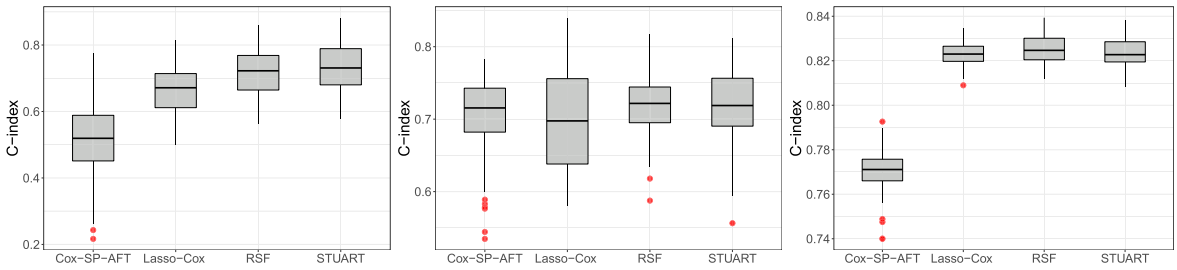
explained at the end of Section I. As a baseline model, we have reported the results of Cox regression with LASSO regularization. Lasso-Cox introduces the $L1$ norm penalty in the Cox log-likelihood function [24]. Since the majority of our used datasets are high-dimensional ($p \gg n$), we have employed Lasso-Cox due to its capability of handling high-dimensional datasets. For the purpose of estimating the generalization capacity of the models, 5-fold cross-validation was performed on each dataset to determine test accuracy, and this process was repeated ten times to obtain reliable results. Throughout the ten iterations of the cross-validation process, the C-index is calculated for each test fold, and the result is the average value across the five folds. In Lasso-Cox, the optimal tuning parameter (λ) is selected by nested cross-validation, whereas no hyperparameter tuning has been applied to the other approaches. For RSF and STUART, the number of trees was set to 500, and the number of candidate variables considered in each tree node was set to $p/3$, where p is the number of variables.

V. RESULTS AND DISCUSSION

In this section, we first describe and compare the results of Lasso-Cox, Cox-SP-AFT, RSF, and STUART on the benchmark datasets described in Table 1. Then, we take a closer

TABLE 2. Performance in terms of concordance index (C-index).

Datasets	Lasso-Cox	Cox-SP-AFT	RSF	STUART
Veteran	70.1 ± 6.5	70.05 ± 6.3	$\mathbf{71}.5 \pm 5.3$	71.48 ± 5.4
Lung	$\mathbf{62}.64 \pm 5.3$	60.82 ± 5.6	61.75 ± 5.1	62.01 ± 5.2
PBC	83.15 ± 3.5	80.51 ± 3.5	$\mathbf{83}.5 \pm 3.1$	82.22 ± 4.5
DrAsGiven	52.53 ± 6.3	52.42 ± 10.5	57.42 ± 4.7	$\mathbf{57}.74 \pm 7.7$
EMTAB386	51.43 ± 6.2	55.42 ± 8.9	50.14 ± 6.9	$\mathbf{59}.11 \pm 5.9$
GSE14764	52.08 ± 8.5	54.63 ± 18.3	56.99 ± 17.3	$\mathbf{66}.82 \pm 20.5$
GSE32062	52.11 ± 4.7	51.90 ± 7.3	50.03 ± 5.6	$\mathbf{56}.12 \pm 6.4$
NSBCD	66.28 ± 8.2	51.05 ± 13.1	71.75 ± 6.5	$\mathbf{73}.2 \pm 12.1$
Veer	62.63 ± 10.1	53.07 ± 10.2	67.4 ± 10.2	$\mathbf{71}.71 \pm 11.6$
DBCD	68.99 ± 7.6	63.13 ± 7.8	73.5 ± 5.7	$\mathbf{74}.15 \pm 6.1$
DLBCL	59.48 ± 6.4	55.74 ± 6.5	$\mathbf{59}.7 \pm 4.4$	59.64 ± 5.9
LungBeer	50.93 ± 10.1	63.40 ± 14.3	67.55 ± 15.8	$\mathbf{72}.34 \pm 11.3$
AML	55.90 ± 9.5	60.37 ± 8.9	60.02 ± 10.3	$\mathbf{64}.99 \pm 3.2$
NHANES I	82.26 ± 0.52	77.06 ± 1.12	$\mathbf{82}.5 \pm 0.6$	82.36 ± 0.6
BRCA	56.61 ± 6.3	56.22 ± 11.1	$\mathbf{63}.62 \pm 5.2$	62.35 ± 5.3
Average	61.81	60.39	65.16	**67.75**



(a) NSBCD (b) Veteran (c) NHANES I

FIGURE 3. Evaluation of the performance of the methods, for three datasets.

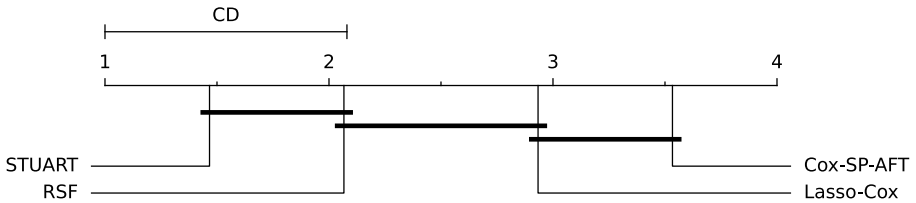
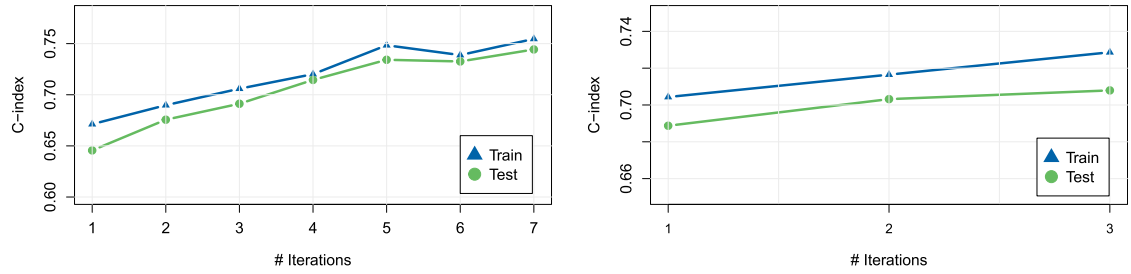


FIGURE 4. Results of the Friedman-Nemenyi test of methods ranking. The methods are compared in terms of their ranking using the evaluation measure, CI.

look at the results obtained for the NSBCD, Veteran, and NHANES I datasets as each represents a different type of dataset in terms of the number of features or observations.

Table 2 shows the means and standard deviations of the c-index on the datasets, as well as the average c-index of each algorithm. Based on the results, we can conclude that STUART is the winning approach, particularly in most high-dimensional datasets ($p \gg n$). More precisely, it can be concluded that on high dimensional datasets with a very small number of samples (e.g., Veer, LungBeer, AML, NSBCD, and GSE14764, all of which contain fewer than 120 observations), STUART is performing the best. In addition, in several datasets with a high percentage of censored instances where very few labeled data are available (e.g., DBCD, GSE14764, EMTAB, and LungBeer, all with a higher than 72% censoring rate), STUART is a much better algorithm than RSF alone. Nevertheless, in the datasets with high censoring rates combined with a large number of observations (NHANES I and BRCA), RSF outperforms STUART, although only by a small margin.



(a) NSBCD (b) Veteran

FIGURE 5. Learning curves of the STUART method for NSBCD (figure a) and Veteran (figure b) datasets.

Although STUART is the best algorithm for datasets with a high censoring rate and a small number of observations, at the other extreme, where the censoring rate is small (such as in Veteran and Lung), it does not perform as well. This may be the result of a lack of sufficient censored data to guide the augmentation process.

In light of the different behavioral patterns seen in different types of data, we selected three datasets that each represent a different type of dataset in terms of having either a very high number of features (NSBCD) or a very high number of observations (NHANES I) and a simple mid-size dataset (Veteran). The results on these datasets are illustrated by box plots in Figure 3. When comparing the range of C-indices (interquartile range), Cox-SP-AFT varies more dramatically and is the last algorithm in most experiments; but overall, STUART acts robustly and behaves like RSF. This robust behavior of STUART could be due to the fact that for censored instances, it compares the predicted survival time with the censoring time, which results in having more confident predictions. However, this should hold for Cox-SP-AFT as well, since it also compares the predicted survival time with the censoring time. However, the reliability and stability of the Cox-SP-AFT model rely heavily on the accuracy of the AFT model and the single AFT model always encounters robustness issues in semi-supervised learning scenarios caused by heavy noise and even outliers [19], [20]. During the experiments, we also noticed that many censored data points always violate the constraint that losses should be smaller than α, restricting the ability to fully exploit the censored data. Therefore, the AFT model does not benefit from a large number of instances in order to be properly trained.

In comparison with the main competitor (RSF), STUART, although with a slight non-statistically significant margin, was ranked in a higher position according to the Friedman-Nemenyi test[1] presented in Figure 4 [46]. STUART outperforms Lasso-Cox and Cox-SP-AFT and manages to be statistically significantly better according to the Friedman-Nemenyi test. As a second-best method, RSF is statistically significantly superior to Cox-SP-AFT and has a slight non-significant lead over Lasso-Cox.

Overfitting is a concern that may arise when self-training approaches are used. Since our algorithms have no tunable hyperparameters, they are not prone to the kind of overfitting that results from the hyperparameter tuning process in other algorithms. Moreover, random forests overall are known to be robust to overfitting due to the fact that by increasing the number of trees, the variance of the error gets reduced. Furthermore, we investigated the learning curve of the STUART algorithm on two datasets with varying censoring rates: NSBCD and Veteran. The train-validation diagrams, illustrated in Figure 5, demonstrate the algorithm's robustness across different levels of censoring. NSBCD's higher censoring rate led to more iterations, while the Veteran dataset's lower censoring rate showed convergence over fewer iterations. These insights reinforce our method's adaptability and underscore its resistance to overfitting, supporting its credibility in time-to-event prediction.

Our findings demonstrate that the self-training technique that uses the information in the censored data points to guide the data augmentation process performs best, resulting in a competitive algorithm compared to RSF.

VI. CONCLUSION

Predicting the time until an event of interest is a common problem encountered in medical applications, and it is traditionally addressed using survival analysis techniques. In this study, we have transformed the time-to-event prediction problem into a semi-supervised regression problem. In our approach, called STUART, censored observations are introduced as partially labeled observations since their (unknown) target values should exceed the censoring time. This property is exploited in the augmentation process of a self-training algorithm for time-to-event prediction. We have evaluated and compared the proposed approach on fifteen real-world survival analysis datasets, including clinical and high-dimensional ones. Our results have shown that our proposed approach especially in high-dimensional settings outperforms the others due to its ability to integrate integrating partial supervision provided by censored data into a semi-supervised learning wrapper approach.

[1] In a critical distance diagram, those algorithms that are not joined by a line (i.e., their rankings differ more than a critical distance (CD)) can be regarded as statistically significantly different [46].

Further research can be carried out in several directions, of which we outline a few below. In this study, we used STUART for survival analysis of right-censored data, but the same approach can be applied easily to left-censored data as well. The concept of the idea that we proposed could be applied using other base learners and semi-supervised learning strategies, but it remains to be investigated whether the results carry over to other learners.

ACKNOWLEDGMENT

The authors would like to thank the Flemish Government (AI Research Program).

REFERENCES

[1] D. R. Cox, "Regression models and life-tables," *J. Roy. Stat. Soc. B, Methodol.*, vol. 34, no. 2, pp. 187–220, 1972. [Online]. Available: http://www.jstor.org/stable/2985181

[2] B. Zupan, J. Demšar, M. W. Kattan, J. R. Beck, and I. Bratko, "Machine learning for survival analysis: A case study on recurrence of prostate cancer," *Artif. Intell. Med.*, vol. 20, no. 1, pp. 59–75, Sep. 2000. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0933365700000531

[3] K. A. Dauda, B. Pradhan, B. U. Shankar, and S. Mitra, "Decision tree for modeling survival data with competing risks," *Biocybern. Biomed. Eng.*, vol. 39, no. 3, pp. 697–708, Jul. 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0208521619300245

[4] L. Vanneschi, A. Farinaccio, G. Mauri, M. Antoniotti, P. Provero, and M. Giacobini, "A comparison of machine learning techniques for survival prediction in breast cancer," *BioData Mining*, vol. 4, no. 1, pp. 1–13, Dec. 2011.

[5] L. Gordon and R. A. Olshen, "Tree-structured survival analysis," *Cancer Treat. Rep.*, vol. 69, no. 10, pp. 1065–1069, 1985.

[6] D. Faraggi and R. Simon, "A neural network model for survival data," *Statist. Med.*, vol. 14, no. 1, pp. 73–82, Jan. 1995.

[7] F. M. Khan and V. B. Zubek, "Support vector regression for censored data (SVRc): A novel tool for survival analysis," in *Proc. 8th IEEE Int. Conf. Data Mining*, Dec. 2008, pp. 863–868.

[8] T. Hothorn, B. Lausen, A. Benner, and M. Radespiel-Tröger, "Bagging survival trees," *Statist. Med.*, vol. 23, no. 1, pp. 77–91, Jan. 2004.

[9] H. Ishwaran, U. B. Kogalur, E. H. Blackstone, and M. S. Lauer, "Random survival forests," *Ann. Appl. Statist.*, vol. 2, no. 3, pp. 841–860, Sep. 2008.

[10] B. Vinzamuri, Y. Li, and C. K. Reddy, "Active learning based survival regression for censored data," in *Proc. 23rd ACM Int. Conf. Conf. Inf. Knowl. Manage.*, Nov. 2014, pp. 241–250.

[11] Y. Li, L. Wang, J. Wang, J. Ye, and C. K. Reddy, "Transfer learning for survival analysis via efficient $l_{2,1}$-norm regularized Cox regression," in *Proc. IEEE 16th Int. Conf. Data Mining (ICDM)*, Dec. 2016, pp. 231–240.

[12] P. Wang, Y. Li, and C. K. Reddy, "Machine learning for survival analysis: A survey," *ACM Comput. Surv.*, vol. 51, no. 6, pp. 1–36, 2017.

[13] X. J. Zhu, "Semi-supervised learning literature survey," University of Wisconsin-Madison Department of Computer Sciences, Tech. Rep., 2005.

[14] J. E. Van Engelen and H. H. Hoos, "A survey on semi-supervised learning," *Mach. Learn.*, vol. 109, no. 2, pp. 373–440, 2020.

[15] E. Bair and R. Tibshirani, "Semi-supervised methods to predict patient survival from gene expression data," *PLoS Biol.*, vol. 2, no. 4, p. e108, Apr. 2004.

[16] B. Roy, T. Stepišnik, C. Vens, and S. Džeroski, "Survival analysis with semi-supervised predictive clustering trees," *Comput. Biol. Med.*, vol. 141, Feb. 2022, Art. no. 105001.

[17] M. Shi and B. Zhang, "Semi-supervised learning improves gene expression-based prediction of cancer recurrence," *Bioinformatics*, vol. 27, no. 21, pp. 3017–3023, Nov. 2011.

[18] H. R. Hassanzadeh, J. H. Phan, and M. D. Wang, "A multi-modal graph-based semi-supervised pipeline for predicting cancer survival," in *Proc. IEEE Int. Conf. Bioinf. Biomed. (BIBM)*, Dec. 2016, pp. 184–189.

[19] Y. Liang, H. Chai, X.-Y. Liu, Z.-B. Xu, H. Zhang, and K.-S. Leung, "Cancer survival analysis using semi-supervised learning method based on cox and AFT models with $L_{1/2}$ regularization," *BMC Med. Genomics*, vol. 9, no. 1, pp. 1–11, Dec. 2016.

[20] H. Chai, Z.-N. Li, D.-Y. Meng, L.-Y. Xia, and Y. Liang, "A new semi-supervised learning model combined with cox and SP-AFT models in cancer survival analysis," *Sci. Rep.*, vol. 7, no. 1, pp. 1–12, Oct. 2017.

[21] D. Yarowsky, "Unsupervised word sense disambiguation rivaling supervised methods," in *Proc. 33rd Annu. Meeting Assoc. Comput. Linguistics*, 1995, pp. 189–196.

[22] M. Li and Z.-H. Zhou, "SETRED: Self-training with editing," in *Proc. Pacific–Asia Conf. Knowl. Discovery Data Mining*. Springer, 2005, pp. 611–621.

[23] O. J. Dunn and V. A. Clark, *Basic Statistics: A Primer for the Biomedical Sciences*. Hoboken, NJ, USA: Wiley, 2009.

[24] R. Tibshirani, "The lasso method for variable selection in the Cox model," *Statist. Med.*, vol. 16, no. 4, pp. 385–395, Feb. 1997.

[25] A. E. Hoerl and R. W. Kennard, "Ridge regression: Biased estimation for nonorthogonal problems," *Technometrics*, vol. 12, no. 1, pp. 55–67, Feb. 1970.

[26] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *J. Roy. Stat. Soc. B, Stat. Methodol.*, vol. 67, no. 2, pp. 301–320, Apr. 2005.

[27] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[28] M. R. Segal, "Regression trees for censored data," *Biometrics*, vol. 44, no. 1, pp. 35–47, 1988.

[29] E. L. Kaplan and P. Meier, "Nonparametric estimation from incomplete observations," *J. Amer. Stat. Assoc.*, vol. 53, no. 282, pp. 457–481, Jun. 1958.

[30] A. Madani, M. Moradi, A. Karargyris, and T. Syeda-Mahmood, "Semi-supervised learning with generative adversarial networks for chest X-ray classification with ability of data domain adaptation," in *Proc. IEEE 15th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2018, pp. 1038–1042.

[31] T. J. Rogers, K. Worden, R. Fuentes, N. Dervilis, U. T. Tygesen, and E. J. Cross, "A Bayesian non-parametric clustering approach for semi-supervised structural health monitoring," *Mech. Syst. Signal Process.*, vol. 119, pp. 100–119, Mar. 2019.

[32] B. Ballinger, J. Hsieh, A. Singh, N. Sohoni, J. Wang, G. H. Tison, G. M. Marcus, J. M. Sanchez, C. Maguire, J. E. Olgin, and M. J. Pletcher, "DeepHeart: Semi-supervised sequence learning for cardiovascular risk prediction," in *Proc. 32nd AAAI Conf. Artif. Intell.*, 2018, pp. 1–8.

[33] R. Collobert and J. Weston, "A unified architecture for natural language processing: Deep neural networks with multitask learning," in *Proc. 25th Int. Conf. Mach. Learn.*, 2008, pp. 160–167.

[34] F. Roli and G. L. Marcialis, "Semi-supervised PCA-based face recognition using self-training," in *Proc. Joint IAPR Int. Workshops Stat. Techn. Pattern Recognit. (SPR) Struct. Syntactic Pattern Recognit. (SSPR)*. Springer, 2006, pp. 560–568.

[35] D. McClosky, E. Charniak, and M. Johnson, "Effective self-training for parsing," in *Proc. Human Lang. Technol. Conf. North Amer. Chapter Assoc. Comput. Linguistics*, 2006, pp. 152–159.

[36] K. Nigam and R. Ghani, "Analyzing the effectiveness and applicability of co-training," in *Proc. 9th Int. Conf. Inf. Knowl. Manage.*, Nov. 2000, pp. 86–93.

[37] F. Nateghi Haredasht and C. Vens, "Predicting survival outcomes in the presence of unlabeled data," *Mach. Learn.*, vol. 111, pp. 4139–4157, Oct. 2022.

[38] R. G. Miller, *Survival Analysis*. Hoboken, NJ, USA: Wiley, 1981.

[39] T. M. Therneau. (2020). *A Package for Survival Analysis in R, R Package Version 3.2-7*. [Online]. Available: https://CRAN.R-project.org/package=survival

[40] National Center for Health Statistics. (2010). *Webpage*. Accessed: Jun. 15, 2022. [Online]. Available: https://wwwn.cdc.gov/nchs/nhanes/nhanes1/

[41] (2010). *SurvLab*. Accessed: Dec. 7, 2014. [Online]. Available: http://user.it.uu.se/~kripe367/survlab/download.html

[42] N. Petrucelli, M. B. Daly, and T. Pal, "BRCA1-and BRCA2-associated hereditary breast and ovarian cancer," in *GeneReviews(R)*, R. A. Pagon et al., Eds. Seattle, WA, USA: Univ. of Washington, May 2022. Accessed: Jun. 30, 2022. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK1247/

[43] F. E. Harrell, "Evaluating the yield of medical tests," *J. Amer. Med. Assoc.*, vol. 247, no. 18, pp. 2543–2546, May 1982.

[44] M. Schmid, M. N. Wright, and A. Ziegler, "On the use of Harrell's C for clinical risk prediction via random survival forests," *Expert Syst. Appl.*, vol. 63, pp. 450–459, Nov. 2016.

[45] H. Steck, B. Krishnapuram, C. Dehing-Oberije, P. Lambin, and V. C. Raykar, "On ranking in survival analysis: Bounds on the concordance index," in *Proc. Adv. Neural Inf. Process. Syst.*, 2008, pp. 1209–1216.
[46] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *J. Mach. Learn. Res.*, vol. 7, pp. 1–30, Dec. 2006.

FATEME NATEGHI HAREDASHT received the B.Sc. degree in electrical engineering from the University of Guilan, the M.Sc. degree in biomedical engineering (bioinformatics) from the Amirkabir University of Technology (Tehran Polytechnic), and the Ph.D. degree in biomedical sciences with emphasis on health and technology from KU Leuven, in 2023. She is currently a Postdoctoral Researcher with the Stanford School of Medicine. Her research interests include semi-supervised learning, machine learning for healthcare, explainable AI, and survival analysis.

KAZEEM ADESINA DAUDA received the B.Sc., M.Sc., and Ph.D. degrees in statistics (biostatistics) from the University of Ilorin, Ilorin, Nigeria, in 2010, 2013, and 2018 respectively. Previously, he was a Research Fellow with the Department of Artificial Intelligence, Indian Statistical Institute, Kolkata, India, through CV Raman International Fellowship for Africa Researcher, in 2017. He is currently a Senior Lecturer with the Department of Mathematics and Statistics, Kwara State University, Malete, Nigeria. His research interests include advanced statistical learning, data science, and machine learning (supervised, unsupervised, and semi-supervised learning).

CELINE VENS received the Ph.D. degree in computer science (machine learning) from KU Leuven, Belgium, in 2007. She is currently an Associate Professor with the Faculty of Medicine and the ITEC-imec Research Group, KU Leuven. Her research interests include multi-label, multi-target, hierarchical prediction, recommender systems, tree ensemble learning, survival analysis, and biological network mining.