Predicting Invasive Disease-Free Survival for Early Stage Breast Cancer Patients Using Follow-Up Clinical Data

Bo Fu, Pei Liu, Jie Lin, Ling Deng, Kejia Hu, and Hong Zheng

Abstract*—*Objective: **Chinese women are seriously threatened by breast cancer with high morbidity and mortality. The lack of robust prognosis models results in difficulty for doctors to prepare an appropriate treatment plan that may prolong patient survival time. An alternative prognosis model framework to predict invasive disease-free survival (iDFS) for early stage breast cancer patients, called MP4Ei, is proposed. MP4Ei framework gives an excellent performance to predict the relapse or metastasis breast cancer of Chinese patients in five years. *Methods:* MP4Ei is built based on statistical theory and gradient boosting decision tree framework. 5246 patients, derived from the clinical research center for breast in West China Hospital of Sichuan University, with early-stage (stage I–III) breast cancer are eligible for inclusion. Stratified feature selection, including statistical and ensemble methods, is adopted to select 23 out of the 89 patient features about the patient' demographics, diagnosis, pathology, and therapy. Then, 23 selected features as the input variables are imported into the XGBoost algorithm, with Bayesian parameter tuning and cross validation, to find out the optimum simplified model for five-year iDFS prediction. *Results:* For eligible data, with 4196 patients (80%) for training, and with 1050 patients (20%) for testing, MP4Ei achieves comparable accuracy with AUC 0.8451, which has a significant advantage ($p < 0.05$). *Conclusion:* This work demonstrates the complete iDFS prognosis model with very competitive performance. *Significance:* The proposed method in this paper could be used in clinical practice to predict patients' prognosis and future surviving state, which may help doctors make treatment plan.**

***Index Terms*—Breast cancer, machine learning, feature selection, prognosis prediction, relapse and metastasis cancer, gradient boosting decision tree.**

I. Introduction

BREAST cancer, one of the largest morbidity and mortality of malignant tumor, seriously endangers Chinese or global female health. It can be medically detected early during a screening examination through mammography or by portable cancer diagnostic tool [1]. Cancerous breast tissues change with the progression of the disease, which can be directly linked to cancer staging. The stage of breast cancer (I–IV) describes how far a patient's cancer has proliferated. Statistical indicators such as tumor size, lymph node metastasis, distant metastasis and so on are used to determine stages [2]. To prevent cancer from spreading, patients have to undergo breast cancer surgery [48], chemotherapy, radiotherapy, endocrine therapy or targeted therapy [3]. These interventions do not completely eradicate the risk of cancer. But patients may still be at the risk of cancer recurrence at any time.

Therefore, it is of utmost importance to help doctors make plan for breast cancer diagnosis and treatment to lessen the suffering of patients. A variety of prognostic prediction methods are used to help breast cancer patients, such as HER-2/neu, angiogenesis, pharmacokinetic tumor heterogeneity[9], BRCA1 and other gene expression [52]. Due to advancement in machine learning and availability of patient electronic medical records, it has been possible to successfully build clinical decision support systems to produce personalized clinical recommendations, and predict patients' prognosis and future surviving state. For example, classifier ensemble [4], Neural Network and Adaboost algorithm [5], and semi-supervised learning [6] were used to predict breast cancer survivability. Discovery engine [7] was demonstrated for breast cancer diagnosis and personalized recommendation [8]. A nonnegative matrix factorization algorithm for dynamic modules [9] was developed to determine the dynamics of pathways associated with cancer progression. Biased minimax probability machine (BMPM) was applied to improve the sensitivity while maintain an acceptable specificity for medical diagnosis [49]. More machine learning algorithms for breast cancer were studied in references [10], [31]–[34], [51].

Breast cancer 5-year prognostic assessment could be targeted toward those patients at high risk to maximize survival benefit.

Manuscript received July 31, 2018; revised November 15, 2018; accepted November 18, 2018. Date of publication November 22, 2018; date of current version June 21, 2019. 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.)*

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

P. Liu and J. Lin are with the Big Data Research Center, and School of Computer Science and Engineering, University of Electronic Science and Technology of China.

L. Deng and K. Hu 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.

H. 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, Chengdu 610041, China (e-mail: hzheng@scu.edu.cn).

Digital Object Identifier 10.1109/TBME.2018.2882867

This is because of highest probability of cancer recurrence happened in early 5 years after initial diagnosis [11]. No recurrence in 5 years as an important indicator reflects the effect of breast cancer treatment. Predict and assess that indicator could help doctors make appropriate treatment plans to prolong patients' survival time. Clinical pathological features of cancer patients can be used to predict 5-year recurrence probability [12]. Wu *et al.* created two models to predict breast cancer on year in advance based on logistic regression (LR) and LASSO logistic regression (LR+Lasso) [13]. 5-year survivability prediction using logistic regression was reported on SEER data [14]. Predictive model for 5-year survivability of breast cancer using decision tree was proposed [15].

However, in China, prognosis prediction for breast cancer patients, especially based on sufficiently large scale of clinical follow-up data, for relapse or metastasis is still not mature. Main reasons to block good models to spring up are due to the lack of clinical patients followed up with so many years, and the difficulty of integrating missing and irregularity data from the clinical management systems that are lagging behind. All these barriers often come from the non-canonical data input, insufficient patient follow-up, lack of patient compliance and other social factors. Therefore, the Clinical Research Center for Breast (CRCB) in West China Hospital of Sichuan University has constructed an information management system for breast cancer. Currently, the system has 12119 breast cancer patients, which are followed up more than 10 years and the annual non-investigation rate is less than 1.1% since 2011.

In this paper, 5246 patients with early-stage (stage I–III) breast cancer are eligible for inclusion to build MP4Ei model to predict Invasive Disease-Free Survival (iDFS). After carefully setting standards and cleaning data, stratified feature selection, including statistical and ensemble methods, is adopted to select 23 out of the 89 patient features about the patient' demographics, diagnosis, pathology and therapy. In statistical method for feature selection, 51 features are selected by using statistical test with significant difference ($p < 0.05$) [51]. Then, further 23 features with scores of variable importance greater than threshold 0.016 are made choice out from 51 features in ensemble method. MP4Ei with or without feature selection is evaluated by AUCs. 23 selected features, as the input variables, are imported into the gradient boosting decision tree (XGBoost) [16] algorithm with Bayesian parameter tuning [17], [18] and cross validation. The excellent and robust model, MP4Ei, is built for 5-year Invasive Disease-Free Survival (iDFS) prediction of early-stage breast cancer patients. Compared with related classical algorithms like SVM, Random Forest, Adaboost and Cox Regression, MP4Ei AUC value (0.8451) has a significant advantage ($p < 0.001$).

By using advanced non-linear machine learning methods and precious follow-up dataset from the Chinese population, MP4Ei incorporates more factors that are significant for breast cancer prognosis. That make us comprehensively consider 89 demographic, diagnostic, pathological and therapic features of a patient, and 23 important features can be selected automatically while the complex prognostic mechanism of a factor is unnecessarily known in advance. It is essential to find new prognostic factors without requiring specialized prior knowledge. For example, we found that besides common pathological factors, some demographic or sociological characteristics should also be considered as important prognostic factors (e.g., education, number of births, medical insurance and so on). It is different from other frameworks like PREDICT and Adjuvant online [35]–[37] also MP4Ei could be used in clinical practice to predict patients' prognosis and future surviving state, and help doctors make treatment plans to reduce the risk of relapse or metastasis cancer.

This paper is organized as follows. Section II describes the data materials. The proposed method is demonstrated in Section III. We then evaluate our method using the CRCB dataset, and discuss the implications of our study and its results, which are described in Section IV. Finally, the concluding remarks are given in Section V.

II. Data Materials

Data materials are provided by the CRCB in the West China Hospital of Sichuan University. Breast cancer data of patients was retrospectively collected from 1989 to 2007, while since 2008 all patients confirmed by pathology in the West China Hospital are 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, *et al.* We classify the information or attributes, also called features in machine learning, in a patient EHR into four categories: demographics, diagnosis, pathology and therapy. There are about 900 new patients each year to be added in. Quality control for missing or un-standardized data is carried out. Therefore, it is important to decrease the annual rate of defaulters to 0.4% recently.

By May 2017, the total number of breast cancer patients is 12,119. All data materials derived from the database in the CRCB are used to build iDFS prognosis models to predict 5-year iDFS events, which may involve relapse or metastasis breast cancer, death, or the second primary tumor. Diagnosed dates of patients are distributed from 1989 to 2017.

In this work, patients with early-stage breast cancer (I–III) are focused. We retain

a) Patients with stage I–III and primary breast cancer;
b) Patients initially diagnosed with unilateral breast cancer.

Then we get a dataset with the total of 11522 cases. On the other hand, patients followed up less than five years, while no clinical events are happened in the meantime, are removed from the dataset to avoid biases and noise.

After filtering data, a total of 5246 patients are eligible for further study, in which there are 1181 iDFS events happened in five years after surgery. In order to intuitively present the distribution of patients with or without iDFS, Principle Component Analysis (PCA) method is adopted to reduce data dimension from 89 features to 2 principal components. And 5-year iDFS patients are displayed as shown in Fig. 1. We can see that patients can't be linearly separable, and a complex mapping exists between the target iDFS variable and 89 explanatory variables.

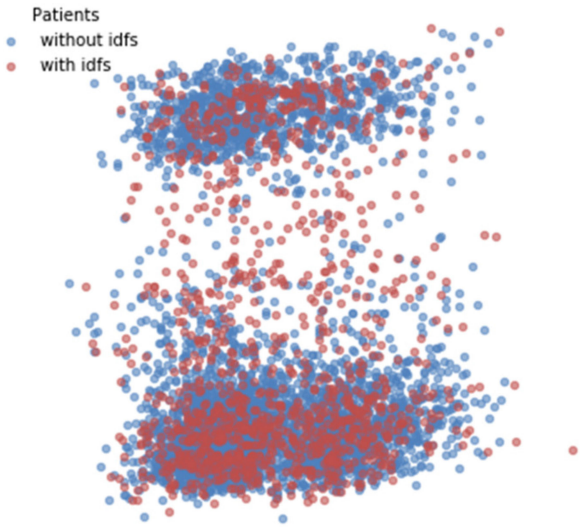


Fig. 1. Distribution of patients with or without iDFS.

III. Methods

In the MP4Ei framework, we retain data materials of 5246 patients to construct the model and then normalize the original records based on the recommendations from medical experts. Firstly, stratified random sampling method is applied to divide data of 5246 patients into training data (4196 cases) and testing data (1050 cases). Then by using the training data, 23 important features are picked out from original 89 patient features in the process of stratified feature selection with two stages. The first stage is Statistical Feature Selection (SFS). One by one, 89 patient features sorted out from data materials are statistically tested based on the type of individual feature. Then 51 attributes or features are selected out. Those features have statistically significant impact on the ending point iDFS. In order to improve the stability of MP4Ei, and enhance the model's application, 23 important features, of which the importance scores are greater than the threshold 0.016, are made choice out from the 51 features at the second stage of stratified feature selection. This stage is named as Ensemble Feature Selection (EFS). After taking 51 features as input and repeating K-fold cross validation on the training data several times, a feature's importance score is calculated by averaging importance scores that are output by gradient boosting decision tree algorithm, XGBoost. After that, 23 features are inputted into the XGBoost algorithm again, with Bayesian parameter tuning and K-fold cross validation to build the 5-year iDFS classifier. Finally, an excellent and robust model is built to predict iDFS for early-stage breast cancer patients. On the other hand, the rest 1050 cases as testing data are used to evaluate the iDFS model after taking 23 features out from 89 original features. Those features are the same as the ones selected in the training procedure. The overall framework of MP4Ei is shown in Fig. 2.

A. Data Processing

We usually run up against obstacles in clinical breast cancer data in real environments, because of massive volumes of the

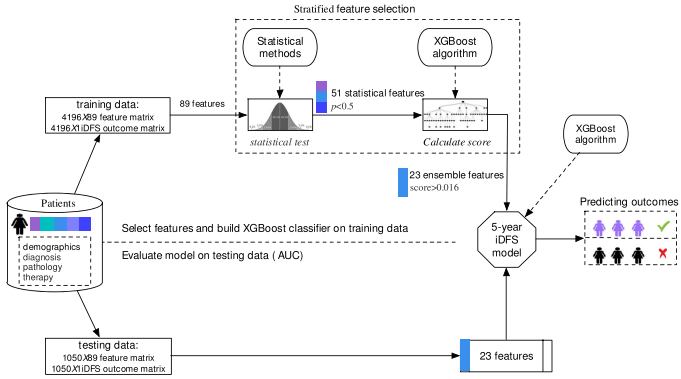


Fig. 2. MP4Ei framework.

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.

Where outliers are evident, necessary information is missing, or data inconsistencies are occurred in the raw data, corrections are made by rechecking and revisiting patients by telephone. On the other hand, because each patient may exist multiple diagnostic, pathological or therapeutic records, we need to merge multiple records of a patient to make the patient correspond to a final description. In this process, patient features and corresponding integration rules are put forward by the professional medical group of breast cancer. Then the specific implementation for data consistence is performed in the database system. The results are reviewed and evaluated at the medical group review meeting for the problems still existing in the consolidated data or cleaning rules. Such an iterative process ensures the data quality.

To get the data for modeling, categorical variables in patient features are encoded by using the numeric coding, while some continuous variables are divided into discretization intervals and also encoded by using the numerical coding based on the advice of the professional medical group. If there are data gaps in variables or some specific variables uninvolved, for example, no breastfeeding history for unmarried female patients, the variables are encoded by maximum classification values, for example, 9 or 10, respectively.

B. Prediction Model

To predict the 5-year iDFS of breast cancer, XGBoost, derived from gradient boosting decision tree, is used to construct the prediction model, which belongs to ensemble learning algorithms. A weak learner in XGBoost depends on others based on the idea of Boosting. Penalty term and node split incorporated into the XGBoost make a model over-fitting be well avoided. On the other hand, XGBoost has realized the underlying code optimization and parallelization, and significantly improve the system efficiency [16]. It has been widely used in all kinds of practical problems, and has a good performance in data mining competitions. Readers can refer to the work of Chen *et al.* [16].

Let breast patients denote by $\mathrm{D} = \{(x_i, y_i)\}$, where $x_i \in \mathbb{R}^m, y_i \in \mathbb{R}$, $|D| = n$, and n is the number of patients, m is

the feature dimension of a patient. To predict the output $\hat{y}_i$, K functions are added up, defined as

$$\hat{y}_{\mathrm{i}} = \phi\left(x_i\right) = \sum_{k=1}^{K} f_k\left(x_i\right), f_k \in \mathcal{F} \tag{1}$$

where $\mathcal{F} = \{f(x) = w_{q(x)}\}$, $q : \mathbb{R}^m \rightarrow T, w \in \mathbb{R}^T$. $\mathcal{F}$ is the space of decision trees, q is a tree structure, T is the number of leaf nodes in the q, and f_k is corresponded to the independent tree structure q with leaf node weight w. For a given test sample, rules of a decision tree determine which leaf node is related to the sample, and the weights of the corresponding leaf node is used to predict. To learn tree structures in a model, we need to minimize the objective function as follows

$$\mathcal{L}\left(\phi\right) = \sum_{i} l\left(\hat{y}_i, y_i\right) + \sum_{k} \Omega\left(f_k\right) \tag{2}$$

where $\Omega(f) = \gamma T + \frac{1}{2}\lambda\|w\|^2$. Penalty term Ω can prevent a model from over-fitting, which makes the model not only be better for prediction, but also be relatively simple. The penalty tem Ω in the objective function is different from that in the ordinary gradient boosting tree. The validation of our method and more details will be discussed in Section IV.

Algorithm 1: Stratified Feature Selection.

Input: original features F_n
Output: selected features F_m

1: **for** F_i in F_n
2: **if** F_i is the Nominal scale
3: p = Chi-square Test for F_i
4: **else if** F_i is the Ordinal scale
5: p = Mann–Whitney U Test for F_i
6: **else if** F_i is the Interval scale
7: **if** F_i is normally distributed(KS Test)
8: p = independent-sample t Test for F_i
9: **else**
10: p = Mann–Whitney U Test for F_i
11: **else** raise error
12: **if** p < 0.05
13: append F_i to F_m
14: F_m scores = xgb-scores(F_m)
15: **for** F_i score in F_m scores
16: **if** F_i score < 0.016
17: delete F_i from F_m
18: return F_m

C. Stratified Feature Selection

Description of a patient, a total of 89 features or variables, consists of demographics, diagnosis, pathology and therapy. The model should predict the ending point variable iDFS whether happened within five years (5-year iDFS). In order to make the model more flexible and practicable, stratified feature selection method is used to choose 23 important features, which are integrated into the follow-up work to build 5-year iDFS model. Just as described in Fig. 2, stratified feature selection is done on the training data. To improve the stability of the model and minimize the impact of irrelevant features, statistical method is carried out on patients with/without iDFS to roughly eliminate the features that are independent of the target variable iDFS. Moreover, to get an accurate score on average for each important feature, XGBoost algorithm is done by repeating 10-fold cross validation for 5 times with hyper-parameter optimized on the training data. Then subsets of features are used to find the cutoff value to determine a feature score is important or not by applying backward selection step by step. The results are shown in Section IV.

The procedure described in Algorithm 1 is adopted. In this procedure, SFS as statistical method is used to determine whether each feature, as a single factor, has a significant impact on the outcome variable. In general, a feature with Interval scale firstly should obey normal distribution by means of the Kolmogorov-Smirnov (KS) statistical test [19]. Then the feature with notable influence on the 5-year iDFS is separated from others by using the independent sample T-test in step 8. At the same time, in step 5, the Wilcoxon Mann-Whitney statistical test is used to test the feature with Ordinal scale, but whose distribution is not normally distributed [20]. For the Nominal scale feature, statistical significance is analyzed by Chi-square test. After SFS in steps 1–13, 51 features are selected from 89 original variables for the EFS feature selection.

In the EFS layer in steps 14–17, to obtain the stable score of a variable, in Function xgb-scores in step 14, we run the XGBoost Classifier over and over again for several times to get the average score to evaluate the importance for each feature. Three methods: *weight, gain*, and *cover* can be used to compute the importance score for a feature in XGBoost, in which *weight* is the number of times a feature appears in a tree. *gain* is the average gain of splits when the feature is used. *cover* is the average coverage of splits, where coverage is defined as the number of samples affected by the split [16]. Since the *gain* method is the main reference factor for splits, it is applied to output importance scores for each feature by performing 10-fold cross validation for 5 times. The final importance score of a feature is calculated by averaging those 5 scores.

After final importance scores of 51 features are all obtained, then the sequential backward selection method is adopted to finally determine which one is important for 5-year iDFS if the importance score is greater than 0.016. Section IV explains how to obtain the importance threshold 0.016 in detail. Unlike the recursive feature elimination algorithm proposed by Guyon *et al.* [21], we should take into account features for modeling with applicability and predictability, which impel us to select target features for the model with (a) an acceptable loss comparing to classical outstanding models; (b) the number of final features as few as possible. Therefore, EFS is proposed to assure the stability of evaluating important features and decrease the complexity of tuning parameters. After selected by the EFS, the feature with a lower average importance score is removed, just as step 17. Also, only one feature is retained if it represents exactly same meaning with others. For example, mlctp1

TABLE I
STRATIFIED FEATURE SELECTION ALGORITHM

		Features				
		Demo.	Diag.	Path.	Ther.	%
Original Features	*N*	sex addr nat edu mar medi urb	meno menoR bearF feed ocs hrt adt adtT fhBC fhGC fhC hisF cmbdt cmUN cmCC cmCH cmMS cmBT cmNR cmUR cmIS cmOT cmBR cmNE cmRS cmDS cmBS cmCS	ER PR HER2 Ki67 CK56 hrmRP mlctp1 mlctp2	sur surT chem chemS chemB chemA radN1 radN2 rad radT surCT druCT edcN edc edcBS smBS aiBS edcAS edcST edcS1 edcS2 traN tra traS	75.3
	O	ageM bmi mediR	ageMR ageMU ageMP preg bearN pregB ageB1 ageB2 ageBM ageME T N as ps	mstp who	-	21.3
	I	Aged	-	ND1 ND2	-	3.4
	%	12.4	47.2	13.5	26.9	-
Selected Features	*N*	edu medi mediR	-	Ki67 CK56 mlctp1	chemB chemA radT edcST edcS1	47.8
	O	Bmi	bearN ageB1 ageMP ageBM ageME T N as ps	-	-	43.5
	I	Aged	-	ND1	-	8.7
	%	21.7	39.2	17.4	21.7	-

Full name of the feature abbreviated form in the above Table is demonstrated in Appendix 1. N = Nominal, O = Ordinal, I = Interval; Demo. = Demographics, Diag. = Diagnosis, Path. = Pathology, Ther. = Therapy.

and mlctp2 follow different standard but all represent molecular typing. Since mlctp2 has lower average importance score, it is eliminated. Finally, 23 important features are selected out to model 5-year iDFS for early-stage breast cancer.

Table I shows a list of overall features according to the category of variables. At the point view of data type, Features selected before (after) have Nominal (*N*) scale variables of 75.3% (47.8%), Ordinal (*O*) scale variables of 21.3% (43.5%), and Interval (*I*) scale variables of 3.4% (8.7%). On the other hand, according to the patient characteristics, there are Demographics (Demo.) variables of 12.4% (21.7%), Diagnosis (Diag.) variables of 47.2% (39.2%), Pathology (Path.) variables of 13.5% (17.4%), and Therapy (Ther.) variables of 26.9% (21.7%). The correlation between two features among those 23 features measured by Pearson correlation coefficient is shown in Fig. 3.

Stratified feature selection identifies and removes unrelated and redundant features that can't help to improve the performance of the iDFS model, or even may actually reduce the model's accuracy. In practice, fewer features not only reduce the complexity, but also make the model be easier to be understood and interpreted [22]. Section IV will discuss the impact of stratified feature selection on the prediction model.

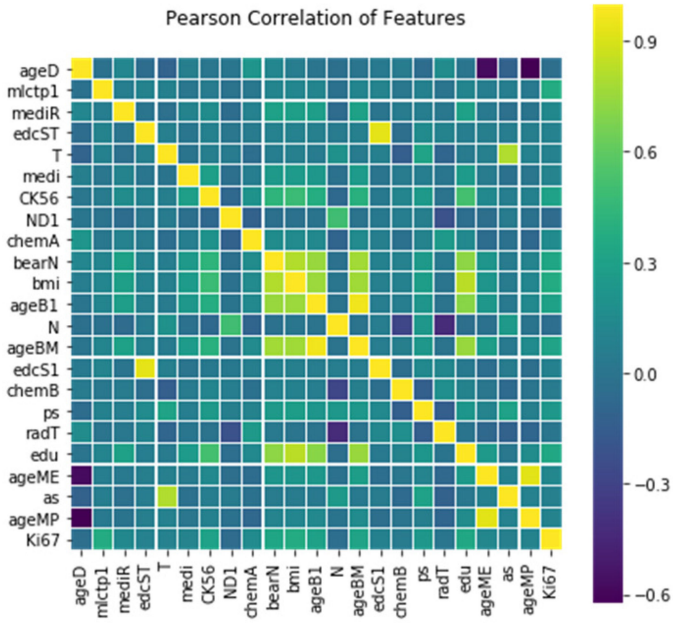


Fig. 3. Pearson correlation coefficients between features.

IV. RESULTS

A total of 5246 cases of early-stage breast cancer patients are analyzed and studied in the MP4Ei framework, of which 1181 (22.95%) patients have iDFS event within 5 years. Patient is represented by 23 features and a binary outcome variable, 5-year iDFS. Holdout method [23] is used to evaluate the performance of the built model by using XGBoost. Stratified random

TABLE II
DIFFERENCE SIGNIFICANCE TEST FOR TRAINING AND TESTING DATA

Category	Features	iDFS event	Train			Test			Description	p>0.05
			Number of cases	Mean	Standard deviation	Number of cases	Mean	Standard Deviation		
Nominal	Edu	0	3251	4.41	3.006	814	4.38	2.957	Education	0.907
		1	945	4.35	3.158	236	4.45	3.213		
	Medi	0	3251	3.18	2.077	814	3.20	2.043	Type of medical insurance	0.723
		1	945	3.16	1.937	236	2.97	1.879		
	mediR	0	3251	5.52	2.141	814	5.49	2.039	Claim ratio of medical insurance	0.156
		1	945	5.25	1.876	236	5.12	1.841		
	Ki67	0	3251	2.35	3.328	814	2.31	3.309	Ki67	0.862
		1	945	2.12	2.995	236	2.53	3.287		
	CK56	0	3251	5.51	4.324	814	5.43	4.334	CK5/6	0.781
		1	945	4.23	4.369	236	4.58	4.398		
	mlctp1	0	3251	10.90	3.980	814	11.11	3.816	Molecular typing	0.935
		1	945	10.83	4.176	236	11.12	4.289		
	chemB	0	3251	9.27	2.368	814	9.14	2.566	Neoadjuvant chemotherapy	0.770
		1	945	7.75	3.756	236	7.64	3.766		
	chemA	0	3251	2.86	2.666	814	2.76	2.579	Postoperative chemotherapy	0.995
		1	945	2.97	2.996	236	3.20	3.152		
	radT	0	3251	6.95	4.177	814	7.24	4.057	Postoperative radiotherapy	0.411
		1	945	6.95	4.245	236	6.73	4.237		
	edcST	0	3251	5.38	3.480	814	5.48	3.438	Type of endocrinotherapy	0.123
		1	945	6.62	3.719	236	6.80	3.729		
	edcS1	0	3251	3.53	4.529	814	3.59	4.551	Adjuvant hormonal therapy (AHT)	0.703
		1	945	5.48	4.794	236	5.81	4.743		
	Bmi	0	3251	4.08	2.964	814	3.99	2.912	BMI	0.307
		1	945	4.46	3.065	236	4.67	3.166		
	ageMP	0	3251	7.59	3.441	814	7.54	3.467	Menopause age	0.460
		1	945	7.31	3.533	236	7.57	3.455		
	bearN	0	3251	3.27	3.328	814	3.15	3.276	Number of births	0.191
		1	945	3.39	3.320	236	3.36	3.339		
	ageB1	0	3251	4.51	3.033	814	4.44	2.980	Age at first childbirth	0.388
		1	945	5.09	3.280	236	5.14	3.370		
	ageBM	0	3251	4.95	3.358	814	4.80	3.324	Duration between menarche and childbirth	0.163
		1	945	5.49	3.529	236	5.54	3.557		
Ordinal	ageME	0	3251	7.87	3.310	814	7.76	3.400	Duration between menarche and menopause	0.463
		1	945	7.41	3.523	236	7.69	3.405		
	T	0	3251	3.02	1.231	814	3.04	1.242	T stage	0.416
		1	945	3.44	1.192	236	3.44	1.178		
	N	0	3251	0.73	0.976	814	0.77	1.012	N stage	0.223
		1	945	1.58	1.197	236	1.58	1.166		
	As	0	3251	2.67	2.299	814	2.71	2.328	Anatomical stage	0.225
		1	945	2.97	1.824	236	3.01	1.770		
	Ps	0	3251	6.53	3.146	814	6.41	3.227	Prognosis stage	0.491
		1	945	7.14	2.376	236	7.38	2.324		
Interval	ageD	0	3251	48.43	10.529	814	48.37	10.508	Age at diagnosis	0.299
		1	945	49.32	11.215	236	47.84	10.685		
	ND1	0	3251	1.10	3.396	814	1.11	3.676	Number of I-II lymph node metastases	0.323
		1	945	3.47	6.992	236	3.29	6.309		

sampling method divides 5248 cases into the training set (4196 cases), and the testing set (1050 cases). There is no significant difference between the training and testing set ($p > 0.05$), as shown in Table II.

Performance is assessed on the CRCB dataset by calculating the area under ROC curve (AUC). AUC is an effective evaluating index, since it balances sensitivity and specificity of a model. Therefore, in the case of data with unbalanced positive and negative samples, AUC index can well reflect model performance for prediction [24]. Its ranges for excellent, very good and good diagnosis accuracy are (0.9–1.0), (0.8–0.9) and (0.7–0.8), respectively [25].

In the process for parameter tuning, the Tree of Parzen Estimators (TPE) algorithm [18], as one of Bayesian hyper-parameter optimizations, is used as the super-parameter optimization algorithm. Compared with grid search and random search for parameter tuning, the TPE searches in a broader space and is not easy to miss the best parameters. But the TPE tends to require longer searching time [26].

Given a set of model parameters, we use the training set to evaluate its performance by repeating 10-fold cross validation for 5 times, which can obtain a good model with fairly comparing with previously-proposed approaches, and make the model with good predictability and stability [27].

TABLE III
SEARCH SPACE FOR SUPER-PARAMETERS

Parameter	Range	Step	Number of values
n_estimators	[80, 150]	10	8
max_depth	[3, 7]	1	5
min_child_weight	[1, 10]	1	10
Subsample	[0.8, 1.0]	0.05	5
colsample_bytree	[0.3, 1.0]	0.1	8
reg_alpha	[0.0, 1.0]	0.05	21

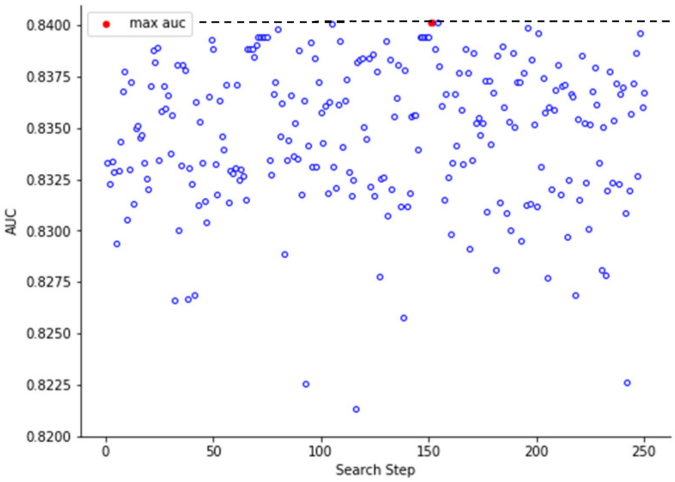


Fig. 4. Searching process for a parameter's AUC.

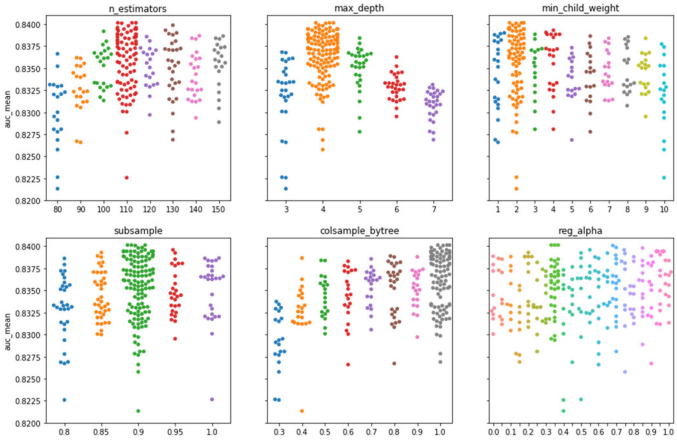


Fig. 5. Searching parameters.

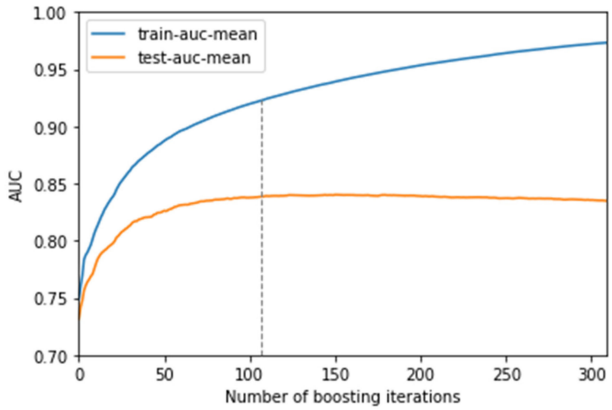


Fig. 6. Model performance at the different number of iterations.

A. Parameter Tuning

Important parameters consist of boosting learning rate (learning_rate), number of boosted trees to fit (n_estimators), maximum tree depth for base learners (max_depth), minimum sum of instance weight needed in a child (min_child_weight), subsample ratio of columns (colsample_bytree), subsample ratio of the training instance (subsample) and L1 regularization term on weights (reg_alpha) in the Xgboost model. In experiment, the parameter optimization toolkit Hyperopt [28] is adopted for parameter tuning, by specifying the parameter searching space, objective function, search algorithm and maximum number of times for searching. The searching space defined in Table III shows that the number of possible combinations of all parameters is greater than 300K (336,000 = 8*5*10*5*8* 21). Parameters input to an objective function are specific values as a set located in the parameters search space. This set of specific values is used to train the predicting model by repeating 5 times with 10-fold cross-validation. Then the average AUC measures the performance of the objective function as a return value. TPE is the optimization algorithm with which maximum number of search times is 250.

By recording 250 times of parameter searches, we can obtain the corresponding AUC means. Each AUC mean in the process of Bayesian optimization for parameters changes as shown in Fig. 4. Red point has the maximum AUC value. Fig. 5 shows the distribution of each parameter for candidate values, where the horizontal axis represents candidate values of each parameter, and the vertical axis represents possible AUC values.

Finally, optimal parameters searched for the model can be obtained as follows: learning_rate = 0.1, n_estimators = 110, max_depth = 4, min_child_weight = 2, subsample = 0.9, colsample_bytree = 1.0, reg_alpha = 0.35. The AUC value after repeated 5 times with 10-fold cross validation is 0.8401 on average.

In order to further verify over-fitting whether existing in the model, we test AUC values at the different number of boosting iterations. It is known that with the increasing number of iterations, the complexity of a model will also increase. That will make the model fit the training data well with a small error. But overly complex model also reduces the model's generalizability with a large error in the testing data. As shown in Fig. 6, with the increasing number of boosting iterations, the AUC values increase at the beginning of both training and testing processes. But we can see the AUC values begin to decrease when testing after several iterations, while the AUC values are still on the rise when training. We can get the best number of boosting iterations is 107, while the optimal number of boosting iteration is 110 (n_estimators = 110) by using the above Bayesian optimizing method. After 107 iterations, the prediction performance of the model is decreased due to the over-fitting happened. Therefore, by using early stopping method (early_stopping_rounds = 10), it is very close to the optimal parameters searched by the Bayesian optimization algorithm.

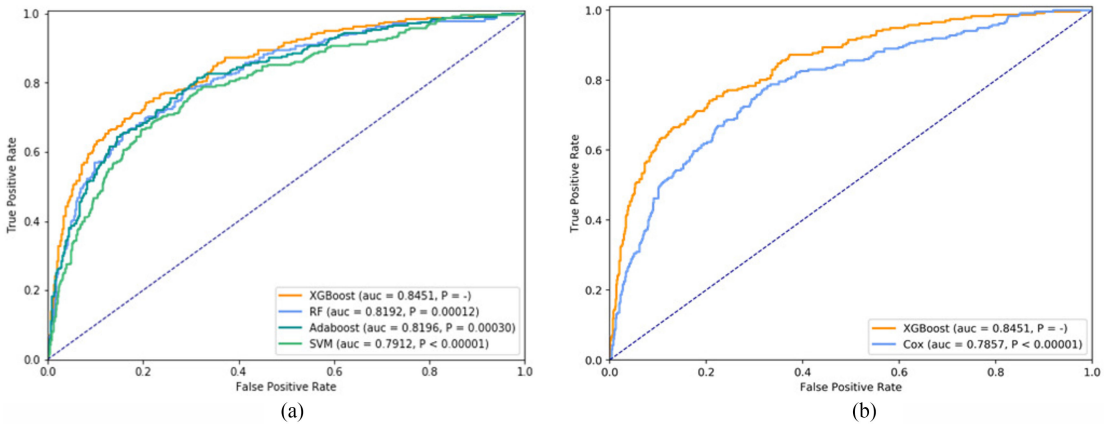


Fig. 7. ROC curves for 5-year iDFS prediction. (a) Comparison with three traditional classification algorithms based on machine learning. (b) Comparison with the Cox regression algorithm based on survival analysis.

TABLE IV
PERFORMANCE OF THE PROPOSED MODEL

	Proposed	RF	Adaboost	SVM	Cox
cut-off	0.235	0.305	0.498	0.231	0.189
Youden index	**0.535**	0.495	0.499	0.469	0.457
Sensitivity	0.742	0.653	**0.814**	0.665	0.784
Specificity	0.794	**0.843**	0.686	0.803	0.673
F-measure	**0.605**	0.595	0.561	0.568	0.539
AUC	**0.8451**	0.8192	0.8196	0.7912	0.7857
P	-	0.00012	0.00030	<0.00001	<0.00001

Specificity=TN/(TN+FP); Sensitivity= TP/(TP+FN); F1= 2TP/(2TP+FP+FN), where TP, FP, FN and TN are true positives, false positives, false negatives and true negatives, respectively.

B. Model Performance

We tried to use other machine learning algorithms, such as Support Vector Machine (SVM), Random Forest (RF) and Adaboost, to build models to compare with our proposed method. Different algorithms are trained on the same training dataset, and then tested in the same testing dataset. In experiments, RF, Adaboost and SVM are taken from the scikit-learn [29] machine learning repository. Bayesian parameter optimization is used as parameter tuning, and K-fold cross-validation approach is also used to get optimal parameters. After that, Parameters of the SVM model are C = 0.3, kernel = 'RBF', gamma = 0.005, class_weight = {1:2.0, 0:1}, probability = True. Parameters of the RF model are n_estimators = 120, max_depth = 14, min_samples_split = 15, min_samples_leaf = 2. And parameters of the Adaboost model are learning_rate = 1.0, n_estimators = 150. Experimental results are shown in Table IV. ROC curves of four models are shown in Fig. 7(a). Just as shown in Table IV, the proposed model has the F-measure of 0.605 and AUC of 0.8451. DeLong [30] method measures the performance differences between SVM, RF, Adaboost and the proposed model, as shown in Table IV also.

After tuning parameters by using Bayesian method, a model with outstanding predictive power is obtained for 5-year iDFS prediction. It has better predicting performance with AUC 0.8451, increased by about 3%. Taking the Youden index as the cut-off point where it is the biggest, we get the cut-off point 0.235 while the Youden index is 0.535. Just as shown in Table IV, the sensitivity, specificity, F-measure and AUC are 0.742, 0.794, 0.605 and 0.8451. It needs to specially explain that F-measure is an important measure in machine learning since it is the harmonic average between precision and recall. A model's performance reaches best at F-measure = 1 and worst at F-measure = 0. Considered from the F-measure perspective. F-measure has also been increased by 0.01, 0.04 and 0.037, compared with RF, Adaboost and SVM.

Cox regression algorithm, as one of the most commonly used models in the field of survival analysis, differs from the traditional classification algorithms mentioned above in that both survival status and time as the outcome variables have to be predicted simultaneously. Since the 5-year iDFS status is focused on by this paper, we can further compare the proposed model with the Cox survival model by setting survival time to 5-year.

After including the follow-up observation time of patients, we can build the Cox regression model by fitting the training set just as used in the previous algorithms. Then the 5-year survival probability in the test set could be predicted by setting survival time to 5-year. Experimental results are shown in Table IV and ROC curves are shown in Fig. 7(b). Just as shown in Table IV, the Cox regression model has the F-measure of 0.539 and AUC of 0.7857 for 5-year iDFS. It also can be seen that the proposed model has a higher AUC value compared with the Cox regression model, and ROC curves between two models have a significant difference ($p < 0.005$), as illustrated in Fig. 7(b).

From above experimental results, it can be seen that the proposed model has a higher AUC value compared with other models. At the same time, ROC curves among those models have a significant difference since p values are all less than 0.005. That illustrates the proposed method is evidently better. This is partly due to regularization items added in XGBoost algorithm to penalize the number of leaf nodes and the values of leaf nodes. That prevents the model at the risk of over-fitting, while the classical algorithms have their disadvantages because of the lack of regularization. Those algorithms are not robust enough, and are very easy to fit well when training the model but poor when testing it.

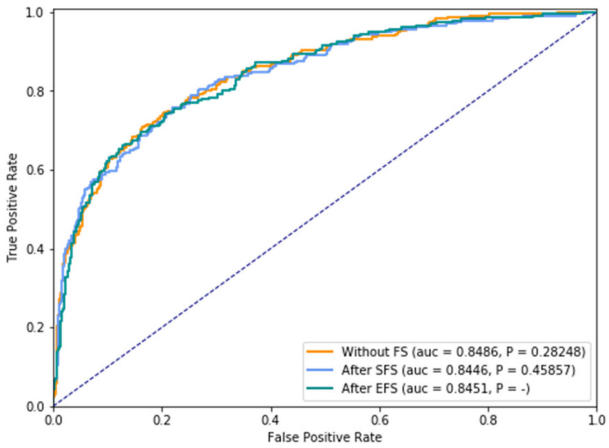


Fig. 8. ROC curves before and after EFS.

TABLE V
PERFORMANCE OF THE MODEL

	Without FS	After SFS	After EFS
Cut-off	0.276	0.234	0.235
Youden index	0.545	0.540	0.535
Sensitivity	0.712	0.758	0.742
Specificity	0.833	0.781	0.794
F1	0.622	0.604	0.605
AUC	0.8486	0.8446	0.8451
P	0.28248	0.45857	-

FS = Feature Selection, SFS = Statistical Feature Selection, EFS = Ensemble Feature Selection

C. Feature Selection Analysis

From the practical point of view, we will be more inclined to choose the model with fewer input variables without significant impact on the performance of the system. At the same time, medical clinical data always has some redundant and sparse features. It is of great significance to identify and eliminate them from clinical data. Through SFS and EFS feature selection, features are reduced from the original 89 to 23. And the performance of the model is not significantly affected.

Each sample or patient in the original dataset contains 89 features and a binary outcome variable. Through the statistical feature selection (SFS), 51 of 89 features with significant effect on the outcome variable ($p < 0.05$) are selected. Then the ensemble feature selection (EFS) method filters out 23 features that have highest scores for measuring variable importance. EFS is realized by comparing the difference between ROC curves calculated before and after features selection under the same dataset.

As shown in Fig. 8, models built by the original 89 features (without FS) and by selected 23 features (with EFS) has no significant difference ($p = 0.28248$) with AUC of 0.8486. There is no significant impact on model performance by SFS also, since the p value is 0.45857 and the AUC is 0.8446. After reducing 66 features, three experimental ROC curves have almost no difference. Therefore, after selecting features by SFS and EFS, we can get the optimum simplified model for 5-year iDFS prediction. More detailed performance comparisons are shown in Table V.

TABLE VI
PERFORMANCE OF MODELS WITH DIFFERENT THRESHOLDS

Threshold	Number of features	AUC mean (cross validation)	AUC (test)	P
0.000	51	0.8447	0.8446	0.39233
0.005	42	**0.8459**	0.8452	0.43784
0.009	35	0.8435	**0.8473**	0.37027
0.012	30	0.8418	0.8426	0.20393
0.014	28	0.8420	0.8413	0.08978
0.016	23	0.8406	0.8459	-
0.020	**21**	0.8336	0.8381	**0.05751**

From Table V, we can see AUC values are 0.8486, 0.8446 and 0.8451 obtained at the different cutoff points, corresponding to models without feature selection (without FS), with statistical feature selection (After SFS), and with ensemble feature selection (After EFS). Relative to the model without feature selection, the AUC loss is only 0.0035 (0.4%) after done ensemble feature selecting. And at the same time, the F1 loss of 2.7% is minor. Therefore, the model is affected very lowly after eliminating some un-important features by using EFS method. On the whole, while a large number of features are deleted by using SFS and EFS methods, the model has had little impact ($p > 0.05$), just as shown in Fig. 8. But the number of features is decreased from 89 to 23 that greatly improve the model's applicability and predictability.

D. EFS Threshold Analysis

As illustrated in the above section, to select important features from 51 features after SFS, the threshold as a cut-off score to judge whether a feature is important or not plays an important role in EFS. Based on the final importance score, and considering both applicability and predictability, we use backward selection method to evaluate features with importance scores.

As shown in Table VI, to get the threshold to cut off an importance score, we try to use 7 different thresholds to screen relatively important features through applying backward selection step by step. Use the selected feature set to establish model, and tune parameters based on the TPE parameter optimization. we can train and test on the same dataset, and obtain the results as shown in Table VI. For example, when the threshold is 0.005, 10-fold cross validation is repeated 5 times. The average AUC value is 0.8459 that is the maximum AUC value when training the model, while the AUC is 0.8452 when testing the model. As shown in the table, with the threshold of 0.016 compared with that of 0.005, AUC loss is about 0.6% (<1.0%) when training the model. And the performance has no significant difference on the same testing set ($p = 0.43784$). But 18 features are decreased. That is a decrease of 42.9%. On the other hand, compared with the one with a threshold of 0.020, the AUC with a threshold value of 0.005 decreased by 1.5% (>1.0%), although the number of features is less, but only 3 features decreased. Therefore, considering the predictive ability and applicative ability, we chose the score 0.016 as the importance threshold to filter features in EFS. From the Fig. 9, we can see that curves go down very quickly from point 0.016 to point 0.02 on the horizontal axis.

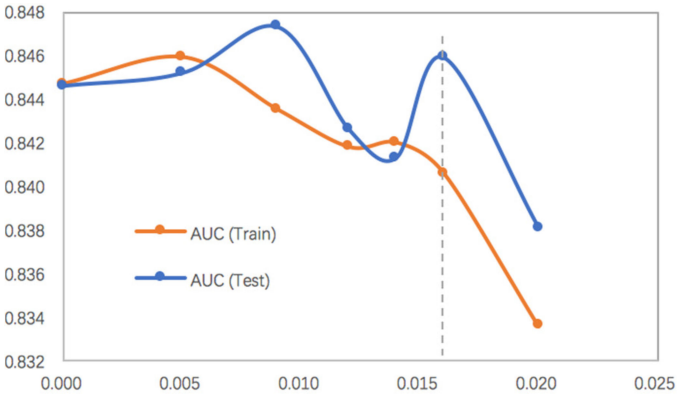


Fig. 9. AUC values at different thresholds.

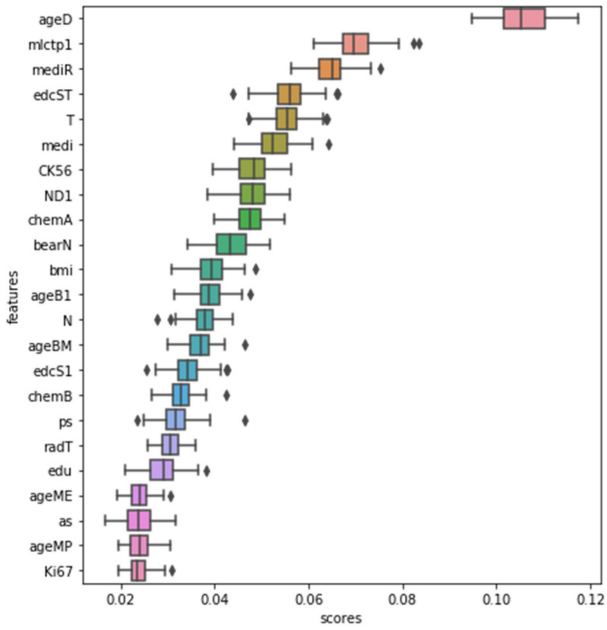


Fig. 10. Feature importance.

E. Feature Importance

In the XGBoost, *gain* method is applied to calculate importance scores. To improve the stability of calculating the importance score, we have done 100 times experiments to get the score distribution of each feature, graphically depicted by the boxplot diagram as shown in Fig. 10.

Seen from Fig. 10, the 5-year iDFS of early-stage breast cancer patients is influenced by top ten factors: age at diagnosis, molecular typing, claim ratio of medical insurance, type of endocrinotherapy, T stage, type of medical insurance, CK56, umber of I–II lymph node metastases, Postoperative chemotherapy, Number of births.

It is usually thought that histopathological factors, such as T stage, N stage and molecular typing, are important to affect the postoperative recurrence and metastasis for breast cancer patients [50]. In order to further understand the model's predictability, we divide the test dataset with 1050 patients into two sets: iDFS event actually happened (236 cases) and not happened (814 cases). we then can investigate the difference of the number of cases by observed and predicted for each feature. For example, Fig. 11 demonstrates the actual number of patients about each feature and the correctly predicted number of patients. Fig. 11(a), (b), (c), illustrate the important clinical features: T stage, N stage and molecular typing. Type of endocrinotherapy, claim ratio of medical insurance, age at diagnosis are demonstrated in Fig. 11(d), (e), (f).

TABLE VII
COMPARISON WITH OTHER FRAMEWORKS

Framework	Proposed	PREDICT	Adjuvant Online
Data sources	CRCB	ECRIC	US SEER
Factors	demographics, diagnosis, pathology and therapy	Age, Comorbidity, ER, HER2, Tumor grade, Tumor size, Positive nodes, Mode of detection, Therapy	Age, Menopausal status, Comorbidity, Tumor size, Positive nodes, ER, Therapy
Method	Gradient boosting decision tree	Cox regression, Multivariable, fractional polynomials	Bayesian method
AUC	0.845	0.805	-

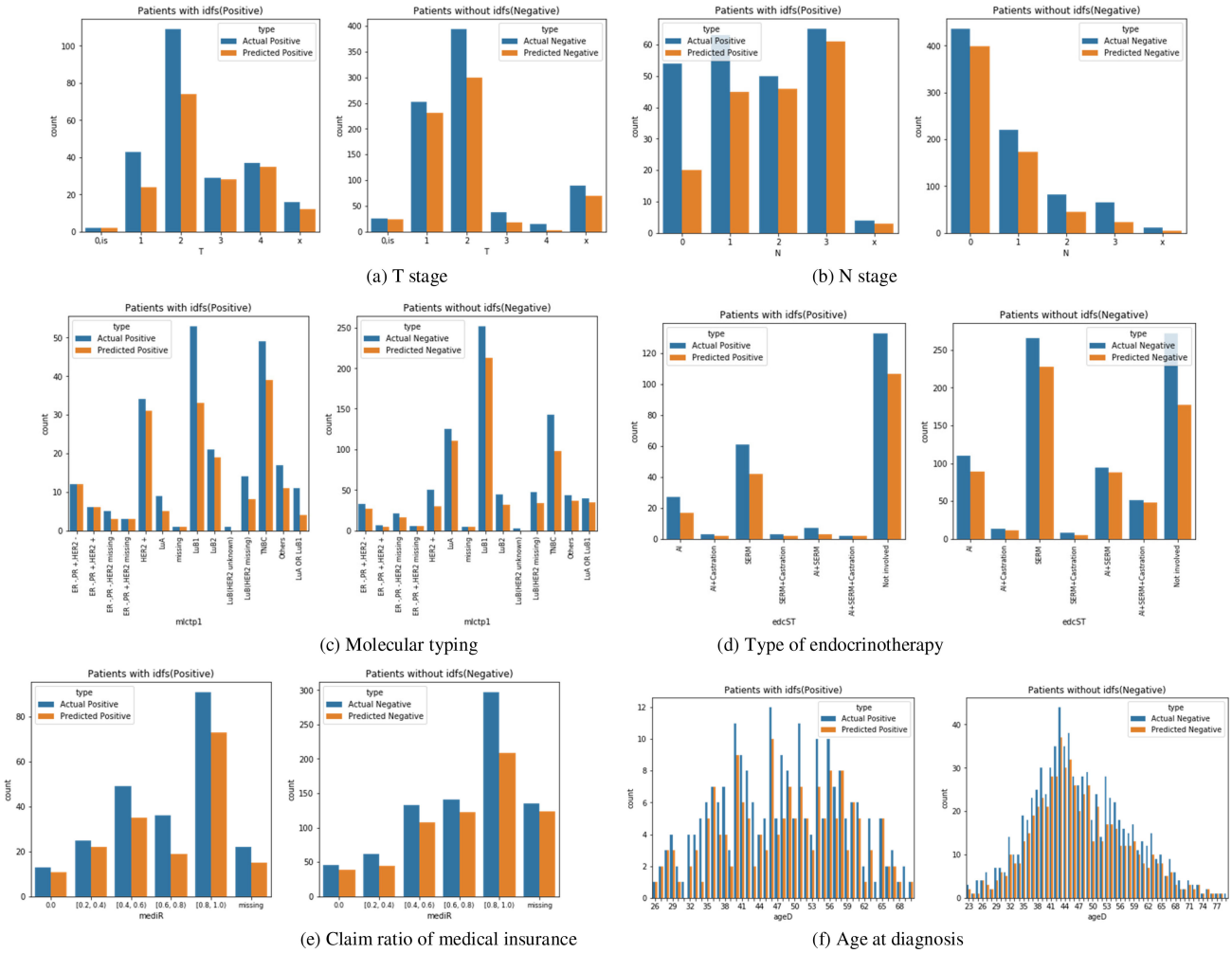
CRCB: Clinical Research Center for Breast in West China Hospital of Sichuan University. ECRIC: Eastern Cancer Registration and Information Centre. US SEER: US Surveillance, Epidemiology, and End Results.

F. Comparison with Other Frameworks

PREDICT and Adjuvant Online are two typical web-based prognosis prediction tools designed to predict survival probability of patients with breast cancer [35]–[37]. Both PREDICT and Adjuvant Online are endorsed by the American Joint Committee on Cancer [42]. PREDICT is based on the data from East Anglia Cancer Registration and Information Centre (ECRIC), and Adjuvant Online is with data from the US Surveillance, Epidemiology, and End Results (SEER) programs [35]–[37]. They are well validated on the cohort of breast cancer patients in both United States and Western Europe, and have been widely used in western countries [43]–[44]. However, validations performed in non-western countries have shown these tools may overestimate or underestimate the patient survival [45]–[47]. Breast cancer prognosis method for population in Asia, especially in China, is urgent to put forward.

Table VII shows the differences in data sources, feature factors, methods and AUC between our proposed method and other two frameworks.

As shown in Table VII, more patient features are included in our proposed frameworks. Features about demographics, diagnosis, pathology and therapy are all considered. For example, menstruation, fertility, urban/rural, claim ratio of medical insurance and so on are used to build the prognosis iDFS model, while they are excluded by PREDICT and Adjuvant Online. Fertility factors, such as parity, age at first birth and breast feeding, are strongly associated with the risk of breast cancer [38], while how to act on the prognosis of breast cancer patients remains unknown. Older age at first birth, overweight and obesity may



(a) T stage (b) N stage

(c) Molecular typing (d) Type of endocrinotherapy

(e) Claim ratio of medical insurance (f) Age at diagnosis

Fig. 11. Comparison between observed and predicted patients.

lead to poor breast cancer prognosis [39], [40]. There is a significant difference in prognosis between urban and rural breast cancer patients [41]. Therefore, the demographic and socioeconomic factors should be taken into account to construct the model also.

Our proposed framework, in which various factors are taken into account, gives an excellent performance to predict the relapse or metastasis breast cancer of Chinese patients in 5 years. As an alternative prognosis framework, it can be used in clinical practice to help doctor make appropriate plan of treatment.

V. Conclusion

This paper mainly introduces the research work that used statistics and machine learning methods to predict the risk of recurrence and metastasis of early breast cancer patients. Based on breast cancer clinical data derived from CRCB in the West China Hospital of Sichuan University, we analyzed and cleaned clinical data. Stratified Feature Selection, SFS and EFS, made the initial number of features decrease from 89 to 23. By comparing models before and after feature selection on the same dataset, it does not significantly affect the performance of the prediction model. After SFS and EFS, we achieved the purpose of eliminate redundant or meaningless variables.

By inputting 23 selected features into XGBoost algorithm, and tuning parameters, we found out an optimum simplified model for 5-year iDFS prediction for early-stage breast patients. we compared the results of SVM, RF and Adaboost with the proposed model, and found that the model constructed by using XGBoost algorithm clearly has a relatively good performance with its AUC 0.8451, and about 3% improved.

With the great development of machine learning, precision medical treatment is put forward to effectively treat early-stage breast cancer patients and reduce recurrence risk. We need to use more effective, accurate, and scientific methods to guide the treatment process. But on the other hand, more clinical or omics data should be introduced to improve the model. In future works, additional clinical data will be collected for improving accuracy of the 5-year iDFS prediction algorithm, and the iDFS system will be developed and tested in hospitals.

References

[1] H. J. Pandya *et al.*, "Toward a portable cancer diagnostic tool using a disposable MEMS-based biochip," *IEEE Trans. Biomed. Eng.*, vol. 63, no. 7, pp. 1347–1353, Jul. 2017

[2] R. Siegel *et al.*, "Cancer statistics," *CA Cancer J. Clin.*, vol. 64, no. 1, pp. 9–29, Jan./Feb. 2014.

[3] R. Lin and P. Tripuraneni, "Radiation therapy in early-stage invasive breast cancer," *Indian J. Surgical Oncology*, vol. 2, no. 2, pp. 101–111, Jun. 2011.

[4] M. A1-Badrashiny and A. Bellaachia, "Breast cancer survivability prediction via classifier ensemble," *Int. J. Comput. Inf. Eng.*, vol. 10, no. 5, pp. 833–837, 2016.

[5] R. K. Kavitha and D. Dorairangasamy, "Breast cancer survivability predictor using Adaboost and CART algorithm," *Int. J. Innovative Res. Sci., Eng. Technol.*, vol. 3, no. 1, pp. 351–353, 2014.

[6] J. kim and H. Shin, "Breast cancer survivability prediction using labeled, unlabeled, and pseudo-labeled patient data," *J. Amer. Med. Inform. Assoc.*, vol. 20, no. 4, pp. 613–618, 2013.

[7] E. Porter *et al.*, "An early clinical study of time-domain microwave radar for breast health monitoring," *IEEE Trans. Biomed. Eng.*, vol. 63, no. 3, pp. 530–539, Mar. 2016.

[8] J. Yoon *et al.*, "Discovery and clinical decision support for personalized healthcare," *IEEE J. Biomed. Health Inform.*, vol. 21, no. 4, pp. 1133–1145, Jul. 2017.

[9] 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.

[10] A. J. Bekker *et al.*, "Multi-view probabilistic classification of breast microcalcifications," *IEEE Trans. Med. Imag.*, vol. 35, no. 2, pp. 645–653, Feb. 2016.

[11] T. Saphner *et al.*, "Annual hazard rates of recurrence for breast cancer after primary therapy," *J. Clin. Oncology*, vol. 14, no. 10, pp. 2738–2746, 1996.

[12] A. M. Gonzalez-Angulo *et al.*, "High risk of recurrence for patients with breast cancer who have human epidermal growth factor receptor 2-positive, node-negative tumors 1 cm or smaller," *J. Clin. Oncology*, vol. 27, no. 34, pp. 5700–5706, Dec. 2009.

[13] Y. Wu *et al.*, "Breast cancer risk prediction using electronic health records," in *Proc. IEEE Inter. Conf. Healthcare Inform.*, 2017, pp. 23–26.

[14] A. Hazra *et al.*, "Predicting lung cancer survivability using SVM and logistic regression algorithms," *Int. J. Comput. Appl.*, vol. 174, no. 2, pp. 19–24, 2017.

[15] K. J. Wang *et al.*, "A hybrid classifier combining SMOTE with PSO to estimate 5-year survivability of breast cancer patients," *Appl. Soft Comput. J.*, vol. 20, no. 7, pp. 15–24, 2014.

[16] 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.

[17] F. Hutter *et al.*, "Sequential model-based optimization for general algorithm configuration," *Learning and Intelligent Optimization*. Berlin, Germany: Springer, 2011, pp. 507–523.

[18] J. Bergstra and Y. Bengio, "Algorithms for hyper-parameter optimization," in *Proc. Inter. Conf. Neural Inf. Process. Syst.*, 2011, pp. 2546–2554.

[19] C. Mballo and E. Diday, "The criterion of kolmogorov-smirnov for binary decision tree: Application to interval valued variables," *Intell. Data Anal.*, vol. 10, no. 4, pp. 325–341, 2006.

[20] J. F. Hilton, "The appropriateness of the Wilcoxon test in ordinal data," *Statist. Med.*, vol. 15, no. 6, pp. 631–645, Mar. 1996.

[21] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *J. Mach. Learn. Res.*, vol. 3, pp. 1157–1182, 2003.

[22] M. Kuhn and K. Johnson, "*An introduction to feature selection*," in *Applied Predictive Modeling*, Berlin, Germany: Springer, 2013, pp. 487–519.

[23] C. Sammut and G. I. Webb, *Encyclopedia of Machine Learning*. Berlin, Germany: Springer, 2011.

[24] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognit. Lett.*, vol. 27, no. 8, pp. 861–874, 2006.

[25] A. Simundic, "Measures of diagnostic accuracy: Basic definitions," *EJIFCC*, vol. 19, no. 4, pp. 203–211, Jan. 2009.

[26] C. Thornton *et al.*, "Auto-WEKA: Combined selection and hyperparameter optimization of classification algorithms," in *Proc. 19th ACM SIGKDD Inter. Conf. Knowledge Discovery Data Mining*, Aug. 2013, pp. 847–855.

[27] J. H. Kim, "Estimating classification error rate: Repeated cross-validation, repeated hold-out and bootstrap," *Comput. Statist. Data Anal.*, vol. 53, no. 11, pp. 3735–3745, 2009.

[28] 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.

[29] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.

[30] E. R. Delong *et al.*, "Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach," *Biometrics*, vol. 44, no. 3, pp. 837–845, 1988.

[31] M. Sehhati *et al.*, "Stable gene signature selection for prediction of breast cancer recurrence using joint mutual information," *IEEE/ACM Trans. Comput. Biol. Bioinform.*, vol. 12, no. 6, pp. 1440–1448, Dec. 2015.

[32] A. M. Alaa *et al.*, "ConfidentCare: A clinical decision support system for personalized breast cancer screening," *IEEE Trans. Multimedia*, vol. 18, no. 10, pp. 1942–1955, Oct. 2016.

[33] J. Xu *et al.*, "Stacked sparse autoencoder (SSAE) for nuclei detection on breast cancer histopathology images," *IEEE Trans. Med. Imag.*, vol. 35, no. 1, pp. 119–130, Jan. 2016.

[34] L. Zhang *et al.*, "Cancer progression prediction using gene interaction regularized Elastic Net," *IEEE/ACM Trans. Comput. Biol. Bioinform.*, vol. 14, no. 1, pp. 145–154, Jan./Feb. 2017.

[35] 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, pp. R1, 2010.

[36] 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, pp. 58, 2017, doi: 10.1186/s13058-017-0852-3.

[37] P. M. Ravdin *et al.*, "Computer program to assist in making decisions about adjuvant therapy for women with early breast cancer," *J. Clin. Oncology*, vol. 19, no. 4, pp. 980–991, 2001.

[38] M. Lambertini *et al.*, "Reproductive behaviors and risk of developing breast cancer according to tumor subtype: A systematic review and meta-analysis of epidemiological studies," *Cancer Treatment Rev.*, vol. 49, pp. 65–76, 2016.

[39] J. S. Lee and O. Minkyung, "Effects of interval between age at first pregnancy and age at diagnosis on breast cancer survival according to menopausal status: A register-based study in Korea," *BMC Women's Health*, vol. 14, no. 1, pp. 113, 2014. [Online]. Available: http://europepmc.org/abstract/MED/25715267

[40] M. D. K. Alsaker *et al.*, "The association of reproductive factors and breastfeeding with long term survival from breast cancer," *Breast Cancer Res. Treatment*, vol. 130, no. 1, pp. 175–182, 2011.

[41] Z. Peng *et al.*, "Diagnosis and treatment pattern among rural and urban breast cancer patients in Southwest China from 2005 to 2009," *Oncotarget*, vol. 7, no. 47, pp. 78168–78179, 2016.

[42] W. M. Lydiatt *et al.*, "Head and neck cancers-major changes in the American joint committee on cancer eighth edition cancer staging manual," *Cancer J. Clinicians*, vol. 67, no. 2, pp. 122–137, 2017.

[43] G. C. Wishart *et al.*, "A population-based validation of the prognostic model PREDICT for early breast cancer," *Eur. J. Surgical Oncology J. Eur. Soc. Surgical Oncology Brit. Assoc. Surgical Oncology*, vol. 37, no. 5, pp. 411–417, 2011.

[44] 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.

[45] H. S. Wong *et al.*, "The predictive accuracy of PREDICT: A personalized decision-making tool for southeast asian women with breast cancer," *Medicine*, vol. 94, no. 8, pp. e593, 2015. [Online]. Available: http://www.biomedcentral.com/1472-6874/14/113

[46] N. Bhoo-Pathy *et al.*, "Adjuvant! Online is overoptimistic in predicting survival of Asian breast cancer patients," *Eur. J. Cancer*, vol. 48, no. 7, pp. 982–989, 2012.

[47] D. Hajage *et al.* "External validation of Adjuvant! online breast cancer prognosis tool. Prioritising recommendations for improvement," *PLoS One*, vol. 6, no. 11, pp. e27446, 2011, doi: 10.1371/journal.pone.0027446.

[48] T. Ungi *et al.*, "Navigated breast tumor excision using electromagnetically tracked ultrasound and surgical instruments," *IEEE Trans. Biomed. Eng.*, vol. 63, no. 3, pp. 600–606, Mar. 2016.

[49] K. Huang *et al.*, "Maximizing sensitivity in medical diagnosis using biased minimax probability machine," *IEEE Trans. Biomed. Eng.*, vol. 53, no. 5, pp. 821–831, May 2006.

[50] H. Fatakdawala *et al.*, "Expectation–maximization-driven geodesic active contour with overlap resolution (EMaGACOR): Application to lymphocyte segmentation on breast cancer histopathology," *IEEE Trans. Biomed. Eng.*, vol. 57, no. 7, pp. 1676–1689, Jul. 2010.

[51] M. R. Mohebian *et al.*, "A hybrid computer-aided-diagnosis system for prediction of breast cancer recurrence (HPBCR) using optimized ensemble learning," *Comput. Structural Biotechnol. J.*, vol. 15, pp. 75–85, 2017.

[52] 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.