IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received March 31, 2019, accepted April 20, 2019, date of publication April 26, 2019, date of current version May 13, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2913428

HitBoost: Survival Analysis via a Multi-Output Gradient Boosting Decision Tree Method

PEI LIU[1], BO FU[1], AND SIMON X. YANG[2], (Senior Member, IEEE)
[1]Big Data Research Center, and School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China
[2]Advanced Robotics and Intelligent Systems (ARIS) Laboratory, School of Engineering, University of Guelph, Guelph, ON N1G2W1, Canada

Corresponding author: Bo Fu (fubo@uestc.edu.cn)

This work was supported by the Science and Technology Department of Sichuan Province of China under Grant 2017SZ0005.

ABSTRACT Survival analysis, in many areas such as healthcare and finance, mainly studies the probability of time to the event of interest. Among various methods that build survival predictive models, a class of methods combining with machine learning techniques make assumptions about hazard functions, while another class of methods directly exploit complex neural networks to learn the latent representation of hazard functions. For the traditional survival predictive models, the assumption about hazard functions restricts their performance to some extends. Similarly, the advanced survival predictive models built by complex neural networks also suffer from fairly poor interpretation in real applications. To solve these problems, in this paper, a novel survival analysis method named HitBoost is proposed to predict the probability distribution of the first hitting time (FHT). Instead of making any assumptions about the underlying stochastic process, the proposed HitBoost adopts the multi-output gradient boosting decision tree to implicitly capture the connections between the static covariate and the underlying stochastic process. Furthermore, in the process of tree boosting, the relevant statistics can be utilized to effectively measure the feature importance. The results of evaluations and case studies on benchmarks show that, in comparison to the classical methods, the proposed HitBoost is superior in prediction performance and risk discrimination. Therefore, the HitBoost can be utilized as an effective method to build survival predictive models or to find the important factors for cause-specific failure.

INDEX TERMS Disease prognosis, first hitting time, gradient boosting decision tree, machine learning, survival analysis.

I. INTRODUCTION

Survival analysis, as a method of studying the probability of time to the event of interest (e.g., death, disease recurrence, or failure of a machine), has a wide range of applications in many fields, such as healthcare and finance. Unlike most classification and regression problems, the target of survival analysis is time to the event, or more than one outcome. In the realm of medicine, the prognostic studies of cause-specific diseases rely on survival predictive models and often combine with relevant statistical methods to predict probabilities that a patient may occur the specific disease at various time points, or to find important disease-related prognostic factors. Throughout this paper, we will mainly focus on the medical setting. It also can be easily extended to other fields.

Traditional methods usually take individual hazard function as the main target, followed by making some assumptions about it to predict the probability of event occurrence at various time points. The Cox proportional hazard model (CoxPH) [1] is the most commonly used predictive model for survival analysis. It assumes that the ratio of an individual hazard function to the population's baseline hazard function is a time-independent constant, which is called the hazard ratio and is also a predictor of the CoxPH model. The first hitting time (FHT) model, as one of the predictive models in survival analysis, is mainly to study the distribution of the first hitting time. Unlike the Cox proportional hazard model, the FHT model assumes that the individual hazard function is a form-fixed stochastic process. More works about the FHT model can be found in Lee and Whitmore [2]. In addition, Stikbakke [32] improved the FHT model by applying the boosting method to estimate model parameters. Both the Cox and FHT methods make strong assumptions about the individual hazard function and deem a linear relationship between the model parameters and static covariates.

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

In some special cases, once the individual hazard function violates the assumptions of the model, the survival predictive models built on the top of these methods will suffer from a decreased prediction accuracy.

In recent years, many classical algorithms in machine learning have been investigated and greatly improved in many different fields, such as support vector machine [5], random forest [6] and gradient boosting machine [3], [4]. These techniques also find their way to survival analysis. For example, Random Survival Forest (RSF) [7], as a classical and popular method, no longer makes any assumptions about the individual hazard function but utilizes statistical methods to estimate the hazard function in the framework of Random Forest [34]. However, this type of nonparametric method without regularization technique is prone to overfitting. Among various survival analysis methods, a branch of methods such as gradient boosting machine (GBM) [9] and deep survival network (DeepSurv) [10] applies machine learning techniques to enhance the capability of representing a complex nonlinear relationship between logarithmic hazard ratio and static covariates. Although those methods can predict the logarithmic hazard ratio as same as the Cox model, but they still follow the hazard ratio assumption. Another branch of methods, however, aims at predicting the distribution of the first hitting time instead of assuming individual hazard function. Such as DeepHit [11] and DRSA [12], they apply deep neural network and recurrent neural network to capture the latent representation between static covariates and the FHT probability distribution, respectively. However, fitting deep learning models often requires a large number of training samples, careful hyper-parameters tuning and iteration training, which could be very time-consuming. Moreover, the complex neural network model is a black box with very poor interpretability, making the algorithm incapable in finding disease-related important prognostic factors which is often required in clinical disease prognosis studies. For example, in breast cancer research, whether a gene related to breast cancer is a dangerous or protective factor can be found in the work of Joseph a Sparano about the gene expressions affecting the risk of breast cancer recurrence [27].

In order to overcome the constraint of assumptions about the underlying stochastic process in traditional survival predictive models, and to solve the problem of poor factor interpretation in complex survival predictive models, we propose a novel survival analysis method, HitBoost, to predict the FHT probability distribution. For method realization, we firstly define the learning objective function and derive the gradient expressions of the learning objective function w.r.t. the predicted value. Then we implement it with the XGBoost framework [8], [13], which is a flexible and scalable GBDT and GBM framework. Experimental results on benchmarks show that the proposed HitBoost method has better prediction performance than classical survival analysis methods, which means that the HitBoost can be utilized as an effective means to build survival predictive models or evaluate feature importance.

In comparison to the previous methods for survival analysis, the proposed HitBoost takes advantage in several aspects. Firstly, no assumptions about the underlying stochastic process are required. The HitBoost exploits the property of Multi-Output Gradient Boosting Decision Tree (GBDT) to directly predict the distribution of the first hitting time, making it possible for the multi-output GBDT to implicitly learn the latent representation between static covariates and the underlying stochastic process. As a result, the model's prediction performance is effectively improved. The second advantage is that the relevant statistics for boosting tree node splitting can be utilized to effectively measure the feature importance, which facilitates the finding of important disease-related factors in survival analysis.

The remainder of this paper is organized as follows. The related work of survival analysis is first introduced in Section II. The proposed method is detailed in Section III. The experimental results, case studies and more are given in Section IV. Finally, some concluding remarks are summarized in Section V.

II. RELATED WORKS

A. SURVIVAL DATA

The survival data can be represented as $\{(\boldsymbol{x}_i, T_i, \delta_i)|i = 1, \ldots, n\}$ where

1) n denotes the number of samples in the survival data.
2) $\boldsymbol{x}_i \in \mathbb{R}^m$, a vector with m-dimension, denotes the static covariates of an individual i.
3) $T_i \in \mathbb{R}^+$, a variable with positive value, denotes the last observed time (follow-up time) of an individual i.
4) $\delta_i \in \{0, 1\}$, an indicator variable, denotes the observed status of an individual i. $\delta_i = 1$ indicates the occurrence of the event of interest (e.g. death, relapse, machine malfunction). $\delta_i = 0$ indicates that an individual i is right-censored (i.e. without observed event occurrence).

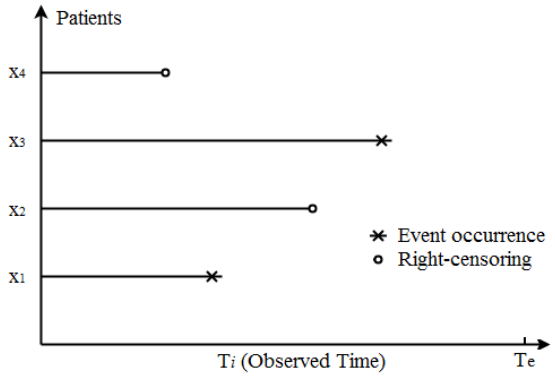


FIGURE 1. Depiction of cases in the survival data.

The depiction of survival data is shown in Fig. 1. We define T_e as the study endpoint for the event of interest, then the set $\{i \,|\, T_i < T_e, \delta_i = 0\}$ denotes the right-censored observations, which is also referred to the lost follow-up in clinical studies.

To analyze survival data, the most popular survival predictive model is the Cox Proportional Hazards (CoxPH) model [1]. It assumes that the ratio of an individual hazard function to the population's baseline hazard function (estimated by the statistical method) is a time-independent constant as

$$e^{f(x)} = \frac{h(t \mid x)}{h_0(t)} \tag{1}$$

where $h_0(t)$ denotes the population's baseline hazard function, and $f(x) = \theta^T x$ denotes logarithmic hazard ratio, $\theta \in \mathbb{R}^m$ is the coefficient of covariates. CoxPH estimates the coefficient θ for survival data without ties via maximizing the partial likelihood function. Many methods based on CoxPH have effectively improved the performance, such as decision tree and deep neural network. All of them [9], [10] can represent nonlinear functions to predict the logarithmic hazard ratio but the assumption of hazard ratio is still followed in them.

B. FIRST HITTING TIME MODEL

Unlike the Cox proportional hazard model, the FHT model assumes that the individual hazard function is a form-fixed stochastic process.

The FHT model mainly studies the probability density function of the first hitting time, i.e., $P(t = t^*, \delta = 1 \mid x = x^*)$, which denotes the true ex-ante probability that an individual with static covariate x^* will experience the event at time t^*. See more examples in [2].

The FHT model takes the individual hazard function as underlying stochastic process $R(t)$

$$R(t) \sim W\left(t|s_0, \mu, \sigma^2 = 1\right), \quad t \geq 0 \tag{2}$$

As given in Equation (3), it assumes that $R(t)$ is a Wiener process with an initial state s_0 and model parameters μ, σ. As given by

$$\begin{aligned} \mu &= \lambda_1^T x \\ \ln(s_0) &= \lambda_2^T x \end{aligned} \tag{3}$$

link functions are used to link the static covariate with the model parameters. The parameters $\lambda_1, \lambda_2 \in \mathbb{R}^m$ are estimated by maximizing the log-likelihood function as

$$\begin{aligned} \ln(L) = \sum_{i=1}^{n} \Big[& \mathrm{I}(\delta_i = 1) * \ln\left(\hat{y}_{T_i}^i\right) + \mathrm{I}(\delta_i = 0) \\ & * \ln\left(1 - \sum\nolimits_{t \leq T_i} \hat{y}_t^i\right) \Big] \end{aligned} \tag{4}$$

where $\hat{y}_t^i$ denotes the estimated probability that the first hitting time of an individual i is t; and $\mathrm{I}(\cdot)$ is an indictor function.

The fhtboost method [33] proposed by Stikbakke improves the performance of the traditional FHT model. However, it still follows the assumption of FHT model to exploit the boosting method for parameter optimization. Some survival predictive models combining with deep learning methods, such as DeepHit [11] and DRSA [12], use complex neural networks to predict the probability distribution of FHT.

Compared to CoxPH model, the FHT model assumes that the hazard function is a stochastic process, which means hazard ratio could be changed over time instead of a time-independent constant. Moreover, the learning objective function of the FHT model is a likelihood function linked with survival function, which is very different from the optimization term of CoxPH. The FHT model is the same with CoxPH in assuming a linear function between model parameters and covariates. As a result, both the CoxPH and FHT models have good interpretability and widespread uses.

III. PROPOSED METHOD

Since the individual survival process is considered as an unfixed and unknown expression, we have to directly learn the function between static covariates and the underlying stochastic process to break away from the constraints imposed by the assumptions. Therefore, we propose a novel survival analysis method that exploits the multi-output gradient boosting decision tree to estimate the probability density function of the first hitting time. This method not only improves the performance of survival prediction but also ensures the interpretability of the model.

A. MODEL DESCRIPTION

The HitBoost is a multi-output gradient boosting decision tree (GBDT) model, as illustrated in Fig. 2. It learns $P(t = t^*, \delta = 1 \mid x = x^*)$ and directly estimates the probability density function of the first hitting time. The HitBoost takes static variates as input, and each output of it is provided by a GBDT (or GBM). As a forward additive model, each GBDT consists of many decision trees. The outputs of multiple GBDTs are transformed into the final predicted value $\hat{y}$ by the softmax layer.

The predicted value $\hat{y}$ is a vector defined a

$$\hat{y} = \left[\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_{T_{max}}, \hat{y}_{T_{max}+1}\right] \tag{5}$$

where T_{max} denotes the longest observed time in the study of interest. Given an individual with the covariate x, an output $\hat{y}_t$ of the model is an estimated probability $\hat{P}(t, \delta = 1 \mid x)$, representing the probability that the event of interest occurs to the individual at time t.

Since

$$\hat{y}_{T_{max}+1} = 1 - \sum\nolimits_{t=1}^{T_{max}} \hat{y}_t \tag{6}$$

the estimated probability $\hat{P}(T_{max} + 1, \delta = 1 \mid x)$ implies that an individual with covariate x never experiences the event in observations. Here, we define the (cause-specific) Cumulative Incidence Function (CIF) [14] of the individual i as

$$\hat{F}(i, t) = P(\tau \leq t, \delta = 1 \mid x_i) = \sum\nolimits_{\tau \leq t} \hat{y}_\tau^i \tag{7}$$

denoting the probability that the individual i experiences an event of interest on or before time t.

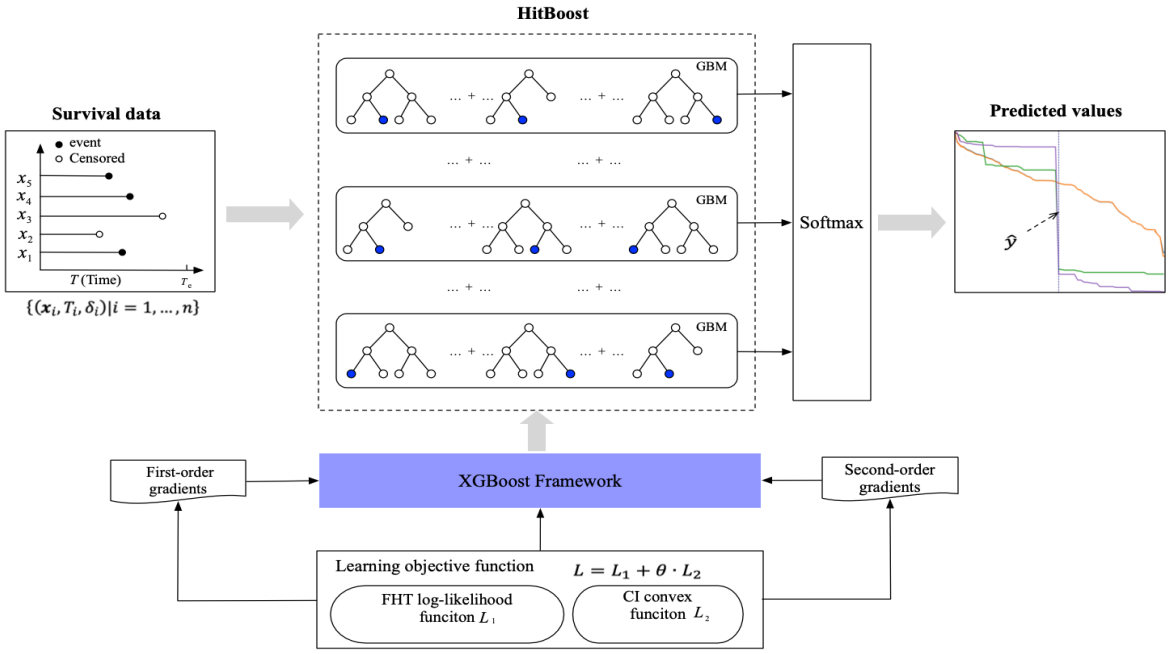


FIGURE 2. Depiction of HitBoost model.

Given the survival data and a specific learning objective function (will be described later), in a parallel way, multiple GBDTs learn the latent relationship between static covariates and hazard functions at each iteration via the gradient boosting method. As a result, the fitted HitBoost model can accurately estimate the probability density function of the first hitting time.

The HitBoost method is able to learn a complex representation of hazard function beyond a proportional hazard or form-fixed stochastic process. Moreover, unlike survival models based on deep learning methods, the HitBoost is constructed by Multi-output GBDTs. It can take advantages of the statistics, e.g., the gradients to find the best feature in the process of tree node splitting, to explore the important factor related to the event of interest, which is also an advantage of HitBoost.

B. LEARNING OBJECTIVE FUNCTION

To train the HitBoost model, we must define the formula of learning objective functions. Here the learning objective functions that we need to optimize are as follows:

1) L_1, the log-likelihood function in FHT models as given by Equation (5);
2) L_2, the C-index [15] approximated by convex function as a part of learning objective functions for risk ranking.

Thus, take into account both L_1 and L_2, we need to minimize the objective function L as

$$L = L_1 + \theta \cdot L_2 \tag{8}$$

where $\theta \in \mathbb{R}$ and $0 \leq \theta \leq 1$. θ is one of hyper-parameters that should be tuned in the model.

As described in Part C of Section II, the formula of L_1 is given as

$$L_1 = -\sum_{i=1}^{n} \Big[\mathrm{I}(\delta_i = 1) \cdot \ln\left(\hat{y}_{T_i}^{(i)}\right) + \mathrm{I}(\delta_i = 0) \cdot \ln\left(1 - \sum\nolimits_{t \leq T_i} \hat{y}_t^{(i)}\right) \Big] \tag{9}$$

For the individual i, the L_1 term ensures that the estimated probability of event occurrence at T_i is maximized if it actually experiences the event at T_i. The estimated probability of event occurrence on or before T_i is minimized if it is right-censored at T_i. More details of FHT models can be found in [2].

For the objective function L_2, the C-index is taken as an optimization term. C-index is a commonly used metric to evaluate model performance in survival analysis. Let us define $\hat{R}_i$ as the estimated risk of the individual i, and Ω as a set of tuples, where $\Omega = \{(i, j)|T_i < T_j, \delta_i = 1\}$. Just as shown in

$$C = \frac{\sum_{(i,j)\in\Omega} \mathrm{I}(\hat{R}_i > \hat{R}_j)}{|\Omega|} \tag{10}$$

the C-index C compares two individuals and thinks that the corresponding risk estimation should be higher for individuals who experience the event formerly.

Since we can't directly optimize the non-convex function C, the indicator function needs to be approximated via convex functions. Inspired by [16], instead of sigmoid function that frequently appears in literature, we adopt the convex function ϕ to approximate the indicator function,

as given by

$$\phi(x, y) = \begin{cases} [-(x - y - \gamma)]^n, & x - y < \gamma \\ 0, & x - y \geq \gamma \end{cases} \tag{11}$$

In the convex function ϕ, $0 < \gamma \leq 1$ and $n > 1$ are the hyper-parameters of the model. After finely tuning the formula of the C-index by using $\hat{F}(i, T_i)$ to represent the risk of the individual i, the expression of L_2 can be obtained as

$$L_2 = \frac{\sum_{(i,j)\in\Omega} w_{i,j} \cdot \phi\left(\hat{F}(i, T_i), \hat{F}(j, T_i)\right)}{\sum_{(i,j)\in\Omega} w_{i,j}} \tag{12}$$

In L_2, the denominator is a normalization factor. w is the weighted value of each tuple in Ω. It indicates the difference between the risk of i and j, as given by

$$w_{i,j} = -\left(\hat{F}(i, T_i) - \hat{F}(j, T_i)\right) \tag{13}$$

For any tuple (i, j) in Ω, the L_2 works as follows.

1) If the individual i experiences the event earlier than j, i.e., the individual i is with the higher risk of failure, minimizing L_2 is equivalent to enlarge the difference between the risk of i and j up to larger or equals than γ.
2) If the difference between the outputs of a tuple in Ω is larger or equals than γ, this tuple of individuals will not have any effects on the learning objective term L_2.

This mechanism can effectively overcome overfitting during training [17].

C. GRADIENT CALCULATION

With the help of the high-performance, flexible and scalable framework *XGBoost* [8], we implement the proposed HitBoost method. Unlike the traditional GBM, the XGBoost requires to derive the first- and second-order gradient of the customized learning objective function w.r.t. the predicted value $\hat{y}$, while the traditional GBM does not require the derivation of second-order gradient. Here, we directly give the result of gradient derivation in the manner of theorem. Due to limited space, the proofs of the theorems are given in the supplementary materials.

1) OBJECTIVE FUNCTION L_1

The theorems for gradients of the objective function L_1 are given as follows.

Theorem 1: For the individual k with the observed indicator variable δ_k and time variable T_k, the *first-order* gradient of L_1 w.r.t. $\hat{y}_t^k$ is

$$\frac{\partial L_1}{\partial \hat{y}_t^k} = \begin{cases} \mathrm{I}(t = T_k) \cdot \dfrac{-1}{\hat{y}_t^k}, & \delta_k = 1 \\ \mathrm{I}(t \leq T_k) \cdot \dfrac{1}{1 - \hat{F}(k, T_k)}, & \delta_k = 0 \end{cases}$$

Theorem 2: For the individual k with the observed indicator variable δ_k and time variable T_k, the *second-order* gradient of L_1 w.r.t. $\hat{y}_t^k$ is

$$\frac{\partial^2 L_1}{\partial \hat{y}_t^k} = \begin{cases} \mathrm{I}(t = T_k) \cdot \dfrac{1}{(\hat{y}_t^k)^2}, & \delta_k = 1 \\ \mathrm{I}(t \leq T_k) \cdot \dfrac{1}{\left[1 - \hat{F}(k, T_k)\right]^2}, & \delta_k = 0 \end{cases}$$

2) OBJECTIVE FUNCTION L_2

We first introduce some related symbolic conventions. Let us define two disjoint subsets related to individual k as

$$\Omega_1 = \{(k, i) | \delta_k = 1, T_k < T_i\}$$

and

$$\Omega_2 = \{(i, k) | \delta_i = 1, T_i < T_k\}$$

which means that the risk with event occurrence of the individual k is higher or lower than the individual i. Then we take the denominator and numerator of L_2 as α and β, respectively.

$$\begin{aligned} \alpha &= \sum\nolimits_{(i,j)\in\Omega} w_{i,j} \\ \beta &= \sum\nolimits_{(i,j)\in\Omega} w_{i,j} \cdot \phi[\hat{F}(i, T_i), \hat{F}(j, T_i)] \end{aligned} \tag{14}$$

Theorems for gradients of the denominator α and numerator β in the objective function L_2 can be obtained as follows.

Theorem 3: For the individual k with the observed indicator variable δ_k and time variable T_k, the *first-order* gradient of α w.r.t. $\hat{y}_t^k$ is

$$\begin{aligned} \frac{\partial \alpha}{\partial \hat{y}_t^k} &= \alpha' \\ &= \begin{cases} \mathrm{I}(t \leq T_k) * \sum\limits_{i:T_i > T_k} (-1) + \sum\limits_{i:\delta_i = 1, T_i < T_k} \mathrm{I}(t \leq T_i), & \delta_k = 1 \\ \sum\limits_{i:\delta_i = 1, T_i < T_k} \mathrm{I}(t \leq T_i), & \delta_k = 0 \end{cases} \end{aligned}$$

and the *first-order* gradients of β w.r.t $\hat{y}_t^k$ is

$$\frac{\partial \beta}{\partial \hat{y}_t^k} = \beta' = \begin{cases} \dfrac{\partial \beta}{\partial \hat{y}_k}|_{\Omega_1} + \dfrac{\partial \beta}{\partial \hat{y}_k}|_{\Omega_2}, & \delta_k = 1 \\ \dfrac{\partial \beta}{\partial \hat{y}_k}|_{\Omega_2}, & \delta_k = 0 \end{cases}$$

where

$$\begin{aligned} \frac{\partial \beta}{\partial \hat{y}_t^k}|_{\Omega_1} &= \mathrm{I}(t \leq T_k) \\ &\quad \cdot \sum_{(k,i)\in\Omega_1} \mathrm{I}\left(-w_{k,i} < \gamma\right) \cdot (w_{k,i} + \gamma)^{n-1} \\ &\quad \cdot \left[-(n+1) \cdot w_{k,i} - \gamma\right] \\ \frac{\partial \beta}{\partial \hat{y}_t^k}|_{\Omega_2} &= \sum_{(i,k)\in\Omega_2} \mathrm{I}(t \leq T_i) \cdot \mathrm{I}\left(-w_{i,k} < \gamma\right) \cdot (w_{i,k} + \gamma)^{n-1} \\ &\quad \cdot \left[\gamma + (n+1) \cdot w_{i,k}\right] \end{aligned}$$

Theorem 4: For the individual k with the observed indicator variable δ_k and time variable T_k, the *second-order* gradient of α w.r.t. $\hat{y}_t^k$ is

$$\frac{\partial^2 \alpha}{\partial \hat{y}_t^k} = \alpha'' = 0$$

And the *second-order* gradient of β w.r.t $\hat{y}_t^k$ is

$$\frac{\partial^2 \beta}{\partial \hat{y}_t^k} = \beta'' = \begin{cases} \frac{\partial^2 \beta}{\partial \hat{y}_k}|_{\Omega_1} + \frac{\partial^2 \beta}{\partial \hat{y}_k}|_{\Omega_2}, & \delta_k = 1 \\ \frac{\partial^2 \beta}{\partial \hat{y}_k}|_{\Omega_2}, & \delta_k = 0 \end{cases}$$

where

$$\frac{\partial^2 \beta}{\partial \hat{y}_t^k}|_{\Omega_1} = \mathrm{I}\left(t \le T_k\right) \cdot \sum_{(k,i)\in\Omega_1} \mathrm{I}\left(-w_{k,i} < \gamma\right) \cdot \left\{(n+1) \cdot \left(w_{k,i}+\gamma\right)^{n-1} - (n-1) \cdot \left(w_{k,i}+\gamma\right)^{n-2} \cdot \left[-(n+1)\cdot w_{k,i} - \gamma\right]\right\}$$

$$\frac{\partial^2 \beta}{\partial \hat{y}_t^k}|_{\Omega_2} = \sum_{(i,k)\in\Omega_2} \mathrm{I}\left(t \le T_i\right) \cdot \mathrm{I}\left(-w_{i,k} < \gamma\right) \cdot \left\{(n+1) \cdot \left(w_{i,k}+\gamma\right)^{n-1} + (n-1) \cdot \left(w_{i,k}+\gamma\right)^{n-2} \cdot \left[\gamma + (n+1)\cdot w_{i,k}\right]\right\}$$

3) OBJECTIVE FUNCTION L

According to Theorems 1, 2, 3 and 4, given

$$L = L_1 + \theta \cdot L_2 = L_1 + \theta \cdot \beta/\alpha$$

we can easily get the gradient of objective function L w.r.t. $\hat{y}_t^k$ using the chain rule as follows.

$$\frac{\partial L}{\partial \hat{y}_t^k} = \frac{\partial L_1}{\partial \hat{y}_t^k} + \theta \cdot \omega(\alpha, \beta)$$

where $\omega(\alpha, \beta) = \frac{\beta' \cdot \alpha - \beta \cdot \alpha'}{\alpha^2}$, and

$$\frac{\partial^2 L}{\partial \hat{y}_t^k} = \frac{\partial^2 L_1}{\partial \hat{y}_t^k} + \theta \cdot \tau(\alpha, \beta)$$

where $\tau(\alpha, \beta) = \frac{\left(\beta'' \cdot \alpha - \beta \cdot \alpha''\right) \cdot \alpha - 2\alpha' \cdot (\beta' \cdot \alpha - \beta \cdot \alpha')}{\alpha^3}$.

We apply vectorization techniques in the implementation of gradient computations, which significantly reduces the running time. The source codes of HitBoost can refer to *https://github.com/liupei101/HitBoost*.

IV. RESULTS

We compare the proposed method with classical survival analysis methods using four public survival datasets and apply the HitBoost model to evaluate the feature importance. Furthermore, case studies are conducted to intuitively compare the HitBoost with classical survival analysis methods in the estimation of hazard functions.

A. DATASETS

Four real-world clinical datasets for experiments are described in Table 1. Throughout the experiments, we take 30 days as 1 month as the basic time unit.

TABLE 1. Description of dataset statistics.

Dataset	M	N	Event Ratio	Survival Time (Months)		
				Min.	Med.	Max.
WHAS	5	1638	42.12%	1	40	67
SUPPORT	14	8873	68.03%	1	8	68
METABRIC	9	1903	57.96%	1	115	356
ROTT2	9	2982	42.66%	2	87	232

M and N: numbers of features and samples, respectively. Min., Med., and Max.: the minimum, median and maximum survival time, respectively.

1) WHAS

The Worcester Heart Attack Study (*WHAS*) [18] studies the survival of acute myocardial infraction (MI). It consists of 1,638 samples and 5 features: age, gender, body-mass-index (bmi), left heart failure complications (chf), and order of MI (miord).

A total of 42.12% patients died during the study. The minimum, median and maximum survival time of patients is 1, 40 and 67 months, respectively.

2) SUPPORT

The Study to Understand Prognoses Preferences Outcomes and Risks of Treatment (*SUPPORT*) [19] researches the survival of seriously ill hospitalized adults on a large scale. It consists of 9,105 samples and 14 features: age, sex, race, number of comorbidities (comorb), presence of diabetes (diabts), presence of dementia (dmt), presence of cancer (cancer), mean arterial blood pressure (m_abp), heart rate (hrt), respiration rate (rsprt), temperature (temp), white blood cell count (cntwbc), serum's sodium (srmsd), and serum's creatinine (srmct).

After excluding patients with missing features, a total of 8,873 samples with an event ratio of 68.03% are eligible. The minimum, median and maximum survival time of patients is 1, 8 and 68 months, respectively.

3) METABRIC

The Molecular Taxonomy of Breast Cancer International Consortium (*METABRIC*) [20] investigates the effect of gene and protein expression profiles on breast cancer survival, and help physicians design better treatment recommendations.

After removing patients with incomplete information from a total of 1,980 patients, the dataset consists of 1,903 patients with an event ratio of 57.96%.

As done in the reference [10], we prepare the same 9 features: MKI67, EGFR, PGR, ERBB2, hormone treatment (hormn), radiotherapy (radiot), chemotherapy (chemot), ER-positive (ER), and age at diagnosis (age). The minimum, median and maximum survival time of patients is 1, 115 and 356 months, respectively.

4) ROTT2

The Rotterdam Tumor Bank (ROTT2) [21] uses patients' pathology and treatment information to study breast cancer. It contains follow-up data of 2,982 women with breast cancer who have gone through breast surgery, 42.66% of which occurred death.

The clinical features in ROTT2 are: age, menopausal status (meno), tumor size (size), tumor grade (grade), the number of positive lymph nodes (nodes), pr, er, hormonal therapy (hormon), chemotheraphy indicator (chemo). The minimum, median and maximum survival time of patients is 2, 87 and 232 months, respectively.

To train and test models, we split each dataset into a training set and a test set by 8:2. We apply statistical methods to test significant difference of data distribution and survival state between training and test sets. Continuous variables are tested by KS-test while categorical variables are tested by chi2-test. As shown in Table 2, there is no significant difference between the training and test set in data distribution and survival state. The P-value is given by log-rank test [22] that tests the difference of survival state between two populations.

TABLE 2. Statistics of training and test sets.

Dataset	Training Set		Test Set		P-value
	N.	R.	N.	R.	
WHAS	1310	42.14%	328	42.07%	0.8632
SUPPORT	7098	67.95%	1775	68.34%	0.9605
METABRIC	1522	58.54%	381	55.64%	0.5091
ROTT2	2385	42.73%	597	42.38%	0.9959

N: number of samples; R: the event ratio.

B. PERFORMANCE

We evaluate the predictive performance of HitBoost on four datasets, and compare it with the classical survival analysis methods, such as methods assuming the hazard ratio (CoxPH and CoxBoost), methods taking the hazard function as a form-fixed stochastic process (ThresReg), and the popular Random Survival Forest (RSF).

(1) *CoxPH* [35], a traditional Cox proportional hazard model as introduced in Section II, takes the hazard ratio as a time-independent constant.
(2) *CoxBoost* [23], a variant of Cox using boosting method to optimize coefficients.

TABLE 3. Models hyper-parameters.

Model	Parameters	WHAS	SUPPORT	METABRIC	ROTT2
HitBoost	eta	0.1	0.02	0.06	0.06
	nrounds	90	150	110	140
	depth	6	4	6	3
	child_weight	0.288	0.902	0.472	0.688
	θ	0.4	0.5	0.8	0.6
	γ	0.01	0.04	0.01	0.06

The eta (learning rate), nrounds (iteration number), depth (max depth) and child_weight (min child weight) are from XGBoost framework. θ and γ are from the objective function. n in objective function is set to 2 by default.

(3) *ThresReg* [24], a traditional FHT model as introduced in Section II, assumes that the hazard function is a Wiener process.
(4) *RSF* [25], a popular and powerful survival analysis method derived from Random Forest, estimates the hazard function without any prior assumptions.

The training and test sets are used to fit and evaluate the models, respectively. The hyper-parameters of models are tuned via 5-fold cross validation and bayesian hyper-parameters optimization [33] on the training sets. As a result, the final tuned hyper-parameters are shown in Table 3. As mentioned before, hyper-parameters of HitBoost come from the custom objective function and XGBoost framework. Since both of CoxPH and ThresReg are essentially linear models, so they don't have any extra hyper-parameters to be tuned. By default, the number of boosting (or iteration) round of RSF and CoxBoost is set to 100. More related details can be seen in the corresponding software packages [23]–[25], [35].

We take *Time-Dependent C-index* (td-CI) [26] as a metric to evaluate the performance of survival predictive models. For observing the stability of HitBoost model, the learning curve on each dataset is shown in Fig. 3.

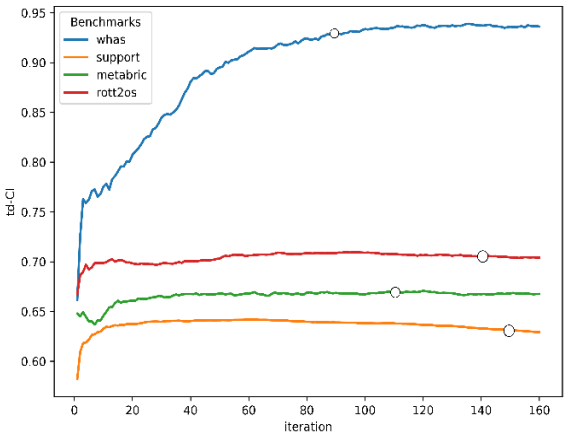


FIGURE 3. Learning curves of HitBoost model.

As shown in Fig. 3, all four models tend to steady in the end, which demonstrates the stability of HitBoost. Among tuned hyper-parameters given in Table 3, *only nrounds* is reset to 150 when models run on each dataset. The hollow dot in Fig. 3 indicates the actual number of iteration obtained

by hyper-parameters tuning. Every point on curves represents the number of iteration and the corresponding model performance on test sets.

TABLE 4. Performance (td-CI).

Method	WHAS	SUPPORT	METABRIC	ROTT2
CoxPH	0.740648	0.593005	0.633109	0.698081
CoxBoost	0.740682	0.590609	0.624546	0.698542
ThresReg	0.732674	0.591483	0.621219	0.658560
RSF	0.913789	0.614945	0.650566	0.675589
HitBoost*	**0.929190**	**0.631281**	**0.668679**	**0.705427**

The asterisk indicates the proposed method.

After tuning hyper-parameters, we evaluate the model performance on independent test sets. As shown in Table 4, HitBoost outperforms all other four methods. For example, HitBoost has better prediction performance with td-CI 0.929190, increased by about 1.7% over RSF in the WHAS dataset. However, RSF has the best performance with td-CI 0.913789 among the four traditional methods. Similarly, td-CI has also been increased by about 2.7% and 2.8% in the SUPPORT and METABRIC, respectively. Only in the ROTT2 dataset, the performance improvement with about 1% is not significant.

It demonstrates that the HitBoost has better prediction performance and more powerful risk discrimination than the classical survival analysis methods. Since the HitBoost no longer makes any assumptions about the hazard function but exploits the higher-performance Multi-Output GBDT to directly learn the latent representation between static covariates and hazard functions, it eventually can be applied in various scenarios.

C. FACTOR ANALYSIS

Besides the superior prediction performance, the HitBoost also can be taken as a tool to find important factors for cause-specific failure in survival analysis, which is indeed essential in clinical research as we have introduced in Section I. But for those deep learning-based survival analysis methods, such as the DeepHit and DRSA, factor analysis is unable to be reached, although they can capture more complex functions. As demonstrated in some pieces of literature and research [28]–[31], Random Forest can be taken as an effective means to extract and rank features in many areas. Therefore, taking Random Forest as a reference, in this section we will inspect the HitBoost from the aspect of feature importance evaluation.

The HitBoost and RSF are all used to evaluate the feature importance in each dataset. The models are built to fit with the training set, followed by measuring the importance scores of features in each dataset. We repeat this procedure for 20 times and obtain the results in Fig. 4.

As illustrated in Fig. 4, in the WHAS, METABRIC and ROTT2 dataset, the importance scores of top 2 features calculated by the HitBoost and RSF are the same. In the SUPPORT dataset, the features except for 'cancer' are of similar

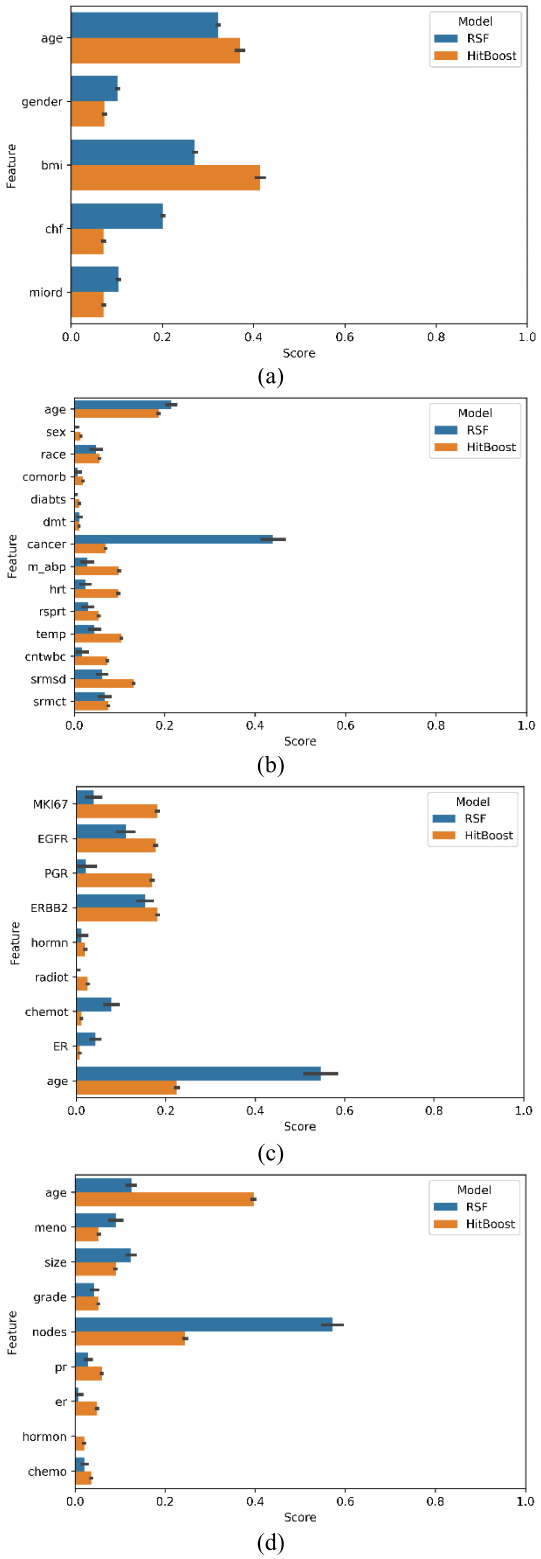


FIGURE 4. Depiction of feature importance evaluations given by HitBoost and RSF on four benchmarks. (a) WHAS. (b) SUPPORT. (c) METABRIC. (d) ROTT2.

importance score. It demonstrates that the HitBoost and RSF are basically consistent in finding important features, although there are minor differences in the quantified metrics.

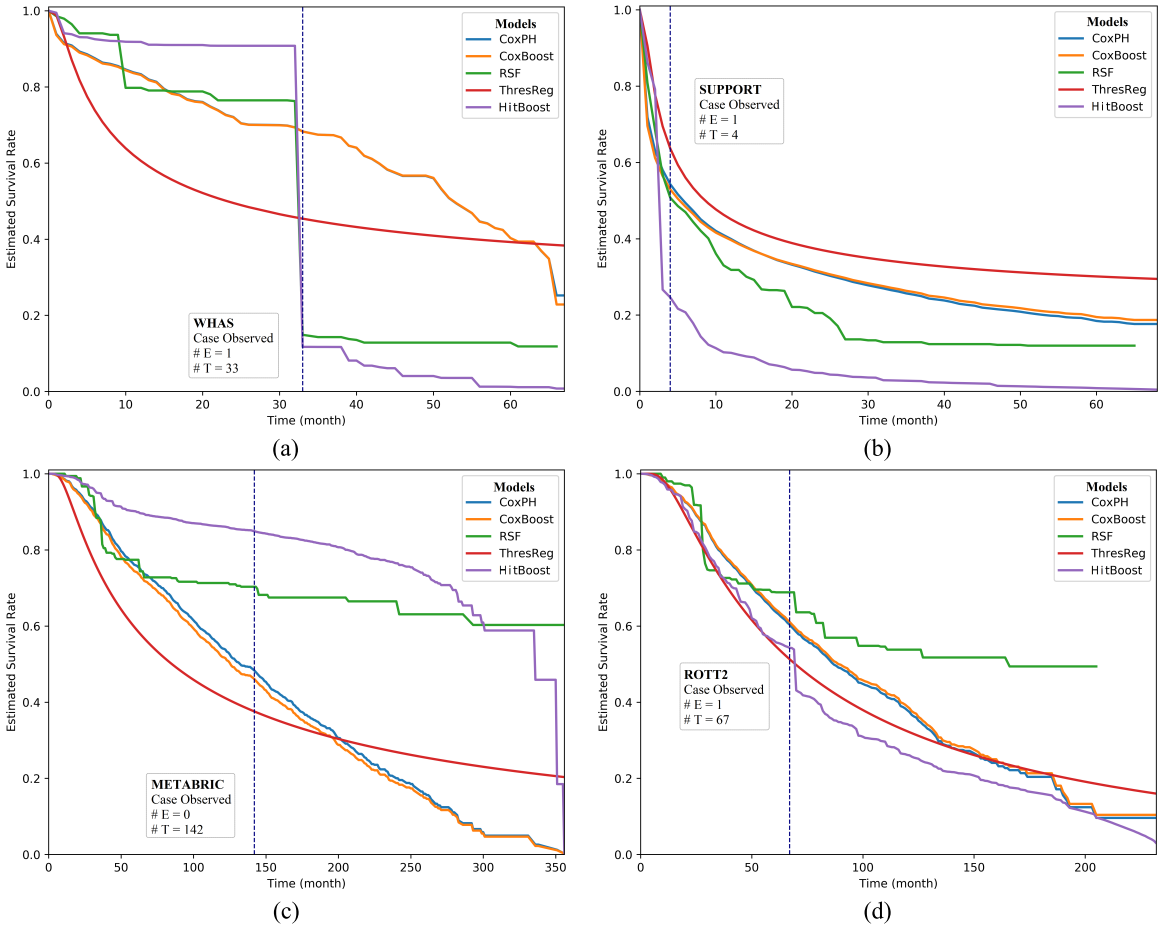


FIGURE 5. Survival curves of the cases randomly selected from datasets. (a) WHAS. (b) SUPPORT. (c) METABRIC. (d) ROTT2.

Therefore, the HitBoost is capable of finding the important factors for cause-specific failure in real-world survival analysis just like RSF.

D. CASE STUDIES

We arbitrarily choose one sample in the test set of each dataset and apply the predictive models to estimate the survival function of them. All the estimated survival functions are visualized in Fig. 5.

As shown in Fig. 5, the curves of survival function estimated by the CoxPH, CoxBoost and ThresReg are relatively smooth, because they all follow the assumption that the individual hazard function is with an explicit and fixed mathematical expression. For samples in the WHAS (Fig. 5a), SUPPORT (Fig. 5b) and ROTT2 (Fig. 5d), all of them occur the event at the corresponding observed time, i.e. $E = 1$. In comparison to the CoxPH, CoxBoost, ThresReg and RSF models, the curve of survival function estimated by the HitBoost model shows a sharp decline at the observed time that is marked by the vertical dashed line in Fig. 5, which implies there could be a higher risk at the observed time point. Moreover, the estimated survival rate of the HitBoost model at the observed time is lower than all other four models, which is consistent with the actual data. For the sample in METABRIC (Fig. 5c), it does not occur the event in the corresponding observed time, i.e. $E = 0$ or right-censored. Unlike the underestimated survival rate of all other four models, the survival rate estimated by the HitBoost model is more in line with the actual situation, i.e. the right-censored patient should have a greater probability of surviving over the observed time.

Case studies intuitively demonstrate that whether it is for patients with event occurrence or right-censored, the HitBoost survival predictive model is more accurate in estimating the probability density function of FHT. Therefore, the HitBoost has a strong power of risk discrimination.

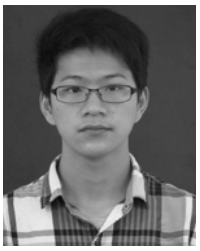
V. CONCLUSION

This paper proposes the HitBoost method that exploits multi-output gradient boosting decision tree to predict the probability density function of the first hitting time. It overcomes the constraint of assuming the hazard function as the potential stochastic process to predict the future survival status and solves the problem of poor factor interpretation in practical applications. Instead of making any assumptions about the underlying stochastic process, the HitBoost uses

the multi-output gradient boosting decision tree to implicitly learn the latent representation between covariates and the underlying stochastic process. That effectively improves the model's prediction performance. Moreover, the proposed method can be utilized to find the important factors for cause-specific failure, which is unreachable for survival predictive models built by complex deep learning methods. Therefore, the HitBoost can provide an important way for survival analysis, and it has a wider range of application scenarios and greater practicality.

REFERENCES

[1] D. R. Cox, "Regression models and life-tables," *J. Roy. Stat. Soc., Ser. B*, vol. 34, no. 2, pp. 187–202, 1972.
[2] M.-L. T. Lee and G. A. Whitmore, "Threshold regression for survival analysis: Modeling event times by a stochastic process reaching a boundary," *Stat. Sci.*, vol. 21, no. 4, pp. 501–513, 2006.
[3] B. Zhang, J. Ren, Y. Cheng, B. Wang, and Z. Wei, "Health data driven on continuous blood pressure prediction based on gradient boosting decision tree algorithm," *IEEE Access*, vol. 7, pp. 32423–32433, 2019.
[4] D. Zhang, L. Qian, B. Mao, C. Huang, B. Huang, and Y. Si, "A data-driven design for fault detection of wind turbines using random forests and xgboost," *IEEE Access*, vol. 6, pp. 21020–21031, 2018.
[5] Y.-D. Zhang *et al.*, "Facial emotion recognition based on biorthogonal wavelet entropy, fuzzy support vector machine, and stratified cross validation," *IEEE Access*, vol. 4, pp. 8375–8385, 2016.
[6] S. F. Abdoh, M. A. Rizka, and F. A. Maghraby, "Cervical cancer diagnosis using random forest classifier with SMOTE and feature reduction techniques," *IEEE Access*, vol. 6, pp. 59475–59485, 2018.
[7] H. Ishwaran, U. B. Kogalur, E. H. Blackstone, and M. S. Lauer, "Random survival forests," *Ann. Appl. Statist.*, vol. 2, no. 3, pp. 841–860, 2008.
[8] T. Chen, T. He, and M. Benesty, *Xgboost: Extreme Gradient Boosting*, document R Package Version 0.4-2, 2015.
[9] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *Ann. Statist.*, vol. 29, pp. 1189–1232, Oct. 2001.
[10] J. L. Katzman, U. Shaham, A. Cloninger, J. Bates, T. Jiang, and Y. Kluger, "DeepSurv: Personalized treatment recommender system using a Cox proportional hazards deep neural network," *BMC Med. Res. Methodol.*, vol. 18, p. 24, Feb. 2018.
[11] C. Lee, W. R. Zame, J. Yoon, and M. van der Schaar, "Deephit: A deep learning approach to survival analysis with competing risks," in *Proc. AAAI*, 2018, pp. 1–8.
[12] K. Ren *et al.*, "Deep recurrent survival analysis," in *Proc. AAAI*, 2019, pp. 1–8.
[13] T. Chen and C. Guestrin, "XGBoost: Reliable large-scale tree boosting system," in *Proc. Conf. Knowl. Discovery Data Mining*, 2016, pp. 1–6.
[14] J. P. Fine and R. J. Gray, "A proportional hazards model for the subdistribution of a competing risk," *J. Amer. Stat. Assoc.*, vol. 94, no. 446, pp. 496–509, 1999.
[15] F. E. Harrell, R. M. Califf, D. B. Pryor, K. L. Lee, and R. A. Rosati, "Evaluating the yield of medical tests," *J. Amer. Med. Assoc.*, vol. 247, no. 18, pp. 2543–2546, 1982.
[16] L. Yan, D. Verbel, and O. Saidi, "Predicting prostate cancer recurrence via maximizing the concordance index," in *Proc. 10th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2004, pp. 479–485.
[17] L. Yan, R. Dodier, M. C. Mozer, and R. Wolniewicz, "Optimizing classifier performance via an approximation to the Wilcoxon-Mann-Whitney statistic," in *Proc. 20th Int. Conf. Mach. Learn.*, 2003, pp. 848–855.
[18] P. V. Rao, D. W. Hosmer, and S. Lemeshow, "Applied survival analysis: Regression modeling of time to event data," *J. Amer. Stat. Assoc.*, vol. 95, no. 450, p. 681, Jun. 2000.
[19] W. A. Knaus *et al.*, "The SUPPORT prognostic model: Objective estimates of survival for seriously ill hospitalized adults," *Ann. Internal Med.*, vol. 122, no. 3, pp. 191–203, 1995.
[20] C. Curtis *et al.*, "The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups," *Nature*, vol. 486, pp. 346–352, Jun. 2012.
[21] J. A. Foekens *et al.*, "The urokinase system of plasminogen activation and prognosis in 2780 breast cancer patients," *Cancer Res.*, vol. 60, no. 3, pp. 636–643, 2000.
[22] J. D. Kalbfleisch and R. L. Prentice, "The statistical analysis of failure data," *IEEE Trans. Rel.*, vol. 35, no. 1, p. 11, 1986.
[23] H. Binder, C. Porzelius, and M. Schumacher, "An overview of techniques for linking high-dimensional molecular data to time-to-event endpoints by risk prediction models," *Biometrical J.*, vol. 53, no. 2, pp. 170–189, 2011.
[24] T. Xiao, *THREG: Threshold Regression*, document R Package version 1.0.3, 2015.
[25] H. Ishwaran and U. B. Kogalur, *RF-SRC: Random Forests for Survival, Regression and Classification*, document R Package version 2.4.2, 2017.
[26] L. Antolini, P. Boracchi, and E. Biganzoli, "A time-dependent discrimination index for survival data," *Statist. Med.*, vol. 24, no. 24, pp. 3927–3944, 2005.
[27] J. A. Sparano *et al.*, "Prospective validation of a 21-gene expression assay in breast cancer," *New England J. Med.*, vol. 373, no. 21, pp. 2005–2014, 2015.
[28] X. Li, Z. Wang, L. Wang, R. Hu, and Q. Zhu, "A multi-dimensional context-aware recommendation approach based on improved random forest algorithm," *IEEE Access*, vol. 6, pp. 45071–45085, 2018.
[29] C. Strobl, A.-L. Boulesteix, T. Kneib, T. Augustin, and A. Zeileis, "Conditional variable importance for random forests," *BMC Bioinf.*, vol. 9, p. 307, Jul. 2008.
[30] K. K. Nicodemus, J. D. Malley, C. Strobl, and A. Ziegler, "The behaviour of random forest permutation-based variable importance measures under predictor correlation," *BMC Bioinf.*, vol. 11, p. 110, Feb. 2010.
[31] F. Miao, Y.-P. Cai, Y.-X. Zhang, X.-M. Fan, and Y. Li, "Predictive modeling of hospital mortality for patients with heart failure by using an improved random survival forest," *IEEE Access*, vol. 6, pp. 7244–7253, 2018.
[32] V. Stikbakke. (2019). *FHTBOOST: Boosting of FHT in Survival Analysis*. Accessed: Feb. 28, 2019. [Online]. Available: https://github.com/vegarsti/fhtboost
[33] J. Bergstra, D. Yamins, and D. D. Cox, "Hyperopt: A python library for optimizing the hyperparameters of machine learning algorithms," in *Proc. 12th Python Sci. Conf.*, 2013, pp. 1–8.
[34] L. Breiman, "Random Forrest," *Mach. Learn.*, vol. 45, pp. 5–32, Oct. 2001.
[35] C. Davidson-Pilon, *Lifelines*, document Python Package version 0.20.0, 2019.



PEI LIU received the B.S. degree from the University of Electronic Science and Technology of China, Chengdu, China, in 2017, where he is currently pursuing the M.S. degree. His current research interests include risk calculation, mathematical modeling, and machine learning in the field of medicine.

BO FU received the Ph.D. degree in communication engineering from the University of Electronic Science and Technology of China, Chengdu, China, in 2009. He joined the School of Computer Science and Engineering, University of Electronic Science and Technology of China, in 2009, where he is currently an Associate Professor. He was a Visiting Scholar with the Advanced Robotics and Intelligent Systems Laboratory, University of Guelph, Guelph, ON, Canada, from 2007 to 2008. He has published a number of research articles in refereed international journals and magazines. His current research interests include intelligent algorithms, machine learning, and medical data analysis.

SIMON X. YANG (S'97–M'99–SM'08) received the B.Sc. degree in engineering physics from Beijing University, Beijing, China, in 1987, the first of two M.Sc. degrees in biophysics from the Chinese Academy of Sciences, Beijing, in 1990, the second M.Sc. degree in electrical engineering from the University of Houston, Houston, TX, USA, in 1996, and the Ph.D. degree in electrical and computer engineering from the University of Alberta, Edmonton, AB, Canada, in 1999.

He is currently a Professor and the Head of the Advanced Robotics and Intelligent Systems Laboratory, University of Guelph, Guelph, ON, Canada. His research interests include robotics, intelligent systems, sensors and multi-sensor fusion, wireless sensor networks, control systems, transportation, and computational neuroscience.

Dr. Yang has been very active in professional activities. Among many of his awards, he was a recipient of the Distinguished Professor Award at the University of Guelph. He was the General Chair of the 2011 IEEE International Conference on Logistics and Automation, and the Program Chair of the 2015 IEEE International Conference on Information and Automation. He serves as the Editor-in-Chief of the *International Journal of Robotics and Automation*, and an Associate Editor of the IEEE TRANSACTIONS ON CYBERNETICS, and several other journals.