Oblique Survival Trees in Discrete Event Time Analysis

Malgorzata Kretowska

***Abstract*—One of the main objectives of survival analysis is to predict the failure time that is usually considered as a continuous variable. In longitudinal studies, the data are often collected at every certain period of time, for example, monthly, quarterly, or yearly. Such data require appropriate techniques to handle the discrete time values that often have incomplete information about the failure occurrence—so-called "censored cases." Tree-based models are common, assumption-free methods of survival prediction. In this paper, the author proposes three recursive partitioning techniques able to cope with discrete-time censored survival data, which, in contrast to already-existing models limited to univariate trees, allow splits to have a form of any hyperplane. The performance of proposed methods, expressed as a mean absolute error, was examined on the basis of both synthetic and real data sets available in the literature and compared with existing tree-based models. To demonstrate the applicability of the methods in identifying subgroups of patients with a similar survival experience and to assess the influence of covariates on the risk of failure, a Veteran's Administration lung cancer data set was used. The results confirm proposed models to be good prediction tools for discrete-time survival data.**

***Index Terms*—Classification tree, piecewise-linear criterion function, oblique splits, survival analysis, discrete-time survival.**

I. Introduction

SURVIVAL analysis is a set of methods for predicting the time of failure occurrence. In the medical domain, the failure usually means death or disease relapse and is measured from a certain initial event, such as surgery or disease diagnosis. In clinical trials, not for all patients, the failures are observed. Often there is a large group of patients who did not experienced the event of interest and their observation finished due to other reasons. For such censored cases, the exact time of failure occurrence is unknown – we only know that it is greater than their follow-up time.

The survival time is usually treated as a continuous variable, and its distribution may be analyzed with the use of nonparametric, semiparametric or parametric survival models [1], [2]. Discrete-time methods are a less common branch of survival analysis, but there is a wide range of problems where their use is reasonable. They involve studies in which data are gathered every certain period of time, for example monthly, quarterly or yearly. Survival time here is divided into a number of intervals for which the survival probability or the hazard function may be calculated. As was pointed out by Tutz and Schmid [3], statistical methods that are designed for discrete event times have a number of advantages compared to continuous-time models. These include a more intuitive interpretation of hazards as conditional probabilities or no problems with ties.

Tree-based models belong to nonparametric methods of data analysis. They are an alternative to statistical models that often require quite restrictive assumptions. In case of survival data, single trees [4]–[11] or ensembles of trees [12]–[18] are mainly used for continuous survival time prediction. Applications of tree-based algorithms for discrete-time survival data analysis are rarer and can be divided into two groups: single and multiple time point models.

In a single time point approach, survival until a determined time point t_D (e.g., $t_D = 5$ years) is analyzed. The problem may be considered a binary classification task, in which the information about the patient's status (e.g., 1 - dead or 0 - alive) at t_D is a target. The problem with target values arises while considering cases censored before t_D since their status at t_D is unknown. Such patients are excluded from the learning process [19] or their target values are imputed [20], [21].

In multiple time point analysis, the time axis is divided into more than two intervals. To cope with such data, Yin and Anderson [22] extended the exponential tree model proposed by Davis and Anderson [23] and modified it to obtain a nonparametric tree, in which no assumption about the distribution of the survival time was needed. The method proposed by Bou–Hamad *et al.* [24] was addressed to survival data with a small number of observed time intervals. The tree was built around a discrete-time proportional odds model [25] with a splitting rule based on the maximum likelihood. The extension of the algorithm to time-varying covariates was done in [26]. It was obtained through replication of patient information for each time period where he/she was at risk. To improve the prediction performance, the authors also applied a bagging technique. A discrete-time survival tree as an extension of [24] was also proposed by Schmid *et al.* [27]. In the approach, a time-interval index was one of the variables of an augmented dataset, and, with the use of the CART algorithm with a Gini impurity measure [28], the tree was inducted to separate the observations

Manuscript received August 21, 2018; revised December 20, 2018 and January 31, 2019; accepted March 27, 2019. Date of publication April 1, 2019; date of current version January 6, 2020. This work was supported under Grant S/WI/2/2018 from the Bialystok University of Technology and by the Polish Ministry of Science and Higher Education.

The author is with the Faculty of Computer Science, Bialystok University of Technology, 15-351 Bialystok, Poland (e-mail: m.kretowska@pb.edu.pl).

Digital Object Identifier 10.1109/JBHI.2019.2908773

with different values of failure indicator (0 or 1) in each time period.

In this paper, I propose three tree-based models able to cope with discrete-time survival data. In contrast to the existing methods that are limited to univariate trees in which a single split tests the value of only one covariate, the presented approaches allow splits to have a form of any hyperplane that need not be parallel to the coordinate axis. An induction process of such oblique trees is based on minimization of convex piecewise-linear (CPL) criterion functions [29] that were previously applied to survival trees for censored [30] and competing risk [31] data. The proposed methods differ in input data, representation of the results and structure, but their induction procedures are similar and utilize the properties of CPL functions. The models are also compared with the Bou–Hamad tree [24] with a pruning based on Akaike or Bayesian information criteria available in the R package DSTree [32] and a tree proposed by Schmid *et al.* [27] with a BIC-based cardinality pruning. The methods' performance is expressed by a mean absolute error. Synthetic and real datasets are used to assess the model's performance, while results on a Veteran's Administration lung cancer dataset [33] present the capabilities of the proposed approaches.

This paper consists of seven sections. Section II introduces a definition and basic concepts of discrete-time survival data. In Section III, the idea of dipolar criterion function is recalled, while a description of the proposed tree-based models for discrete time event data is presented in Section IV. A validation measure is presented in Section V, and Section VI shows the results of the experiments on simulated and real datasets. Section VII discusses and concludes the results.

II. Discrete-Time Survival Data

In discrete-time data, the survival time is divided into K distinct time intervals $I_1, I_2, \ldots, I_K$ where $I_1 = (0, t_1], I_2 = (t_1, t_2], \ldots, I_K = (t_{K-1}, \infty)$ and $0 < t_1 < t_2 < \ldots < t_{K-1} < \infty$. The learning set, LS, is defined as $LS = (\mathbf{x}_i, k_i, \delta_i)$, $i = 1, 2, \ldots, M$, where $\mathbf{x}_i = [x_{i1}, x_{i2}, \ldots, x_{iN}]^T$ is N-dimensional covariate vector, $k_i \in \{1, 2, \ldots K\}$ is an index of the time interval (I_{k_i}) in which the failure occurred or the observation was lost to follow-up, and $\delta_i \in \{0, 1\}$ is a failure indicator that is equal to 1 for uncensored cases and to 0 otherwise.

The distribution of the random variable T representing the survival time may be described by the discrete-time hazard, defined as the conditional failure probability

$$\lambda_k = \lambda(k) = P(T \in I_k | T \geq t_{k-1}), \tag{1}$$

the survival function

$$S_k = S(k) = P(T > t_k) = \prod_{i=1}^{k} (1 - \lambda_i), \tag{2}$$

where $S_0 = 1$, and the probability function

$$f_k = f(k) = P(T \in I_k) = S_{k-1} - S_k = \lambda_k \prod_{i=1}^{k-1} (1 - \lambda_i). \tag{3}$$

The likelihood function calculated for survival data is as follows:

$$L = \prod_{i=1}^{M} f_{ik_i}^{\delta_i} S_{ik_i}^{(1-\delta_i)}, \tag{4}$$

where k_i denotes the last time interval in which the ith subject was observed. For uncensored observations, for which the failure occurred in the interval I_{k_i}, the contribution to the likelihood function is $f_{ik_i} = f(k_i|\mathbf{x}_i)$; the contribution of censored observations is $S_{ik_i} = S(k_i|\mathbf{x}_i)$, since the failure time is unknown for them.

If we replicate the information about each subject ($i = 1, \ldots, M$) for each time interval in which he/she was observed ($k = 1, \ldots, k_i$) and for each newly-created observation we add a failure indicator d_{ik} equal to 1 for the last time interval I_{k_i} for uncensored patients and 0 otherwise, the likelihood function (4) takes the following form

$$L = \prod_{i=1}^{M} \prod_{k=1}^{k_i} \lambda_{ik}^{d_{ik}} (1 - \lambda_{ik})^{1-d_{ik}} \tag{5}$$

where $\lambda_{ik} = \lambda(k|\mathbf{x}_i)$. The maximum-likelihood estimator of the hazard function λ_{ik} is $\hat{\lambda}_{ik} = d_{ik}$.

Assuming a homogeneous population, equation (5) may be presented in the binomial form [34]:

$$L = \prod_{k=1}^{K} \binom{m_k}{m_{k1}} \lambda_k^{m_{k1}} (1 - \lambda_k)^{m_k - m_{k1}} \tag{6}$$

where m_k and m_{k1} are the number of subjects at risk and the number of events in the kth time interval. The maximum-likelihood estimator of the hazard function λ_k is

$$h_k = \hat{\lambda}_k = \frac{m_{k1}}{m_k} \tag{7}$$

and, using the equations (2) and (3), we obtain the estimators of S_k and f_k denoted as $\hat{S}_k$ and $\hat{f}_k$. Therefore, the observed log-likelihood calculated for equation (4) with homogeneity assumption is given by

$$ll = \sum_{k=1}^{K} (m_{k1} \ln \hat{f}_k + m_{k0} \ln \hat{S}_k) \tag{8}$$

where m_{k0} is a number of censored subjects in kth time interval (i.e., subjects for whom $\delta_i = 0$ and $k_i = k$).

III. Dipolar Tree

A dipolar tree is a binary oblique tree that divides the feature space into disjoint areas represented by terminal nodes. The splits defined in internal nodes take a form of any hyperplane $H(\mathbf{w}, \theta) = \{\mathbf{x} : \mathbf{w}^T \mathbf{x} = \theta\}$ [30] obtained by minimization of a convex piecewise-linear criterion function, the so-called dipolar criterion [29].

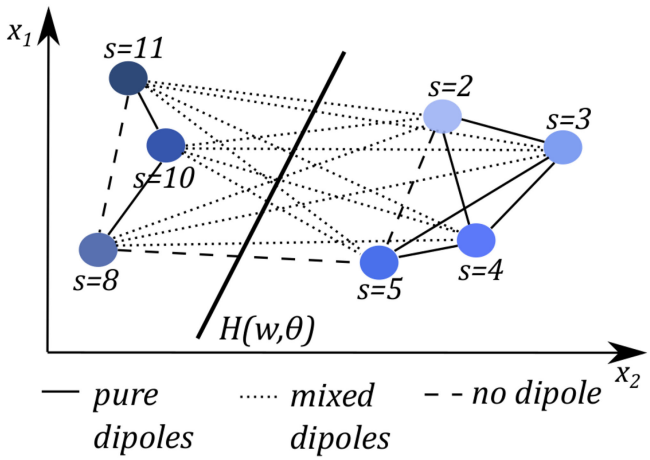


Fig. 1. A splitting hyperplane, $H(\mathbf{w}, \theta)$, that separates points with different color intensity represented by s values. A pair of points (feature vectors) creates: a mixed dipole, if the absolute value of the difference between their s values is greater than 3; a pure dipole, if the absolute value of the difference between their s values is less than 3; no dipole, if the absolute value of the difference between their s values is equal to 3.

A classical top-down tree induction process starts from a root node, in which the dipolar criterion is calculated for the whole learning set and then minimized in order to estimate the parameters of the splitting hyperplane (i.e., $\mathbf{w}$ and θ). It divides the learning set into two subsets, thereby generating two child-nodes. The observations situated on the positive side of the hyperplane reach the left child-node of the root, while those on the negative side reach the right one. The splits in newly-created nodes are obtained by minimization of the dipolar criterion function calculated on the basis of the data that reach the node. Such a procedure is repeated until a stopping condition is fulfilled. Then the node becomes a leaf and does not split the data, but represents the obtained feature space area. The representation depends on the analyzed problem – in classification, it is a decision class; in regression, it is usually the mean value of a predicted variable; and, in survival analysis, it is the Kaplan–Meier survival function.

One of the basic stages of a tree induction algorithm is to decide how to divide the feature space in order to receive the best outcome from the point of view of the criterion used. Applied here, the dipolar criterion function is based on dipoles. They are pairs of feature vectors – $(\mathbf{z}_j, \mathbf{z}_{j'})$, $j \neq j'$, $j = 1, 2, \ldots, M$ and $j' = 1, 2, \ldots, M$ – which can be either mixed or pure. If two vectors are different from the point of view of a given problem (e.g., belong to two different classes), they should be separated. For this purpose, we create a mixed dipole between them. Pure dipoles are formed between similar observations that should not be separated (e.g., belong to one class). If we cannot decide whether two vectors are similar or not, we do not create any dipole between them. Having a set of mixed and pure dipoles, we are interested in calculating a hyperplane that divides as many mixed dipoles as possible and avoids splitting the pure ones (Figure 1).

For this purpose, for any augmented, feature vector $\mathbf{z}_j = [1, x_{j1}, x_{j2}, \ldots, x_{jN}]^T$, $j = 1, 2, \ldots, M$ from the learning set LS, we define two types of CPL penalty functions:

$$\varphi_j^+(\mathbf{v}) = \begin{cases} \gamma_j - \mathbf{v}^T \mathbf{z}_j & \text{if } \mathbf{v}^T \mathbf{z}_j \leq \gamma_j \\ 0 & \text{if } \mathbf{v}^T \mathbf{z}_j > \gamma_j \end{cases} \tag{9}$$

and

$$\varphi_j^-(\mathbf{v}) = \begin{cases} \gamma_j + \mathbf{v}^T \mathbf{z}_j & \text{if } \mathbf{v}^T \mathbf{z}_j \geq -\gamma_j \\ 0 & \text{if } \mathbf{v}^T \mathbf{z}_j < -\gamma_j \end{cases} \tag{10}$$

where $\mathbf{v} = [-\theta, w_1, w_2, \ldots, w_N]^T$ is an augmented weight vector, and $\gamma_j \geq 0$ is a margin usually equal to 1.

With each mixed dipole, we associate the sum of two different penalty functions, denoted by $\varphi_{jj'}^m$ ($\varphi_{jj'}^m(\mathbf{v}) = \varphi_j^+(\mathbf{v}) + \varphi_{j'}^-(\mathbf{v})$ or $\varphi_{jj'}^m(\mathbf{v}) = \varphi_j^-(\mathbf{v}) + \varphi_{j'}^+(\mathbf{v})$). Its minimization causes the feature vectors $\mathbf{z}_j$ and $\mathbf{z}_{j'}$ to be separated by the hyperplane $H(\mathbf{v})$. On the other hand, the sum of two penalty functions of the same type associated with the pure dipoles denoted by $\varphi_{jj'}^p$ ($\varphi_{jj'}^p(\mathbf{v}) = \varphi_j^+(\mathbf{v}) + \varphi_{j'}^+(\mathbf{v})$ or $\varphi_{jj'}^p(\mathbf{v}) = \varphi_j^-(\mathbf{v}) + \varphi_{j'}^-(\mathbf{v})$) prevents them from being divided. The dipolar criterion function is the sum of penalty functions over the whole set of dipoles and takes the following simplified form:

$$\Psi_d(\mathbf{v}) = \sum_{(j,j') \in I^p} \alpha_{jj'} \varphi_{jj'}^p(\mathbf{v}) + \sum_{(j,j') \in I^m} \alpha_{jj'} \varphi_{jj'}^m(\mathbf{v}) \tag{11}$$

where $\alpha_{jj'}$ is a price (importance) of the dipole $(\mathbf{z}_j, \mathbf{z}_{j'})$, p and m stand for pure and mixed dipoles, respectively, and I^p and I^m are sets of pure and mixed dipoles. In my previous experiments, the value of $\alpha_{jj'}$ was constant and set to 1 for pure dipoles and to 100 for mixed dipoles. Such a difference forces the algorithm to divide mixed dipoles, even when the penalty functions associated with pure dipoles are large compared to the mixed dipoles. The dipolar criterion is calculated for each internal node on the basis of the data that reach the node and minimized with the use of basis exchange algorithm [35]. As a result, we obtain the parameters of the splitting hyperplane $H(\mathbf{v})$. A node becomes a leaf when it contains no more than a determined number of observations (e.g., 10), or when we cannot create any mixed dipoles for it. A general flowchart of dipolar tree induction procedure is shown in Figure 2. A more detailed description of the dipolar criterion function is given in [30].

IV. Discrete-Time Survival Tree Models

Dipolar trees were successfully used for different types of survival data. Kretowska applied them to right-censored survival data [30] and to analyze the data with competing risks [31]. In this paper, I propose three new approaches to deal with discrete-time survival data.

A. Discrete Survival Tree

A discrete survival tree (*DST*) aims to divide the feature space into regions that represent individual time intervals. Observations with the failure occurring in the same time period should reach the same terminal nodes. On the other hand, the observations with failures occurring in different time intervals should reach different terminal nodes, i.e., they should be separated. This leads to the following dipoles formation rules:

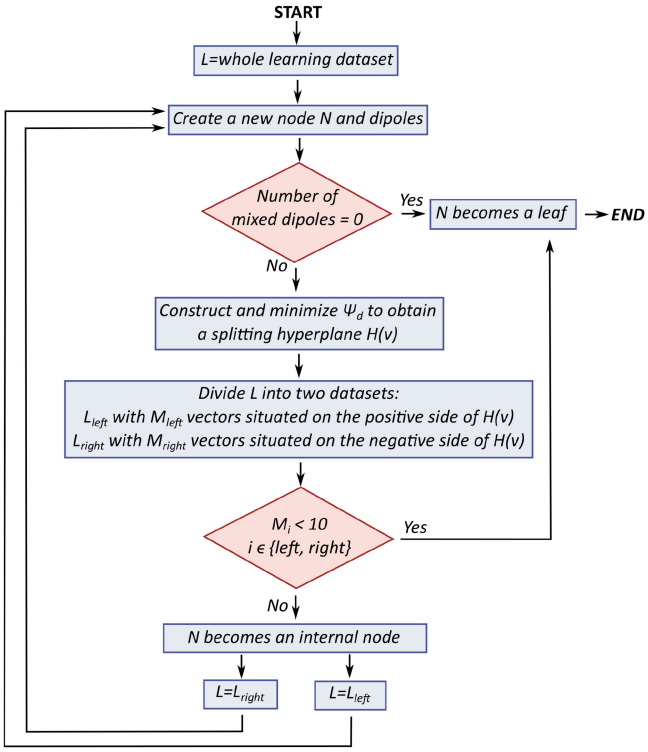


Fig. 2. A flowchart of dipolar tree induction procedure.

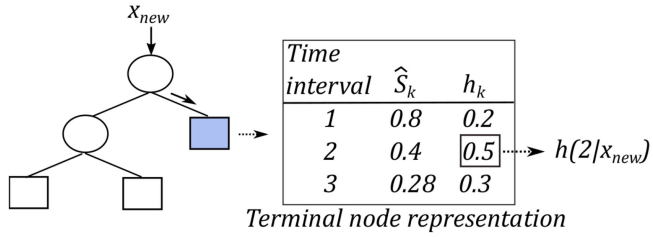


Fig. 3. Discrete survival tree, *DST*.

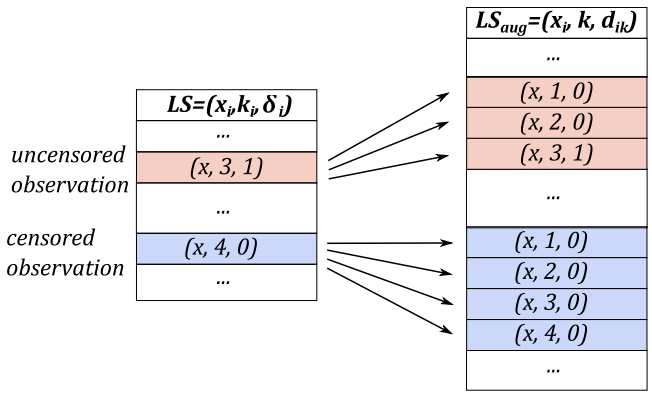


Fig. 4. Augmented learning dataset, LS_{aug}.

1) a pair of feature vectors $(\mathbf{x}_i, \mathbf{x}_j)$ forms the pure dipole, if $\delta_i = \delta_j = 1$ and $k_i = k_j$
2) a pair of feature vectors $(\mathbf{x}_i, \mathbf{x}_j)$ forms the mixed dipole, if
   a) $\delta_i = \delta_j = 1$ and $k_i \neq k_j$
   b) $\delta_i = 0, \delta_j = 1$ and $k_i \geq k_j$

As we can see, the censored observations are used only in the mixed dipoles construction. If we assume that for a censored subject i that was lost to follow-up in the time interval I_{k_i}, the failure may occur in the time period I_k, $k > k_i$, the subject can create mixed dipoles only with these uncensored cases for which the failure occurred in previous intervals $I_1, \ldots, I_{k_i}$. Only then, we are sure that the failure of censored and uncensored observation will be in different time periods.

Although the target value is the time interval index (k_i), as a result we obtain a tree, in which each terminal node contains the estimated hazard rate and the survival probability for all time intervals [24]. Hence, for a new feature vector $\mathbf{x}_{new}$, the tree predicts the risk of failure over all considered periods (Figure 3). The time interval with the highest hazard value is considered the most probable period of failure.

B. Classification Tree

A classification tree (*CT*) requires an additional initial phase of data preparation. Each observation, i, from the learning set, LS, should be replicated for each time interval, I_k, in which he/she was observed [27]. Hence, we receive a new augmented learning set $LS_{aug} = (\mathbf{x}_i, k, d_{ik}) = (\mathbf{y}_{ik}, d_{ik})$, where $i = 1, \ldots, M$, $k = 1, \ldots, k_i$, $d_{ik} \in \{0, 1\}$ is a failure indicator that is equal to 1 only for the observed last time interval of uncensored subjects and to 0 otherwise, and $\mathbf{y}_{ik} = [x_{i1}, x_{i2}, \ldots, x_{iN}, k]^T$ (Figure 4).

For each observation from the augmented dataset, we know the value of d_{ik} that informs us about patient status in a given time interval. It may be also considered as an estimator of the hazard function λ_{ik} (5). Our aim is, therefore, to separate the observations with different values of d_{ik} (a target value). Hence, the problem becomes a binary classification task and a tree is inducted to divide the input space into areas with similar values of failure indicators, 1 or 0. This is obtained by the minimization of the dipolar criterion function calculated on the basis of dipoles formed according to the following rules:

1) a pair of input vectors $(\mathbf{y}_{ik}, \mathbf{y}_{jk'})$ forms the pure dipole, if $d_{ik} = d_{jk'}$.
2) a pair of input vectors $(\mathbf{y}_{ik}, \mathbf{y}_{jk'})$ forms the mixed dipole, if $d_{ik} \neq d_{jk'}$.

Although the target is a binary variable, terminal nodes can be characterised by the hazard rate, h_k, calculated as the ratio of the number of failures ($d_{ik} = 1$) and the number of all observations that reach the node.

To calculate the conditional probabilities of failure for a new patient $\mathbf{x}_{new}$, for each time interval ($k = 1, \ldots, K$) we should create an input vector $[\mathbf{x}_{new}^T, k]^T$. Dropping the vectors down the classification tree, we obtain the hazards for corresponding time intervals (Figure 5).

C. Modular Tree

A modular tree (*MT*) is a set of K trees T_k, $k = 1, \ldots, K$; each single tree T_k aims at predicting the risk of failure in the kth

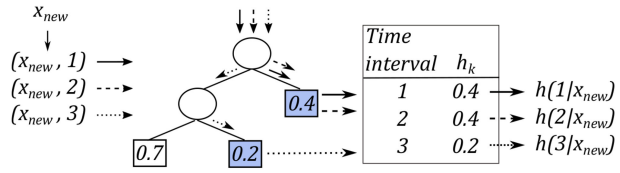


Fig. 5. Classification tree, *CT*.

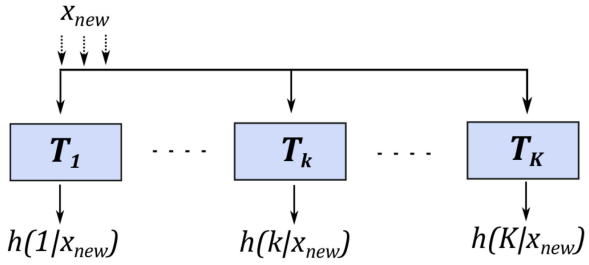


Fig. 6. Modular tree, *MT*.

time interval. A tree $T_{k'}$ is inducted on the basis of observations that are at risk in the time period $I_{k'}$. It means all the cases from the augmented learning set LS_{aug} with $k = k'$. They create a new training dataset $LS^{k'}_{aug} = (\mathbf{x}_i, k', d_{ik'}), i \in M_{k'}$, where $M_{k'}$ denotes a set of indexes of feature vectors being at risk in the time period k'.

Considering the induction process of the single tree T_k, we are interested in such a division of the feature space that will separate observations with different values of d_{ik}, i.e., the patients who experienced a failure from those who are still alive in the kth time interval. Hence, to create the tree T_k, we use the learning set LS^k_{aug} with the following definition of pure and mixed dipoles:

1) a pair of input vectors $(\mathbf{x}_i, \mathbf{x}_j)$ forms the pure dipole, if $d_{ik} = d_{jk}$.
2) a pair of input vectors $(\mathbf{x}_i, \mathbf{x}_j)$ forms the mixed dipole, if $d_{ik} \neq d_{jk}$.

For each terminal node of the kth tree, we calculate the conditional probability of failure in the kth time interval, h_k. Having a new patient with the covariate vector $\mathbf{x}_{new}$, we obtain the hazards $h(1|\mathbf{x}_{new}), h(2|\mathbf{x}_{new}), \ldots, h(K|\mathbf{x}_{new})$ as the outcomes of all the composite trees T_k (Figure 6).

The induction process of the oblique tree-based models is similar and is based on the methodology described in Section III. The stopping criteria defined for *DST*, *CT*, and all the single trees T_k of MT cover two conditions. The node becomes a leaf when 1) we cannot form any mixed dipole for it or 2) it contains no more than 10 observations.

The comparison of discrete-time dipolar survival trees is given in Table I.

D. Pruning

The discrete-time survival trees are usually deep and have a poor generalization. To improve the results, an additional pruning phase is often applied. It removes irrelevant tree branches and replaces them with terminal nodes. One of the most common pruning techniques is a cost-complexity method [28] that was extended by LeBlanc and Crowley [36] as a split-complexity algorithm applied to survival data.

For a given tree T, the cost-complexity technique minimizes a total cost

$$C_\alpha(T) = R(T) + \alpha|\tilde{T}| \tag{12}$$

where α is a nonnegative complexity parameter, $|\tilde{T}|$ is the number of terminal nodes, and $R(T)$ is a resubstitution error calculated as a mean squared error (MSE) between target values d_{ik} and the estimated hazard ratios h_{ik}.

The choice of the tree that minimizes (12) is strictly related to the value of α. Although α is a continues variable, the number of subtrees of T is finite. Therefore, only one subtree is the best for the whole interval of α values. The first phase of the algorithm is then to create the sequence of subtrees associated with increasing values of α,

$$\begin{array}{ccccccc} T_0 & \succ & T_1 & \succ & \ldots & \succ & T_m \\ 0 & < & \alpha_1 & < & \ldots & < & \infty \end{array} \tag{13}$$

where T_0 is an unpruned tree, T_m is the root node, and $T_i \succ T_{i+1}$ means that T_{i+1} is a subtree of T_i that is the best subtree for $\alpha \in [\alpha_i, \alpha_{i+1})$. The choice of a subtree with the best generalization is narrowed to the set $(T_0, T_1, \ldots, T_m)$ and is conducted with the use of the 10-fold cross-validation technique.

The split-complexity algorithm applied to DST maximizes a split-complexity measure defined as

$$G_\alpha(T) = G(T) - \alpha|T| \tag{14}$$

where $|T|$ is the number of internal nodes of T and $G(T)$ is the sum of the splitting statistics calculated over all internal nodes of T. For *DST*, $G(T)$ is the sum of the likelihood ratio statistics given as

$$G(T) = \sum_{t \in In(T)} G(t) = \sum_{t \in In(T)} -2(ll_t - (ll_{t_l} + ll_{t_r})) \tag{15}$$

where $In(T)$ is a set of internal nodes of T and t_l, t_r denote the left and the right child node of t, respectively, and ll_t is the observed log-likelihood (8) calculated for the node t.

Although in contrast to the former algorithm, the split-complexity method maximizes (14), the main stages of the procedure are similar to those of the cost-complexity algorithm. A difference is an additional penalty term α_c that should be taken into account during the final tree selection, in which the quality of each subtree from the sequence (13) is calculated with the use of the split-complexity measure (14) with $\alpha = \alpha_c$. In this paper, I use $\alpha_c = 2$, according to suggestions made by LeBlanc and Crowley [36].

In this paper, the split-complexity method is used to prune the *DST*, while the *CT* and the component trees of the *MT* use the cost-complexity algorithm.

V. Model Validation

The discrete-time dipolar survival tree models aim at a proper estimation of the hazard rates or the survival functions for all analyzed time intervals. While assessing the predictive ability of the method, for a given observation we should take into account

TABLE I
COMPARISON OF DISCRETE-TIME DIPOLAR SURVIVAL TREES

	DST – Discrete survival tree	*CT* – Classification tree	*MT* – Modular tree
Structure	Binary tree	Binary tree	Set of K binary trees T_k
Learning data	$LS = (\mathbf{x}_i, k_i, \delta_i)$, $i = 1, \ldots, M$	$LS_{aug} = (\mathbf{x}_i, k, d_{ik})$, $i = 1, \ldots, M$, $k = 1, \ldots, k_i$	for T_k: $LS^k_{aug} = (\mathbf{x}_i, k, d_{ik})$, $i \in M_k$
Input data	$\mathbf{x}_i$	$\mathbf{y}_{ik} = [x_{i1}, x_{i2}, \ldots, x_{iN}, k]^T$	$\mathbf{x}_i$
Target value	k_i	d_{ik}	for T_k: d_{ik}
Leaf representation	$\hat{S}_k$, h_k, $k = 1, \ldots, K$	h_k	for T_k: h_k
Pruning method	Split-complexity method	Cost-complexity method	Cost-complexity method

the estimated and the true predicted value over all time intervals in which such a true value can be calculated. Thus, for synthetic data, in which a true model is known, we can obtain the mean absolute error (MAE) as

$$MAE = \frac{1}{M}\sum_{i=1}^{M}\frac{1}{K}\sum_{k=1}^{K}|S_{ik} - \hat{S}_{ik}| \tag{16}$$

where $S_{ik} = S(k|\mathbf{x}_i)$ and $\hat{S}_{ik} = \hat{S}(k|\mathbf{x}_i)$ are the true and the estimated survival function for the kth interval given $\mathbf{x}_i$, respectively.

In case of real datasets, the true values of survival functions are unknown. The MAE can be based on the hazard function, since in minimizing the equation (5) we obtain $\hat{\lambda}_{ik} = d_{ik}$. For a given feature vector $\mathbf{x}_i$, d_{ik} is known only for these time intervals in which the patient was at risk $(I_1, \ldots, I_{k_i})$. Hence, the MAE has the following form:

$$MAE = \frac{1}{M}\sum_{i=1}^{M}\frac{1}{k_i}\sum_{k=1}^{k_i}|d_{ik} - h_{ik}| \tag{17}$$

where $h_{ik} = h(k|\mathbf{x}_i)$ is the estimated hazard function for the kth interval given $\mathbf{x}_i$.

To estimate the performance measure, a bootstrap cross-validation procedure [37] was applied. In this approach, in each of B runs of the procedure, the training set $L^b_{tr}, b = 1, 2, \ldots, B$ is drawn with replacement from the original learning set, LS. The test set L^b_{ts} contains those feature vectors that are not sampled in L^b_{tr}. In the bth run of the experiment, we train a prediction model with L^b_{tr} and test it using the bth test set, L^b_{ts}. The performance measure estimate is calculated as a mean value over B iterations:

$$MAE = \frac{1}{B}\sum_{b=1}^{B}\frac{1}{M_b}\sum_{i\in In(L^b_{ts})}\frac{1}{k_i}\sum_{k=1}^{k_i}|d_{ik} - h_{ik}| \tag{18}$$

where M_b is the number of observations in the bth test set, L^b_{ts}, and $In(L^b_{ts})$ is a set of indexes of observations belonging to L^b_{ts}.

VI. EXPERIMENTAL RESULTS

The results are divided into three parts. The first part is based on the experiments conducted for synthetic datasets to show the models' performance when true survival function is known. The second part contains the comparison of the performance of the new methods with the existing techniques and is based on several real datasets available in the literature. In the third part, the Veteran's Administration lung cancer dataset is analyzed to present the capabilities of the proposed methods.

The proposed models, *DST*, *CT*, and *MT* were compared with the method proposed by Bou–Hamad *et al.* [24] implemented in R package DSTree [32] with the pruning strategies based on Akaike (AIC) or Bayesian (BIC) information criterion and with the tree proposed by Schmid *et al.* [27] inducted with the use of R package rpart [38] with the BIC-based cardinality pruning. The model performance (the mean absolute error) was calculated with the use of the bootstrap cross-validation with $B = 100$. To evaluate a statistical significance of the results, I used the Friedman test and the Dunn's multiple comparison test. A 0.05 significance level was applied for all comparisons.

The dipolar criterion function contains several parameters, which may be established separately for each dataset to obtain better prognosis. In my experiments, I set them constant. The margins $\gamma_j, j = 1, 2, \ldots, M$ (equations (9) and (10)) were equal to 1 and the dipole prices $\alpha_{jj'}$ (equation (11)) were equal to 1 and 100 for the pure and the mixed dipoles, respectively.

A. Synthetic Data

The synthetic datasets were generated with the use of R package survsim [39]. Each training dataset (SDvar2, SDvar5, SDvar10, SDvar20, and SDvar50) contained 400 cases with 30.2% of censored observations. Two variables, x_1 and x_2, were drawn from a uniform distribution on [1,5], and the failure, as well as the censoring time, followed a Weibull distribution with the survival function $S(t|x_{j1}, x_{j2}) = \exp(-\lambda_j t^p)$, where $\lambda_j = \exp(-p(\beta_0 + \beta_1 x_{j1} + \beta_2 x_{j2}))$. In the simulation process, the parameters for the failure time distribution were fixed as: $p = 1.1$, $\beta_0 = 1.0$, $\beta_1 = -0.6$, and $\beta_2 = 0.9$, and, for censoring, $p = 1.3$, $\beta_0 = 2.6$, $\beta_1 = 0$, and $\beta_2 = 0$. The maximum follow-up time was equal to 12 years. Additionally, in synthetic data SDvar5 – 3, SDvar10 – 8, SDvar20 – 18, and SDvar50 – 48, independent non-informative predictors were generated from the uniform distribution on the interval [1,2]. Test datasets contained 2,000 observations.

The experiments were conducted for five synthetic datasets with survival time divided into 4 ($K = 4$) or 6 ($K = 6$) equal time intervals. The results are presented as mean and standard deviation of MAE calculated over 100 runs of a procedure, in which tree-based models are inducted on the basis of a bootstrap samples drawing with replacement from the training data.

In Table II we can see the results obtained for five synthetic datasets with two different time divisions. For SDvar2 through SDvar20 for $K = 4$ and $K = 6$, the best results were obtained

TABLE II
MEAN ABSOLUTE ERRORS OBTAINED FOR SYNTHETIC DATA SETS (SDVAR2, SDVAR5, SDVAR10, SDVAR20, SDVAR50) WITH FOUR ($K = 4$) AND SIX ($K = 6$) TIME INTERVALS OF FIVE MODELS: *BHTree* – BOU–HAMAD TREE WITH PRUNING BASED ON AIC OR BIC, *SchTree* – SCHMID TREE; *DST* – DISCRETE SURVIVAL TREE, *CT* – CLASSIFICATION TREE, AND *MT* – MODULAR TREE. MEAN ABSOLUTE ERRORS (MAE) ARE PRESENTED AS MEAN (STANDARD DEVIATION) OVER 100 RUNS OF EXPERIMENT MULTIPLIED BY 100. SIZE IS CALCULATED AS THE MEDIAN NUMBER OF NODES, FOR *MT* – THE MEDIAN NUMBER OF NODES FOR ALL K COMPONENT TREES

Data set	*BHTree* AIC		*BHTree* BIC		*SchTree*		*DST*		*CT*		*MT*	
	MAE	Size	MAE	Size	MAE	Size	MAE	Size	MAE	Size	MAE	Size
$K = 4$												
SDvar2	12.8 (1.1)	29	11.8 (1.0)	11	11.0 (1.0)	19	10.2 (1.7)	5	7.6 (0.3)	39	**7.1** (0.4)	58
SDvar5	15.0 (1.3)	33	12.8 (1.3)	11	12.2 (1.2)	24	12.4 (1.4)	3	11.1 (0.6)	27	**9.5** (0.3)	42
SDvar10	16.5 (1.5)	35	13.3 (1.5)	11	13.2 (1.5)	26	13.6 (0.8)	3	**12.2** (0.4)	21	**12.2** (0.8)	29
SDvar20	17.8 (1.5)	37	14.1 (1.5)	13	14.7 (1.7)	31	15.5 (2.5)	3	**11.5** (2.1)	9	13.9 (0.1)	21
SDvar50	19.5 (1.8)	39	**15.3** (1.9)	15	17.0 (1.8)	35	23.1 (4.6)	5	20.5 (3.3)	17	17.5 (0.4)	12
$K = 6$												
SDvar2	12.0 (1.0)	23	12.7 (1.3)	9	11.2 (0.9)	19	9.3 (1.4)	5	**6.8** (0.7)	35	7.1 (0.3)	62
SDvar5	13.9 (1.4)	27	13.0 (1.3)	9	12.2 (1.2)	23	11.9 (1.4)	5	9.5 (0.9)	25	**9.4** (0.4)	50
SDvar10	15.6 (1.4)	31	13.1 (1.3)	9	13.1 (1.3)	27	13.5 (0.7)	3	**11.4** (1.3)	27	11.8 (0.6)	36
SDvar20	17.0 (1.6)	33	13.5 (1.4)	9	14.4 (1.4)	32	15.0 (2.1)	3	**13.2** (2.8)	5	13.7 (0.5)	23
SDvar50	18.9 (1.4)	35	**14.1** (1.5)	9	16.1 (1.7)	34	23.2 (4.5)	5	16.1 (4.5)	5	22.6 (0.2)	18

for *CT* and *MT*. In the case of SDvar50, the best performance was for the Bou–Hamad tree with BIC-based pruning.

Additional non-informational variables seem to usually impair the performance of the models; the more variables, the larger the MAE. The only exception is the CT result for SDvar20 and $K = 4$, where $MAE = 11.5$ is less than $MAE = 12.2$ for SDvar10. You can also see the effect of additional variables on the size of *SchTree*, $BHTree$ AIC, and MT. Increasing the number of variables causes an increase in the size of *SchTree* and *BHTree* AIC and a decrease in the size of *MT*. The effect of K on the MAE is ambiguous; it varies depending on the method and the number of variables. Increasing the number of time intervals affects the size of *MT*. This dependency is related to the way the *MT* is built and is not observed for other tree models presented in Table II.

The experiments were performed on laptop with Intel Core i7-5500U CPU, 2.40GHz. *BHTree* BIC was the fastest algorithm with the time of tree induction for $K = 4$ ($K = 6$) from a range $0.4 - 3.3$ ($0.5 - 3.6$) seconds for SDvar2 – SDvar50, respectively. For SDvar2 – SDvar10, the Schmid tree was the slowest method with induction times of $16 - 20$ seconds for $K = 4$ and 23 seconds for $K = 6$, for SDvar20, the *CT* had the longest induction times equal to 20 and 31 seconds for $K = 4$ and $K = 6$, respectively, and for SDvar50, the *MT* was the slowest algorithm with 67 and 73 seconds for $K = 4$ and $K = 6$, respectively.

Figure 7 shows the influence of x_1 and x_2 on the survival function for four distinguished time intervals (I_1, I_2, I_3, I_4) calculated by the true model, *DST*, *CT*, and *MT* inducted on the basis of synthetic data SDvar2 ($K = 4$). Please note that the shape of survival functions obtained by *CT* and *MT* are more similar to the true model than those obtained by *DST*.

B. Real Datasets

The comparison of the proposed methods against the Bou–Hamad and Schmid trees was conducted with the use of twelve real datasets from the literature (Table III). In order to obtain discrete-time survival data, the publicly available continuous-time datasets were discretized. For competing risk data, an event of interest (death in pbc, relapse in follic dataset) or overall survival (prostate) was taken into account. The analyzed datasets differ in terms of the observation number (90–2843), the attribute number (2–13), and the percentage of censored cases (3.8–72.2). Since the results of synthetic datasets showed that the influence of K for the model performance exists, but is difficult to predict, I decided to choose a varied number of time periods (3–6).

Table IV shows the results of five tree-based models obtained for real datasets. As we can see, the performance of *SchTree* is better for the four datasets (follic, GBSG2, larynx, and melanoma). For other datasets, the mean absolute error is lower for oblique dipolar trees. Among them, the *CT* and the *MT* performance is better than the *DST*. Statistical analysis shows that the models' performance differs significantly ($p = 0.006$), and statistically significant differences exist between *BHTree* BIC and *MT*. In Figure 8 we can see the differences between the mean absolute error of *SchTree* and *BHTree* with pruning based on AIC or BIC and the error of *DST*, *CT*, and *MT*. The median value (horizontal line) of differences between MAE of *BHTree* BIC and *MT*, equal to 4.75, is the most distant from 0 in comparison with other median values. Since the medians are usually greater than 0, the mean absolute errors of the proposed methods are lower than the MAE of existing algorithms. The only exception is *SchTree* while compared to the results of *DST*.

The tree sizes reported in Table IV give us only the rough information about the problem since only the median values are presented. We can see that *BHTree* BIC and *DST* have the smallest number of nodes, while the structures of *MT* and *BHTree* AIC are the most complex.

C. Veteran's Administration Lung Cancer

The Veteran's Administration lung cancer dataset (valung) contains the data from a trial in which male patients with advanced inoperable tumors were randomized to either standard (69 subjects) or test chemotherapy (68 subjects) – Trt (0 and 1, respectively). Only 9 subjects out of 137 were censored. Information on cell type – $CellType$ (0 – squamous (35 patients), 1 – small cell (48), 2 – adeno (27), 3 - large (27)), a

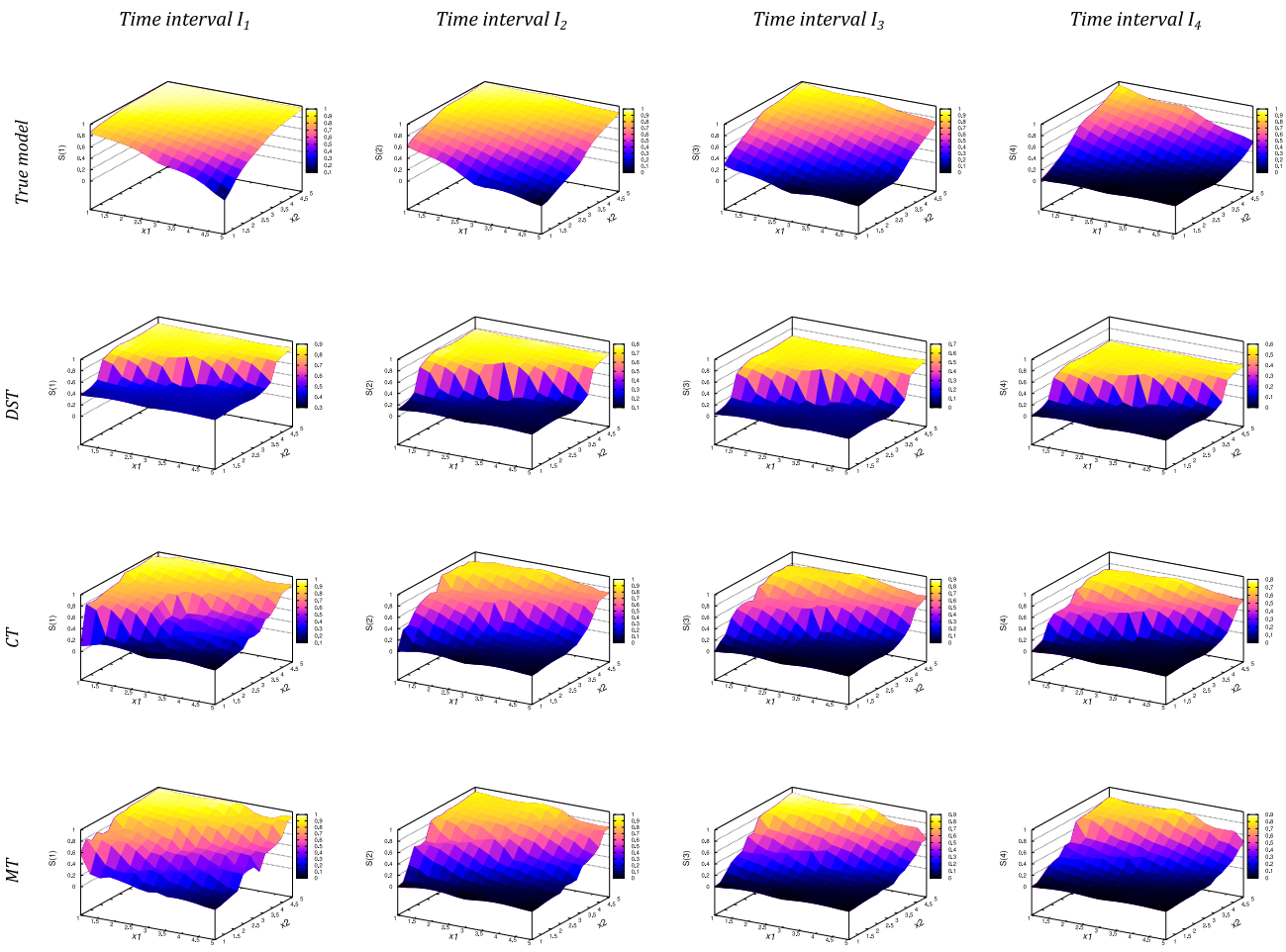


Fig. 7. The influence of x_1 and x_2 on the survival function calculated for four time intervals (I_1, I_2, I_3, I_4) by the true model, discrete survival tree *DST*, classification tree *CT*, and modular tree *MT* inducted on the basis of synthetic data SDvar2 with $K = 4$.

TABLE III
DESCRIPTION OF DATA SETS; #OBS DENOTES THE NUMBER OF OBSERVATIONS, #ATTR - THE NUMBER OF ATTRIBUTES, K - THE NUMBER OF TIME INTERVALS, %CENS - PERCENT OF CENSORED CASES

Data set	#obs	#attr	K	%cens	Description
aids2 [40]	2843	7	6	38.1	AIDS in Australia
bfeed [1], [41]	927	8	5	3.8	weaning of breast-fed newborns
cml [42]	507	7	5	21.3	chronic myelogenous leukemia (simulated)
cost [37], [43]	518	13	6	22	Copenhagen stroke study
follic [44]	541	4	4	49.7	follicular cell lymphoma
GBSG2 [45], [46]	686	8	5	56.4	German breast cancer
larynx [1], [41]	90	3	4	44.4	laryngeal cancer
mayo [47]	312	2	4	59.9	primary biliary cirrhosis of the liver
melanoma [48]	205	4	3	72.2	malignant melanoma
pbc [49]	418	8	4	61.5	primary biliary cirrhosis of the liver
prostate [50]	483	8	4	28.8	prostate cancer
valung [33]	137	6	4	6.5	lung cancer

prior therapy – $PriorTh$ (0 – no (97), 1 – yes (40)), performance status at baseline (Karnofsky rating) – KPS ($Q_1 = 40$; $Me = 60$, $Q_3 = 75$), months from diagnosis to randomisation – $DiagTime$ ($Q_1 = 3$; $Me = 5$, $Q_3 = 11$), and age in years at randomization – Age ($Q_1 = 51$; $Me = 62$, $Q_3 = 66$) was available. Survival time (days) was divided into 4 time intervals: $(0, 80], (80, 160], (160, 240], (240, \infty)$.

The *DST* inducted based on the whole dataset is shown in Figure 9. The hyperplane in the root node, $H_0 : -0.394 \cdot Trt + 0.045 \cdot CellType + 1.125 \cdot PriorTh + 0.079 \cdot KPS + 0.007 \cdot DiagTime - 0.013 \cdot Age = 4.61$, divides the feature space into two subspaces (leaf nodes tn_1 and tn_2) containing 56 and 81 observations, respectively. We can see, that the node tn_2 gives the worse prediction. The hazard function in the first time interval is equal to 0.73; that is much more than the risk of failure in the first time interval obtained by the node tn_1 – 0.18. Thus, although the hazards in other time intervals are comparable between the two leaves, the

TABLE IV

MEAN ABSOLUTE ERROR OF THE FIVE METHODS: *BHTree* AIC – BOU–HAMAD TREE WITH PRUNING BASED ON AIC; *BHTree* BIC – BOU–HAMAD TREE WITH PRUNING BASED ON BIC; *SchTree* – SCHMID TREE; *DST* – DISCRETE SURVIVAL TREE; *CT* – CLASSIFICATION TREE; *MT* – MODULAR TREE. MEAN ABSOLUTE ERRORS (MAE) ARE PRESENTED AS MEAN (STANDARD DEVIATION) OVER 100 RUNS OF EXPERIMENT MULTIPLIED BY 100. SIZE IS CALCULATED AS THE MEDIAN NUMBER OF NODES, FOR *MT* – THE MEDIAN NUMBER OF NODES FOR ALL K COMPONENT TREES

Data set	*BHTree* AIC		*BHTree* BIC		*SchTree*		*DST*		*CT*		*MT*	
	MAE	Size	MAE	Size	MAE	Size	MAE	Size	MAE	Size	MAE	Size
aids2	38.7 (0.5)	87	38.9 (0.5)	3	36.1 (0.2)	1	35.9 (0.4)	7	35.8 (0.3)	37	**35.5** (0.3)	72
bfeed	38.2 (1.1)	43	38.5 (0.8)	3	40.7 (0.5)	3	31.2 (0.3)	5	37.1 (0.3)	3	**30.2** (0.4)	59
cml	41.2 (1.7)	48	42.4 (1.3)	5	37.6 (1.0)	18	35.6 (0.9)	11	**30.5** (2.9)	59	34.3 (1.1)	37
cost	39.7 (1.6)	55	41.5 (1.7)	5	35.4 (1.0)	13	39.6 (1.3)	10	34.9 (2.1)	37	**32.1** (0.9)	50
follic	38.0 (1.7)	51	38.0 (1.2)	5	**29.7** (0.6)	9	44.0 (0.7)	5	39.4 (1.6)	33	41.7 (0.9)	57
GBSG2	29.0 (1.4)	51	29.4 (1.1)	5	**25.6** (0.8)	11	36.9 (1.3)	9	34.2 (1.3)	29	32.9 (1.1)	41
larynx	38.0 (4.4)	11	38.3 (3.8)	5	**34.1** (2.7)	16	41.0 (1.1)	5	39.8 (2.7)	9	38.4 (1.7)	18
mayo	22.6 (1.9)	25	23.4 (1.8)	5	20.2 (1.7)	21	21.9 (1.5)	7	20.5 (1.4)	19	**19.3** (1.4)	30
melanoma	23.1 (2.4)	21	22.5 (2.0)	6	**20.6** (2.2)	19	28.2 (2.8)	3	26.2 (2.2)	11	24.6 (2.2)	21
pbc	23.0 (1.8)	41	24.7 (1.8)	11	21.5 (1.3)	27	25.5 (1.9)	3	**19.3** (1.8)	25	**19.3** (1.5)	32
prostate	42.7 (2.0)	69	45.2 (1.2)	9	40.2 (1.5)	21	41.7 (1.0)	6	38.8 (1.2)	15	**35.7** (1.5)	42
valung	37.9 (3.3)	10	40.2 (2.6)	3	38.7 (3.1)	23	27.0 (2.0)	3	**22** (3.6)	17	22.6 (1.9)	16

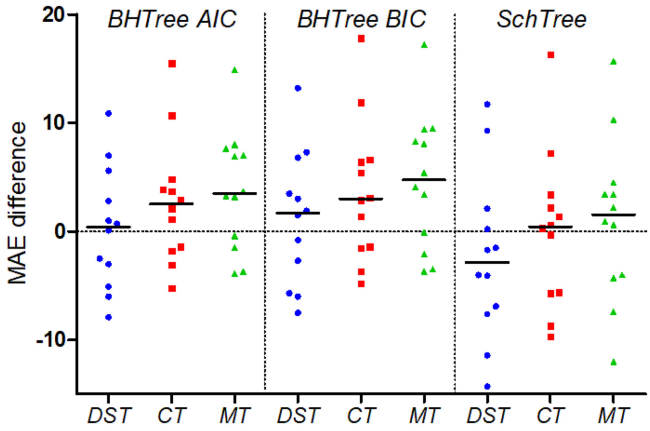


Fig. 8. Differences between the MAE of Schmid tree (*SchTree*) and Bou–Hamad tree (*BHTree*) with pruning based on AIC or BIC and the MAE of discrete survival tree (*DST*), classification tree (*CT*), and modular tree (*MT*). Horizontal lines represent median values.

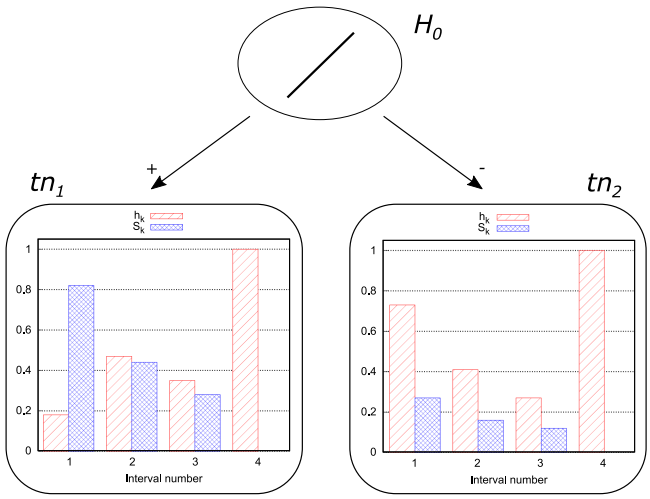
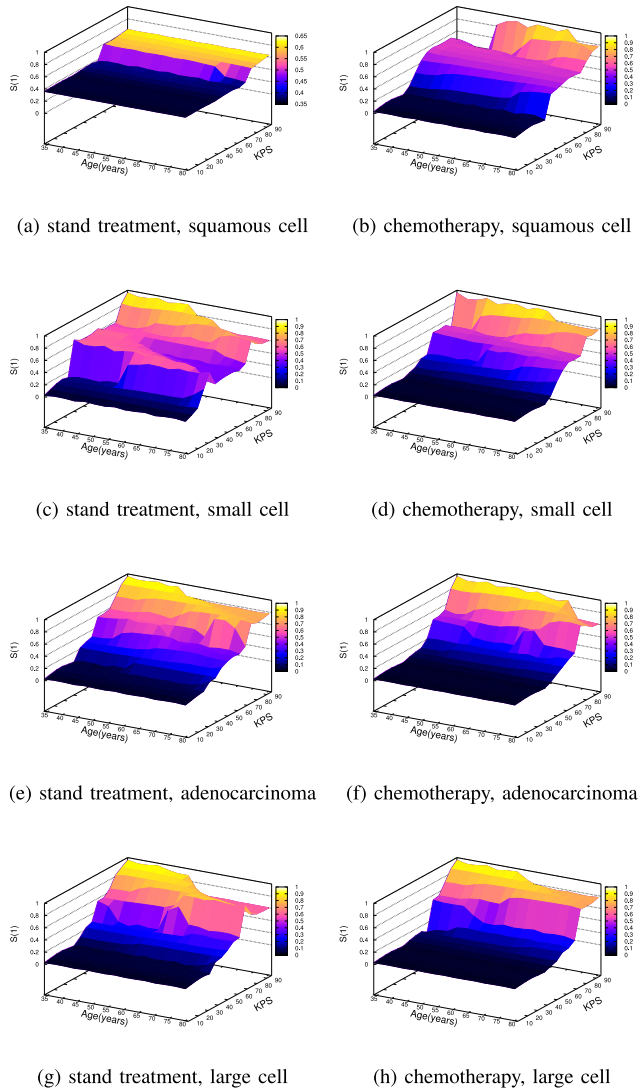


Fig. 9. Discrete survival tree (*DST*) for Veteran's Administration lung cancer data.



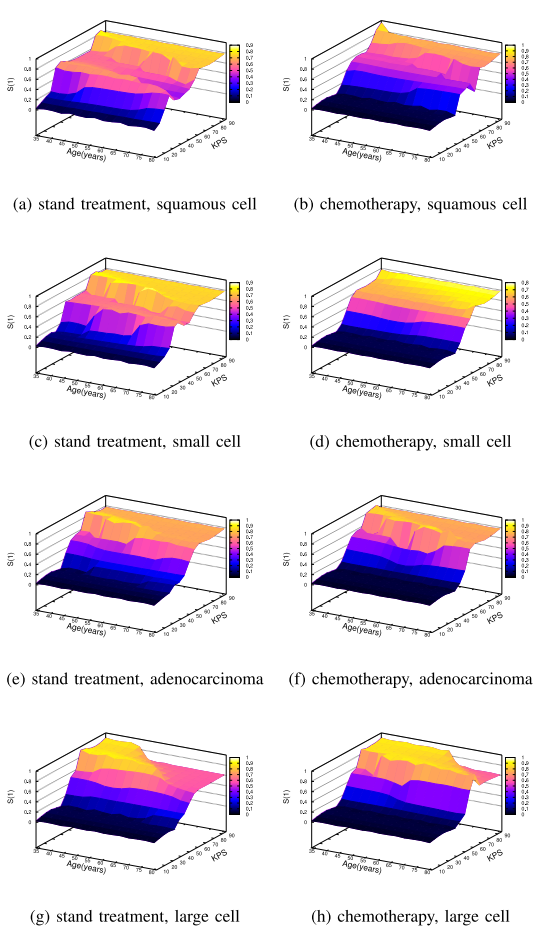
(a) stand treatment, squamous cell (b) chemotherapy, squamous cell

(c) stand treatment, small cell (d) chemotherapy, small cell

(e) stand treatment, adenocarcinoma (f) chemotherapy, adenocarcinoma

(g) stand treatment, large cell (h) chemotherapy, large cell

Fig. 10. The influence of KPS and age on survival function in the first time interval for standard treatment or chemotherapy and different cell types (squamous, small, adenocarcinoma, large) obtained for patients without prior therapy and $DiagTime = 5$.



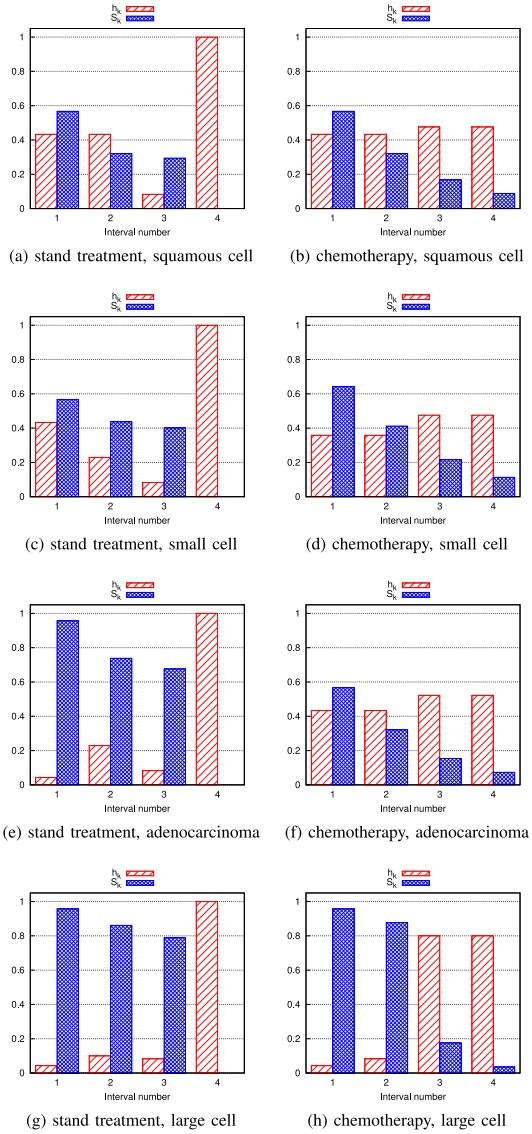
(a) stand treatment, squamous cell (b) chemotherapy, squamous cell

(c) stand treatment, small cell (d) chemotherapy, small cell

(e) stand treatment, adenocarcinoma (f) chemotherapy, adenocarcinoma

(g) stand treatment, large cell (h) chemotherapy, large cell

Fig. 11. The influence of KPS and age on survival function in the first time interval for standard treatment or chemotherapy and different cell types (squamous, small, adenocarcinoma, large) obtained for patients with prior therapy and $DiagTime = 5$.



(a) stand treatment, squamous cell (b) chemotherapy, squamous cell

(c) stand treatment, small cell (d) chemotherapy, small cell

(e) stand treatment, adenocarcinoma (f) chemotherapy, adenocarcinoma

(g) stand treatment, large cell (h) chemotherapy, large cell

Fig. 12. The hazard (h_k) and survival probability (S_k) in four time intervals (I_1, I_2, I_3, I_4) for Veteran's Administration lung cancer data calculated for standard treatment or chemotherapy and four cell types (squamous, small, adenocarcinoma, large) with prior therapy. Other variables were set on their medians: $Age = 62$, $KPS = 60$, and $DiagTime = 5$.

survival function has higher values and, therefore, the better prognosis for tn_1.

The *CT* is built from 19 nodes (10 leaves), while the size of the component trees of *MT* is as follows: T_1 – 9 nodes, T_2 – 5 nodes, T_3 – 3 nodes, T_4 – 1 node. The tree T_4 consists of only one node, since all the observations in the last time interval are uncensored, so their target value is always equal to 1.

Since *CT* has the best prediction performance for the valung dataset (Table IV), the Figures 10 and 11 are based on its results. They present the influence of age and KPS on the survival probability in the first time interval for patients with two types of treatment (standard or chemotherapy), four types of cells (squamous, small, adenocarcinoma, large), and with or without prior therapy (Figures 11 and 10, respectively). Time from diagnosis to randomization was set to 5 months (median). The charts presented in Figures 11 and 10 show a strong influence of KPS on survival probability. Low values of KPS are usually associated with small survival probability and, hence, a shorter survival time. The influence of age is not so straightforward and is visible mainly while analysing the higher values of KPS. For large cell lung cancer with and without prior therapy (Figures 10g, 10h, and 11g, 11h, respectively), and for small cell lung cancer (Figure 10c) and lung adenocarcinoma (Figure 10e) in standard treatment without prior therapy, the survival probability for older people with high KPS is worse than for younger people.

Chemotherapy improves survival probability of older people with high KPS with large cell lung cancer (Figures 10g, 10h and 11g, 11h). The results obtained for patients with prior therapy are usually better. It is mainly visible for all cell types treated with chemotherapy, but also for standard treatment of squamous cell lung cancer with KPS greater than 30 (Figures 10a and 11a).

In Figure 12, we can see the hazards and survival probability calculated for 62-year old patients treated with a prior therapy and KPS equal to 60. The graphs are calculated for standard treatment and chemotherapy, and four cell types. There are only small differences between survival functions obtained for standard treatment and chemotherapy for patients with squamous cell lung cancer (Figures 12 a and 12 b). For small cell lung can-

cer and standard treatment, the risk of failure (h_k) is decreasing to the third time interval (Figure 12 c) and is equal to 1 in the last interval; the hazard for patients treated with chemotherapy is similar across the intervals (Figure 12 d). Analogous hazard values are for adenocarcinoma treated with chemotherapy (Figure 12 f), while for standard therapy, the survival probability is high in the intervals I_1, I_2, and I_3, and equals 0 in I_4 (Figure 12 e). For large cell lung cancer, the difference between standard treatment and chemotherapy is visible for I_3, in which the risk of failure for chemotherapy is higher while compared to standard therapy (Figures 12 g and 12 h).

VII. Discussion and Conclusion

In this paper, three types of tree-based models for discrete event time analysis were proposed. In contrast to the existing discrete-time univariate trees, the new methods create splits in a form of any hyperplane that does not have to be parallel to the axis of the coordinate system. This allows to construct more flexible predictive tools that take into account more complex relationships between covariates and a predicted value, and is a novelty in discrete event time analysis.

The induction of the proposed survival trees is based on the minimization of the dipolar criterion function. Its basic elements, dipoles, are created according to the rules defined separately for each problem. Its flexibility is visible in the solutions presented in this paper, in which three different approaches to the analysis of discrete-time survival data have a common feature - a set of properly defined rules to create dipoles. One of the most important elements of these rules is the ability to use incomplete information from censored data.

Each of the proposed models allows us to analyse discrete-time survival data from a different point of view. The *DST* aims at dividing the feature space into regions that contain observations homogeneous from the point of view of failure time. The *CT* partitions the augmented feature space into areas containing the cases with the same failure indicator $d_{ik} \in \{0, 1\}$, which is equivalent to a binary classification problem. The more complicated structure of the *MT* consists of K classification trees, T_k. Each of them divides the feature space of observations being at risk in the kth time interval into regions with similar values of failure indicator in kth interval.

The methods' performance was validated based on synthetic and real datasets. For synthetic data with known survival time distribution, the influence of the number of time intervals and additional non-informative variables was investigated. In the majority of experiments, the proposed methods outperformed the existing tree models. The only exceptions were the results for 50-dimensional feature space with 48 noisy variables, where the MAE for the Bou–Hamad tree with BIC-based pruning was the best.

The comparison of MAE obtained for 12 real datasets confirmed good predictive capabilities of the proposed algorithms. For eight datasets, their results were better than those of Bou–Hamad and Schmid trees. The improvement of the predictive quality of the oblique trees is mainly visible for valung, bfeed, and cml datasets, where the mean absolute error of the best method decreased by 15.9, 8 and 7.1, respectively, compared to the univariate solutions. The performance of the modular tree was significantly better than the performance of *BHTree* BIC.

The practical application of the proposed methods was presented on the basis of the Veteran's Administration lung cancer data. The analysis of consecutive splits of the *DST* allows us to see the feature space division into regions with different survival experiences expressed by terminal node descriptions. The results of the *CT* and the *MT* give us the opportunity to discover the dependencies between covariates and survival or hazard function in separate time intervals. Additionally, through the detailed analysis of the *MT* component trees structure, we can distinguish the feature space areas responsible for the higher risk of failure in the kth time period, $k = 1, \ldots, K$.

The presented results confirm proposed models to be good predictive tools for discrete-time survival data and, through the use of dipolar criterion function, they are able to use knowledge from censored cases. The algorithms are dedicated to the data that are already discretized at the stage of their collection. If the original data are not in a discrete form, we should use models for continuous survival time.

The splitting hyperplanes in internal nodes of dipolar trees are calculated by the minimization of dipolar criterion function that is the sum of simple penalty functions associated with dipoles, i.e., pairs of feature vectors. The penalty functions punish an inappropriate position of the hyperplane in relation to the dipole-forming vectors, and their form is similar to the hinge loss used in soft margin SVM [51]. Dipole formation rules of *DST* are also related to the C–index [52], which is one of the measures of survival model performance and is calculated as a fraction of concordant pairs among all evaluable pairs of feature vectors.

Discrete-time survival data often contain covariates whose values are changed over time [3]. Due to the learning set preparation process, the *CT* and *MT* models have the ability to include time-varying covariates in the learning process, which is an important feature of discrete-time methods, and the prediction problem becomes a binary classification task. The way the *MT* works is similar to fitting a binary model for each time interval, and in the case of *CT*, to fitting one binary model for the entire augmented dataset.

The *DST* and the *MT* require each time interval to have uncensored observations – the more, the better. With too few uncensored cases, the number of dipoles will be low, causing the model to have a poor generalization ability. Hence, these two methods work better when the number of intervals is rather small. The same assumption was made for the Bou–Hamad tree [24]. However, this does not apply to the *CT*.

The future work will concern modifications of dipolar criterion function in order to perform a feature selection in high-dimensional survival data analysis. Such a solution was already used in [53] for genomic data. An interesting extension of the proposed models will be ensemble methods that usually have better generalization ability combined with a lower interpretability of the results [31]. An outcome of the ensemble may be calculated as a mean value of results returned by single trees or as estimators obtained directly from aggregated data as proposed by Hothorn *et al.* [12].

REFERENCES

[1] J. Klein and M. Moeschberger, *Survival Analysis, Techniques for Censored and Truncated Data*. New York, NY, USA: Springer, 1997.

[2] D. Hosmer, S. Lemeshow, and S. May, *Applied Survival Analysis: Regression Modeling of Time to Event Data*, 2nd ed. New York, NY, USA: Wiley, 2008.

[3] G. Tutz and M. Schmid, *Modeling Discrete Time-to-Event Data* Springer Series in Statistics. New York, NY, USA: Springer, 2016.

[4] A. Ciampi, "Generalized regression trees," *Comput. Statist. Data Anal.*, vol. 12, pp. 57–78, 1991.

[5] M. LeBlanc and J. Crowley, "Relative risk trees for censored survival data," *Biometrics*, vol. 48, pp. 411–425, 1992.

[6] S. Keleş and M. S. Segal, "Residual-based tree-structured survival analysis," *Statist. Med.*, vol. 21, pp. 313–326, 2002.

[7] H. J. Cho and S.-M. Hong, "Median regression tree for analysis of censored survival data," *IEEE Trans. Syst., Man, Cybern. A, Syst., Humans*, vol. 38, no. 3, pp. 715–726, May 2008.

[8] H. Seibold, A. Zeileis, and T. Hothorn, "Model-based recursive partitioning for subgroup analyses," *Int. J. Biostatist.*, vol. 12, no. 1, pp. 45–63, 2016.

[9] J. A. Steingrimsson, L. Diao, A. M. Molinaro, and R. L. Strawderman, "Doubly robust survival trees," *Statist. Med.*, vol. 35, no. 20, pp. 3595–3612, 2016.

[10] A. Linden and P. R. Yarnold, "Modeling time-to-event (survival) data using classification tree analysis," *J. Eval. Clin. Practice*, vol. 23, no. 6, pp. 1299–1308, 2017.

[11] R. Mokarram and M. Emadi, "Classification in non-linear survival models using cox regression and decision tree," *Ann. Data Sci.*, vol. 4, no. 3, pp. 329–340, 2017.

[12] T. Hothorn, B. Lausen, A. Benner, and M. Radespiel-Tröger, "Bagging survival trees," *Statist. Med.*, vol. 23, pp. 77–91, 2004.

[13] T. Hothorn, P. Bühlmann, S. Dudoit, A. Molinaro, and M. van der Laan, "Survival ensembles," *Biostatistics*, vol. 7, pp. 355–373, 2006.

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

[15] I. Yosefian, E. M. Farkhani, and M. R. Baneshi, "Application of random forest survival models to increase generalizability of decision trees: A case study in acute myocardial infarction," *Comput. Math. Methods Med.*, vol. 2015, 2015, Art. no. 576413.

[16] L. Zhou, Q. Xu, and H. Wang, "Rotation survival forest for right censored data," *PeerJ*, vol. 3, pp. 1–12, 2015.

[17] H. Wang and L. Zhou, "Random survival forest with space extensions for censored data," *Artif. Intell. Med.*, vol. 79, pp. 52–61, 2017.

[18] J. B. Nasejje, H. Mwambi, K. Dheda, and M. Lesosky, "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. Methodology*, vol. 17, no. 1, 2017, Art. no. 115.

[19] I. Gómez, N. Ribelles, L. Franco, E. Alba, and J. M. Jerez, "Supervised discretization can discover risk groups in cancer survival analysis," *Comput. Methods Programs Biomed.*, vol. 136, pp. 11–19, 2016.

[20] P. J. García-Laencina, P. H. Abreu, M. H. Abreu, and N. Afonoso, "Missing data imputation on the 5-year survival prediction of breast cancer patients with unknown discrete values," *Comput. Biol. Med.*, vol. 59, pp. 125–133, 2015.

[21] H. Nilsaz-Dezfouli, M. R. Abu-Bakar, J. Arasan, M. B. Adam, and M. A. Pourhoseingholi, "Improving gastric cancer outcome prediction using single time-point artificial neural network models," *Cancer Informat.*, vol. 16, pp. 1–11, 2017.

[22] Y. Yin and S. J. Anderson, "Nonparametric tree-structured modeling for interval-censored survival data," in *Proc. Biometrics Section, Joint Statist. Meetings*, New York, NY, USA, Aug. 11–15, 2002, pp. 3877–3882.

[23] R. Davis and J. Anderson, "Exponential survival trees," *Statist. Med.*, vol. 8, pp. 947–961, 1989.

[24] I. Bou-Hamad *et al.*, "Discrete-time survival trees," *Can. J. Statist.*, vol. 37, pp. 17–32, 2009.

[25] J. Singer and J. Willett, *Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence*. Oxford, U.K.: Oxford Univ. Press, 2003.

[26] I. Bou-Hamad, D. Larocque, and H. Ben-Ameur, "Discrete-time survival trees and forests with time-varying covariates: Application to bankruptcy data," *Statist. Model.*, vol. 11, no. 5, pp. 429–446, 2011.

[27] M. Schmid, H. Küchenhoff, A. Hoerauf, and G. Tutz, "A survival tree method for the analysis of discrete event times in clinical and epidemiological studies," *Statist. Med.*, vol. 35, no. 5, pp. 734–751, 2016.

[28] L. Breiman, J. Friedman, R. Olshen, and C. Stone, *Classification and Regression Trees*. Belmont, CA, USA: Wadsworth, 1984.

[29] L. Bobrowski, M. Kretowska, and M. Kretowski, "Design of neural classifying networks by using dipolar criterions," in *Proc. 3rd Conf. Neural Netw. Their Appl.*, 1997, pp. 689–694.

[30] M. Kretowska, "Piecewise-linear criterion functions in oblique survival trees induction," *Artif. Intell. Med.*, vol. 75, pp. 32–39, 2017.

[31] M. Kretowska, "Tree-based models for survival data with competing risks," *Comput. Methods Programs Biomed.*, vol. 159, pp. 185–198, 2018.

[32] P. Mayer, D. Larocque, and M. Schmid, "DStree: Recursive partitioning for discrete-time survival trees," 2016, R package version 2.4.1. [Online]. Available: https://CRAN.R-project.org/package=DStree

[33] J. Kalbfleisch and R. Prentice, *The Statistical Analysis of Failure Time Data*. New York, NY, USA: Wiley, 1980.

[34] E. Biganzoli, P. Boracchi, L. Mariani, and E. Marubini, "Feed forward neural networks for the analysis of censored survival data: A partial logistic regression approach," *Statist. Med.*, vol. 17, no. 10, pp. 1169–1186, 1998.

[35] L. Bobrowski and W. Niemiro, "A method of synthesis of linear discriminant function in the case of nonseparability," *Pattern Recognit.*, vol. 17, pp. 205–210, 1984.

[36] M. LeBlanc and J. Crowley, "Survival trees by goodness of split," *J. Am. Statist. Assoc.*, vol. 88, no. 422, pp. 457–467, 1993.

[37] U. Mogensen, H. Ishwaran, and T. Gerds, "Evaluating random forests for survival analysis using prediction error curves," *J. Statist. Softw.*, vol. 50, no. 11, pp. 1–23, 2012. [Online]. Available: https://www.jstatsoft.org/v050/i11

[38] T. Therneau, B. Atkinson, and B. Ripley, "rpart: Recursive partitioning and regression trees," 2018, R package version 4.1-13. [Online]. Available: http://CRAN.R-project.org/package=rpart

[39] D. Morina and A. Navarro, "The R package survsim for the simulation of simple and complex survival data," *J. Statist. Softw.*, vol. 59, no. 2, pp. 1–20, 2014.

[40] W. N. Venables and B. D. Ripley, *Modern Applied Statistics With S*, 4th ed. New York, NY, USA: Springer, 2002.

[41] J. Klein, M. Moeschberger, and J. Yan, "KMSurv: Data sets from Klein and Moeschberger (1997)," 2012, R package version 0.1-5. [Online]. Available: http://CRAN.R-project.org/package=KMSurv

[42] T. Hothorn, F. Bretz, P. Westfall, R. M. Heiberger, A. Schuetzenmeister, and S. Scheibe, "Multcomp: Simultaneous inference in general parametric models," 2017, R package version 1.4-8. [Online]. Available: https://cran.r-project.org/package=multcomp

[43] H. S. Jørgensen, H. Nakayama, J. Reith, H. O. Raaschou, and T. S. Olsen, "Acute stroke with atrial fibrillation. The copenhagen stroke study," *Stroke*, vol. 27, no. 10, pp. 1765–1769, 1996.

[44] M. Pintilie, *Competing Risks: A Practical Perspective*, vol. 58. New York, NY, USA: Wiley, 2006.

[45] G. Schumacher *et al.*, "German breast cancer study group: Randomized 2×2 trial evaluating hormonal treatment and the duration of chemotherapy in nodepositive breast cancer patients," *J. Clinical Oncology*, vol. 12, pp. 2086–2093, 1994.

[46] T. Hothorn, "TH.data: TH's data archive," 2018, R package version 1.0-9. [Online]. Available: https://cran.r-project.org/package=TH.data

[47] P. J. Heagerty and Y. Zheng, "Survival model predictive accuracy and ROC curves," *Biometrics*, vol. 61, no. 1, pp. 92–105, 2005.

[48] P. Andersen, O. Borgan, and R. Gill, *Statistical Models Based on Counting Processes*. New York, NY, USA: Springer, 1995.

[49] T. Fleming and D. Harrington, *Counting Processes and Survival Analysis*. New York, NY, USA: Wiley, 1991.

[50] D. Byar and S. Green, "The choice of treatment for cancer patients based on covariate information: Application to prostate cancer," *Bull. du Cancer*, vol. 67, no. 4, pp. 477–490, 1980.

[51] V. Vapnik, *Statistical Learning Theory*. New York, NY, USA: Wiley-Interscience, 1998.

[52] E. L. Korn and R. Simon, "Measures of explained variation for survival data," *Statist. Med.*, vol. 9, pp. 487–503, 1990.

[53] J. Krawczuk and T. Łukaszuk, "The feature selection bias problem in relation to high-dimensional gene data," *Artif. Intell. Med.*, vol. 66, pp. 63–71, 2016.