




# Optimizing Survival Analysis of XGBoost for Ties to Predict Disease Progression of Breast Cancer

Pei Liu , Bo Fu , Simon X. Yang , Ling Deng, Xiaorong Zhong, and Hong Zheng

**Abstract—Objective:** Some excellent prognostic models based on survival analysis methods for breast cancer have been proposed and extensively validated, which provide an essential means for clinical diagnosis and treatment to improve patient survival. To analyze clinical and follow-up data of 12119 breast cancer patients, derived from the Clinical Research Center for Breast (CRCB) in West China Hospital of Sichuan University, we developed a gradient boosting algorithm, called EXSA, by optimizing survival analysis of XGBoost framework for ties to predict the disease progression of breast cancer. **Methods:** EXSA is based on the XGBoost framework in machine learning and the Cox proportional hazards model in survival analysis. By taking Efron approximation of partial likelihood function as a learning objective for ties, EXSA derives gradient formulas of a more precise approximation. It optimizes and enhances the ability of XGBoost for survival data with ties. After retaining 4575 patients (3202 cases for training, 1373 cases for test), we exploit the developed EXSA method to build an excellent prognostic model to estimate disease progress. Risk score of disease progress is evaluated by the model, and the risk grouping and continuous functions between risk scores and disease progress rate at 5- and 10-year are also demonstrated. **Results:** Experimental results on test set show that the EXSA method achieves competitive performance with concordance index of 0.83454, 5-year and 10-year AUC of 0.83851 and 0.78155, respectively. **Conclusion:** The proposed EXSA method can be utilized as an effective method for survival analysis. **Significance:** The proposed method in this paper can provide an important means for follow-up data of breast cancer or other disease research.

**Index Terms—**Breast cancer, survival analysis, machine learning, prognostic prediction, relapse and metastasis cancer, gradient boosting machine, proportional hazard model.

## I. INTRODUCTION

IN THE realm of medicine, survival analysis of follow-up observations, such as the study of disease progress, has been an important topic. Survival analysis based on multiple factors mainly focuses on estimating the survival probability of a particular time of interest for a patient. It can provide the estimation of a fully individualized survival and hazard function for each case. Multifactor approach for survival analysis, such as the Cox proportional hazards model, predicts risk scores by using a set of patient data to regress features or covariates to the time of failure while ignoring censored data. It is very different from the more familiar classification and regression problems in machine learning. In most instances, the survival predictive model can give the personalized survival function that describes the changes in risk scores at different times. But the binary classification model can only give a probability of the interested outcome at a specific time.

The Cox proportional hazard (CPH) model [1], [2], also known as the Cox regression model, is a linear prognostic prediction model commonly used in survival analysis, and has a good interpretability. When there is a nonlinear function between the log-hazard ratio and static covariates, it will result in a decreased prediction accuracy. In other words, the CPH model cannot adequately represent the complex nonlinear relationship between the log-hazard ratio and static covariates. Many strict restrictions have to be imposed. Therefore, many methods that are widely used in machine learning also find its way in the field of survival analysis [3]. Based on the Random Forest method [4], Ishwaran *et al.* [5] proposed Random Survival Forest (RSF) for survival analysis without proportional hazard assumption. RSF accounts for nonlinear interactions for features. Examples of its applications can be found in modeling complex metabolomics data [6].

The Gradient Boosting Machine (GBM) [7], based on the Gradient Boosting Decision Tree (GBDT) method, is one technique applied in many fields. GBDT has shown excellent results on many standard classification benchmarks [8]. XGBoost (eXtreme Gradient Boosting) [11] as a variant of GBM, proposed by Chen *et al.* in 2016, dramatically extends the original GBM method. It exploits more precise approximation of learning objective, and regularization term that effectively avoids overfitting.

Manuscript received November 13, 2019; revised April 1, 2020; accepted May 5, 2020. Date of publication May 11, 2020; date of current version December 21, 2020. This work was supported by the Science and Technology Department of Sichuan Province of China under Grant 2017SZ0005. (Corresponding authors: Bo Fu and Hong Zheng.)

Bo Fu is with the Big Data Research Center, and School of Computer Science and Engineering, University of Electronic Science and Technology of China, Xiyuan Ave, 611731 Chengdu, China (e-mail: fubo@uestc.edu.cn).

Pei Liu is with the Big Data Research Center, and School of Computer Science and Engineering, University of Electronic Science and Technology of China.

Simon X. Yang is with the Advanced Robotics and Intelligent Systems (ARIS) Laboratory, School of Engineering, University of Guelph.

Ling Deng and Xiaorong Zhong are with the Laboratory of Molecular Diagnosis of Breast, Clinical Research Center for Breast, State Key Laboratory of Biotherapy, National Collaborative Innovation Center for Biotherapy, West China Hospital, Sichuan University.

Hong Zheng is with the Laboratory of Molecular Diagnosis of Breast, Clinical Research Center for Breast, State Key Laboratory of Biotherapy, National Collaborative Innovation Center for Biotherapy, West China Hospital, Sichuan University, Guo Xue Xiang, 610041 Chengdu, China (e-mail: hzheng@scu.edu.cn).

Digital Object Identifier 10.1109/TBME.2020.2993278

overfitting. Thus, XGBoost is a novel algorithm for tree node splitting and leaf node estimation. Combining with other techniques, XGBoost achieves state-of-the-art results in many classification and regression problems. To realize the non-linear machine learning for survival analysis, the CPH model is introduced into the GBM by Ridgeway and Binder *et al.* [9], [10], which still follows the hypothesis of proportional hazard, but it is no longer limited to linear function between the log-hazard ratio and static covariates. Although its interpretability is not as good as the CPH model, it has better performance to predict individual hazard ratio and survival status in practical applications [3]. The XGBoost-based survival analysis method has been implemented in the library *xgboost* [11], [35]. If a large number of ties exist in survival dataset, its approximation of partial likelihood function is not precise enough in the implementation. Some events happened at the same time greatly affect the performance of survival prediction to some extent.

Among cancer diseases, breast cancer is one of the largest morbidity and mortality of malignant tumor, which seriously endangers female health. Consequently, mathematical models and techniques have been developed to approach problems related to breast cancer prognosis from a theoretical perspective. Machine learning [37], [38], convolutional neural networks [39], deep neural networks [40], and Bayesian networks [41] have shown considerable success recently among other promising methods. Prognostic prediction model, built by using the clinical data of breast cancer patients, can be exploited to assess the risk and status of patients with relapse and metastasis. It can assist doctors to carry out appropriate treatment plans to lessen the suffering of patients. Based on the structured data, various prognostic prediction methods are used in breast cancer, such as electromechanical coupling factor of breast tissue [30] and pharmacokinetic tumor heterogeneity [31] as new prognostic biomarkers, BRCA1 and other gene expressions [32]. Based on the unstructured or semi-structured data like the microwave and histological images, some tumor detection methods are used for treatment in breast cancer, such as automated classification of breast cancer stroma maturity [33] and RAR algorithm of an ultra-wideband imaging system for breast cancer detection [34].

PREDICT and Adjuvant Online [12]–[14], as famous prognostic prediction models of breast cancer, mainly collect pathological and treatment characteristics of breast cancer patients from electronic health records (EHRs) to predict their survival probability. The Cox regression method, as a classical multivariate analysis method for survival, is adopted to build these famous prognostic prediction models. Based on the Cox regression and related statistical analysis methods [27], 21gene [17], another popular prognostic prediction model for evaluating the risk score of recurrence in breast cancer, collects breast cancer-related gene expression data to evaluate the risk score of distant recurrence and chemotherapy benefit in breast cancer patients. All the above models have been well validated in Western countries and widely used in clinical practice [15], [16], [28]. However, due to the lack of clinical follow-up data, population heterogeneity, and other reasons, prognosis models suitable for the Chinese population have not been developed and widely used in clinical practice.

In this paper, in order to build a prognostic prediction model for Chinese breast cancer patients, 12119 patients in EHRs, derived from the Clinical Research Center for Breast (CRCB) in West China Hospital of Sichuan University, including personal information, diagnosis, medical history, pathology, treatment, and follow-up characteristics, are included. Breast cancer patients are followed up for more than ten years and the annual non-investigation rate is less than 1.1% since 2011. We developed a gradient boosting algorithm, called EXSA, by optimizing survival analysis of XGBoost framework for ties to predict the disease progress of breast cancer. EXSA is based on the XGBoost framework in machine learning and the Cox proportional hazards model in survival analysis. By taking Efron approximation of partial likelihood function as a learning objective for ties, EXSA derives gradient formulas of a more precise approximation. It optimizes and enhances the ability of XGBoost for survival data with ties. This method fully takes advantage of XGBoost to apply more precise approximation of learning objective relative to GBM. And EXSA can add regularization terms into the model to effectively avoid overfitting. Further, after defining a more precise approximation of partial likelihood function as a learning objective, we can derive the gradient of the learning objective for XGBoost framework. That makes it be able to build a more excellent and robust prognostic prediction model, especially for real-world data with tied events. Accordingly, the established model precisely estimates a fully individualized survival and hazard function for each patient, which can effectively reveal the patient's future survival status.

Using appropriate exclusion criteria, we include 4575 eligible patients from 12119 diagnosed patients in EHRs. We then use the proposed EXSA method to build a prognostic model towards the prediction of disease progress, in which 24 features out of the 89 patient variables about patient's demographics and clinical treatment including diagnosis, pathology and therapy are selected. With 3202 patients (70%) for training, and 1373 patients (30%) for test, the model with 24 features as input variables achieves comparable accuracy. In the independent test set, the model achieves performance with the concordance index of 0.83454, 5-year AUC of 0.83851 and 10-year AUC of 0.78155, respectively. Compared with the XGBoost, Cox, RSF, and GBM models for survival prediction, the EXSA model is of the best performance. Moreover, we also evaluate risk scores of disease progress of the prediction model and demonstrate continuous functions between risk score and disease progression rate at 5- and 10-year. The cut-off values for risk grouping are calculated to divide breast cancer patients into the following categories: low risk (risk score between 0 and 21.5), mid-low risk (risk score between 21.6 and 29.6), mid-high risk (risk score between 29.7 and 38.8), and high risk (risk score between 38.9 and 100). The results show that the cut-off values for risk grouping have a strong discriminative power and practical value.

The remainder of this paper is organized as follows. In Section II, we first introduce the related works of survival analysis. Section III presents the proposed method and related derivation. Section IV mainly describes the data materials. Then, in Section V we show the experimental results and model applications. Finally, in Section VI we give some concluding remarks.

## II. RELATED WORKS

### A. Problem Formulation

Survival data of patients is represented as  $\{(x_i, T_i, \delta_i) | i = 1, \dots, n\}$ ,  $x_i \in \mathbb{R}^m$ , where  $n$  denotes the number of patients;  $x_i$  denotes a vector of covariates;  $m$  denotes the dimension of covariates;  $T_i \in \mathbb{R}^+$  denotes the last observed time of patient  $i$ , *i.e.*, last follow-up time; indicator variable  $\delta_i = 1$  denotes occurrence of the event of interest,  $\delta_i = 0$  denotes censored event of interest. 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, referred to as the loss of follow-up in clinical studies. If no tied events occur at a certain time in survival data, *i.e.*, the observed survival data has no ties, the set  $\{t_i | i = 1, \dots, k\}$  consists of  $k$  different time points where only one event is observed.  $R(t) = \{i | T_i \geq t\}$  is the set in that patients are still in the observational study at the time  $t$ . That is to say patients are at risk still.

If observed time is measured in a perfectly continuous scale, survival data with ties would never exist. In real applications, observed time is usually recorded in a discrete manner that results in the existence of ties in survival data. When the survival data has ties, we define  $D$ ,  $N(t)$ ,  $q(t)$ , and  $C_t$  associated with the survival data, where  $D = \{t_i | i = 1, \dots, k\}$ ,  $t_1 < t_2 < \dots < t_k$  denotes  $k$  different time points of event occurrence;  $N(t) = \{i | T_i = t\}$  denotes patients observed at the time  $t$ ;  $q(t) = \{i | T_i = t, E_i = 1\}$  denotes patients with events observed at the time  $t$ ;  $C_t = |q(t)|$  denotes the number of patients in  $q(t)$ .

### B. Cox Proportional Hazard Model

In survival analysis, survival function  $S(t) = \Pr(T > t)$  denotes the survival probability till the time  $t$ . Then the corresponding hazard function  $h(t)$  is given as

$$h(t) = \lim_{\Delta t \rightarrow 0} \frac{\Pr(t \leq T \leq t + \Delta t | T > t)}{\Delta t} \quad (1)$$

Cox Proportional Hazards (CPH) model [1], [2] assumes that the ratio of an individual hazard to a population-level baseline hazard is a time-invariant scalar factor, *i.e.*,  $h(t|x) = h_0(t)e^{f(x)}$ , where  $f(x) = \theta^T x$  denotes log-hazard ratio.  $\theta \in \mathbb{R}^m$  denotes the coefficient vector of covariates. The CPH estimates the coefficient vector  $\theta$  for no ties survival data via maximizing the partial likelihood function given as

$$\mathcal{L}_C = \prod_{i=1}^k \frac{e^{\hat{y}_{q(t_i)}}}{\sum_{j \in R(t_i)} e^{\hat{y}_j}} \quad (2)$$

where  $\hat{y}_i = f(x_i) = \hat{\theta}^T x_i$  denotes the predicted log-hazard ratio of individual  $i$ , and  $q(t_i)$  denotes an individual with an event observed at time  $t_i$ . When the survival data has ties, the Breslow approximation [18] of the partial likelihood function is exploited as

$$\mathcal{L}_B = \prod_{t \in D} \frac{e^{\sum_{j \in q(t)} \hat{y}_j}}{\left[ \sum_{j \in R(t)} e^{\hat{y}_j} \right]^{C_t}} \quad (3)$$

### C. Random Survival Forest

Random Survival Forest (RSF) is derived from Random Forest. RSF uses log-rank test [29] for splitting survival trees, and estimates survival function associated with a terminal node by the Kaplan-Meier (KM) estimator. It calculates the cumulative hazard function associated with a terminal node by the Nelson-Aalen estimator. For an individual, RSF estimates the cumulative hazard function by averaging over the terminal node statistic of all trees.

Based on the nonparametric estimation method of survival and risk function in survival analysis, RSF is no longer limited to the hypothesis of CPH.

### D. Gradient Boosting Machine

Gradient Boosting Machine (GBM) [7] generates a new decision tree at each iteration to learn the “residual” of the prediction from the previous model. GBM calculates the sum of the predicted values on the fitted tree of each round as the output. At each iteration, a decision tree model  $\mathcal{F}_t(x)$ , as presented in (4), is built to minimize the loss function  $l$ .

$$\mathcal{F}_t = \operatorname{argmin}_{\mathcal{F}_t} \sum_{i=1}^n l(y_i, \hat{y}_i^{(t-1)} + \mathcal{F}_t(x_i)) \quad (4)$$

GBM has been implemented in R package *gbm* [36]. Under the framework of GBM, the algorithm optimizes (3), and takes the negative gradient of loss function with respect to the prediction of the previous model as the approximation of the “residual” [26]. The negative gradient, as a working response, is fitted using a regression tree at each iteration. The optimal estimation  $\rho_k$  of the terminal node  $k$  is computed by

$$\rho_k = \operatorname{argmin}_{\rho} \sum_{i \in S_k} l(y_i, \hat{y}_i^{(t-1)} + \rho) \quad (5)$$

where  $S_k$  is the set of individuals that defines terminal node  $k$ . (5) can be solved using the Newton-Raphson algorithm.

For an individual, GBM estimates the log-hazard ratio by adding up the prediction of each regression tree and then estimates the hazard and survival function. The GBM survival model still follows the assumption of hazard ratio, while its hypothesis space is based on regression trees instead of linear space. Therefore, it can capture the complex relationship between log-hazard ratio and covariates.

### E. XGBoost

The XGBoost uses a more precise approximation of learning objective and adds regularization term to avoid overfitting effectively. It proposes a novel algorithm for tree node splitting and leaf node estimation. The regularized learning objective at the  $t$ -th iteration in XGBoost is

$$\mathcal{L}^{(t)} = \sum_{i=1}^n l(y_i, \hat{y}_i^{(t-1)} + \mathcal{F}_t(x_i)) + \Omega(\mathcal{F}_t) \quad (6)$$

The second-order approximation can be used to quickly optimize the objective in the general setting, as shown in (7), where  $g_i = \frac{\partial l(y_i, \hat{y}_i^{(t-1)})}{\partial \hat{y}_i^{(t-1)}}$  and  $h_i = \frac{\partial^2 l(y_i, \hat{y}_i^{(t-1)})}{\partial \hat{y}_i^{(t-1)^2}}$  are first- and



second-order gradient statistics on the loss function.

$$\mathcal{L}^{(t)} \simeq \sum_{i=1}^n \left( l \left( y_i, \hat{y}^{(t-1)} \right) + g_i \mathcal{F}_t(x_i) + \frac{1}{2} h_i \mathcal{F}_t^2(x_i) \right) + \Omega(\mathcal{F}_t) \quad (7)$$

After modifying and transforming the above (7) from the perspective of the regression tree, the node splitting strategy and the terminal node estimation formula can be obtained. Their expressions contain  $g_i$  and  $h_i$ . After calculating first- and second-order gradient statistics, the XGBoost algorithm will use it to find the node splitting which minimizes the loss function, and calculate the terminal node estimation when fitting the new regression tree at each iteration (let  $M$  be the total number of iterations). As same as the procedure of GBM, the final model adds the estimated quantities of all regression trees at the terminal node to predict the log-hazard ratio using (8).

$$\hat{y}_i = f(x_i) = \sum_{t=1}^M \mathcal{F}_t(x_i) \quad (8)$$

The GBM method performs the first-order expansion of the loss function and employs a negative gradient as an optimization strategy. The XGBoost method uses a more precise approximation to get utterly different optimization strategies. The regularization items in XGBoost can effectively avoid overfitting. Same as the GBM method, XGBoost is still based on the hazard ratio assumption, taking the Breslow approximation as the loss function for ties, which can represent complex nonlinear functions.

### III. EXSA METHOD

In the proposed method, we take a more precise approximation of partial likelihood function as a learning objective, then derive the formula of the gradient for the new learning objective. It is provided to optimize the original XGBoost method for survival analysis. As demonstrated in Part E of Section II, we must customize a specific loss function based on the survival data and derive the first- and second-order gradient of the loss function. In this paper, we take the Efron approximation of partial likelihood function as a custom loss function and derive the simplified mathematical expression of the gradients. At the same time, we propose a theorem related to gradient derivation to represent the core optimization part of EXSA. The implementation of EXSA method, named EfnBoost in our libsurv package on Github, is available at <https://github.com/liupeil01/libsurv>.

#### A. Efron Approximation for Ties

When the number of tied events for any point of time is relatively large, Efron method [19] can give a better approximation than Breslow method. The loss function of the Efron method is as

$$\mathcal{L} = \prod_{t \in D} \frac{e^{\sum_{j \in q(t)} \hat{y}_j}}{\prod_{l=1}^{C_t} \left[ \sum_{j \in R(t)} e^{\hat{y}_j} - \frac{l-1}{C_t} \sum_{j \in q(t)} e^{\hat{y}_j} \right]} \quad (9)$$

We define  $SR(t) = \sum_{j \in R(t)} e^{\hat{y}_j}$  as the sum of estimated hazard ratio of individuals who are still at risk at time  $t$ , and  $SD(t) = \sum_{k \in q(t)} e^{\hat{y}_k}$  as the sum of estimated hazard ratio of individuals whose events are observed at time  $t$ . Take negative logarithms on both sides of (9), and then we get its loss function as

$$L_E = \sum_{t \in D} \left\{ \sum_{j \in q(t)} [\ln(SR(t) - w_j * SD(t)) - \hat{y}_j] \right\} \quad (10)$$

where  $w_j = (l-1)/C_t, l=1, \dots, C_t$ , which denotes the assigned unique weight of each individual in  $q(t)$ .

To calculate the first- and second-order gradients of  $L_E$ , we give two definitions. Definition 1 is given to express the form of the first-order gradient.

*Definition 1:* Let  $\alpha(t), \beta(t)$  be two functions with respect to time  $t$ .

$$\alpha(t) = \sum_{j \in q(t)} 1/[SR(t) - w_j * SD(t)]$$

$$\beta(t) = \sum_{j \in q(t) w_j} 1/[SR(t) - w_j * SD(t)]$$

To derive the form of the second-order gradient, we give Definition 2 as follows.

*Definition 2:* Let  $\varphi(t), \omega(t)$  be two functions with respect to time  $t$ .

$$\varphi(t) = \sum_{j \in q(t)} 1/[SR(t) - w_j * SD(t)]^2$$

$$\omega(t) = \sum_{j \in q(t)} [1 - (1 - w_j)^2] / [SR(t) - w_j * SD(t)]^2$$

Based on Definition 1, we can get Theorem 1.

*Theorem 1:* For individual  $i$ , observed indicator variable  $\delta_i$ , then the first-order gradients of  $L_E$  with respect to  $\hat{y}_i$  is

$$g_i = \begin{cases} e^{\hat{y}_i} \{ [\sum_{t \leq T_i} \alpha(t)] - \beta(T_i) \} - 1, & \text{if } \delta_i = 1 \\ e^{\hat{y}_i} \sum_{t \leq T_i} \alpha(t), & \text{if } \delta_i = 0 \end{cases}$$

*Proof:* Referring to (10), for individual  $i$ , we derive the gradient with respect to individual prediction  $\hat{y}_i$ . On the one hand,  $\hat{y}_i$  occurs in all  $SR(t)$  if the individual  $i$  is still at risk at a time point  $t$  ( $t \in D$ ). Therefore, we only need to consider the part of  $L_E$  where  $t \leq T_i$  when deriving gradient of the individual  $i$ ; On the other hand,  $\hat{y}_i$  also occurs in  $SD(t)$  if individual  $i$  is with  $\delta_i = 1$  and  $T_i = t$ . Therefore, we discuss the following two cases.

1) For individual  $i$ , If an event occurs, i.e.,  $\delta_i = 1$ .

We consider the two parts of  $L_E$  separately, where  $t = T_i$  and  $t < T_i$ . Using the chain rule, the first-order gradient of them are as follows, respectively.

$$g_i|_{t=T_i, \delta_i=1} = e^{\hat{y}_i} \sum_{j \in q(T_i)} \frac{1 - w_j}{SR(T_i) - w_j * SD(T_i)} - 1$$

$$g_i|_{t < T_i, \delta_i=1} = e^{\hat{y}_i} \sum_{t < T_i} \left[ \sum_{j \in q(t)} \frac{1}{SR(t) - w_j * SD(t)} \right]$$

2) For individual  $i$ , if an event isn't observed, *i.e.*,  $\delta_i = 0$ .

Since the prediction  $\hat{y}_i$  does not occur in  $SD(T_i)$ , the first-order gradient is similar to  $g_i|_{\delta_i=1}$ , which is

$$g_i|_{\delta_i=0} = e^{\hat{y}_i} \sum_{t \leq T_i} \left[ \sum_{j \in q(t)} \frac{1}{SR(t) - w_j * SD(t)} \right]$$

From Definition 1, the first-order gradient of  $L_E$  with respect to  $\hat{y}_i$  can be summarized as

$$\frac{\partial L_E}{\partial \hat{y}_i} = g_i = \begin{cases} e^{\hat{y}_i} \{ [\sum_{t \leq T_i} \alpha(t)] - \beta(T_i) \} - 1, & \text{if } \delta_i = 1 \\ e^{\hat{y}_i} \sum_{t \leq T_i} \alpha(t), & \text{if } \delta_i = 0 \end{cases}$$

□

Then, based on Definition 2 and the above  $g_i$ , we can derive the second-order gradient and give Theorem 2.

**Theorem 2:** For individual  $i$ , observed indicator variable  $\delta_i$ , then the second-order gradient of  $L_E$  with respect to  $\hat{y}_i$  is

$$h_i = \begin{cases} g_i|_{\delta_i=1} - (e^{\hat{y}_i})^2 \{ [\sum_{t \leq T_i} \varphi(t)] - \omega(T_i) \} + 1, & \text{if } \delta_i = 1 \\ g_i|_{\delta_i=0} - (e^{\hat{y}_i})^2 \sum_{t \leq T_i} \varphi(t), & \text{if } \delta_i = 0 \end{cases}$$

**Proof:** From Theorem 1, we get the first-order gradient of  $L_E$ . Then, we can derive the second-order gradient of  $g_i$ . Similarly, we discuss the following two cases.

1) If  $\delta_i = 1$ , using the chain rule, the second-order gradient of  $L_E$  with respect to  $\hat{y}_i$  is

$$\begin{aligned} h_i|_{\delta_i=1} &= g_i|_{\delta_i=1} + 1 - (e^{\hat{y}_i})^2 \\ &\quad \times \sum_{j \in q(T_i)} \frac{(1 - w_j)^2}{SR(T_i) - w_j * SD(T_i)} - (e^{\hat{y}_i})^2 \\ &\quad \times \sum_{t < T_i} \left[ \sum_{j \in q(t)} \frac{1}{[SR(t) - w_j * SD(t)]^2} \right] \end{aligned}$$

2) If  $\delta_i = 0$ , we can get

$$\begin{aligned} h_i|_{\delta_i=0} &= g_i|_{\delta_i=0} - (e^{\hat{y}_i})^2 \sum_{t \leq T_i} \\ &\quad \times \left[ \sum_{j \in q(t)} \frac{1}{[SR(t) - w_j * SD(t)]^2} \right] \end{aligned}$$

With Definition 2, the second-order gradient of  $L_E$  with respect to  $\hat{y}_i$  is written as following simplified form.

$$\frac{\partial^2 L_E}{\partial \hat{y}_i^2} = h_i = \begin{cases} g_i|_{\delta_i=1} - (e^{\hat{y}_i})^2 \{ [\sum_{t \leq T_i} \varphi(t)] - \omega(T_i) \} + 1, & \text{if } \delta_i = 1 \\ g_i|_{\delta_i=0} - (e^{\hat{y}_i})^2 \sum_{t \leq T_i} \varphi(t), & \text{if } \delta_i = 0 \end{cases}$$

□

## B. Algorithm Implementation

The algorithm pseudo-code for computing the loss function is shown in Algorithm 1. For each time  $t$  in all different time points

### Algorithm 1: Computing Loss Function.

**Input:**  $\hat{y}$ , predictions of individuals on log hazard ratio scale;  $\delta, T$ , the survival data of individuals.

**Output:**  $loss$

```

1: loss = 0
2: for  $t$  in  $D$ 
3:    $SR_t = \sum_{k \in R(t)} e^{\hat{y}_k}$ 
4:    $SD_t = \sum_{k \in q(t)} e^{\hat{y}_k}$ 
5:    $C_t = |q(t)|$ 
6:    $l = 1$ 
7:   for  $j$  in  $q(t)$ 
8:      $w_j = (l - 1)/C_t$ 
9:      $l = l + 1$ 
10:   loss = loss +  $\ln(SR_t - w_j * SD_t) - \hat{y}_j$ 
11: end
12: end
13: return loss

```

of event occurrence ( $t \in D$ ), we compute the value of  $SR(t)$ ,  $SD(t)$ , and  $C_t$ . Then for each individual with events observed at the time  $t$ , we add the individual's contribution to the loss value. Finally, we obtain the value of loss function for the model prediction.

For Theorem 1 and 2, we give the algorithm pseudo-code for computing the gradient of loss function in Algorithm 2. Firstly, we get the array  $A_t$  by sorting the time  $T$  in survival data and removing duplicates. Then for each time  $t$  in  $A_t$ , we compute the value of  $SR(t)$ ,  $SD(t)$ ,  $\alpha_t$ ,  $\beta_t$ ,  $\varphi_t$  and  $\omega_t$  for the following gradient computation. For each individual whose observed time equals to  $t$ , we compute the gradients of  $L_E$  with respect to its prediction by using the formulas shown in Theorem 1 and 2.

## C. Algorithm Performance

In order to deal with tied events in CPH model estimation, it is necessary to adjust partial likelihood function appropriately. Kalbfleisch and Prentice [24] proposed the most natural way to consider all possible orders of event happened at the same survival time for individuals. The exact expression is shown in (11), where  $Q_t$  denotes the set of  $C_t!$  permutations for  $C_t$  events observed at time  $t$ .  $P = \{p_1, p_2, \dots, p_{C_t}\}$  is one element of  $Q_t$ . Maximization of (11) might be time-consuming, especially where there is a large number of ties, *i.e.*, a certain  $C_t$  is large.

$$\mathcal{L} = \prod_{t \in D} \frac{e^{\sum_{j \in q(t)} \hat{y}_j}}{\sum_{P \in Q_t} \prod_{l=1}^{C_t} \left[ \sum_{j \in R(t) - \{P_1, \dots, P_{l-1}\}} e^{\hat{y}_j} \right]} \quad (11)$$

In comparison to the exact expression, Breslow approximation does not consider orders of event occurrences for individuals having the same survival time, but directly take the sum of their estimated hazard ratio as the denominator. Efron approximation also does not consider the order, but it assigns different weights to the estimated hazard ratio of tied individuals. By comparing Efron, Breslow approximation with exact expression of the partial likelihood function (as shown in (3), (9), and (11), respectively), we can find Efron approximation is more precise

**Algorithm 2:** Computing the Gradient of Loss Function.

**Input:**  $\hat{y}$ , predictions of individuals on log hazard ratio scale;  $\delta, T$ , the survival data of individuals.

**Output:**  $g, h$

```

1:  $A_t = \text{sort}(\text{unique}(T))$  # sort all time and remove
   duplicates
2:  $C_1 = 0$ 
3:  $C_2 = 0$ 
4: for  $t$  in  $A_t$ 
5:    $SR_t = \sum_{k \in R(t)} e^{\hat{y}_k}$ 
6:    $SD_t = \sum_{k \in q(t)} e^{\hat{y}_k}$ 
7:   # pre-calculation for the first-order gradient
8:    $\alpha_t = \sum_{k \in q(t)} 1/(SR_t - w_k * SD_t)$ 
9:    $\beta_t = \sum_{k \in q(t)} w_k/(SR_t - w_k * SD_t)$ 
10:   $C_1 = C_1 + \alpha_t$ 
11:  # pre-calculation for the second-order gradient
12:   $\varphi_t = \sum_{k \in q(t)} 1/(SR_t - w_k * SD_t)^2$ 
13:   $\omega_t = \sum_{k \in q(t)} [1 - (1 - w_k)^2]/(SR_t - w_k * SD_t)^2$ 
14:   $C_2 = C_2 + \varphi_t$ 
15:  # all patients whose observed time equals to t
16:  for  $j$  in  $N(t)$ 
17:    if  $\delta_j = 0$ 
18:       $g_j = e^{\hat{y}_j} * C_1$ 
19:       $h_j = g_j - e^{\hat{y}_j} * e^{\hat{y}_j} * C_2$ 
20:    else if  $\delta_j = 1$ 
21:       $g_j = e^{\hat{y}_j} * (C_1 - \beta_t) - 1$ 
22:       $h_j = g_j + 1 - e^{\hat{y}_j} * e^{\hat{y}_j} * (C_2 - \omega_t)$ 
23:    end
24:  end
25: end
26: return  $g, h$ 

```

than Breslow approximation, and the deviation of Breslow approximation will be larger as the number of ties increases, which results in worse model fit. If the survival data has no ties, then three expressions of the partial likelihood function are reduced to the same form as shown in (2).

#### IV. DATA MATERIALS

##### A. Clinical Data Studies

Clinical data materials are provided by the CRCB in the West China Hospital of Sichuan University. Breast cancer data of patients were retrospectively collected from 1989 to 2007, while since 2008 all patients confirmed by pathology in the West China Hospital were recorded in electronic health records (EHRs) and followed up to obtain prognosis and surviving state. Those EHRs include the basic information, such as gender, education, and medical insurance, diagnosis information, personal medical history, menstruation, pathology, surgery, chemotherapy, radiotherapy, endocrine therapy and targeted therapy, etc. The outcomes including invasive Disease-Free Survival (iDFS), Breast Cancer-Specific Survival (BCSS) and Overall Survival (OS) were followed up and recorded in the Breast Cancer Information Management System (BCIMS). We take iDFS as the outcome

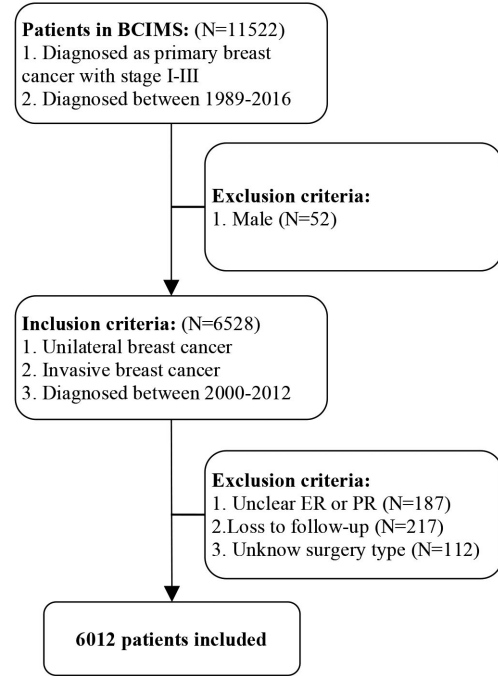


Fig. 1. Description of the included cohorts.

to build the model and test the proposed algorithms. Therefore, the disease progress refers to iDFS in this paper.

By May 2017, the total number of breast cancer patients in BCIMS is 12119. In order to ensure the quality of the clinical data, we carry out a manual inspection for the missing or unstandardized data. Moreover, we regularly follow up patients and reduce the loss to follow-up rate to be 0.4% in 2017. In this work, to get the survival data for modeling, we only considered 11522 patients diagnosed as primary breast cancer with stage I-III between 1989-2016 in BCIMS, and retained: (a) Patients with stage I-III, primary, and invasive breast cancer; (b) Patients initially diagnosed with unilateral breast cancer; (c) Patients with known receptor status and other information.

The total of 6012 patients were eligible for further study. The detailed screening criteria in this study can refer to Fig. 1.

##### B. Data Preprocessing

We usually run up against obstacles in clinical breast cancer data in real environments, because of massive volumes of the missing, abnormal, duplicate, and inconsistent data. Therefore, to ensure data quality, we have to take appropriate measures to reduce data problems as much as possible, which is the same as our previous work referring to [20].

After cleaning the data, we get 6012 patients with 89 features including personal basic information, diagnosis, medical history, pathology, treatment, and follow-up. Then we select 23 important features out of 89 features by means of MP4Ei framework [20]. Based on recommendations of breast cancer clinical medical experts, we tweak a few features that may not be in line with medical practice. Four features (Residence, Age at marriage, Histological grade and Surgery type) are included

TABLE I  
STATISTICS FOR MISSING FIELDS

No. of Missing fields	Excluded patients	Included patients	Log-rank statistics	Log-rank p-value
$\geq 13$	12	6000	0.000	0.98505
$\geq 12$	64	5948	0.001	0.97788
$\geq 11$	288	5724	0.095	0.75732
$\geq 10$	673	5339	0.397	0.52871
$\geq 9$	1127	4885	0.196	0.65766
$\geq 8$	<b>1437</b>	<b>4575</b>	<b>0.340</b>	<b>0.55963</b>
$\geq 7$	1526	4486	2.762	0.09654
$\geq 6$	1612	4400	12.285	0.00046
$\geq 5$	1684	4328	21.928	<0.0001
$\geq 4$	1767	4245	30.032	<0.0001
$\geq 3$	1937	4075	35.975	<0.0001
$\geq 2$	2484	3528	43.814	<0.0001
$\geq 1$	3718	2294	48.685	<0.0001

The log-rank method is used to test the difference between the survival state of 6012 patients and included patients.

in while three features (CK5/6, Number of I-III lymph node metastases and anatomical stage) are excluded. Therefore, the final 24 features related to the tumor, basic information, and treatment are screened out to the model, as shown in Table II.

As a prognostic prediction model based on the classification algorithm, MP4Ei framework aims to predict the probability of disease progress in early breast cancer patients at 5-year. A clinical dataset with 4196 patients is adopted to train the classification model of MP4Ei framework. However, the survival prediction model established in this paper is mainly to predict the survival status of disease progress in breast cancer patients at various time points after the initial diagnosis. It is important to evaluate the risk score of disease progress via survival analysis methods. Hence, we focus on disease progress in breast cancer after initial diagnosis by taking into account the status of disease progress  $\delta$  and the time of disease progress or lost follow-up (in quarters)  $T$ . A dataset of follow-up cohort should be used to train the survival prediction model to predict disease progress or survival state.

To reduce the effect of noise and deviation in the dataset to impact the performance and stability of the survival prediction model, we exclude some patients who have too many features with missing values. The log-rank method [29] under the null hypothesis is used to test whether two intensity processes are similar. The total number of missing fields for patients and related statistics in the dataset are shown in Table I, and the missing fields refer to the missing predictor variables. For example, if we take the number of missing fields greater than or equal to 7, a total of 1526 patients are excluded. And the corresponding p-value of the log-rank test will increase rapidly relative to the number of missing fields greater than or equal to 8. Therefore, we take eight as a cutoff value for missing fields, and a total of 1437 patients with the number of missing fields greater than or equal to 8 are excluded. The remaining 4575 patients are eligible for further study. The survival curves of patients before and after exclusion are shown in Fig. 2. The log-rank test [29] shows that there is no significant difference between curves ( $p > 0.05$ ). And  $T_e = 50$  is taken as the study endpoint of disease progress after initial diagnosis in breast cancer.

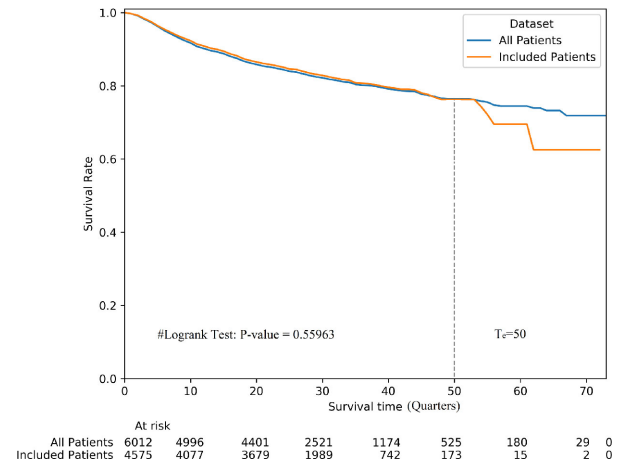


Fig. 2. Survival curves of patients before and after exclusion. The number of patients at risk at various time points is demonstrated under the figure. The dotted line indicates the study endpoint.

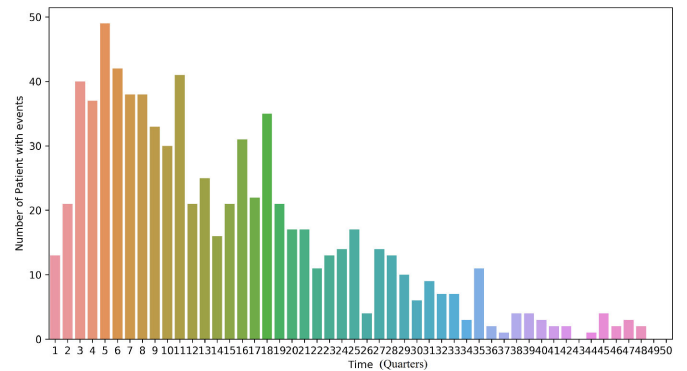


Fig. 3. The number of ties at various time points.

Since 4575 cases are taken from the periodic follow-up data, multiple patients may occur events at the same time. Fig. 3 shows the number of ties at various time points in the cohort data. We can find that the number of ties reaches a maximum of 49 in 5 quarters, and it is greater than 10 at most time points.

The random sampling method is adopted to divide 4575 cases into the training set and test set by 7:3. There is no significant difference between the training and test set ( $p > 0.05$ ), as shown in Table II. In Fig. 4, the training set contains 3202 cases, and the rates of freedom from disease progress at 5-year, 10-year, and study endpoint are 86.51%, 79.47%, and 75.69%, respectively. The test set contains 1373 cases, and the rates of freedom from disease progress at 5 years, 10 years, and study endpoint are 86.59%, 79.92%, and 77.83%, respectively. Kaplan-Meier method is adopted to estimate the rate of freedom from disease progress of patients.

## V. RESULTS

We build a prognostic model for predicting disease progress after initial diagnosis in breast cancer and compare it with the original XGBoost method, the classical GBM, RSF, and Cox methods in the realm of survival analysis. Our proposed



TABLE II  
DIFFERENCE SIGNIFICANCE TEST FOR THE TRAINING AND TEST SET

Columns	Training Set (3202 cases)		Test Set (1373 cases)		Implication	P-value (>0.05)
	Mean	Std.	Mean	Std.		
Age_at_diagnosis	48.28	10.38	48.08	10.17	age at diagnosis	0.92
Education_level	2.31	1.34	2.28	1.34	education level	0.80
Age_at_marriage	2.17	1.20	2.18	1.34	age at marriage	0.10
Reimbursement_rate	4.17	2.20	4.19	2.13	reimbursement rate	0.49
BMI	2.22	1.00	2.16	1.04	body mass index	0.29
Residence	1.22	0.43	1.24	0.43	residence	0.46
Menopausal_status	0.39	0.49	0.39	0.49	menopausal status at diagnosis	0.20
N_of_pregnancies	2.53	1.29	2.50	1.30	number of pregnancies	0.97
N_of_abortions	1.56	1.83	1.55	1.81	number of abortions	0.96
Age_at_birth_of_first_child	2.53	1.85	2.53	1.86	age at birth of first child	0.47
Gap_menarch_birth	2.63	1.91	2.64	1.93	the interval between menarche and birth	0.50
Gap_menarche_menopause	6.86	4.03	6.90	4.04	the interval between menarche and menopause	0.57
Tumor_stage	2.49	1.21	2.49	1.23	clinical or pathological tumor stage	0.90
Node_stage	1.83	1.03	1.84	1.06	clinical or pathological lymph node stage	0.75
Ki67	0.63	0.61	0.63	0.61	Ki67 index	0.98
Tumor_grade	1.68	1.63	1.70	1.63	histological grade	0.65
Receptor_status	2.90	1.24	2.90	1.22	status of ER, PR, and HER2	0.32
Surgery	2.99	0.36	2.99	0.35	surgery type	0.91
CT_compliance	1.40	2.01	1.36	1.96	chemotherapy compliance	0.12
NACT	8.74	3.03	8.81	2.97	neoadjuvant chemotherapy	0.38
ACT	2.32	2.16	2.24	2.14	adjuvant chemotherapy	0.13
Radiotherapy	0.37	0.48	0.36	0.48	Radiotherapy	0.67
AHT	5.38	3.56	5.40	3.53	adjuvant hormonal therapy	0.28
AHT_compliance	5.34	3.71	5.32	3.72	adjuvant hormonal therapy compliance	0.91
dp_bin	0.17	0.38	0.17	0.38	status of disease progress	0.96
dp_time	27.34	12.16	27.77	11.80	time of disease progress	0.96

Std. = Standard Deviation. The status and time of disease progress are combined and tested by log-rank method in survival analysis. For continuous variables and discrete categorical variables, the Kolmogorov-Smirnov test and Chi-square test are used to calculate the p-value, respectively. Except for the variable "Age\_at\_diagnosis", all the features in the table are all discretized and encoded as categorical variables.

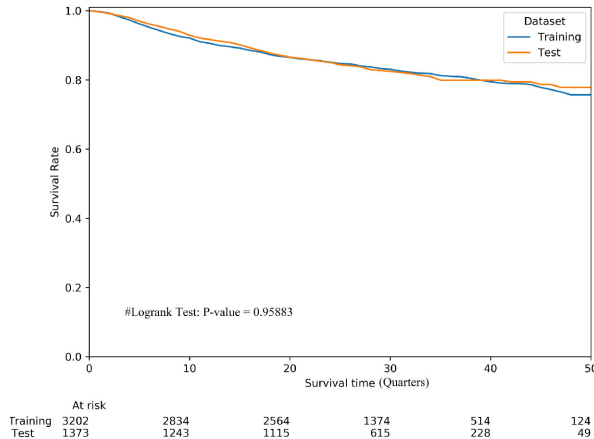


Fig. 4. Survival curves of the training set and test set. The number of patients at risk at various time points is under the figure.

EXSA method provides highly competitive performance on the independent test set. Further, considering the application of the model in clinical practice, we exploit the prognostic model to evaluate risk scores of disease progress and to demonstrate risk grouping and continuous function between risk score and rate of disease progress.

#### A. Method Performance

In the experiment, hyper-parameters are tuned by the scientific Bayesian hyper-parameter optimization algorithm, which

is running in the python package *hyperopt* [21], [22]. In the realm of survival analysis, concordance index (CI or C-index) and time-dependent AUC [23] are important metrics to evaluate the performance of survival prediction model. C-index involves the prediction of right-censored patients to measure the overall performance of the survival prediction model, which ranges in an interval [0.0, 1.0]. Time-dependent AUC defines an ending time (denoted as  $t$ ) of interest, and then compare the predicted probabilities with the actual binary survival status at the interested time  $t$ , while the right-censored patient may not be involved in. It measures the power of discriminant, which has a range of (0.9–1.0), (0.8–0.9), and (0.7–0.8) for excellent, very good, and good diagnosis accuracy, respectively.

In the experiment, we compare the proposed EXSA method with Cox, RSF, GBM, and XGBoost. The process of the hyper-parameters tuning is conducted on the training set (3202 cases) using 10-fold cross-validation repeated three times. Hyper-parameters of each model are carefully tuned by Bayesian hyper-parameter optimization algorithm. The search space of each model's hyper-parameter is given in Table III. Each set of parameters of a model is evaluated by the average C-index on the cross-validation set. The result of each parameter in Table III is the value obtained by maximizing the average C-index in the search space. For the meaning of each parameter, readers can refer to the R package: *randomForestSRC*, *gbm*, and *xgboost*.

After getting the optimal parameters of each model, we fit each model with the training set and then evaluate them in the independent test set (1373 cases) by using the metrics of C-index and Time-dependent AUC. The evaluation results of



TABLE III  
HYPER-PARAMETERS SEARCH SPACE

Model	Parameters	Range	Step	Result
RSF	ntree	[80, 150]	10	140
	mtry	[12, 24]	1	12
	nodesize	[5, 40]	5	25
	shrinkage	[0.001, 0.01]	0.001	0.008
GBM	n.trees	[1000, 5000]	1000	2000
	interaction.depth	[1, 5]	1	5
	n.minobsinnode	[5, 50]	5	10
	eta	[0.01, 0.10]	0.01	0.07/0.07
	nrounds	[80, 150]	10	140/140
	max_depth	[2, 6]	1	3/3
XGBoost/EXSA	min_child_weight	[0.0, 1.0]	-	0.81/0.66
	subsample	[0.4, 1.0]	0.1	0.9/0.9
	colsample_bytree	[0.4, 1.0]	0.1	0.5/0.5
	lambda	[0.0, 1.0]	-	0.95/0.88
	gamma	[0.0, 1.0]	-	0.49/0.92

Meaning of each parameter can refer to R package: randomForestSRC, gbm, and xgboost.

TABLE IV  
METHOD PERFORMANCE

Method	CI $\pm$ Std.	AUC at 5-Year	AUC at 10-Year
Cox	0.75912 $\pm$ 0.03289	0.76018	0.71543
RSF	0.81435 $\pm$ 0.02633	0.81198	0.76363
GBM	0.82992 $\pm$ 0.02432	0.82563	0.77513
XGBoost	0.83448 $\pm$ 0.02355	0.83368	0.77713
EXSA	<b>0.83454 <math>\pm</math> 0.02361</b>	<b>0.83851</b>	<b>0.78155</b>

Std. = Standard Deviation.

each model are shown in Table IV. We can see that the C-index of EXSA is 0.83454, and the 5-year and 10-year AUC are 0.83851 and 0.78155, respectively. The C-index of original XGBoost is 0.83448, and the 5-year and 10-year AUC are 0.83368 and 0.77713, respectively. Therefore, from the point of view of the C-index, the overall performance of EXSA and XGBoost model are very close, while the optimized EXSA model is of more powerful discriminant than XGBoost model from the point of view of the time-dependent AUC. That experimentally demonstrates the dominance and effectiveness of the proposed method. Similarly, compared with the other three classical models, the EXSA model is optimal on all three metrics. Time-dependent ROC curves of each model at 5-year and 10-year are shown in Fig. 5. In the further study about feature importance and model's applications, we will exploit the best survival prediction model based on EXSA method for experimentation.

To observe the fitting of the EXSA model in the gradient boosting process, we represent the loss function value and C-index of the model at each step of iterations, as shown in Fig. 6. The loss function of the model is  $L_E$  given by (10) in Section III. We set the results shown in Table III as the hyper-parameters of the model. The number of iterations is changed from 140 to 200. As illustrated by Fig. 6, we can see that when the 140-th iteration is completed, the loss function value and C-index of the model is stable when testing, which implies the model with a good generalization property effectively fits the training set. After the 140-th iteration, loss function values and C-index of the model are increased due to over-fitting. Therefore, the number of iterations of 140 given by Bayesian hyper-parameter

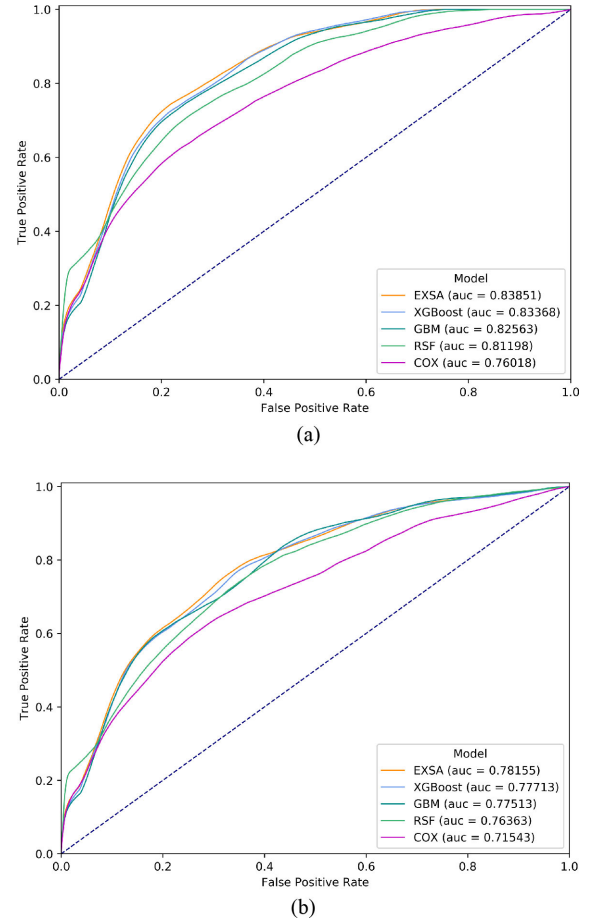


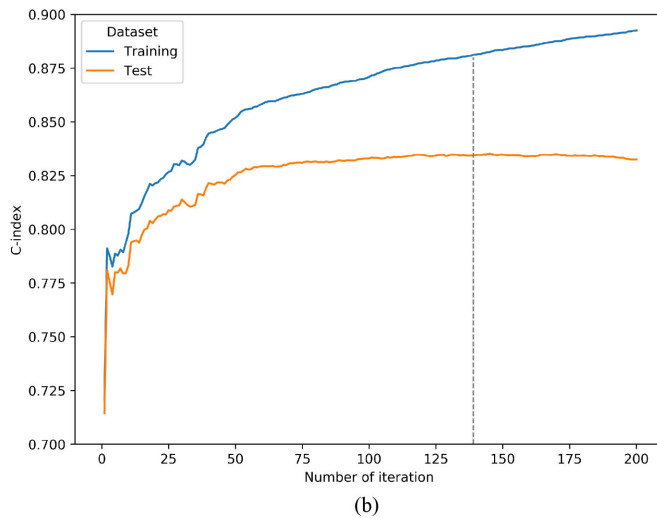
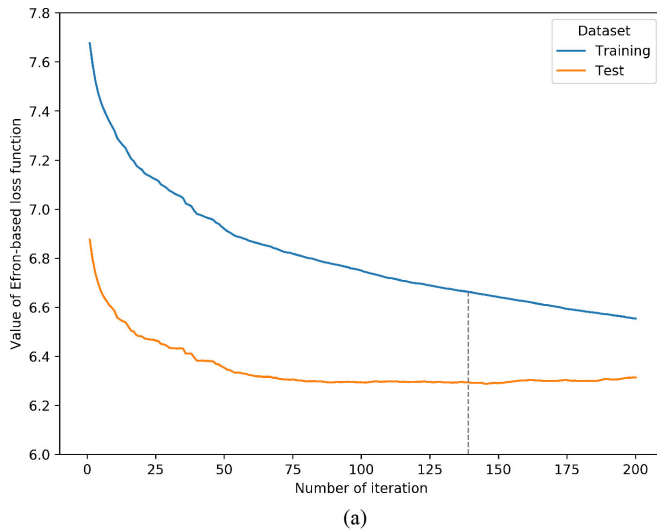
Fig. 5. Time-dependent ROC curves. AUC = Area Under Curve. (a) At 5-year. (b) At 10-year.

optimization algorithm is optimal, and the EXSA model does not over-fit in the test set.

## B. Evaluation of Feature Importance

To find features that have an important effect on disease progress of breast cancer patients, we evaluate the importance of 24 features (in Table II) using the EXSA model. As described in the previous section, we set the parameters of the EXSA model to be the same as the optimal searching results, then repeat model to fit the training set for 30 times. Finally, average scores of features are computed by the "weight" method provided in the library *xgboost*. As shown in Fig. 7, 24 features are ranked according to the average scores of 30 experiments, where larger values indicate a more important effect.

The top 5 important features are adjuvant hormonal therapy compliance, age at diagnosis, reimbursement rate, clinical or pathological lymph node stage, and adjuvant hormonal therapy, respectively. Among them, the clinical or pathological lymph node stage is generally considered as an important clinical factor. In addition to the tumor and treatment features, host-related features should also be concerned, such as reimbursement rate and the interval between menarche and birth. In our previous paper [20], the proposed MP4Ei framework focused on 5-year

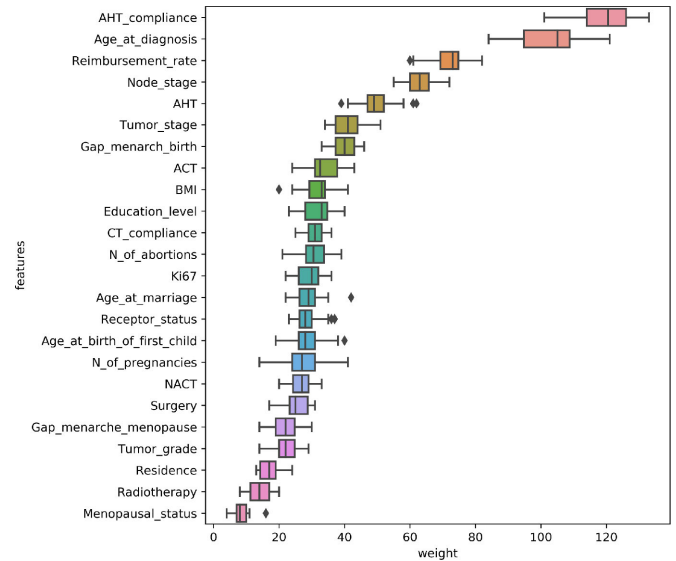


**Fig. 6.** Loss values and C-index of the EXSA model on the training and test set at the different number of iterations. The dotted line in figure indicates the optimal iteration, provided by the Bayesian hyper-parameter optimization algorithm. (a) Loss value. (b) C-index.

disease progress of breast cancer gives the top 5 important features are: age at diagnosis, the status of ER&PR&HER2, reimbursement rate, adjuvant hormonal therapy, and clinical or pathological tumor stage. Those are roughly consistent with the results of feature importance ranking obtained by the EXSA model.

### C. Application of Survival Prediction Model

We study the predictive value (log-hazard ratio) of the EXSA model for disease progress prediction in training set (3202 breast cancer patients), with a minimum of  $-2.3393297$ , a maximum of  $5.618587$ , and a mean of  $0.141283$ . Referring to the risk grouping approach in Predicting Prostate Cancer Recurrence Risk [25], we use the quartiles of the predictive value in training set to classify 3202 patients into four risk groups, which are defined as low risk group, mid-low risk group, mid-high risk group, and high risk group.



**Fig. 7.** Scores of feature importance evaluated by the EXSA model. Fit the training set for 30 times. Larger values of weight indicate more important effect.

Considering the application of the EXSA model in clinical practice, the risk score is rescaled to (0, 100) from the predictive value in the training set according to Min-Max Scale. We obtain the corresponding cut-off points of the risk score for above four risk groups: 0–21.5 for low risk group, 21.6–29.6 for mid-low risk group, 29.7–38.8 for mid-high risk group, and 38.9–100 for high risk group. The survival curves of the four risk groups are shown in Fig. 8(a), and the statistics of them are shown in Table V. The median risk scores of risk groups from low to high are 15.3, 26.4, 33.5, and 48.4, respectively. The survival rates (rates of freedom from disease progress) at 5-year of risk groups from low to high are 99.87%, 97.37%, 92.89%, and 54.65%, respectively. The survival rates (rates of freedom from disease progress) at 10-year of risk groups from low to high are 97.63%, 93.94%, 83.09%, and 36.24%, respectively. The survival rate is estimated by the Kaplan-Meier method. It can be seen there is a significant difference in the survival status between all two adjacent risk groups ( $p < 0.001$ ), and larger risk score indicates higher risk and more prone to disease progress. That demonstrates the EXSA prognostic model is of a strong discriminative power for the patients in the training set.

To test the generalization ability of the model, we apply the cut-off points of risk grouping obtained from the training set to check patients in the independent test set (1373 breast cancer patients). Firstly, we get the predicted value (log-hazard ratio) in the test set using the trained model. Then the predicted value is scaled to risk score in the same way. We finally divide them into four risk groups according to the cut-off points. The survival curves of the four risk groups for testing are shown in Fig. 8(b), and the statistics of them are shown in Table VI. The number of patients of risk groups from low to high is 345, 326, 335, and 367, respectively. The proportion of each risk group is comparable as shown in the training set. The median risk score of risk groups

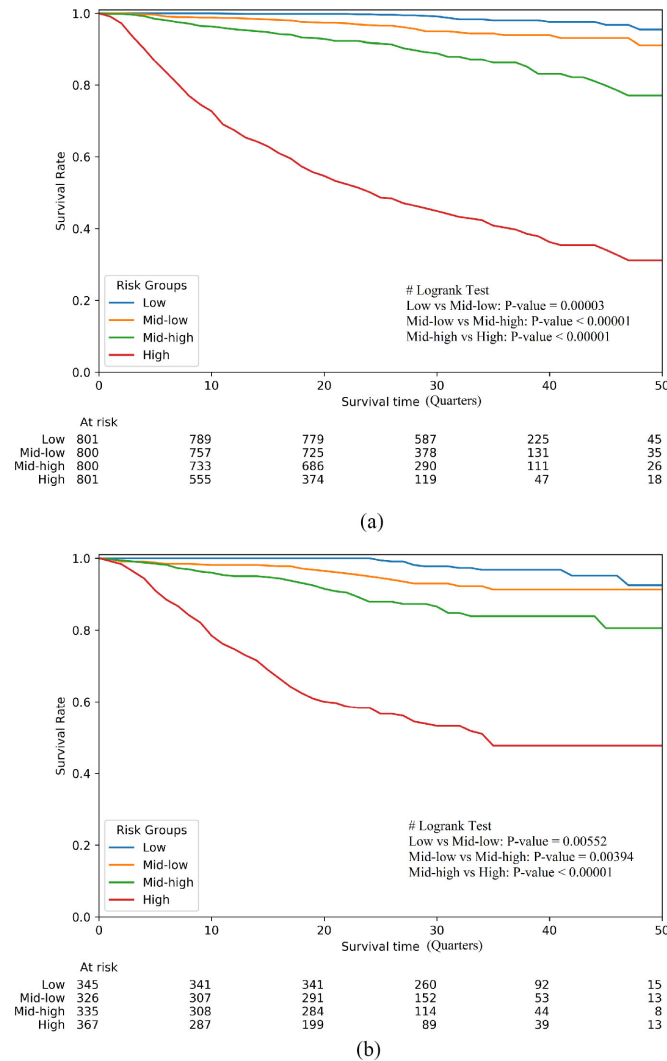


Fig. 8. Survival curves of four risk groups. The bottom of figure demonstrates the number of patients at risk at various time points. (a) Survival curves of the training set. (b) Survival curves of the test set.

TABLE V  
RISK GROUPING ON THE TRAINING SET

Risk Group	Risk Score	Median Risk Score	Number of Patients	Survival Rate at 5 Years	Survival Rate at 10 Years
Low	0-21.5	15.3	801	99.87%	97.63%
Mid-low	21.6-29.6	26.4	800	97.37%	93.94%
Mid-high	29.7-38.8	33.5	800	92.89%	83.09%
High	38.9-100	48.4	801	54.65%	36.24%

TABLE VI  
RISK GROUPING ON THE TEST SET

Risk Group	Risk Score	Median Risk Score	Number of Patients	Survival Rate at 5-Year	Survival Rate at 10-Year
Low	0-21.5	15.4	345	100.0%	96.77%
Mid-low	21.6-29.6	26.6	326	96.47%	91.30%
Mid-high	29.7-38.8	33.8	335	91.51%	83.93%
High	38.9-100	47.7	367	60.01%	47.72%

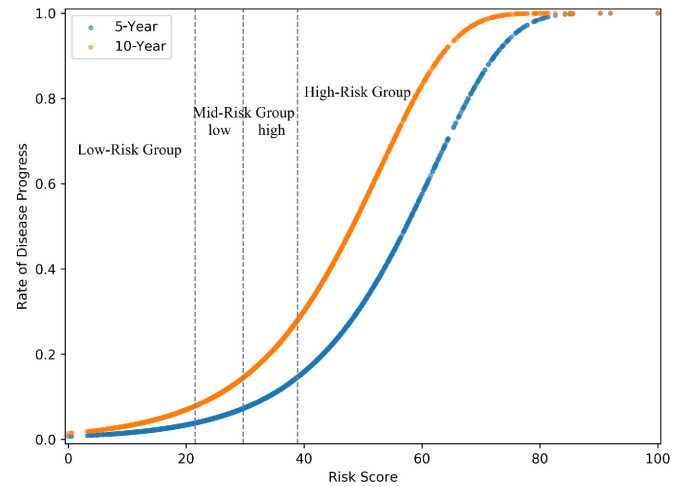


Fig. 9. Rate of disease progress as a continuous function of the risk score at 5-year and 10-year. A point on a curve represents an individual patient in the training set. Dotted lines indicate the cut-off values for risk grouping.

from low to high is 15.4, 26.6, 33.8, and 47.7, respectively. At 5-year, the survival rates (rates of freedom from disease progress) of risk groups from low to high are 100.0%, 96.47%, 91.51%, and 60.01%, respectively. At 10-year, the survival rates (rates of freedom from disease progress) of risk groups from low to high are 96.77%, 91.30%, 83.93%, and 47.72%, respectively. It can be seen that rates of freedom from disease progress between each risk group are distinguishable at 5-year and 10-year. The log-rank test shows that there is a significant difference between the survival status of all two adjacent risk groups in the test set ( $p < 0.01$ ). It demonstrates that the cut-off points for risk grouping based on the training set are very discriminative and informative, and the EXSA prognostic model has a strong generalization ability.

Fig. 9 shows continuous function between risk score and rate of disease progression at 5-year and 10-year. The figure is a scatter plot of the risk score and its corresponding rate of disease progress at 5-year and 10-year. The dotted lines indicate the cut-off values for risk grouping and divide the patients into four risk groups. In clinical practice, the EXSA prognostic model can predict the risk score of disease progress for a patient with breast cancer. Doctors can refer to the continuous function to estimate the rate of the patient occurring disease progress. Thereby, it can assist doctors to make a more appropriate plan in treatment.

#### D. Subgroup Analysis

Tumor stage, an important and widely used prognostic factor, has a huge impact on breast cancer patients' future survival state. To comprehensively test the effectiveness of the model, we analyze and compare the distribution of risk score calculated by the model under different tumor stage subgroups. The tumor stage can be one of the five classes 0, 1, 2, 3 and 4. For tumor stage 0, a total of 237 and 107 patients are occult breast cancer in the training and test set, respectively. Since there are very few patients with tumor stage 0, patients with tumor stage 0 and 1

TABLE VII  
ESTIMATED RISK SCORES

Tumor stage	1	2	3	4
Num. of Patients	1059/441	1619/701	176/75	111/49
Mean	28.0/27.8	31.7/32.1	43.5/43.1	49.3/50.6
Std.	12.5/12.8	14.2/14.3	16.9/18.1	17.4/17.3
Median	27.4/27.3	30.0/30.4	42.5/40.5	46.9/48.0

Std. = Standard Deviation.

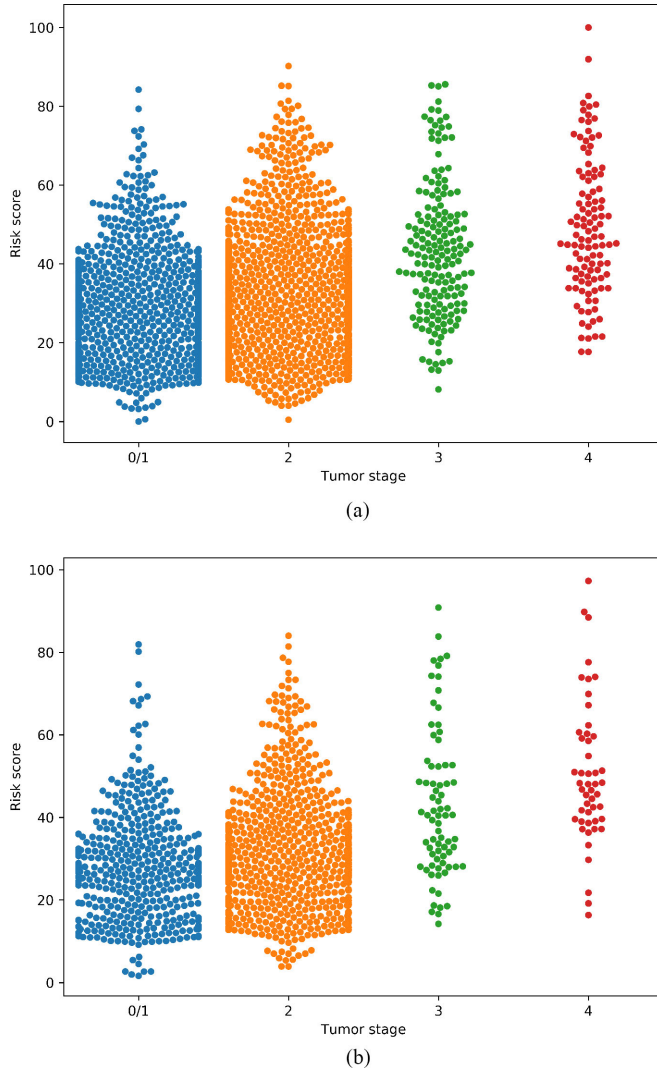


Fig. 10. Swarm plotting of risk score distribution. (a) Risk scores in the training set. (b) Risk scores in the test set.

are merged into one subgroup as tumor stage 1. Therefore, we focus on 4 tumor stage subgroups. The estimated risk scores in each subgroup are presented in Table VII. From the tumor stages 1 to 4, we can see that the larger the tumor stage is, a higher risk score is estimated by the model, which is in line with clinical experience.

Utilizing the swarm plotting and KDE (Kernel Density Estimation) density plotting, we can visualize the distribution of risk scores of each tumor stage subgroup. As shown in Figs. 10 and 11, the model can capture and quantify the difference of disease progress risk between each tumor stage subgroup. Although the

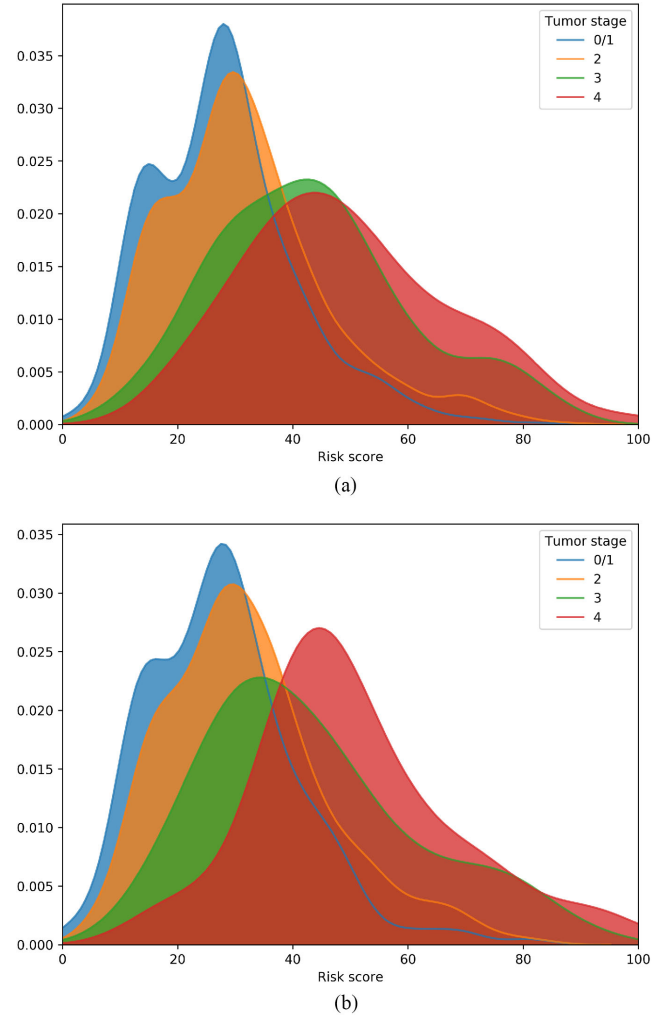


Fig. 11. KDE density plotting of risk score distribution. (a) KDE density in the training set. (b) KDE density in the test set.

tumor stage is an import factor for the prognosis, the overlap of areas under the density curve in Fig. 11 illustrates that the interaction of multiple factors including tumor stage may lead to a complex prognostic prediction.

## VI. CONCLUSION

In this paper, we mainly introduced the optimized survival analysis of XGBoost. We applied it to predict disease progress in breast cancer. The proposed EXSA method was based on XGBoost in machine learning and the CPH model in survival analysis. We took the more precise approximation of partial likelihood function as learning objective and derived the corresponding mathematical expression used for XGBoost, which extremely optimized and enhanced the ability of XGBoost to analyze survival data with a large number of ties.

Clinical data of breast cancer was derived from the Clinical Research Center for Breast (CRCB) in West China Hospital of Sichuan University. After cleaning data, using the EXSA survival analysis method, we included 4575 patients with 24 import features (with 3202 patients for training, 1373 patients for test)



to build a prognostic prediction model of disease progress after diagnosis based on Chinese breast cancer patients. The model achieved comparable performance with C-index of 0.83454, 5-year and 10-year AUC of 0.83851 and 0.78155, respectively. Compared with the original XGBoost, the classical GBM, RSF, and Cox survival analysis methods, the proposed EXSA method had a relatively better performance.

To apply the prognostic model into clinical practice, we evaluated risk scores of disease progress and obtained the cut-off values of risk grouping and continuous functions between risk scores and disease progression rate at 5- and 10-year. The prognostic model divided breast cancer patients into the following categories: low risk (risk score between 0 and 21.5), mid-low risk (risk score between 21.6 and 29.6), mid-high risk (risk score between 29.7 and 38.8), and high risk (risk score between 38.9 and 100). The experimental result demonstrated that the EXSA prognostic model was of a strong discriminative power in disease progress risk of breast cancer. It can assist doctors in diagnosis and treatment.

In future works, additional clinical data of breast cancer and more follow-up data will be collected to improve the EXSA prognostic model. The EXSA prognostic model will be further validated on a large scale in hospitals.

## REFERENCES

- [1] D. R. Cox, "Partial likelihood," *Biometrika*, vol. 62, no. 2, pp. 269–276, 1975.
- [2] D. R. Cox *et al.*, "Regression models and life tables," *J. Roy. Statistical Soc.: Ser. B (Methodological)*, vol. 34, no. 2, pp. 187–220, 1972.
- [3] L. Yang and K. Pelckmans, "Machine learning approaches to survival analysis: Case studies in microarray for breast cancer," *Int. J. Mach. Learn. Comput.*, vol. 4, no. 6, Dec. 2014.
- [4] C. Adele *et al.*, "Random forests," *Mach. Learn.* 45. vol. 1 (2004): pp. 157–176.
- [5] H. Ishwaran *et al.*, "Random survival forests," *The Ann. Appl. Statist.*, pp. 841–860, 2008.
- [6] S. Dietrich *et al.*, "Random survival forest in practice: A method for modelling complex metabolomics data in time to event analysis," *Int. J. Epidemiol.*, 2016, pp. 1406–1420.
- [7] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *Ann. Statist.* vol. 5, no. 29, pp. 1189–1232, 2001.
- [8] L. Ping, "Robust logitboost and adaptive base class (ABC) logitboost," *Comput. Sci.* (2012).
- [9] G. Ridgeway, "The state of boosting," *Comput. Sci. Statist.*, pp. 172–181, 1999.
- [10] B. Harald and M. Schumacher, "Allowing for mandatory covariates in boosting estimation of sparse high-dimensional survival models," *BMC Bioinf.* 9. vol. 1 (2008): pp. 14–0.
- [11] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2016, pp. 785–794.
- [12] G. C. Wishart *et al.*, "PREDICT: A new UK prognostic model that predicts survival following surgery for invasive breast cancer," *Breast Cancer Res.*, vol. 12, no. 1, p. R1, 2010.
- [13] F. J. C. dos Reis *et al.*, "An updated PREDICT breast cancer prognostication and treatment benefit prediction model with independent validation," *Breast Cancer Res.*, vol. 19, no. 1, p. 58, 2017.
- [14] P. M. Ravdin *et al.*, "Computer program to assist in making decisions about adjuvant therapy for women with early breast cancer," *J. Clin. Oncol.*, vol. 19, no. 4, pp. 980–991, 2001.
- [15] G. C. Wishart *et al.*, "A population-based validation of the prognostic model PREDICT for early breast cancer," *Eur. J. Surgical Oncol. J. Eur. Soc. Surgical Oncol. Brit. Assoc. Surgical Oncol.*, vol. 37, no. 5, pp. 411–417, 2011.
- [16] H. E. Campbell *et al.*, "An investigation into the performance of the adjuvant! Online prognostic program in early breast cancer for a cohort of patients in the United Kingdom," *Brit. J. Cancer*, vol. 101, no. 7, pp. 1074–1084, 2009.
- [17] 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.
- [18] N. Breslow, "Covariance analysis of censored survival data," *Biometrics* 30. vol. 1 (1974): pp. 89–99.
- [19] B. Efron, "The efficiency of cox's likelihood function for censored data," *Publications Amer. Statistical Assoc.* 72. vol. 359 (1977): pp. 9.
- [20] B. Fu *et al.*, "Predicting invasive disease-free survival for early stage breast cancer patients using follow-up clinical data," *IEEE Trans. Biomed. Eng.*, preprint, 2018.
- [21] J. Bergstra and Y. Bengio, "Algorithms for hyper-parameter optimization," in *Proc. Inter. Conf. Neural Inf. Process. Syst.*, 2011, pp. 2546–2554.
- [22] J. Bergstra *et al.*, "Hyperopt: A python library for model selection and hyperparameter optimization," *Comput. Sci. Discovery*, vol. 8, no. 1, 2015, Art. no. 014008.
- [23] P. Heagerty *et al.*, "Time-dependent ROC curves for censored survival data and a diagnostic marker," *Biometrics*, vol. 56, no. 2, pp. 337–344, Jun. 2000.
- [24] J. D. Kalbfleisch and R. L. Prentice, *The statistical analysis of failure time data*, Second Edition. 2011.
- [25] L. Yan *et al.*, "Predicting prostate cancer recurrence via maximizing the concordance index," in *Proc. KDD'04*, Aug. 22–25, 2004, Seattle, WA, USA.
- [26] G. Ridgeway, "Generalized boosted models: A guide to the gbm package," 2007.
- [27] J. A. Sparano *et al.*, "Adjuvant chemotherapy guided by a 21-gene expression assay in breast cancer," *N. Engl. J. Med.* 379. pp. 2, 2018.
- [28] J. A. Sparano *et al.*, "Prospective validation of a 21-gene expression assay in breast cancer," *N. Engl. J. Med.* 373. vol. 373, no. 21, pp. 2005–2014, 2015.
- [29] H. Torsten and B. Lausen, "Bagging tree classifiers for laser scanning images: A data- and simulation-based strategy," *Artif. Intell. Med.*, vol. 27, no. 1, pp. 65–79, 2003.
- [30] K. Park *et al.*, "Electromechanical coupling factor of breast tissue as a biomarker for breast cancer," *IEEE Trans. Biomed. Eng.*, vol. 65, no. 1, pp. 96–103, Apr. 2017.
- [31] M. Mahrooghy *et al.*, "Pharmacokinetic tumor heterogeneity as a prognostic biomarker for classifying breast cancer recurrence risk," *IEEE Trans. Biomed. Eng.*, vol. 62, no. 6, pp. 1585–1593, Jun. 2015.
- [32] U. Maulik *et al.*, "Gene-expression-based cancer subtypes prediction through feature selection and transductive SVM," *IEEE Trans. Biomed. Eng.*, vol. 60, no. 4, pp. 1111–1117, Apr. 2013.
- [33] S. Reis *et al.*, "Automated classification of breast cancer stroma maturity from histological images," *IEEE Trans. Biomed. Eng.*, vol. 64, no. 10, pp. 2344–2352, Feb. 2017.
- [34] T. Yin *et al.*, "A robust and artifact resistant algorithm of ultrawideband imaging system for breast cancer detection," *IEEE Trans. Biomed. Eng.*, vol. 62, no. 6, pp. 1514–1525, Jan. 2015.
- [35] T. Chen *et al.*, "Scalable, portable and distributed gradient boosting (GBDT, GBRT or GBM) library, for python, R, java, scala, C++ and more," 2014. [Online]. Available: <https://github.com/dmlc/xgboost>. Accessed: Jan. 28, 2019.
- [36] B. Greenwell *et al.*, "GBM3: Gradient boosted models," 2013. [Online]. Available: <https://github.com/gbm-developers/gbm3>. Accessed: Sep. 16, 2018.
- [37] R. Turkki *et al.*, "Breast cancer outcome prediction with tumour tissue images and machine learning," *Breast Cancer Res. Treatment*, vol. 177, no. 1, pp. 41–52, May 2019.
- [38] M. Darshini *et al.*, "Predicting factors for survival of breast cancer patients using machine learning techniques," *BMC Med. Inform. Decis. Making*, vol. 19, no. 1, pp. 48, Mar. 2019.
- [39] F. Ting *et al.*, "Convolutional neural network improvement for breast cancer classification," *Expert Syst. Appl.*, vol. 120, pp. 103–115, Nov. 2018.
- [40] N. Wu *et al.*, "Deep neural networks improve radiologists' performance in breast cancer screening," *IEEE Trans. Med. Imag.*, vol. 39, no. 4, pp. 1184–1194, Apr. 2020.
- [41] H. Vundavilli *et al.*, "Bayesian inference identifies combination therapeutic targets in breast cancer," *IEEE Trans. Biomed. Eng.*, vol. 66, no. 9, pp. 2684–2692, Sep. 2019.