ARTICLE TEMPLATE

Improving the within-Node Estimation of Survival Trees while Retaining Interpretability

Haolin Li*, Yiyang Fan*, and Jianwen Cai

*contributed equally to this work.
Department of Biostatistics, University of North Carolina at Chapel Hill, NC, USA

ARTICLE HISTORY
Compiled March 28, 2024

ABSTRACT
In statistical learning for survival data, survival trees are favored for their capacity to detect complex relationships beyond parametric and semiparametric models. Despite this, their prediction accuracy is often suboptimal. In this paper, we propose a new method based on super learning to improve the within-node estimation and overall survival prediction accuracy, while preserving the interpretability of the survival tree. Simulation studies reveal the proposed method's superior finite sample performance compared to conventional approaches for within-node estimation in survival trees. Furthermore, we apply this method to analyze the North Central Cancer Treatment Group Lung Cancer Data and cardiovascular medical records from the Faisalabad Institute of Cardiology.

KEYWORDS
Survival Analysis, Censored Data, Decision Trees, Interpretable Machine Learning, Nonparametric Statistics

1. Introduction

Survival trees have gained popularity in statistical learning literature for censored data due to their ability to detect complex relationships beyond parametric and semiparametric models [1–3]. However, a recognized limitation of survival trees is their instability, resulting in suboptimal accuracy for survival prediction [4, 5]. To address this issue, researchers often turn to tree-based ensembles, such as the random survival forest [6], bagging survival trees [7], recursively imputed survival trees [4], and conditional inference survival forest [8]. While these ensembles enhance prediction accuracy, they lead to "black-box" models, lacking interpretability and insight into the underlying predictive process [9, 10]. In recent years, the issue of interpretability has become increasingly important in the development and implementation of statistical learning models [11], particularly in application settings where interpretability is crucial, such as defining prognosis and risk subgroups based on disease-free survival for cervical carcinoma patients [12] or offering practical silvicultural guidance to minimize losses in oak forest mortality [13]. In such contexts, tree-based ensembles may be impractical, and it is of interest to investigate ways to improve the prediction accuracy of survival trees while preserving their interpretability.

CONTACT Jianwen Cai. Email: cai@bios.unc.edu

Current methodological research on improving survival tree prediction accuracy primarily focuses on refining splitting rules. Commonly employed splitting rules include log-rank-type statistics [1, 14], likelihood-based measurements [2, 15, 16], and various residuals [17, 18]. The review papers by Bou-Hamad, Larocque, and Ben-Ameur (2011) [19] and Wang and Li (2017) [20] provide comprehensive summaries of the existing splitting rules of survival trees. Recent advancements in survival tree methodology involve optimizing the tree construction procedure through mixed-integer optimization and local search techniques, yielding globally optimized survival trees [21]. In this paper, we propose a new approach, focusing on improving the within-node estimation to enhance the prediction accuracy. Our strategy adopts a super learning approach [22, 23], stacking estimates of nested partitions to improve the within-node estimation and overall survival prediction accuracy, while retaining the interpretability of the survival tree.

The paper is structured as follows: In Section 2, we provide detailed descriptions of the proposed method. In Section 3, we present simulation studies assessing its finite sample performances. In Section 4, we apply the method to analyze two real-world datasets: (1) North Central Cancer Treatment Group (NCCTG) Lung Cancer Data [24] and (2) cardiovascular medical records from the Faisalabad Institute of Cardiology [25]. In Section 5, we provide concluding remarks on the method's significance and potential implications in survival tree methodological research.

2. Proposed Method

In this section, we start with a conceptual overview (2.1) and subsequently provide a rigorous description of the algorithm (2.2).

2.1. Conceptual Overview

Our goal is to improve the within-node estimation of a survival tree in estimating the survival distribution (i.e., the survival function or the cumulative hazard function). Using Figure 1, we graphically illustrate the proposed method using an example with two covariates (denoted as Z_1 and Z_2, respectively). Generalization to higher dimensions is straightforward. Imagine in a practical setting, a survival tree algorithm partitions the 2-dimensional covariate space into terminal nodes A, B, C, and D (see left panel, top row of Figure 1). The conventional within-node estimation method involves separately estimating the survival distribution in each terminal node. Specifically, a Nelson-Aalen estimator, based on all observations within a terminal node, is employed to estimate the survival of the corresponding terminal node. Suppose the conventional method yields event rates of 0.9, 0.3, 0.4, and 0.88 for terminal nodes A, B, C, and D, respectively.

Then intuitively, subjects in node A exhibit survival patterns similar to those in node D, and subjects in node B are akin to those in node C. The proposed method aims to construct a sequence of nested partitions by merging similar nodes and then combining the estimates based on these nested partitions to obtain a final estimate. Specifically, we start by merging nodes A and D to form a partition, in which a shared survival distribution of nodes A and D is estimated using all subjects in both nodes (see right panel, top row of Figure 1). Then, nodes B and C are merged to form another partition, and the survival distribution in nodes B and C is estimated from all subjects in nodes B and C (see left panel, second row of Figure 1). Further merging

then produces a one-sample estimate based on all subjects in the entire data set (see right panel, second row of Figure 1). Subsequently, a final estimate is obtained by taking a weighted average of the estimates based on all four partitions demonstrated in Figure 1 using a data-driven approach (i.e., super learning).

By doing so, different terminal nodes with similar survival patterns are allowed to "share information", consequently improving the prediction accuracy. Note that the partitions in the sequence are not necessarily in a tree-based structure, since widely separated terminal nodes may be merged during the process. Nevertheless, since all partitions are nested to the original survival tree (left panel, top row of Figure 1), their weighted combination will also retain the interpretation of the original tree.

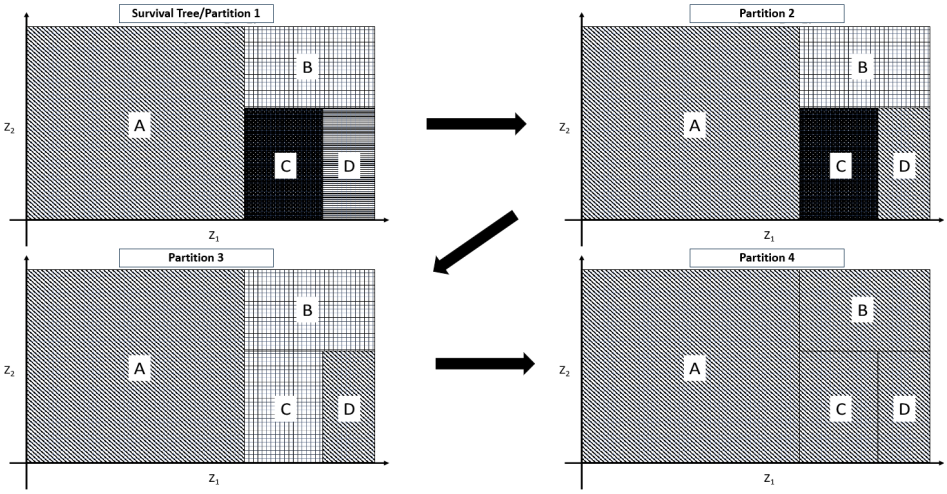


Figure 1. Graphical Illustration of the Proposed Method in 2-Dimensional Covariate Space

2.2. Algorithm

In the standard right-censored setting, we let T denote the survival time, C denote the censoring time, and $\boldsymbol{Z}$ denote the p dimensional covariate vector. The observed time is $X = \min(T, C)$ and the event indicator is $\Delta = I(T \leq C)$. The training data based on n i.i.d. observations are denoted as $\mathcal{L} = \{(X_i, \Delta_i, \boldsymbol{Z}_i), i = 1, \ldots, n\}$. Let $N_i(t) = \Delta_i I(X_i \leq t)$ and $Y_i(t) = I(X_i \geq t)$ denote the counting process and at-risk process of subject i, respectively. The objective is to estimate $S(t|\boldsymbol{Z}) = P(T > t|\boldsymbol{Z})$, or equivalently, $\Lambda(t|\boldsymbol{Z}) = -\log S(t|\boldsymbol{Z})$, using a survival tree. Without loss of generality, we assume that the survival tree is grown by recursively choosing a split in the covariate space that maximizes the value of a particular splitting rule. Let $R_1, \ldots, R_M$ denote the M terminal nodes of the survival tree.

The conventional method of within-node estimation involves calculating the Nelson-Aalen estimator for each terminal node based on all observations in the node. i.e.,

$$\widehat{\Lambda}^M(t|\boldsymbol{Z}_i) = \sum_{m=1}^{M} \widehat{\Lambda}^m(t|\boldsymbol{Z}_i) I(\boldsymbol{Z}_i \in R_m), \tag{1}$$

where $\widehat{\Lambda}^m(t|\boldsymbol{Z}_i) = \dfrac{1}{n_m} \sum_{i=1}^{n} \int_0^t \dfrac{I(\boldsymbol{Z}_i \in R_m) dN_i(u)}{n_m^{-1} \sum_{j=1}^{n} I(\boldsymbol{Z}_j \in R_m) Y_j(u)}$, and $n_m = \sum_{i=1}^{n} I(\boldsymbol{Z}_i \in R_m)$

denotes the number of subjects in node R_m. Then given the new covariate information $\boldsymbol{z}_0$, the predicted cumulative hazard function is $\widehat{\Lambda}^M(t|\boldsymbol{z}_0)$ and the predicted survival function is $\widehat{S}^M(t|\boldsymbol{z}_0) = \exp(-\widehat{\Lambda}^M(t|\boldsymbol{z}_0))$.

To improve the efficiency of the within-node estimation, we propose the follow estimation method based on super learning using K-fold cross-validation:

1. Randomly split $\mathcal{L}$ into K folds (a recommended choice of K is 10), denoted as $\mathcal{L}_{(1)}, \ldots, \mathcal{L}_{(K)}$.
2. For $k = 1, \ldots, K$, do:
   i. Let $\mathcal{L}^{(k)} = \mathcal{L} - \mathcal{L}_{(k)}$. Use observations in $\mathcal{L}^{(k)}$ to recalculate the within-nodes estimation of the survival tree (1), denoted as $\widehat{\Lambda}^M_{(k)}(t|\boldsymbol{Z}_i)$:

$$\widehat{\Lambda}^M_{(k)}(t|\boldsymbol{Z}_i) = \sum_{m=1}^{M} \widehat{\Lambda}^m_{(k)}(t|\boldsymbol{Z}_i) I(\boldsymbol{Z}_i \in R_m),$$

   where $\widehat{\Lambda}^m_{(k)}(t|\boldsymbol{Z}_i) = \dfrac{1}{n^{(k)}_m} \displaystyle\sum_{i=1}^{n} \int_0^t \frac{I(\boldsymbol{Z}_i \in R_m, i \in \mathcal{L}^{(k)}) dN_i(u)}{(n^{(k)}_m)^{-1} \sum_{j=1}^n I(\mathbf{Z}_j \in R_m, j \in \mathcal{L}^{(k)}) Y_j(u)}$,

   and $n^{(k)}_m = \displaystyle\sum_{i=1}^{n} I(\boldsymbol{Z}_i \in R_m, i \in \mathcal{L}^{(k)})$ denotes the number of subjects in $\mathcal{L}^{(k)}$ and in node R_m.

   ii. Merge two terminal nodes with the smallest value of the splitting rule among all pairwise comparisons of all terminal nodes, say R_p and R_q, where $p, q \in \{1, \ldots, M\}$ and $p \neq q$, to obtain a partition where R_p and R_q are combined. This partition is nested in the original survival tree. The estimate based on this partition is calculated as

$$\widehat{\Lambda}^{M-1}_{(k)}(t|\boldsymbol{Z}_i) = \sum_{m \neq p,q} \widehat{\Lambda}^m_{(k)}(t|\boldsymbol{Z}_i) I(\boldsymbol{Z}_i \in R_m) + I(\boldsymbol{Z}_i \in R_p \cup R_q)$$
$$\times \frac{1}{n^{(k)}_p + n^{(k)}_q} \sum_{i=1}^{n} \int_0^t \frac{I(\boldsymbol{Z}_i \in R_p \cup R_q, i \in \mathcal{L}^{(k)}) dN_i(u)}{(n^{(k)}_p + n^{(k)}_q)^{-1} \sum_{j=1}^n I(\boldsymbol{Z}_j \in R_p \cup R_q, j \in \mathcal{L}^{(k)}) Y_j(u)}.$$

   iii. Treat R_p and R_q as a single node. Continue merging two nodes with the smallest value of the splitting rule and calculate the corresponding estimate of the cumulative hazard function, $\widehat{\Lambda}^{M-2}_{(k)}(t|\boldsymbol{Z})$. Then, continue the procedure to obtain a sequence of partitions and calculate the corresponding estimates, denoted as $\widehat{\Lambda}^M_{(k)}(t|\boldsymbol{Z}), \ldots, \widehat{\Lambda}^1_{(k)}(t|\boldsymbol{Z})$, where

$$\widehat{\Lambda}^1_{(k)}(t|\boldsymbol{Z}_i) = \frac{1}{n^{(k)}} \sum_{i=1}^{n} \int_0^t \frac{I(i \in \mathcal{L}^{(k)}) dN_i(u)}{(n^{(k)})^{-1} \sum_{j=1}^n I(j \in \mathcal{L}^{(k)}) Y_j(u)},$$

   where $n^{(k)} = \displaystyle\sum_{i=1}^{n} I(i \in \mathcal{L}^{(k)})$.

   iv. Calculate the out-of-fold predicted cumulative hazard function for $\mathcal{L}_{(k)}$ based on the sequence of partitions.
3. Denote the out-of-fold predicted cumulative hazard function based on the sequence of partitions as $\widetilde{\Lambda}^M(t|\boldsymbol{Z}), \ldots, \widetilde{\Lambda}^1(t|\boldsymbol{Z})$. Compute the optimal stacking

weights $\widehat{w}_1, \ldots, \widehat{w}_M$ by maximizing cross-validated C-index [26]. Specifically,

$$\{\widehat{w}_1, \ldots, \widehat{w}_M\} = \text{argmax}_{w_1,\ldots,w_M} C\bigg(\mathcal{L}, \sum_{m=1}^{M} w_m \widetilde{\Lambda}^m(\tau|\boldsymbol{Z})\bigg), \tag{2}$$

where τ denotes the end-of-study time, and

$$C(\mathcal{L}, \mathcal{S}(\boldsymbol{Z})) = \bigg\{ \sum_{i=1}^{n} \Delta_i \sum_{j=i+1}^{n} \big[I(X_i < X_j) + (1-\Delta_j) I(X_i = X_j)\big] \big[I(\mathcal{S}(\boldsymbol{Z}_i) > \mathcal{S}(\boldsymbol{Z}_j)) + \frac{1}{2} I(\mathcal{S}(\boldsymbol{Z}_i) = \mathcal{S}(\boldsymbol{Z}_j))\big] \bigg\} \bigg/ \bigg\{ \sum_{i=1}^{n} \Delta_i \sum_{j=i+1}^{n} \big[I(X_i < X_j) + (1-\Delta_j) I(X_i = X_j)\big] \bigg\}$$

denote the C-index, where $\mathcal{S}(\boldsymbol{Z})$ is some risk score in which a high value of $\mathcal{S}(\boldsymbol{Z})$ indicates a greater probability of developing the event of interest. The optimization in (2) can be conducted by the augmented Lagrangian adaptive barrier minimization algorithm [27] under the constrains such that $\sum_{m=1}^{M} w_m = 1$ and $w_m \geq 0$ for all $m = 1, \ldots, M$.

4. The final estimation of the survival tree is

$$\widehat{\Lambda}(t|\boldsymbol{Z}) = \sum_{m=1}^{M} \widehat{w}_m \widehat{\Lambda}^m(t|\boldsymbol{Z}),$$

where $\widehat{\Lambda}^m(t|\boldsymbol{Z})$ is calculated based on the sequence of partitions formed by sequentially merging the terminal nodes based on all observations in $\mathcal{L}$.

Then given the new covariate information $\boldsymbol{z}_0$, the proposed method of within-node estimation yields the predicted cumulative hazard function as $\widehat{\Lambda}(t|\boldsymbol{z}_0)$ and the predicted survival function is $\widehat{S}(t|\boldsymbol{z}_0) = \exp(-\widehat{\Lambda}(t|\boldsymbol{z}_0))$.

3. Simulation Studies

We perform simulation studies to evaluate the finite sample performance of the proposed methods across various scenarios. Using a sample size of 100 or 300, we generate the covariate vector $\boldsymbol{Z} = (Z_1, \ldots, Z_{10})^\top$ from a 10-dimensional multivariate normal distribution with a mean vector $\mathbf{0}$ and a variance-covariance matrix Σ, where the (i,j)-th component $\Sigma_{i,j} = 0.6^{|i-j|}$, for all $i, j = 1, \ldots, p$. For the sample size of 100, we consider event rates of 50% and 80%; for the sample size of 300, we consider event rates of 30% and 50%. Survival times are generated from a Weibull proportional hazards model with three types of covariate effects: (1) tree-based structure, (2) linear covariate effect, and (3) nonlinear covariate effect. For tree-based structure, the survival time is generated from

$$\Lambda(t|\boldsymbol{Z}) = t^{0.7} \exp(3(Z_1 > 0) - 2(Z_2 < 0) + 3(Z_4 < 0)(Z_6 < 0) - 0.4(Z_7 > 0)(Z_3 > 0)). \tag{3}$$

For linear covariate effect, the survival time is generated from

$$\Lambda(t|\boldsymbol{Z}) = t^{0.7}\exp(3Z_1 - 4Z_2 + 2.5Z_3 + 3Z_4 - Z_6 - 4Z_7 + 1.5Z_8 + 3Z_9 - 2Z_{10}). \quad (4)$$

For nonlinear covariate effect, the survival time is generated from

$$\Lambda(t|\boldsymbol{Z}) = t^{0.7}\exp(0.3Z_1 - 0.2Z_2Z_3 + 0.3Z_4Z_6 - 4Z_7^3 + Z_3Z_5 - 0.25Z_8 - 3Z_1Z_9Z_{10}). \quad (5)$$

Censoring times are generated from a Uniform[0, c_{max}] distribution truncated at τ_{max}, where $\tau_{max} < c_{max}$. c_{max} and τ_{max} are chosen based on the targeted event rates.

When fitting survival trees, we use the log-rank test as the splitting rule and apply the default stopping rules for tree growing. When applying the proposed method for within-node estimation, we use the 10-fold cross-validation (i.e., $K = 10$). Prediction accuracy is compared between survival trees with conventional within-node estimation (i.e., solely based on observations within each terminal node) and survival trees with the proposed within-node estimation method. For each scenario, a test dataset is generated using the same mechanism as the training dataset, and the C-index based on the test dataset is used to measure the goodness of prediction.

We perform 500 simulations per scenario and present results in Table 1. Reported are the median, lower quartile (Q1), and upper quartile (Q3) of C-indices for survival trees, comparing conventional and proposed methods for within-node estimation. Additionally, we calculate the percentage improvement in median C-index by the proposed method compared to the conventional one. Results in Table 1 consistently show the proposed method outperforming, with median C-index improvement ranging from 2% to 10% across most scenarios. Notably, for tree-based covariate effect (equation 3) with a sample size of 100 and a 80% event rate, the proposed method enhances median C-index from 0.55 to 0.60, a 10.4% improvement. In linear covariate effect (equation 4) with a sample size of 300 and a 50% event rate, the proposed method increases median C-index from 0.64 to 0.66, a 3.4% rise. In the case of nonlinear covariate effect (equation 5) with a sample size of 100 and a 50% event rate, the proposed method raises median C-index from 0.64 to 0.68, a 6.5% increase. Similarly, Q1 and Q3 of the C-index using the proposed method consistently surpass those of survival trees with conventional within-node estimation.

Table 1. Median (Q1, Q3) of C-Indices of Survival Trees with Conventional and Proposed Methods of within-Node Estimation in Various Simulation Settings (% Improvement measured by the % Increase in Median C-Index by the Proposed Method)

			within-Node Estimation method		
Covariate Effect	Sample Size	Event Rate	Conventional	Proposed	% Improvement
Tree-Based	100	80%	0.55 (0.49, 0.60)	0.60 (0.56, 0.64)	10.4%
		50%	0.62 (0.57, 0.66)	0.65 (0.61, 0.69)	4.3%
	300	50%	0.67 (0.65, 0.70)	0.71 (0.69, 0.73)	5.6%
		30%	0.69 (0.67, 0.72)	0.71 (0.68, 0.74)	2.9%
Linear	100	80%	0.54 (0.49, 0.57)	0.59 (0.55, 0.62)	10.1%
		50%	0.60 (0.56, 0.64)	0.62 (0.59, 0.66)	3.7%
	300	50%	0.64 (0.62, 0.65)	0.66 (0.64, 0.68)	3.4%
		30%	0.66 (0.64, 0.68)	0.67 (0.65, 0.69)	1.7%
Nonlinear	100	80%	0.57 (0.50, 0.64)	0.60 (0.55, 0.64)	5.8%
		50%	0.64 (0.59, 0.69)	0.68 (0.63, 0.73)	6.5%
	300	50%	0.63 (0.57, 0.67)	0.67 (0,63, 0.70)	5.5%
		30%	0.72 (0.67, 0.75)	0.74 (0.70, 0.77)	2.7%

4. Applications

4.1. North Central Cancer Treatment Group Lung Cancer Data

The North Central Cancer Treatment Group (NCCTG) Lung Cancer Data [24] includes 228 patients, aiming to evaluate the survival probabilities of lung cancer patients after diagnosis. The dataset encompasses demographic information and performance assessments, including sex, age, Eastern Cooperative Oncology Group (ECOG) performance score [28] (a score ranging from 0 to 5, with higher scores indicating worse performances), Karnofsky performance score [29] (a score ranging from 0 to 100, with higher scores indicating better performances), caloric intake at meals, and weight loss in the last six months (in pounds). The ECOG performance score and the Karnofsky performance score are two commonly used scales to measure the functional status of a patient. Both scales are used to classify patients according to their functional impairment, assess the prognosis, and compare the effectiveness of therapies. Among the 228 patients, 63 cases are right-censored.

To compare the proposed method against the conventional within-node estimation approach, we divide the dataset by a 6:4 ratio. The training set includes 60% of the original data, and the test set includes the remaining 40%. Utilizing the training data, we constructed a survival tree with the log-rank test as the splitting rule and used both the conventional within-node estimation and the proposed method with 10-fold cross-validation (i.e., $K = 10$) for obtaining the weights. Subsequently, the test set were employed to evaluate their predictive accuracy. Results indicate that the proposed within-node estimation method yields a C-index of 0.581, surpassing the conventional method's C-index of 0.534, representing a 9% improvement.

Then we conduct the analysis based on the entire North Central Cancer Treatment Group Lung Cancer Data. In Figure 2, analysis results are presented. The upper segment illustrates the survival tree structure, while the lower segment presents the estimated survival function utilizing the proposed within-node estimation method for each terminal node. The survival tree is interpretable through sequential "if-then" clauses:

(i) If patient caloric intake at meals is ≤ 500, then terminal node 1 is reached, with a median survival time of approximately 250 days and a survival probability of around 0.10 at 600 days.

(ii) If (i) is false and the patient's ECOG performance score is ≤ 0.5, then terminal node 2 is reached, with a median survival time of about 450 days and a survival probability of approximately 0.35 at 600 days.

(iii) If neither (i) nor (ii) is true and the patient's weight loss in the past six months exceeds 14.5 pounds, then terminal node 5 is reached, with a median survival time of around 350 days and a survival probability of about 0.30 at 600 days.

(iv) If none of the conditions (i)-(iii) hold and the patient is male, then terminal node 3 is reached, with a median survival time of approximately 250 days and a survival probability of around 0.15 at 600 days.

(v) If none of the conditions (i)-(iv) hold, the patient reaches terminal node 4, with a median survival time of about 350 days and a survival probability of around 0.30 at 600 days.

4.2. Cardiovascular Medical Records from the Faisalabad Institute of Cardiology

We also analyze the cardiovascular medical records obtained from the Faisalabad Institute of Cardiology [25], including data from 299 patients with heart failure. The cohort consists of 105 females and 194 males, aged between 40 and 95 years. All subjects exhibited left ventricular systolic dysfunction, placing them in New York Heart Association (NYHA) classes III or IV due to prior heart failures. The dataset contains 11 covariates: age, anemia, high blood pressure, creatinine phosphokinase, diabetes, ejection fraction, sex, platelets, serum creatinine, serum sodium, and smoking. Notably, anemia, high blood pressure, diabetes, sex, and smoking are binary variables. Additionally, 96 out of 299 patients experienced the event of interest (death).

To facilitate method evaluation, we partition the dataset into training and test sets using a 6:4 ratio and assess the performance of conventional and proposed within-node estimation methods with 10-fold cross-validation (i.e., $K = 10$) for obtaining the weights. Again, the survival tree is grown using the log-rank test as the splitting rule. The proposed method demonstrated a C-index of 0.671, surpassing the conventional method, which yielded a C-index of 0.625. This represents a 7% improvement in the prediction accuracy.

Figure 3 presents the analysis results based on the entire dataset. The survival tree can be interpreted as follows:

(i) If the patient's ejection fraction is lower than 32.5%, then terminal 1 is reached, with a survival probability of 0.30 at 250 days.
(ii) If (i) is false and if the patient's creatinine exceeds 1.25 mg/dL, then terminal 4 is reached, with a survival probability of 0.47 at 250 days.
(iii) If (i) and (ii) do not hold and if the creatinine phosphokinase of the patient is less than 577 mcg/L, then terminal 2 is reached, with a survival probability of 0.82 at 250 days.
(iv) If none of the conditions (i)-(iii) hold, then terminal 3 is reached, with a survival probability of 0.77 at 250 days.

5. Discussion

In this paper, we propose a super learning strategy to improve the within-node estimation and overall prediction accuracy of survival trees. Previous research indicates that achieving satisfactory performance in survival trees necessitates a relatively large sample size because the survival estimation for each terminal node relies solely on observations within that node [5]. Our method promotes "information sharing" among terminal nodes exhibiting similar survival patterns, thus allowing more efficient utilization of information for a specified sample size. The degrees of information shared among terminal nodes are also determined by a data-driven process. Our method offers distinct advantages. First, the proposed method helps the survival tree improve within-node estimation and overall prediction accuracy while maintaining its original interpretation. Second, the proposed method has versatile applicability, as it can be integrated into survival trees using any splitting rules.

The proposed method offers advantages over other tree-based ensembles (e.g., random survival forest, bagging survival trees, recursively imputed survival trees, and conditional inference survival forest) due to its straightforward and transparent in-

terpretation. The proposed method retains the simple structure of the survival tree, which represents sequential "if-then" clauses, directly illustrating the decision-making process. Each terminal node represents a subgroup based on recursive partitioning of the covariate space, facilitating an easy understanding of how the survival tree arrives at its predictions. In contrast, other tree-based ensembles aggregate estimations from numerous survival trees with very different structures, complicating interpretation by obscuring the direct path from input to output and making interpretation less straightforward.

The application of the proposed method that improves the within-node estimation and overall prediction accuracy of survival trees will make a significant impact on biomedical and healthcare research, especially in advancing personalized medicine and optimizing healthcare delivery. Accurate predictions enable tailored treatment plans, efficient resource allocation, and informed decision-making for patients. This can help enhance making medical prognosis, support early intervention, and aid in risk stratification for public health planning. For example, in oncology, a survival tree with improved prediction accuracy can help identifying patients with specific genetic markers or tumor characteristics that influence response to treatment. This information guides oncologists in tailoring therapies, minimizing side effects, and optimizing the chances of successful outcomes. Additionally, accurate survival predictions aid in the selection of eligible participants for clinical trials, accelerating the development of targeted therapies and fostering advancements in cancer research. Overall, increased prediction accuracy contributes to better patient outcomes, resource efficiency, and the ongoing evolution of precision medicine in the healthcare landscape.

There are also future directions for further research. First, it will be of interest to extend the proposed method to the classification and regression trees (CART) [30] for modeling continuous and categorical outcomes and investigating their finite sample performances. Additionally, extending the method to more complex survival analysis settings, such as competing risks [31, 32], recurrent events [33, 34], left truncation [5, 35], and multivariate survival analysis [36, 37], would be another prospective direction. Finally, in many epidemiological studies and biomedical research, the observed data are not i.i.d. (e.g., complex survey sampling [38], outcome dependent sampling [39]). Further research is needed to incorporate the study design into the proposed method in these contexts.

Funding Details

This work was supported by the National Institute of Environmental Health Science (NIEHS) under Grant T32ES007018.

Disclosure Statement

The authors report there are no competing interests to declare.

Data Availability Statement

All data and code are publicly available at `https://github.com/LiLeoHaolin/ImprovedST`.

References

[1] Segal M. Regression trees for censored data. *Biometrics*. 1988;44(1):35. doi:10.2307/2531894

[2] Fan J, Nunn ME, Su X. Multivariate Exponential Survival Trees And Their Application to Tooth Prognosis. *Comput Stat Data Anal*. 2009;53(4):1110-1121. doi:10.1016/j.csda.2008.10.019

[3] Steingrimsson JA, Diao L, Molinaro AM, Strawderman RL. Doubly robust survival trees. *Stat Med*. 2016;35(20):3595-3612. doi:10.1002/sim.6949

[4] Zhu R, Kosorok MR. Recursively Imputed Survival Trees. *J Am Stat Assoc*. 2012;107(497):331-340. doi:10.1080/01621459.2011.637468

[5] Fu W, Simonoff JS. Survival trees for left-truncated and right-censored data, with application to time-varying covariate data. *Biostatistics*. 2017;18(2):352-369. doi:10.1093/biostatistics/kxw047

[6] Ishwaran H, Kogalur UB, Blackstone EH, Lauer MS. Random survival forests. *The Annals of Applied Statistics*. 2008;2(3). doi:10.1214/08-aoas169

[7] Hothorn T, Lausen B, Benner A, Radespiel-Tröger M. Bagging survival trees. *Stat Med*. 2004;23(1):77-91. doi:10.1002/sim.1593

[8] Nasejje JB, Mwambi H, Dheda K, Lesosky M. A comparison of the conditional inference survival forest model to random survival forests based on a simulation study as well as on two applications with time-to-event data. *BMC Med Res Methodol*. 2017;17(1):115. doi:10.1186/s12874-017-0383-8

[9] Castelvecchi D. Can we open the black box of AI?. *Nature*. 2016;538(7623):20-23. doi:10.1038/538020a

[10] Samek W, Müller, KR. Towards explainable artificial intelligence. *Explainable AI: interpreting, explaining and visualizing deep learning*. 2019.

[11] Gilpin LH, Bau D, Yuan BZ, Bajwa A, Specter M, Kagal L. Explaining explanations: An overview of interpretability of machine learning. *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*. 2018. doi:10.1109/dsaa.2018.00018

[12] Sevin BU, Lu Y, Bloch DA, Nadji M, Koechli OR, Averette HE. Surgically defined prognostic parameters in patients with early cervical carcinoma. A multivariate survival tree analysis. *Cancer*. 1996;78(7):1438-1446. doi:10.1002/(SICI)1097-0142(19961001)78:7¡1438::AID-CNCR10¿3.0.CO;2-0

[13] Fan Z, Kabrick JM, Shifley SR. Classification and regression tree based survival analysis in oak-dominated forests of Missouri's Ozark highlands. *Canadian Journal of Forest Research*. 2006;36(7):1740-1748. doi:10.1139/x06-068

[14] Ciampi A, Thiffault J, Nakache J-P, Asselain B. Stratification by stepwise regression, correspondence analysis and recursive partition: A comparison of three methods of analysis for survival data with covariates. *Computational Statistics & Data Analysis*. 1986;4(3):185-204. doi:10.1016/0167-9473(86)90033-2

[15] Ciampi A, Chang C-H, Hogg S, McKinney S. Recursive partition: A versatile method for exploratory-data analysis in Biostatistics. *Biostatistics*. Published online 1987:23-50. doi:10.1007/978-94-009-4794-8_2

[16] Davis RB, Anderson JR. Exponential survival trees. *Stat Med*. 1989;8(8):947-961. doi:10.1002/sim.4780080806

[17] Therneau TM, Grambsch PM, Fleming TR. Martingale-based residuals for survival models. *Biometrika*. 1990;77(1):147. doi:10.2307/2336057

[18] Keleş S, Segal MR. Residual-based tree-structured survival analysis. *Stat Med*. 2002;21(2):313-326. doi:10.1002/sim.981

[19] Bou-Hamad I, Larocque D, Ben-Ameur H. A review of Survival Trees. *Statistics Surveys*. 2011;5(none). doi:10.1214/09-ss047

[20] Wang H, Li G. A Selective Review on Random Survival Forests for High Dimensional Data. *Quant Biosci*. 2017;36(2):85-96. doi:10.22283/qbs.2017.36.2.85

[21] Bertsimas D, Dunn J, Gibson E, Orfanoudaki A. Optimal survival trees. *Machine Learn-*

ing. 2022;111(8):2951-3023. doi:10.1007/s10994-021-06117-0
[22] Breiman L. Stacked regressions. *Machine Learning*. 1996;24(1):49-64. doi:10.1007/bf00117832
[23] van der Laan MJ, Polley EC, Hubbard AE. Super learner. *Stat Appl Genet Mol Biol*. 2007;6:Article25. doi:10.2202/1544-6115.1309
[24] Loprinzi CL, Laurie JA, Wieand HS, et al. Prospective evaluation of prognostic variables from patient-completed questionnaires. North Central Cancer Treatment Group. *J Clin Oncol*. 1994;12(3):601-607. doi:10.1200/JCO.1994.12.3.601
[25] Chicco D, Jurman G. Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone. *BMC Med Inform Decis Mak*. 2020;20(1):16. doi:10.1186/s12911-020-1023-5
[26] Harrell FE Jr, Califf RM, Pryor DB, Lee KL, Rosati RA. Evaluating the yield of medical tests. *JAMA*. 1982;247(18):2543-2546.
[27] Madsen K, Nielsen HB, & Tingleff O. Optimization with constraints. 2004
[28] Zubrod CG, Schneiderman M, Frei E, et al. Appraisal of methods for the study of chemotherapy of cancer in man: Comparative therapeutic trial of nitrogen mustard and triethylene thiophosphoramide. *Journal of Chronic Diseases*. 1960;11(1):7-33. doi:10.1016/0021-9681(60)90137-5
[29] Karnofsky DA. The clinical evaluation of chemotherapeutic agents in cancer. *Evaluation of chemotherapeutic agents*. 1949;191-205.
[30] Breiman L, Friedman JH, Olshen RA, Stone CJ. Regression trees. *Classification And Regression Trees*. 2017:216-265. doi:10.1201/9781315139470-8
[31] Andersen PK, Geskus RB, de Witte T, Putter H. Competing risks in epidemiology: possibilities and pitfalls. *Int J Epidemiol*. 2012;41(3):861-870. doi:10.1093/ije/dyr213
[32] Kretowska M. Tree-based models for survival data with competing risks. *Comput Methods Programs Biomed*. 2018;159:185-198. doi:10.1016/j.cmpb.2018.03.017
[33] Cai J, Schaubel DE. Analysis of Recurrent Event Data. *Handbook of Statistics*. 2003:603-623. doi:10.1016/s0169-7161(03)23034-0
[34] Sparapani RA, Rein LE, Tarima SS, Jackson TA, Meurer JR. Non-parametric recurrent events analysis with BART and an application to the hospital admissions of patients with diabetes. *Biostatistics*. 2020;21(1):69-85. doi:10.1093/biostatistics/kxy032
[35] Guo G. Event-history analysis for left-truncated data. *Sociol Methodol*. 1993;23:217-243.
[36] Gill RD. Multivariate survival analysis. *Theory of Probability & Its Applications*. 1993;37(2):284-301. doi:10.1137/1137061
[37] Su X, Fan J. Multivariate survival trees: a maximum likelihood approach based on frailty models. *Biometrics*. 2004;60(1):93-99. doi:10.1111/j.0006-341X.2004.00139.x
[38] Lohr SL. *Sampling: Design and Analysis*. CRC Press, Taylor & Francis Group; 2022.
[39] Ding J, Lu TS, Cai J, Zhou H. Recent progresses in outcome-dependent sampling with failure time data. *Lifetime Data Anal*. 2017;23(1):57-82. doi:10.1007/s10985-015-9355-7

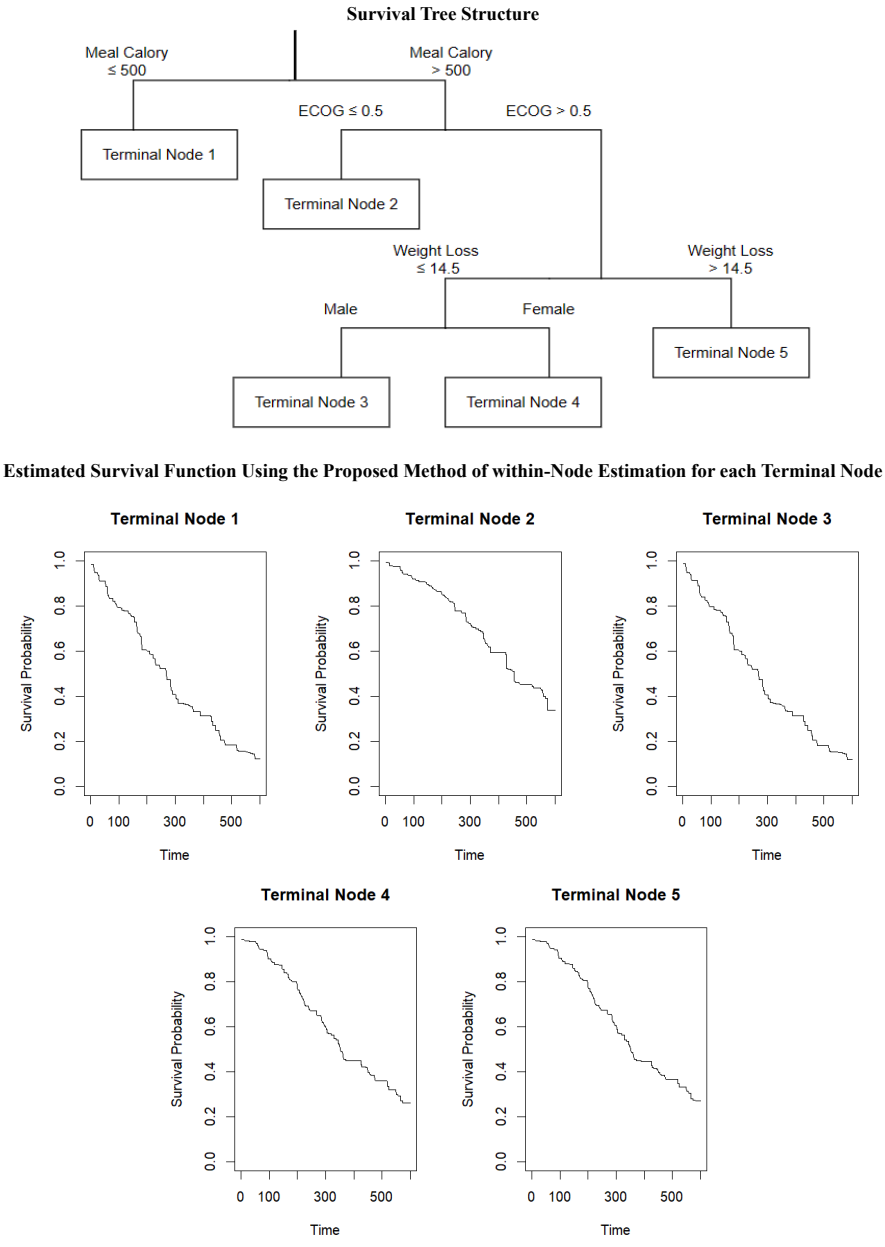
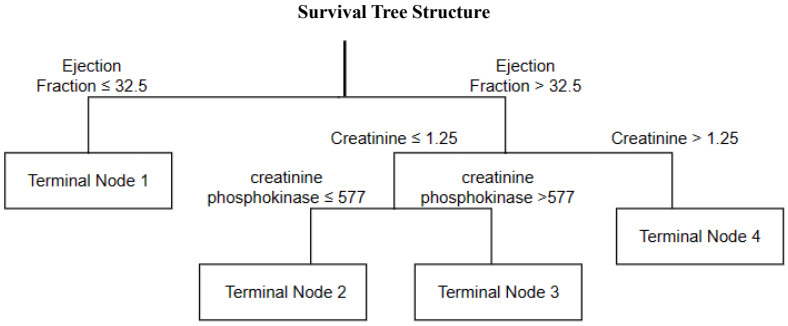


Figure 2. Analysis Results for North Central Cancer Treatment Group Lung Cancer Data



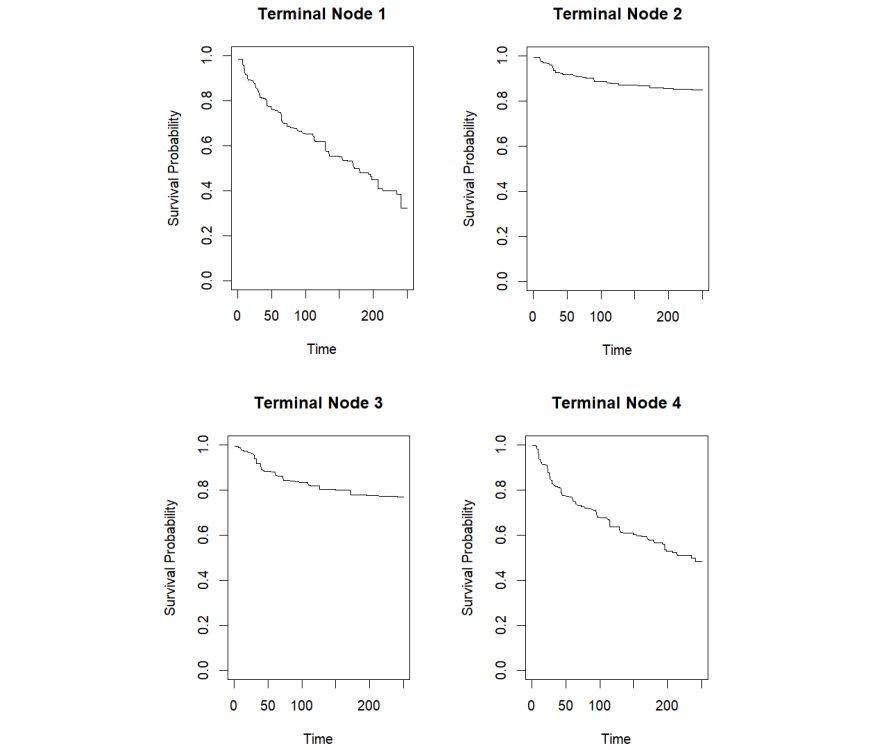


Figure 3. Analysis Results for Cardiovascular Medical Records from the Faisalabad Institute of Cardiology