Applied Intelligence (2023) 53:7041–7055
https://doi.org/10.1007/s10489-022-03846-0

Deep survival forests for extremely high censored data

Xuewei Cheng[1] · Sizheng Wang[1] · Hong Wang[1] · Shu Kay Ng[2]

Accepted: 1 June 2022 / Published online: 13 July 2022
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2022

Abstract

The Cox proportional hazard model and random survival forests (RSF) are useful semi-parametric and non-parametric methods in modeling time-to-event data. However, both approaches may fail in case of small sample size and/or high censoring rate. In this research, we want to tackle such problems within the random forests framework using semi-supervised data transduction techniques and a layer-by-layer processing similar to deep forest. Experiments from both extensive simulated data and real-world benchmark datasets have shown that the proposed deep survival forests (DSF) outperforms Cox, RSF by a noticeable margin and also work better than several state-of-art survival ensembles including Cox boosting models and latest survival forest extensions on a variety of scenarios. The superiority of DSF stands out when small sample-sized and highly censored data are confronted.

Keywords Deep survival forests · Semi-supervised learning · Highly censored data · Data transduction · Cascade Forest · Few shots learning

1 Introduction

Machine learning methods have been highly successful in data-intensive survival analysis and applications [34, 36, 37] but it often hampered when the data set is small. Furthermore, in medical studies, there is often a portion of patients who did not experience the event of interest when the study ends and for these observations, we have incomplete or censored time-to-event data [15, 49]. In the past few decades, a large amount of parametric, semi-parametric and non-parametric survival models have been developed for modeling time-to-event survival data. Among them, the most popular is the regression-based semi-parametric Cox model [12], and its extensions [45, 52, 60]. When the underlying proportional hazard assumption is not satisfied, nonparametric machine learning-based approaches are useful alternatives [6, 19, 24, 31]. However, regardless of whether it is semi-parametric or non-parametric, all the aforementioned methods require an adequate number of both censored and uncensored observations [43]. When a small sample size and/or highly censored data are present, these methods may face severe difficulties [50].

Among all machine learning based survival methods, the most popular one is the Random survival forests (RSF) model [28] inherited from Random forest (RF) [7]. Different from entropy or Gini-index based RF, log-rank split rules are generally adopted for RSF. There are many extensions of RSF in the past decades and one can refer to [56] for more detailed information. More recently, there are also quite a few improvements on the traditional RSF method. For example, linear combination of input variables are used for recursively partition and a higher prognostic value is achieved in [30]. In [48], a novel paradigm for building regression trees is proposed for survival analysis. In [53], the standard procedure of simple average is replaced by a weighted average for hazard function estimation in RSF. In UST (uni.survival.tree) [17], a stabilized score test is suggested to select significant covariates first to reduce time complexity.

✉ Hong Wang
wh@csu.edu.cn

Xuewei Cheng
xwcheng@csu.edu.cn

Sizheng Wang
wszhchina@csu.edu.cn

Shu Kay Ng
s.ng@griffith.edu.au

[1] School of Mathematics and Statistics, Central South University, Changsha, Hunan, China

[2] School of Medicine and Dentistry, Menzies Health Institute Queensland, Griffith University, Brisbane, Queensland, Australia

One may notice that all the above random forest based approaches are one-layered, i.e. only the raw features are exploited to train the forests model and predictions are immediately made, neglecting the fact that multiple intermediate trained layers may produce better representations of the training data [42]. Moreover, as a machine learning approach, tree-based methods usually rely on a large or medium number of samples to obtain a satisfactory predictive performance [44]. However, this may not be satisfied in real practices as small sample size survival data are commonplace in clinical studies [29, 61]. Furthermore, a large proportion of censored observations under such circumstances will make the survival modeling process more complicated, if not impossible.

In light of the above discussions, we propose to address the problems mentioned above using a layer-by-layer deep random forests framework, where perceptions in traditional deep neural networks are replaced by random forests. To alleviate the high censoring problem, semi-supervised learning [21] and data transformation techniques [3] are also adopted in the proposed deep survival forests (DSF) method. The superior empirical performance of the proposed method is illustrated by simulation examples and real data applications.

The major contributions of this paper are summarized as follows:

- A non-NN (Neural Networks) style deep learning method is proposed for survival prediction.
- We provide an effective approach to model highly censored survival data with a small sample size.

The rest of the paper is organized as follows. Section 2 introduces the motivation to modeling censored data in the case of small sample size, and then we propose a novel deep forest structure in Section 3. Experimental analysis and real data applications are described in Sections 4 and 5. Finally, we discuss and conclude the paper in Sections 6 and 7.

2 Preliminaries

In this section, we first discuss the highly censoring problem, then we give a short description of semi-supervised learning and the deep forest model. Later on in Section 3, we will develop a novel semi-supervised framework using deep forest to deal with the highly censoring problem.

2.1 The highly censoring problem

In survival analysis, an instance can be presented by a triplet $(\boldsymbol{x_i}, \delta_i, y_i)$, $i = 1, 2, ..., n.$, where $\boldsymbol{x_i} = (x_{i1}, x_{i2}, ..., x_{ip})$ is the feature vector. In the case of right censored data, $y_i = min\{C_i, T_i\}$, where T_i is the truly survival time, C_i is the censoring time, and $\delta_i = I(T_i \leq C_i)$ the censoring indicator. In biomedical studies, such survival data are often characterized by a small sample size with a high dimension. Take GEO (Gene Expression Omnibus) (https://www.ncbi.nlm.nih.gov/geo/) genomics data repository as an example. So far, this database contains 4348 data sets, all of which are high-dimensional with sample sizes ranging from from 2 to 202. When a highly censoring rate is married with a small sample size, the uncensored samples may not be sufficient for predictive modeling. In such cases, the parameter estimation of the Cox model may not converge in the optimization procedure and the RSF model may fail due to the constraint that a leaf node must have some unique samples with events [62].

We will illustrate this problem with the popular RSF model. In RSF, the Nelson-Aalen (NA) estimator is used to predict the cumulative hazard function(CHF). The CHF for terminal node h is

$$\widehat{H}_h(t) = \sum_{t_{l,h} \leq t} \frac{d_{l,h}}{Y_{l,h}}, \tag{1}$$

where $d_{l,h}$ and $Y_{l,h}$ are the number of deaths and individuals at risk at time $t_{l,h}$. Obviously, all cases within node h have the same CHF. Suppose in one terminal node, we have only one death instance and nine censored instances and the detailed survival times are (2+, 3+, 5+, 7+, 8+, 10, 13+, 14+, 18+, 25+). In this case, the NA's estimator can only show that the risk is 0% at $T < 10$ and 20% at $T \geq 10$, which is extremely vague and inaccurate. Consequently, the resulting RSF model may face an under-fitting problem.

One may notice that, in calculating CHF, only the number of censored samples are used and other censoring information such as the specific values of censoring times are ignored. In case of a small sample size with a highly censoring rate, one may consider improving the model's predictive capability by exploiting such information.

2.2 Semi-supervised learning

In classification and regression problems, semi-supervised learning can make use of unlabelled data to gain more information about the underlying marginal data distribution $p(x)$, and thereby obtain more accurate inference about the posterior distribution $p(y \mid x)$ [26].

However, semi-supervised learning for survival analysis so far is still underdeveloped. In survival analysis, instances that have experienced the event of interest can be regarded as labeled data. But censored data are not the same as unlabeled data in that censored data always imply that the truly survival times are within some intervals specified by survival times and hence carry more information than the unlabeled data.

Here, we consider a toy example in Fig. 1 with which we can observe how censored information may help us in classification problem. In this example, we have two classes of 34 instances: eight of them (squared dots) denotes event (uncensored) data and 26 others (circled dots) are censored. If we only use 8 uncensored instances (E1 to E8 in Fig. 1) in model training, the decision boundary may be the densely dotted line. However, from semi-supervised learning, we know that the dotted line violates the smoothness and low-density assumptions as the decision boundary of a classifier should preferably pass through low-density regions in the input space [54]. Hence, if both censored and uncensored samples are dealt with properly, the solid optimal decision hyper-plane may be found.

2.3 Deep forest

Deep learning based approaches find vast applications in a variety of fields. The mystery behind the success of deep learning may lie in three characteristics, i.e., layer-by-layer processing, in-model feature transformation and sufficient model complexity [63]. However, training of deep neural networks requires a large number of samples [1], which is often difficult to be satisfied in medical practice. In 2017, a deep forest framework with a cascade random forest structure is proposed to hold the strengths from the deep leaning [63].

One may observe that both deep neural network (DNN) and deep forest (gcForest) have a layer-by-layer structure for representational learning. As stated in [42], for any $g \in (\mathcal{C}_r[0,1]^r, \beta, H)$, there exists a deep neural network $f \in \mathcal{F}(l, \{d_j\}_{j=0}^{L+1}, s, V)$ such that

$$\|f-g\|_\infty \le (2H+1)6^r \cdot (1+r^2+\beta^2) \cdot N2^{-m} + H \cdot 3^\beta \cdot N^{-\beta/r}$$
$$m \propto depth;\ N \propto width. \tag{2}$$

According to (2), the generalization error upper bound decreases exponentially with the increase of the model depth. Later on in the next section, we shall develop a different deep forest framework for highly censored survival data using a semi-supervised learning technique.

3 The DSF approach

In this section, we first show how pseudo survival times can be approximated using censored times through a semi-supervised learning approach called data transduction technique. Then we propose a deep survival forest(DSF) approach that can utilize both censored and uncensored sample information.

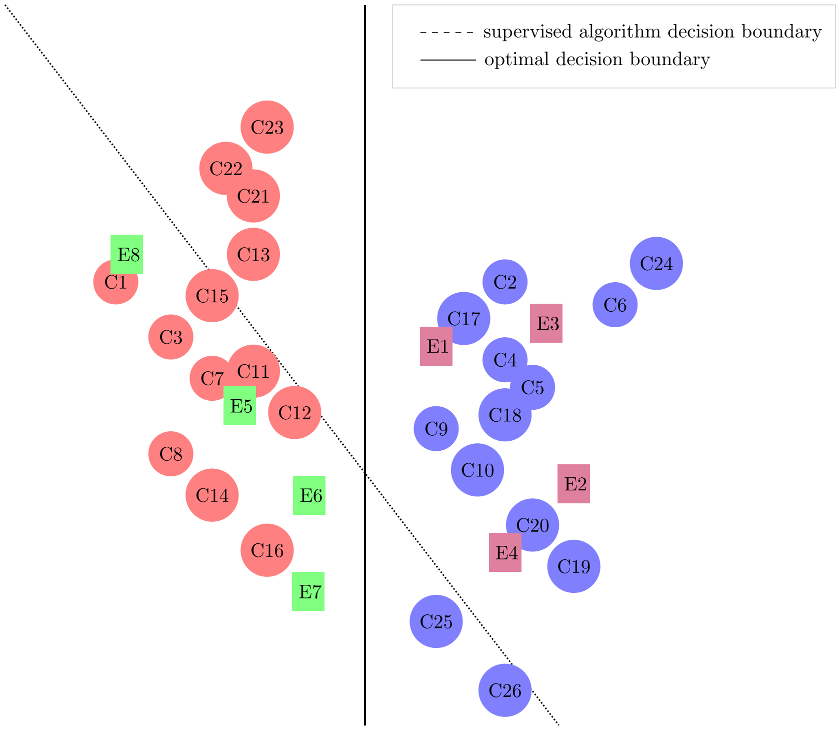


Fig. 1 A basic example to explain semi-supervised learning

3.1 Data transduction

Given m labeled samples $(x_1, y_1), ..., (x_m, y_m)$ as well as $n - m$ unlabeled samples $x_{m+1}, x_{m+2}, ..., x_n$, the purpose of semi-supervised learning is to predict the remaining unknown labels $y_{m+1}, y_{m+2}, ..., y_n$ [11]. However, most semi-supervised learning approaches are born for regression or classification problems and cannot be extended to survival modeling directly [32]. Here, we try to employ a transductive semi-supervised algorithm [54] in dealing with censored observations.

Formally, given a supervised loss function ℓ for the labeled data and unsupervised loss function ℓ_U for the pairs of labeled or unlabeled data, transductive methods attempt to obtain a pseudo labeling $\hat{y}$ that minimizes

$$\lambda \cdot \sum_{i=1}^{l} \ell(\hat{y}_i, y_i) + \sum_{i=1}^{n} \sum_{j=1}^{n} W_{ij} \cdot \ell_U(\hat{y}_i, \hat{y}_j) \tag{3}$$

where W_{ij} contains the edge weights for all pairs of nodes and λ governs the supervised term's relative importance.

In the most typical right censored survival case, censored instances' actual survival times are unknown but are greater than observed censoring times. Based on this fact, we can assign an optimal class label (or possible target label) to each censored instance via data transduction. In other words, we attempt to infer a pseudo-label for ith right censored instance by

$$\widehat{T_i} = f(\boldsymbol{x_i}, \boldsymbol{x_j}, C_i, T_j), \ \ j \neq i \text{ and } \delta_j = 1 \tag{4}$$

where δ_j is the censoring indicator, $\delta_j = I(T_j \leq C_j)$.

We assume that the distribution of survival time possesses the memoryless property, that is $P(T > t) = P(T > s + t \mid T \geq s)$. For one censored instance at time C, we suppose that the longest survival time of this instance is $C + \tau$ where τ is the maximum event time in the training set. As a result, an effective way to obtain an optimal target is data transduction via exhaustive searching from the censored time to maximum pseudo time $C+\tau$. The transduced time can be further formulated as

$$\widehat{T_i} = C_i + k\frac{\tau}{\zeta} = C_i + ks \ \ (k = 1, 2, \cdots, \zeta) \tag{5}$$

In (5), s is an iteration stride which determines the iteration times ζ. The larger ζ, the more time the algorithm will takes and a higher accuracy will be obtained.

To avoid noise accumulation from pseudo-labels of the censored data and ensure the robustness of the whole data transduction, the proposed method makes the censored samples enter the model one by one, and carry out collaborative training with the uncensored data. Once an instance obtains the transduced pseudo-label, it becomes a new "uncensored" instance in the next training process. That is to say, a pseudo-label for the ith ($m + 1 \leq i \leq n - m$) censored sample is transduced, and the $m + 1, \cdots, m + i$ censored observations are regarded as uncensored samples in the subsequent training process.

3.2 Deep survival forests

In the proposed deep survival forests (DSF) approach, we attempt to apply a similar methodology to survival data in the hope to have a smaller error upper bound. However, the censored data problem makes the popular cascade structure in trouble. Hence, the cascade structure is redesigned to cope with the challenges from highly censored data.

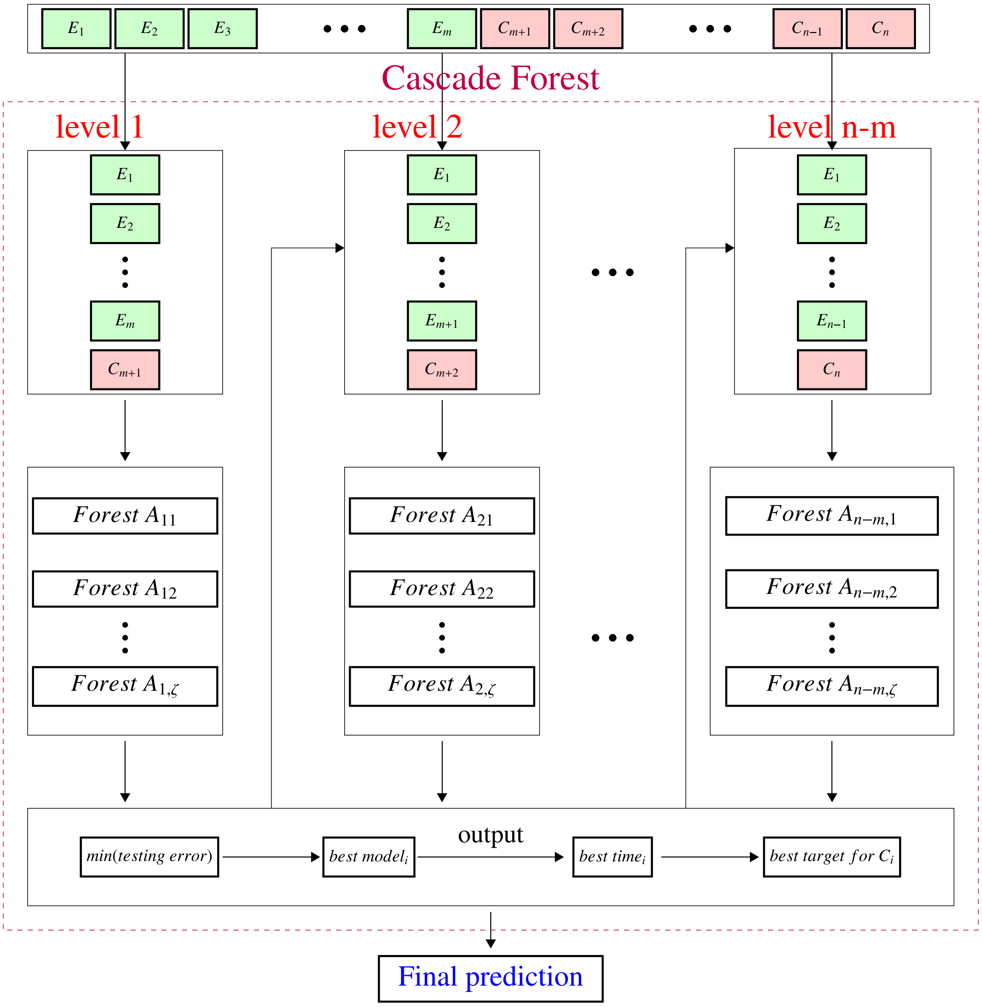
As illustrated in Fig. 2, DSF is distinct from deep forest in that each level of cascade receives new transduced censored samples from its preceding level, and transduce its processing result to the next level. Compared to a hierarchical framework extracting complex non-linear features in deep forest, a sequential framework expanding the training sample size by transducing survival time for the censored samples is applied in the proposed approach.

In each layer of cascade forest, the stride or the number of forests should be determined in advance. In practice, these settings depend on the accuracy requirements. For example, if the censored time for one instance is 300, and if the maximum survival time in the training set is 600, and if our stride is set to 50, we need six random forests in this layer to search the optimal target time for this censored sample. If higher accuracy is required, a smaller stride (such as 10) can be set and more (30) random forests are involved.

We then will replace the censored time with a transduced label for each censored sample that yields the best performance improvement on out of bag (OOB) data. In other words, the pseudo-label of this instance is obtained based on the minimum error criterion on the testing set. The corresponding status is also transduced from censored to uncensored. Instead of other imputation methods, we design a cascade forest to realize data transduction. This procedure minimizes the corresponding estimation bias and extracts more effective feature information than one-of f procedure.

Finally, once optimal pseudo-values are transduced for all censored instances, a learning model such as random

Fig. 2 The overall procedure of DSF



forest can be built with these pseudo-label censored samples and uncensored samples. As we can see, the final model is actually a cascade of cascades, where each cascade consists of multiple levels.

In general, the DSF can be formulated as the following optimization problem:

$$\begin{aligned}&\underset{k}{Min}\ \sum_{b=1}^{B}\sum_{i=1}^{n} I_{i,b}\cdot\delta_{i,b}[y_{i,b}-\hat{f}^{*}(x_{i,b})]^2\\ &s.t.\ y^{*}_{i,b}\le\tau+y_{i,b}\cdot(1-\delta_{i,b})\end{aligned} \tag{6}$$

where $y^{*}_{i,b}=y_{i,b}+k_{i,b}\cdot(1-\delta_{i,b})\cdot s$, for $b=1,2,...,B$ and B is number of bootstrap samples from the original data; for $i=1,2,...,n$; $I_{i,b}=1$ if i is an OOB sample for bth bootstrap sampling.

In our study, random forests are chosen as the base learners. And for each random forest, a default of 100 regression trees using the most common Mean Squared Error (MSE) loss function are built. The Gini-index criteria is adopted as the splitting rules in these regression trees. Like deep forest and other deep learning approaches, we have to make a trade-off between computational efficiency and predictive performance. If computational resource is enough, we suggest setting large iteration times ζ such as 1000, 5000 to gain more accurate transduced labels and more reliable prediction performances.

The pseudo-code of the proposed DSF is presented in Algorithm 1:

Since the “while” parts based on iteration times ζ can be executed concurrently, thus in case of big survival data under the larger ζ, DSF can be trained on a multi-core CPU or computer clusters in parallel to save time.

Algorithm 1 Deep survival forest.

```
Input:
    Training data: $D = (\boldsymbol{x_i}, \delta_i, y_i), i = 1, ..., n$, where $\boldsymbol{x_i}$ is $p$-dimensional
    Testing data: $l$ testing examples $\mathcal{X} = \mathcal{X}_{l \times p}$
    $\zeta$: iteration times
    $trees$: the number of tree in each level of cascade
    $Trees$: the number of tree in the level of final prediction
Output:
    The predicted labels of testing examples $\mathcal{X}$
procedure DSF (Training)
    Impute the number $m$ of censored examples based on $\delta$
    Impute the stride $s$
    while $i$ in $1 : m$ do
        while $j$ in $1 : \zeta$ do
            Choose a pseudo-label $T_i = C_i + \zeta s$ for $i$-th censored sample
            Randomly select $\lceil \sqrt{p} \rceil$ covariates from $p$-dimensional data
            Save the corresponding covariate indexes into matrix $\mathcal{A}[i, j]$.
            Generate a bootstrap sample $D_{ib}$ with $\lceil \sqrt{p} \rceil$ selected covariates from $D_i$.
            Using $D_{ib}$ as the training set, train random forests model $\mathcal{M}_{ij}$
        end while
    return The best pseudo-label for $i$-th censored sample according to the validation error of OOB
    end while
end procedure
procedure DSF (Training after data transduction)
    Train random forests model $\mathcal{M}$ based on transductive data
end procedure
procedure DSF (Testing)
    while $i$ in $1 : l$ do
    Predict the above testing samples $\mathcal{X}_i$ with $\mathcal{M}$
    end while
    return The predicted labels of testing examples $\mathcal{X}$
end procedure
```

4 Simulation studies

In this section, we use simulation studies to evaluate the effectiveness of the proposed method for survival prediction on a variety of scenarios.

4.1 The comparing models

We will compare our method with several popular semi-parametric and nonparametric models widely used in real applications. We do not consider the deep survival models such as DeepCox [33] and DeepSurv [31] as competitors, because these methods require more training samples and are not applicable in scenarios of small sample size.

- Cox proportional hazards model [12, 13] is a popular semi-parametric model and the most commonly-used survival analysis method.
- GlmBoost [9] is a generalized linear model which is fitted using a boosting algorithm based on component univariate linear models.
- RSF (Random Survival Forests) [28] extended random forest [7] to model right-censored data, which is the most popular nonparametric method in the field of survival analysis.
- CoxBoost [4] is one of the few methods that allow the implementation of popular boosting techniques in conjunction with the Cox model.
- ORSF (Oblique Random Survival Forests) [30] is a tree-based ensemble for right-censored survival data that uses linear combinations of input variables to recursively partition a set of training data.
- OSTE (Optimal Survival Trees Ensemble) [20] is tree-based ensemble method, which is initiated with the survival tree which stands first in rank, then further tress are tested one by one ba adding them to the ensemble in order of rank.
- UST [17] construct a survival tree by a novel matrix-based algorithm in order to tests a number of nodes simultaneously via stabilized score tests [59].

Comparisons with these models are conducted with corresponding “survival”, “mboost”, “randomForestSRC”, “CoxBoost”, “obliqueRSF”, “OSTE” and “uni.survival.tree” packages in R. The default settings of these methods in packages are adopted for ensemble tree methods and the number of trees is set to 500. For the proposed DSF method, we set $trees = 100$, iteration times $\zeta = 200$ for each level. Here, these values are relatively small to make a trade-off between accuracy and efficiency. In the last level of DSF, we set $Trees = 500$.

4.2 Performance comparison metrics

To evaluate the predictive accuracy of survival models, we adopt the concordance index (C-index) measure [22, 23], which is also the most popular criteria for survival predictions. The C-index metric has an attractive feature that does not depend on a single event time for evaluation and more precisely accounts for censored time. The C-index value is calculated as follows:

- Calculate all possible pairs of cases over the data.
- Omit those pairs whose shorter survival time is censored. Omit pairs i and j if $y_i = y_j$. Let π denote the total number of permissible pairs.
- For each permissible pair where $y_i \neq y_j$, count 1 if the longer survival time has a better predicted outcome; count 0 if predicted outcomes have opposite results. Let ω denote the sum over all permissible pairs.
- $C = \omega / \pi$ defines C-index.

In our experiments, 5∗2 fold cross-validation [57] is used for all datasets. To be specific, each trial randomly divided the dataset into two halves, 50% for training and 50% for testing and vice versa. This process is repeated five times for each dataset and all the compared methods.

4.3 Simulation scenario settings

The simulation settings reported here are very similar to settings [48, 64]. The five settings considered are, respectively, described below:

Scenario 1. In this basic scenario, each simulated dataset is created using 90 independent observations, where the covariate vector $(x_1, x_2, ..., x_{10})$ is multivariate normal with $\mu = 0$ and a covariance matrix having elements equal to $0.9^{|i-j|}$. Survival times are simulated from an exponential distribution with $\mu_T = e^{0.1 \sum_{i=5}^{8} x_i}$(i.e.,a proportional hazards model) and censoring distribution is exponential with $\mu_C = 0.8 e^{0.1 \sum_{i=5}^{8} x_i}$ to get an approximately 66% censoring rate (CR for short when necessary).

Scenario 2. In this nonlinear scenario, the proportional hazards assumption is mildly violated by our settings. Each simulated dataset is created using 90 independent observations, where the covariate vector $(x_1, x_2, ..., x_{10})$ consists of 10 independent and identically distributed uniform random variables on the interval [0,1]. The survival times follow an exponential distribution with $\mu_T = sin(x_1 \pi) + 2 \mid x_2 - 0.5 \mid + x_3^3$. Censoring has a uniform distribution over [0,2], which results in approximately 58% censoring rate.

Scenario 3. In this nonproportional hazard scenario, the proportional hazards assumption is strongly violated by our settings. Each simulated dataset is created using 90 independent observations, where the covariate vector $(x_1, x_2, ..., x_{10})$ is multivariate normal with $\mu = 0$ and a covariance matrix having elements equal to $0.9^{|i-j|}$. Survival times are gamma-distributed with shape parameter $\mu_T = 0.5 + 0.3 \mid \sum_{i=5}^{8} x_i \mid$ and scale parameter 2. Censoring time has a uniform distribution over [0,25], which results in approximately 71% censoring rate.

Scenario 4. In this dependent censoring scenario, the underlying censoring distribution is conditionally dependent on covariates by our settings. Each simulated dataset is created using 90 independent observations, where the covariate vector $(x_1, x_2, ..., x_{10})$ is multivariate normal with $\mu = 0$ and a covariance matrix having elements equal to $0.9^{|i-j|}$. Survival times are simulated according to a log-normal distribution with $\mu_T = 0.1 \mid \sum_{i=1}^{2} x_i \mid + 0.1 \mid \sum_{i=6}^{7} x_i \mid$. Censoring times are log-normal with $\mu_C = \mu_T - 1.5$ and scale parameter 1, which results in approximately 62% censoring rate.

Scenario 5. In this more complicated scenario [27], the log-rank test may have a significant loss of power when the hazard function crosses each other. Each simulated dataset is created using 90 independent observations, where the covariate vector $(x_1, x_2, ..., x_{10})$ is uniformly distributed on the interval [0,1]. Survival time is only related to x_1. Censoring time is uniformly distributed on the interval [0,10], which results in approximately 42% censoring rate. The hazard function is

$$\begin{cases} 0.27t, & x_1 \leq 0.5, t \leq 2 \\ 0.27(t-2) + 5.4, & x_1 \leq 0.5, t > 2 \\ 0.1t, & x_1 > 0.5, t \leq 6 \\ 5.5(t-6) + 0.6, & x_1 > 0.5, t > 6 \end{cases} \tag{7}$$

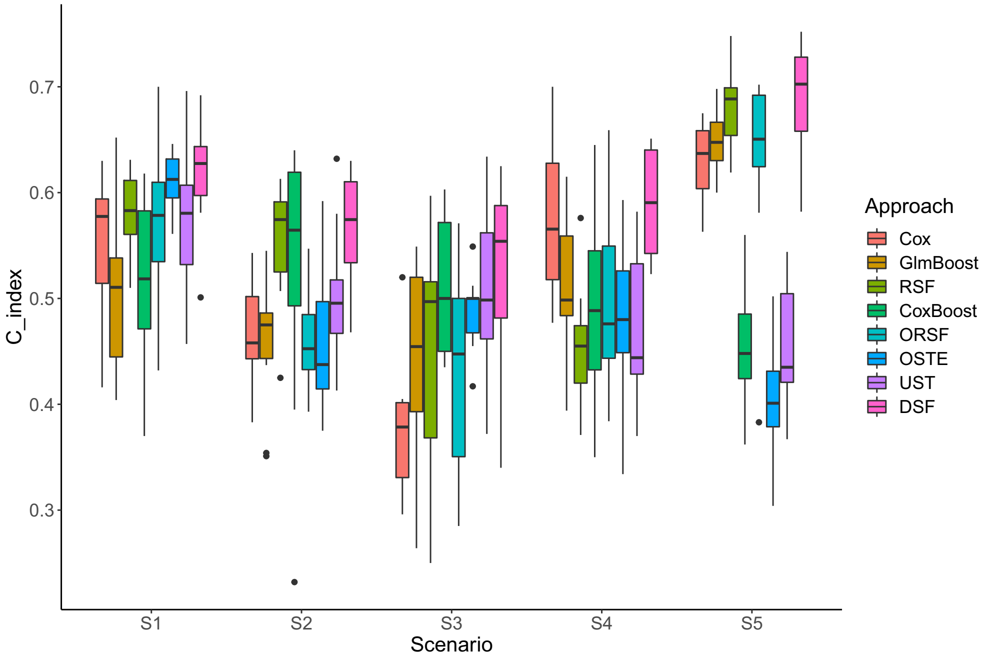
4.4 Simulation results

The following Table 1 and Fig. 3 present the performance of all methods in terms of C-index on five simulation datasets.

Table 1 The result of C-index with other advanced competing approaches

Scenario	Cox	GlmBoost	RSF	CoxBoost	ORSF	OSTE	UST	DSF
S1	0.548	0.512	0.580	0.514	0.567	0.610	0.573	0.618
S2	0.465	0.456	0.554	0.527	0.462	0.463	0.500	0.563
S3	0.376	0.445	0.448	0.511	0.430	0.488	0.502	0.526
S4	0.572	0.508	0.453	0.497	0.500	0.477	0.467	0.591
S5	0.627	0.649	0.680	0.459	0.631	0.405	0.455	0.690

Fig. 3 Boxplots of performance in terms of C-index



In S1 (the most basic proportional hazards case), all eight methods perform relatively well and all output C-index values over 0.5. And the proposed DSF outperforms the other seven methods by a noticeable margin. For S2-S5 (non-proportional hazards cases), all competing methods shows a degrade of performance. In these cases, OSTE and UST often fail and make predictions worse than random guessing. However, DSF works strikingly well in all these scenarios and outperforms the other seven methods by large margins. In the highest censoring rate case (S3), only DSF gives adequate predictions with a mean C-index value of 0.526 and all the other methods fails.

These simulated results indicate that when small sample sized ($n = 90$ in our simulations) and highly censored data (from 58% to 71% in our cases) are confronted and if other approaches fail, one can resort to DSF for help.

5 Real applications

In this section, we verify the potential of the proposed DSF based on real data applications. We will demonstrate its effectiveness using both low dimensional and high dimensional benchmark datasets.

5.1 Applications to low dimensional data

Here, eight real survival datasets with different censoring rates ranging from 24% to 88% are employed and all these datasets are publicly available in corresponding R packages. These datasets are further preprocessed by eliminating all columns having the same values or data with missing values. Short descriptions of the benchmark datasets are given below.

- The *breast* [25] is a breast cancer dataset with 74% censoring rate, containing information on 100 breast cancer patients, including survival time, survival status, Tumor stage, Nodal status, Grading and Cathepsin-D tumor expression. The data can be obtained from the R package “coxphf”.
- The *DLBCL* [2] contains gene expression data from diffuse large B-cell lymphoma (DLBCL) patients. This dataset contains 34 samples and 14 covariates with 47% censoring rate, which is available in R package “ipred”.
- The *leukemia* [18] describes the treatment results for leukemia patients and contains 51 samples and nine covariates with 88% censoring rate. The data can be obtained from the R package “Stat2Data”.
- The *WPBC* [10] exhibits invasive breast cancer cases and contains 194 samples and 32 covariates with 76% censoring rate. The data is available in R package “TH.data”.
- The *ovarian* [16] is a randomized trial comparing two treatments for ovarian cancer with 26 samples and six covariates with 54% censoring rate. The data can be found in R package “survival”.
- The *colon* [35] is from one of the first successful trials of adjuvant chemotherapy for colon cancer. This dataset contains 1858 samples and 14 covariates with 50% censoring rate, which can be obtained from the R package “survival”.

Table 2 The result of C-index on low dimensional datasets

Dataset	Cox	GlmBoost	RSF	CoxBoost	ORSF	OSTE	UST	DSF
breast	0.789	0.770	0.763	0.642	0.768	0.629	0.686	0.804
DLBCL	0.579	0.638	0.523	0.659	0.622	0.554	0.505	0.667
leukemia	0.386	0.335	0.543	0.550	0.500	0.500	0.507	0.820
WPBC	0.571	0.641	0.604	0.629	0.634	0.582	0.610	0.645
ovarian	0.669	0.709	0.751	0.572	0.719	0.525	0.623	0.792
colon	0.664	0.661	0.703	0.664	0.692	0.667	0.703	0.707
kidney	0.755	0.765	0.729	0.763	0.735	0.657	0.722	0.743
pbc	0.802	0.814	0.821	0.813	0.829	0.686	0.819	0.829

- The *kidney* [38] is a kidney patients data. It represents the recurrence times to infection at the point of insertion of the catheter for kidney patients using portable dialysis equipment. This dataset contains 76 samples and six covariates with 24% censoring rate, which can be obtained from the R package “survival”.
- The *pbc* [51] is from the Mayo Clinic trail in primary biliary cirrhosis (pbc) of liver conducted between 1974 and 1984. A total of 276 pbc patients and 17 covariates with 60% censoring rate, referred to Mayo Clinic during that ten-year interval, met eligibility criteria for the randomized placebo controlled trail of the drug D-penicillamine.

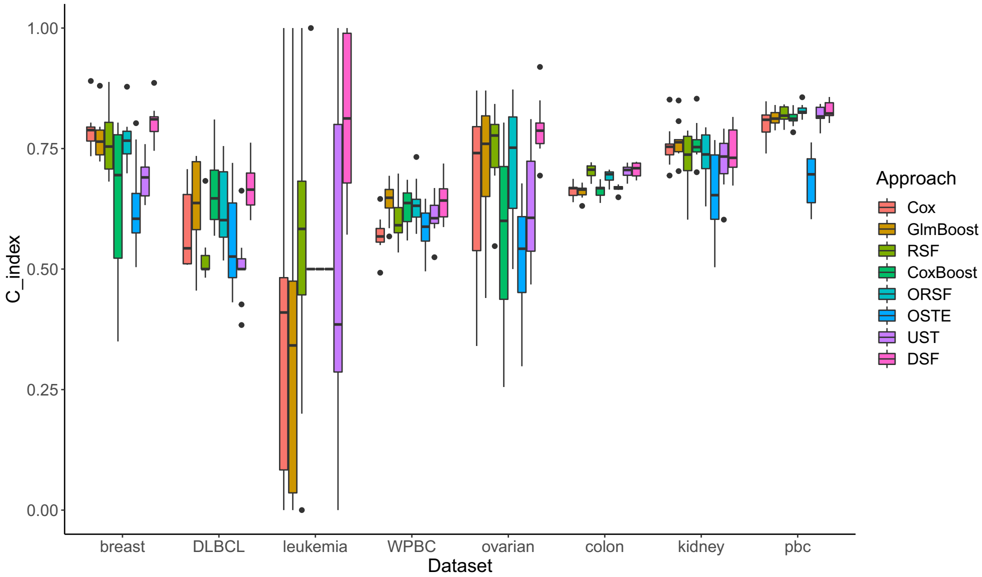
The prediction performance in terms of C-index on low dimensional datasets is summarized in Table 2 and Fig. 4. From these results, one may find that DSF works remarkably well on almost all these datasets and outperforms most methods by big margins in most cases. When extremely high censoring rate is encountered such as the case of *leukemia* dataset, most competitors lose their predictive power with low C-index values(0.335, 0.386, etc.) and CoxBoost manages to give predictions just above 0.5 sometimes but also fails in most runs. However, this mission impossible case is made possible with our proposed DSF method. On the same *leukemia* dataset, DSF performs strikingly well and achieves an average C-index of 0.820.

We also find that, when low censoring rate is present as in the case of *kidney* dataset(24% censoring rate), DSF performs not as good as some of competing methods(GlmBoost or CoxBoost), but its performance is still comparable to other competing approaches.

5.2 Applications to high dimensional data

Next, we will verify the validity of DSF on high-dimensional datasets. Here, for efficiency purpose, a two-stage strategy is adopted for high dimension survival analysis. In the first stage, irrelevant features are filtered out using an effective screening procedure and in the subsequent stage, different competing models come into play. To ensure the fairness of comparison, the same model-free screening

Fig. 4 Boxplots of performance in terms of C-index on low dimensional datasets



method (Ball correlation sure independent screening, BCorSIS) [41] is applied in all first stages. Short descriptions of the benchmark higher-dimensional datasets are given below.

- The *GSE12945* describes an expression module of WIPF1-coex pressed genes that identifies patients. This dataset, with 82 % censoring rate, has 61 patients. For each instance, 14 clinical covariates and 12985 gene features are provided. The data can be obtained from the R package “curatedCRCData” of “Bioconductor”.
- The *NSBCD* [46] contains repeated observations of breast tumor subtypes in independent gene expression datasets. This dataset, with 67 % censoring rate, has 115 patients. For each observation, 549 “intrinsic” genes are provided and can be downloaded from http://user.it.uu.se/~liuya610/.
- The *vdv* [55] is a gene expression profiling for predicting the clinical outcome of breast cancer. This dataset, with 56 % censoring rate, contains 4705 expression values on 78 patients, which is available in R package “randomForestSRC”.
- The *Veer* [5] represents the Circulating Breast Tumor Cells by Differential Expression of Marker Genes. This dataset, with 56 % censoring rate, has 78 patients. For each observation, 4571 gene features are provided, which can be downloaded from https://clincancerres.aacrjournals.org/.
- The *unt* [47] contains the gene expression, annotations and clinical data on breast cancer. This dataset, with 83 % censoring rate, has 62 patients. For each observation, five clinical covariates and 44928 gene features are provided. The data can be obtained from the R package “breastCancerUNT” of “Bioconductor”.
- The *vdx* [40, 58]contains the gene expression, annotations and clinical data. This dataset, with 64 % censoring rate, has 197 patients. For each observation, three clinical covariates and 22283 gene features are provided. The data can be obtained from the R package “breastCancerVDX” of “Bioconductor”.
- The *transbig* [14] contains the gene expression information for lymph node-negative (N-) breast cancer patients. This dataset, with 68 % censoring rate, has 196 patients. For each observation, 22292 gene features are provided. The data can be obtained from the R package “breastCancerTRANSBIG” of “Bioconductor”.
- The *upp* [39] contains transcript profiles of 251 p53-sequenced primary breast tumors. This dataset, with 78 % censoring rate, has 197 patients. For each observation, 44938 gene features are provided. The data can be obtained from the R package “breastCancerUPP” of “Bioconductor”.

From Table 3 and Fig. 5, one can observe that DSF significantly outperforms all seven competing methods on all these high dimensional datasets. On datasets that most methods achieves relatively good predictive performance, such as *GSE12945*, *NSBCD*, *vdv* ,*Veer*, *unt* and *vdx* datasets, the proposed DSF is the best performer. On datasets that most methods may have hard times such as *transbig* and *upp* datasets, DSF performs reasonably well. Thus, from the results shown above, similar to its performance on low dimensional datasets, DSF also achieves good predictive capability in terms of C-index on these high dimensional survival datasets.

Hence, according to the results from both low and high dimensional real datasets with different censoring rates, the proposed DSF method generally obtains a good predictive performance and its superiority stands out if heavily censoring is present.

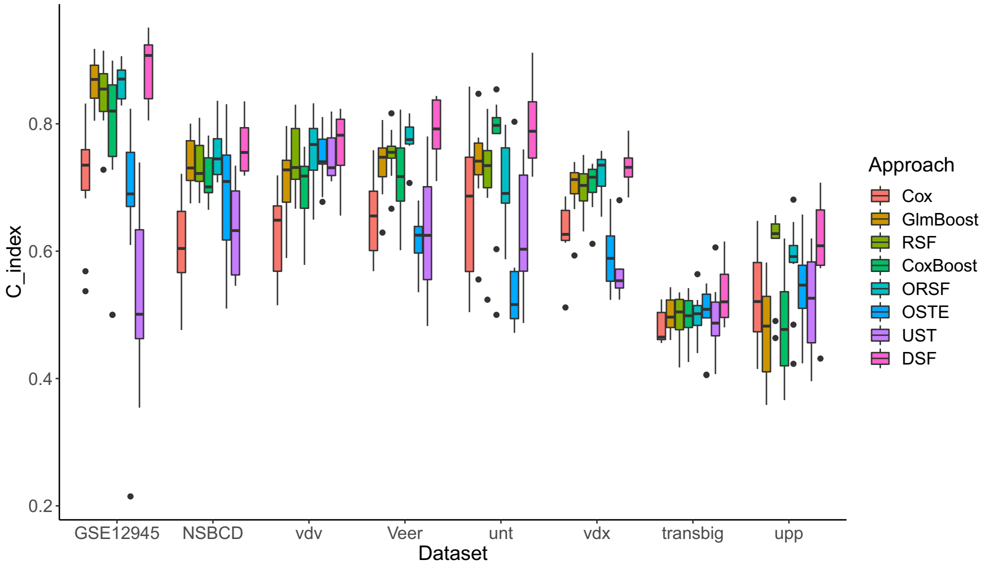
6 Discussions

In the previous two sections, we have demonstrated the effectiveness of the proposed DSF method using extensive simulated scenarios and real benchmark datasets. The success of DSF is probably due to the combination of both the data transduction technique and the cascade

Table 3 The result of C-index on high dimensional datasets

Dataset	Cox	GlmBoost	RSF	CoxBoost	ORSF	OSTE	UST	DSF
GSE12945	0.713	0.866	0.846	0.769	0.867	0.668	0.537	0.886
NSBCD	0.605	0.739	0.738	0.715	0.752	0.683	0.631	0.764
vdv	0.628	0.712	0.746	0.700	0.759	0.746	0.749	0.765
Veer	0.653	0.736	0.753	0.720	0.777	0.614	0.631	0.792
unt	0.663	0.735	0.723	0.757	0.706	0.547	0.631	0.797
vdx	0.630	0.699	0.697	0.702	0.723	0.592	0.565	0.732
transbig	0.480	0.502	0.496	0.496	0.500	0.497	0.495	0.532
upp	0.527	0.476	0.604	0.485	0.580	0.539	0.520	0.610

Fig. 5 Boxplots of performance in terms of C-index on high dimensional datasets



forest structure. The former has shown to exploit more censored information while the latter can achieve a better representation of the original features. When adequate uncensored samples (lower censoring rates and/or large sample sizes) are given, the proposed method may perform worse than other competitors as noises may be introduced in transducting the censored data.

Here, we have conducted additional experiments to verify the above conjectures. First, we test the effectiveness of the deep cascade structure. For this experiment, mean square error (testing_mse) is used for the prediction evaluation in each layer. Figure 6 shows the error rates in terms of testing_mse for each layer in scenarios 1-5. According to Fig. 6, one may observe that as censored data is transduced layer by layer, the testing_mse is generally on the decrease. Hence, if high precision in prediction is required, we can set a larger ζ value to obtain a deeper cascade.

Next, we test the effect of sample sizes on the proposed method. For simplicity, here we only consider the most popular semi-parametric Cox model and non-parametric RSF

Fig. 6 Performance improvement in terms of the influence of cascade structure

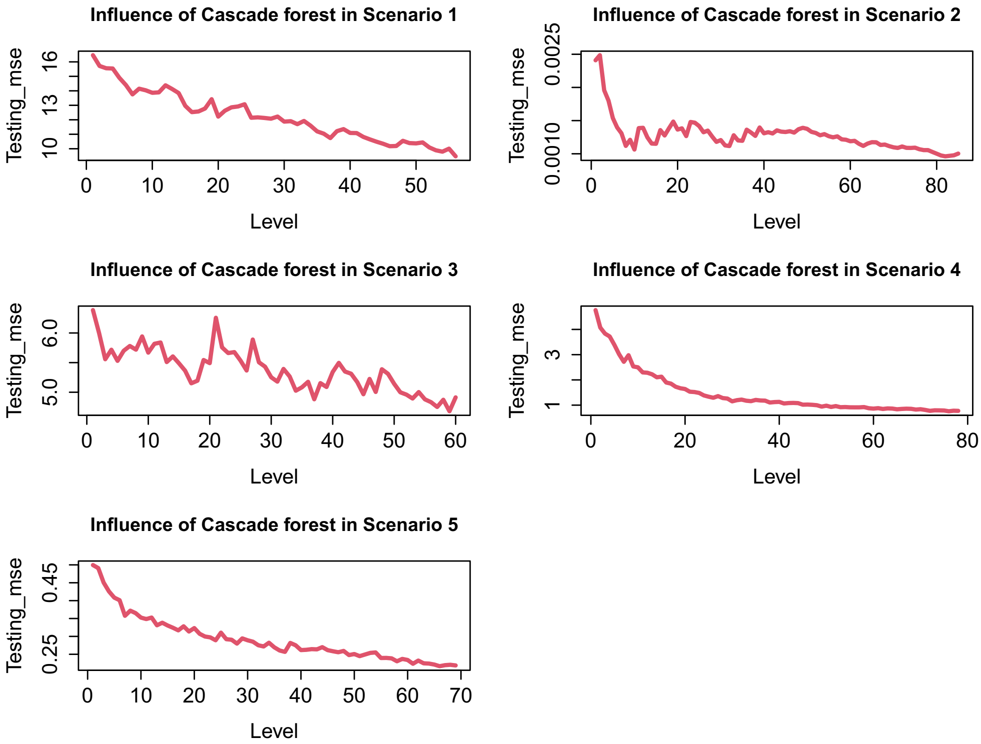


Table 4 Predictive result in terms of average C-index with different sample sizes

Dataset	CR	Sample size	Cox	RSF	DSF
S1	48%	60	0.544	0.497	0.619
S4	92%	60	0.450	0.507	0.627
S1	61%	70	0.497	0.499	0.532
S4	90%	70	0.360	0.491	0.788
S1	56%	80	0.386	0.487	0.531
S4	85%	80	0.485	0.488	0.548
S1	53%	90	0.545	0.499	0.560
S4	81%	90	0.505	0.510	0.620
S1	55%	100	0.609	0.648	0.560
S4	86%	100	0.407	0.780	0.751
S1	56%	120	0.537	0.702	0.600
S4	88%	120	0.407	0.766	0.705
S1	54%	150	0.460	0.649	0.506
S4	81%	150	0.524	0.723	0.573
S1	54%	200	0.526	0.674	0.530
S4	91%	200	0.411	0.842	0.598
S1	51%	300	0.516	0.694	0.554
S4	85%	300	0.460	0.775	0.629
S1	56%	1000	0.498	0.730	0.512
S4	84%	1000	0.513	0.777	0.588

model as the comparing methods. Here, we vary the sample size from 60 to 1000 in two scenarios, one satisfies the proportional hazards assumption (Scenario 1) while the other violates the proportional hazards assumption (Scenario 4). Moreover, to make the comparisons more challenging, all simulated data generated with higher censoring rates. Summary information of different simulated datasets and corresponding comparison results can be found in Table 4 and Fig. 7.

It can be observed that DSF is somewhat sensitive to sample size. When the sample size is less than 100, DSF performs better than the other two competing approaches and the censoring rates seem to have little influence on the predictive performance on DSF. Cox and RSF, however, usually get a bad performance under such scenarios. In contrast, when there is a large sample size with a lower censoring rate, DSF is not as good as RSF, but it still achieves comparable results and outperforms the Cox model by a large margin.

Similar to other deep learning approaches, the computing time of DSF is rather long in the current implementation. But this limitation is counterbalanced by the ability to model small sample sized and highly censored survival data and hence remarkable gains in the predictive capability. Furthermore, the computational issue can be alleviated by parallel computing framework and fast C++ routines in future implementations.

7 Conclusions

In this research, we have proposed a non-neural network like algorithm deep survival forests (DSF) for modelling highly censored survival data, which is prevalent in biomedical

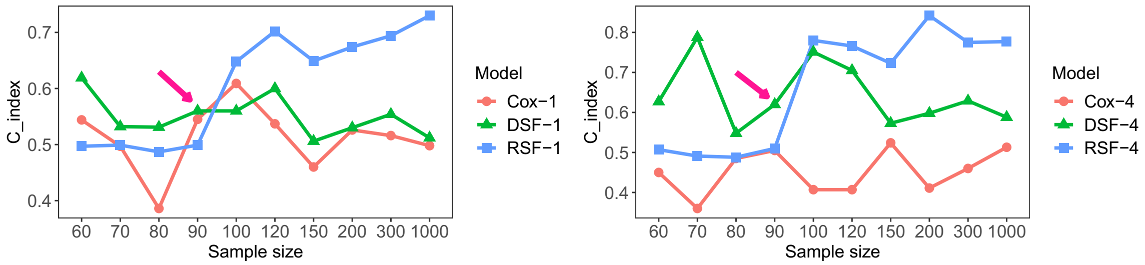


Fig. 7 Performance in terms of C-index on different sample sizes. The deep pink arrow indicates the maximum sample size when the DSF usually performs well

studies. Extensive numerical studies from both simulated and real data have shown that the proposed algorithm outperforms popular Cox, RSF and other state-of-the-art survival ensembles in terms of predictive performance. These results also indicates that the proposed DSF works best with small sample sized survival data with heavy censored rates when sufficient samples are not available in training workable Cox and RSF models.

Potential future research include extending the cascade forest structure to more complex survival data such as interval-censored data or competing risks data. Meanwhile,we also want to study the performance of other transduction techniques, such as Buckley-James [8] and censoring unbiased transduction [48] to make better utilization of censoring information.

Acknowledgements We are grateful for resources from the High Performance Computing Center of Central South University. This work is support in part by the National Social Science Foundation of China (No.17BTJ019), the Postgraduate Innovation Project in Hunan Province (No.150110019) and Changsha Municipal Natural Science Foundation (No.kq2202080).

The authors want to thank all reviewers for their valuable and constructive comments, which greatly improves the quality of this paper.

References

1. Acharya J, Basu A (2020) Deep neural network for respiratory sound classification in wearable devices enabled by patient specific model tuning. IEEE Trans Biomed Circ Syst 14(3):535–544
2. Alizadeh AA, Eisen MB, Davis RE et al (2000) Distinct types of diffuse large b-cell lymphoma identified by gene expression profiling. Nature 403(6769):503–511
3. Bahri M, Bifet A, Maniu S et al (2020) Survey on feature transformation techniques for data streams. In: International joint conference on artificial intelligence (IJCAI)
4. Binder H, Benner A, Bullinger L et al (2013) Tailoring sparse multivariable regression techniques for prognostic single-nucleotide polymorphism signatures. Stat Med 32(10):1778–1791
5. Bosma AJ, Weigelt B, Lambrechts AC et al (2002) Detection of circulating breast tumor cells by differential expression of marker genes. Clin Cancer Res 8(6):1871–1877
6. Bou-Hamad I, Larocque D, Ben-Ameur H et al (2011) A review of survival trees. Stat Surv 5:44–71
7. Breiman L (2001) Random forests. Mach Learn 45(1):5–32
8. Buckley J, James I (1979) Linear regression with censored data. Biometrika 66(3):429–436
9. Buehlmann P et al (2006) Boosting for high-dimensional linear models. Ann Stat 34(2):559–583
10. Bühlmann P, Hothorn T (2007) Boosting algorithms: regularization, prediction and model fitting. Stat Sci 22(4):477–505
11. Ciano G, Rossi A, Bianchini M et al (2021) On inductive–transductive learning with graph neural networks. IEEE Trans Pattern Anal Mach Intell 44(2):758–769
12. Cox DR (1972) Regression models and life-tables. J R Stat Soc: Series B (Methodological) 34(2):187–202
13. Cox DR (1975) Partial likelihood. Biometrika 62(2):269–276
14. Desmedt C, Piette F, Loi S et al (2007) Strong time dependence of the 76-gene prognostic signature for node-negative breast cancer patients in the transbig multicenter independent validation series. Clin Cancer Res 13(11):3207–3214
15. Ding S, Qian W, Wang L (2020) Double-slicing assisted sufficient dimension reduction for high-dimensional censored data. Ann Stat 48(4):2132–2154
16. Edmonson JH, Fleming TR, Decker D et al (1979) Different chemotherapeutic sensitivities and host factors affecting prognosis in advanced ovarian carcinoma versus minimal residual disease. Cancer Treatment Reports 63(2):241–247
17. Emura T, Hsu WC, Chou WC (2021) A survival tree based on stabilized score tests for high-dimensional covariates. J Appl Stat, 1–27
18. Everitt BS (1994) Data come from statistical analysis using s-plus. Chapman & Hall
19. Fard MJ, Wang P, Chawla S et al (2016) A bayesian perspective on early stage event prediction in longitudinal data. IEEE Trans Knowl Data Eng 28(12):3126–3139
20. Gul N (2018) Optimal survival trees ensemble. PhD thesis, University of Essex
21. Guo LZ, Zhang ZY, Jiang Y et al (2020) Safe deep semi-supervised learning for unseen-class unlabeled data. In: International conference on machine learning, PMLR, pp 3897–3906
22. Harrell FE, Califf RM, Pryor DB et al (1982) Evaluating the yield of medical tests. J Amer Med Assoc 247(18):2543–2546
23. Harrell FE Jr, Lee KL, Califf RM et al (1984) Regression modelling strategies for improved prognostic prediction. Stat Med 3(2):143–152
24. Heinze G, Dunkler D (2008) Avoiding infinite estimates of time-dependent effects in small-sample survival studies. Stat Med 27(30):6455–6469
25. Heinze G, Schemper M (2001) A solution to the problem of monotone likelihood in cox regression. Biometrics 57(1):114–119
26. Hoffmann F, Hosseini B, Ren Z et al (2020) Consistency of semi-supervised learning algorithms on graphs: probit and one-hot methods. J Mach Learn Res 21:1–55
27. Moradian H, Larocque D, Bellavance F (2017) l_1 splitting rules in survival forests. Lifetime Data Anal 23(4):671–691
28. Ishwaran H, Kogalur UB, Blackstone EH et al (2008) Random survival forests. Annals Appl Stat 2(3):841–860
29. Ishwaran H, Kogalur UB, Gorodeski EZ et al (2010) High-dimensional variable selection for survival data. J Am Stat Assoc 105(489):205–217
30. Jaeger BC, Long DL, Long DM et al (2019) Oblique random survival forests. Annals Appl Stat 13(3):1847–1883
31. Katzman JL, Shaham U, Cloninger A et al (2018) Deepsurv: personalized treatment recommender system using a cox proportional hazards deep neural network. BMC Medical Res Methodol 18(1):1–12
32. Khan FM, Liu Q (2011) Transduction of semi-supervised regression targets in survival analysis for medical prognosis. In: 2011 IEEE 11th International conference on data mining workshops. IEEE, pp 1018–1025
33. Kvamme H, Borgan Ø, Scheel I (2019) Time-to-event prediction with neural networks and cox regression. J Mach Learn Res 20:1–30
34. Li Z, Liu H, Zhang Z et al (2021) Learning knowledge graph embedding with heterogeneous relation attention networks. IEEE Transactions on Neural Networks and Learning Systems
35. Lin D (1994) Cox regression analysis of multivariate failure time data: the marginal approach. Stat Med 13(21):2233–2247
36. Liu H, Zheng C, Li D et al (2021) Edmf: efficient deep matrix factorization with review feature learning for industrial recommender system. IEEE Transactions on Industrial Informatics

37. Liu H, Liu T, Zhang Z et al (2022) Arhpe: asymmetric relation-aware representation learning for head pose estimation in industrial human-machine interaction. IEEE Transactions on Industrial Informatics
38. McGilchrist C, Aisbett C (1991) Regression with frailty in survival analysis. Biometrics, 461–466
39. Miller LD, Smeds J, George J et al (2005) An expression signature for p53 status in human breast cancer predicts mutation status, transcriptional effects, and patient survival, vol 102
40. Minn AJ, Gupta GP, Padua D et al (2007) Lung metastasis genes couple breast tumor size and metastatic spread. Proc Natl Acad Sci 104(16):6740–6745
41. Pan W, Wang X, Xiao W et al (2019) A generic sure independence screening procedure. J Am Stat Assoc 114(526):928–937
42. Schmidt-Hieber J et al (2020) Nonparametric regression using deep neural networks with relu activation function. Ann Stat 48(4):1875–1897
43. Schoenberg MB, Bucher JN, Koch D et al (2020) A novel machine learning algorithm to predict disease free survival after resection of hepatocellular carcinoma. Ann Transl Med 8:7
44. Shen W, Guo Y, Wang Y et al (2019) Deep differentiable random forests for age estimation. IEEE Trans Pattern Anal Mach Intell 43(2):404–419
45. Sit T, Ying Z, Yu Y (2021) Event history analysis of dynamic networks. Biometrika 108(1):223–230
46. Sorlie T, Tibshirani R, Parker J et al (2003) Repeated observation of breast tumor subtypes in independent gene expression data sets. Proc Natl Acad Sci USA 100(14):8418–8423
47. Sotiriou C, Wirapati P, Loi S et al (2006) Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis. J Natl Cancer Inst 98(4):262–272
48. Steingrimsson JA, Diao L, Strawderman RL (2019) Censoring unbiased regression trees and ensembles. J Am Stat Assoc 114(525):370–383
49. Tang N, Yan X, Zhao X (2020) Penalized generalized empirical likelihood with a diverging number of general estimating equations for censored data. Ann Stat 48(1):607–627
50. Tang W, Ma J, Mei Q et al (2022) Soden: a scalable continuous-time survival model through ordinary differential equation networks. J Mach Learn Res 23(34):1–29
51. Therneau TM, Grambsch PM (2000) The cox model. In: Modeling survival data: extending the Cox model. Springer, pp 39–77
52. Tibshirani R (1997) The lasso method for variable selection in the cox model. Stat Med 16(4):385–395
53. Utkin LV, Konstantinov AV, Chukanov VS et al (2019) A weighted random survival forest. Knowl-Based Syst 177:136–144
54. Van Engelen JE, Hoos HH (2020) A survey on semi-supervised learning. Mach Learn 109(2):373–440
55. Van't Veer LJ, Dai H, Van De Vijver MJ et al (2002) Gene expression profiling predicts clinical outcome of breast cancer. Nature 415(6871):530–536
56. Wang H, Li G (2017) A selective review on random survival forests for high dimensional data. Quant Bio-Sci 36(2):85
57. Wang H, Li G (2019) Extreme learning machine cox model for high-dimensional survival analysis. Stat Med 38(12):2139–2156
58. Wang Y, Klijn JG, Zhang Y et al (2005) Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. The Lancet 365(9460):671–679
59. Witten DM, Tibshirani R (2010) Survival analysis with high-dimensional covariates. Stat Methods Med Res 19(1):29–51
60. Yang S, Pieper K, Cools F (2020) Semiparametric estimation of structural failure time models in continuous-time processes. Biometrika 107(1):123–136
61. Zhang P, Ma J, Chen X et al (2020) A nonparametric method for value function guided subgroup identification via gradient tree boosting for censored survival data. Stat Med 39(28):4133–4146
62. Zhou H, Hanson T (2018) A unified framework for fitting bayesian semiparametric models to arbitrarily censored survival data, including spatially referenced data. J Am Stat Assoc 113(522):571–581
63. Zhou ZH, Feng J (2017) Deep forest: towards an alternative to deep neural networks. In: Twenty-Sixth international joint conference on artificial intelligence, pp 3553–3559
64. Zhu R, Kosorok MR (2012) Recursively imputed survival trees. J Am Stat Assoc 107(497):331–340

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Xuewei Cheng is currently a Ph.D. student at Central South University. His current research interests focus on machine learning for survival analysis and feature screening in ultrahigh-dimensional data.

Sizheng Wang was born in 1997. He received his BSc degree from Fudan University, Shanghai, China in 2019. Since 2020, he has been working toward a master's degree in statistics at Central South University, supervised by Prof. Hong Wang. His research interests are mainly on variable selection, feature screening, and survival analysis. He is currently working on the variable selection with FDR control on survival data.

Hong Wang was born in 1977. He received his PhD from Central South University, China in 2015. Currently, he is an associate professor and PhD supervisor at Central South University. He was a visiting post-doc at University of California, Los Angeles,USA during 2017–2018. His current research interests are survival analysis, big data mining and biostatistics.

Shu Kay Ng is a senior biostatistician at the School of Medicine and Dentistry, Griffith University. He was awarded his PhD degree in statistics from the University of Queensland in 1999. Professor Ng is an experienced researcher, with expertise in biostatistics, statistical modelling and computation, image analysis, machine learning, and survival analysis. He has more than 160 publications. In the field of mixture model-based cluster analysis, he has pioneered the theoretical development of random-effects models for the analysis of complex heterogeneous and correlated data. His research also contributes to solving real-world problems in multidisciplinary fields including bioinformatics, oncology, comorbidity research, medical imaging, and health economics, through modelling and assessment of randomised trials, cohort and longitudinal studies.