Original Paper

The Application and Comparison of Machine Learning Models for the Prediction of Breast Cancer Prognosis: Retrospective Cohort Study

Jialong Xiao[1,2,3], BS, MS; Miao Mo[2,3], BS, MS; Zezhou Wang[2,3], BS, MS; Changming Zhou[2,3], BS, MS, PhD; Jie Shen[2,3], BS, MS, PhD; Jing Yuan[2,3], BS; Yulian He[1,2], BS, MS; Ying Zheng[2,3,4], BS, MS

[1]Department of Epidemiology, School of Public Health, Fudan University, Shanghai, China

[2]Department of Cancer Prevention, Fudan University Shanghai Cancer Center, Shanghai, China

[3]Department of Oncology, Shanghai Medical College, Fudan University, Shanghai, China

[4]Shanghai Engineering Research Center of Artificial Intelligence Technology for Tumor Diseases, Shanghai, China

Corresponding Author:
Ying Zheng, BS, MS
Department of Cancer Prevention
Fudan University Shanghai Cancer Center
270 Dong 'an Road, Xuhui District
Shanghai, 200000
China
Phone: 86 21 64175590
Fax: 86 21 64175590
Email: zhengying@fudan.edu.cn

Abstract

Background: Over the recent years, machine learning methods have been increasingly explored in cancer prognosis because of the appearance of improved machine learning algorithms. These algorithms can use censored data for modeling, such as support vector machines for survival analysis and random survival forest (RSF). However, it is still debated whether traditional (Cox proportional hazard regression) or machine learning-based prognostic models have better predictive performance.

Objective: This study aimed to compare the performance of breast cancer prognostic prediction models based on machine learning and Cox regression.

Methods: This retrospective cohort study included all patients diagnosed with breast cancer and subsequently hospitalized in Fudan University Shanghai Cancer Center between January 1, 2008, and December 31, 2016. After all exclusions, a total of 22,176 cases with 21 features were eligible for model development. The data set was randomly split into a training set (15,523 cases, 70%) and a test set (6653 cases, 30%) for developing 4 models and predicting the overall survival of patients diagnosed with breast cancer. The discriminative ability of models was evaluated by the concordance index (C-index), the time-dependent area under the curve, and D-index; the calibration ability of models was evaluated by the Brier score.

Results: The RSF model revealed the best discriminative performance among the 4 models with 3-year, 5-year, and 10-year time-dependent area under the curve of 0.857, 0.838, and 0.781, a D-index of 7.643 (95% CI 6.542, 8.930) and a C-index of 0.827 (95% CI 0.809, 0.845). The statistical difference of the C-index was tested, and the RSF model significantly outperformed the Cox-EN (elastic net) model (C-index 0.816, 95% CI 0.796, 0.836; P=.01), the Cox model (C-index 0.814, 95% CI 0.794, 0.835; P=.003), and the support vector machine model (C-index 0.812, 95% CI 0.793, 0.832; P<.001). The 4 models' 3-year, 5-year, and 10-year Brier scores were very close, ranging from 0.027 to 0.094 and less than 0.1, which meant all models had good calibration. In the context of feature importance, elastic net and RSF both indicated that TNM staging, neoadjuvant therapy, number of lymph node metastases, age, and tumor diameter were the top 5 important features for predicting the prognosis of breast cancer. A final online tool was developed to predict the overall survival of patients with breast cancer.

Conclusions: The RSF model slightly outperformed the other models on discriminative ability, revealing the potential of the RSF method as an effective approach to building prognostic prediction models in the context of survival analysis.

(JMIR Med Inform 2022;10(2):e33440) doi: 10.2196/33440

KEYWORDS

breast cancer; machine learning; survival analysis; random survival forest; support vector machine; medical informatics; prediction models

Introduction

Breast cancer is a leading cause of morbidity and mortality in women worldwide, and the prediction of breast cancer prognosis is crucial for decision-making. Accurate outcome prediction can assist doctors with providing appropriate treatment plans for patients, which in turn could improve their chances of survival and lessen the suffering. Several prognostic prediction models have already been developed. PREDICT and Adjuvant! Online are 2 famous prognostic prediction tools for breast cancer based on clinical and pathological characteristics [1,2]. These models have been validated by external data set and are commonly used in the United States and Western Europe. However, several external validations that were made in Asian countries revealed a less-than-optimal predictive ability [3-6].

For survival analysis of follow-up observations, the most important challenge is dealing with censored data. The Cox proportional hazard regression is a classical modeling method used to analyze right-censored data in survival analysis with good interpretability. Typically, the Cox proportional hazard regression imposes proportional hazard assumption and the assumption that continuous covariates have a linear effect on the logarithm of the hazard, which the real-world data may not satisfy [7]. Compared with the Cox proportional hazard regression, machine learning methods do not make any parametric or semiparametric assumptions and have the ability to detect and account for higher-order interactions as well as nonlinear relationships [8]. While there have been some attempts to use machine learning to build cancer prognosis prediction models [6,9-13], currently, there is no consensus on whether traditional or machine learning-based prognostic prediction models have a better predictive performance.

Here, we discuss two main types of prognostic prediction models using machine learning algorithms. The first types are the binary classification models, which give a probability of the interested outcome at a specific time. Several studies have used machine learning methods to generate prognostic prediction models based on classification. The outcome variable of these models is the status of survival at 5 years [14-17] or at the time of data collection [18,19]. The limitation of these models is that they are not able to include right-censored observations that were censored before the specified time, because the outcome of these observations is unknown. Moreover, using the classification outcome (survival status at a specific time) instead of the survival outcome (survival time and status of the censor) can lead to a loss of information. The second types are models using improved algorithms of original machine learning algorithms to enable modeling and analysis of censored data, such as support vector machines (SVM) for survival analysis [20] and random survival forest (RSF) [21]. These methods can describe probability (RSF) and risk scores (SVM and RSF) of the interested outcomes at different time points rather than at a specific time point and can consider both the survival time and the status of the censor.

In this study, traditional (Cox) and machine learning-based (SVM and RSF) prognostic prediction models were developed for patients with breast cancer based on a large cohort of Chinese patients diagnosed with breast cancer and hospitalized in Fudan University Shanghai Cancer Center. We aimed to compare the performance of different models to pick the optimal predictive model and provide a reference for the development of machine learning in the prognosis prediction of breast cancer.

Methods

Study Design and Ethical Considerations

This retrospective cohort study included all patients diagnosed with breast cancer and subsequently hospitalized in Fudan University Shanghai Cancer Center between January 1, 2008, and December 31, 2016. Data containing demographic and clinicopathologic features were obtained from the hospital information system. Overall Survival, defined as the duration between the time of first treatment and the date of death, was taken as the outcome to build the predictive models. The outcome information was derived from medical visit records, telephone visits, and death certificate data linkage with the cancer registry system or death certificate system run by the provincial Centers for Disease Control and Prevention.

By March 1, 2021, medical information and follow-up information were collected from 25,629 patients. After excluding male patients, patients with bilateral breast cancer (362 cases), and patients with ≥3 missing features, 22,176 cases with 21 features were eligible for further analysis. Patients were followed for a median follow-up time of 68.9 months (95% CI 68.42, 69.33). The data set was then randomly split into a training set (15,523 cases, 70%) and a test set (6653 cases, 30%). The statistical description of features and the survival curves of patients in the training and test set are shown in Table S1 and Figure S1 in Multimedia Appendix 1.

This study was approved by the Fudan University Shanghai Cancer Center Institutional Review Board (Registration YF-2021-01).

Preprocess of Missing Data

Since the data were generated and collected in a real medical environment, there were many observations with missing features. As the SVM and RSF methods do not support the analysis of data sets with missing values, we performed a 2-step process in order to reduce the impact of missing values on the training process of developing prediction models. Firstly, we excluded patients with too many missing features. The number of missing features of patients and the log-rank test results are shown in Table S1 in Multimedia Appendix 2. The log-rank method was used to test the difference between the survival state of 25,267 patients and the remaining patients. Based on the results of the log-rank test, when we excluded patients with ≥3 missing features, there was no significant difference between the survival of the remaining patients (22,176 cases) and the

survival of the overall patients (25,267 cases; *P*=.17). Therefore, 3 was taken as the cut-off value, and patients with ≥3 missing features were excluded. The statistics for missing features before and after the first step of processing are shown in Table S2 in Multimedia Appendix 2, and the remaining 22,176 cases are eligible for further analysis. Secondly, the remaining missing data were imputed by the missForest algorithm using library "missingpy" (0.2.0) in Python (Python Software Foundation). MissForest is a nonparametric imputation method that could be applied for both continuous and categorical variables and does not make explicit assumptions about the functional form of the data [22]. In the process of imputing the missing values, the outcome data were not involved in case imputed data were affected and falsely related to the outcome data.

Statistical Analysis

The objective outcome in the study was time to event, which is right-censored survival data. Therefore, the following 3 survival modeling approaches were used to predict the survival time of patients diagnosed with breast cancer: Cox proportional hazard regression [23], SVM [24], and RSF [21]. Elastic Net (EN) was used as the feature selection method to screen important features to train the 3 models. Technical implementation details, including the libraries and the process of hyperparameter tuning, are provided in the Multimedia Appendix 3. Moreover, we have open sourced the Python and R code that we developed for generating the models and evaluating the performance of the models in the GitHub repository [25].

The Cox proportional hazard regression is a classical modeling method for survival analysis. The model predicts the probability that the event of interest has occurred at a given time for given values of the predictor variables [23]. We added a traditional feature selection method for the Cox model, where univariate Cox analysis was performed before significant (*P*<.1) and clinically relevant features were forced into multivariate Cox regression analysis. The Cox model using the EN method was named "Cox-EN," and the one using the traditional variable selection method was named "Cox."

Usually, the predictors should satisfy the proportional hazard assumption in the Cox model. However, the main goal of modeling in this study was survival prediction and maximizing concordance index (C-index) and time-dependent area under the curve (AUC), regardless of how predictions are generated. Therefore, we did not perform the test for proportional hazards in the process of modeling [26].

SVM is a supervised machine learning algorithm, which can be used for both classification and regression challenges. An SVM model is basically a representation of different classes in a hyperplane in multidimensional space. The hyperplane is generated iteratively by SVM so that the error can be minimized. The goal of SVM is to divide the data sets into classes to find a maximum marginal hyperplane [24].

Several extensions of SVM to survival analysis were proposed. Shivaswamy et al [27] introduced an approach for censored targets by casting survival analysis as a regression problem. Van Belle et al [24,28] proposed the ranking approach and the hybrid approach combining the regression and ranking approach for survival outcomes. As an objective function of the ranking-based technique depends on a quadratic number of constraints with respect to the number of training samples, which makes training intractable with medium to large-sized data sets, we chose an approach of efficient training of linear survival SVM [20].

RSF, which was developed by Ishwaran et al [21], is an ensemble of tree-based learners for survival analysis of right-censored data and an extension of the random forest method. Using independent bootstrap samples, each tree in RSF is grown by randomly selecting a subset of features for each node and then splitting the node using a survival criterion involving information of survival time and censoring status [21].

EN is a feature screening technique that uses the penalties L1 and L2 from both the least absolute shrinkage and selection operator (LASSO) and ridge techniques to regularize regression models. The EN method is improved based on the shortcomings of both ridge and LASSO methods. The ridge method keeps all the features and cannot perform the function of feature screening. When it comes to multiple correlated features, the LASSO method randomly picks one of these features from such groups and entirely ignores the rest, while the EN method is likely to pick a few at once [29].

Evaluation of Model Performance

The discriminative ability of models was evaluated by the C-index [30], time-dependent AUC [31], and D-index [32]. C-index measures the overall discriminative ability of models, while time-dependent AUC measures the discriminative ability of models by comparing the predicted probabilities with the actual binary survival status and the probability estimation of a death outcome of censored observations at an interested time. C-index and time-dependent AUC both range in an interval from 0 to 1, and a value of 0.5 is comparable to random guessing, while a value of 1 means perfect discrimination. D-index was used to measure the separation between patients from equally sized high-risk and low-risk groups divided according to the risk score obtained from different models. Higher values of D-index indicate a more remarkable discriminative ability of the model. The survival curves of high-risk and low-risk groups was estimated using the Kaplan-Meier method, and the log-rank test was used to compare survival curves. The calibration ability of models was evaluated by the Brier score [33], which varies between 0 and 1, while a lower Brier score was indicative of a better-calibrated prediction. A value of 0.25 is comparable to random guessing, while a value of 0 means perfect discrimination.

Results

User and Model Statistics

A total of 22,176 patients with 68.9 months (95% CI 68.42, 69.33) of median follow-up were included in this study. We fitted 4 prognostic models (Cox, Cox-EN, RSF, and SVM) for predicting the overall survival of breast cancer patients with the training set and then used C-index, time-dependent AUC, D-index, and Brier score to evaluate them in the independent

test set. All models showed good calibration, and RSF outperformed other models on discriminative ability with a C-index of 0.827 (95% CI 0.809, 0.845).

Evaluation of Feature Importance

In order to screen out features with a large contribution to predicting the prognosis of breast cancer, the EN was first used to select important features, resulting in a total of 21 features. The ways the coefficients changed for varying α is shown in Figure 1, and the coefficient of each feature corresponding to the optimal α is shown in Figure 2. The top 5 important features were TNM staging, neoadjuvant therapy, number of lymph node metastases, age, and diameter of the tumor. RSF was used to rank the importance of the 11 features selected by the EN, and the results are shown in Figure 3. The top 5 important features were the number of lymph node metastasis, age, tumor diameter, neoadjuvant therapy, and TNM staging.

The results of univariate and multivariate Cox analysis are shown in Multimedia Appendix 4. Except for cases of the side of the tumor, multiple tumors, adjuvant chemotherapy, and targeted therapy, all features had a *P* value of less than .1 in the univariate analysis. Considering that adjuvant chemotherapy and targeted therapy could be confounding factors, multivariate analysis was performed using adjuvant chemotherapy, targeted therapy, and the significant factors (*P*<.1) from univariate analysis. The results of the multivariate analyses showed that age, menopause, invasive, diameter, lymph node metastasis, TNM, Ki 67, estrogen receptors, progesterone receptors, breast surgery, axillary surgery, adjuvant chemotherapy, targeted therapy, adjuvant radiotherapy, adjuvant endocrine therapy, and neoadjuvant therapy had a *P* value of less than .05, and the Cox model was built by these features.

Figure 1. The coefficients of features change for varying α.

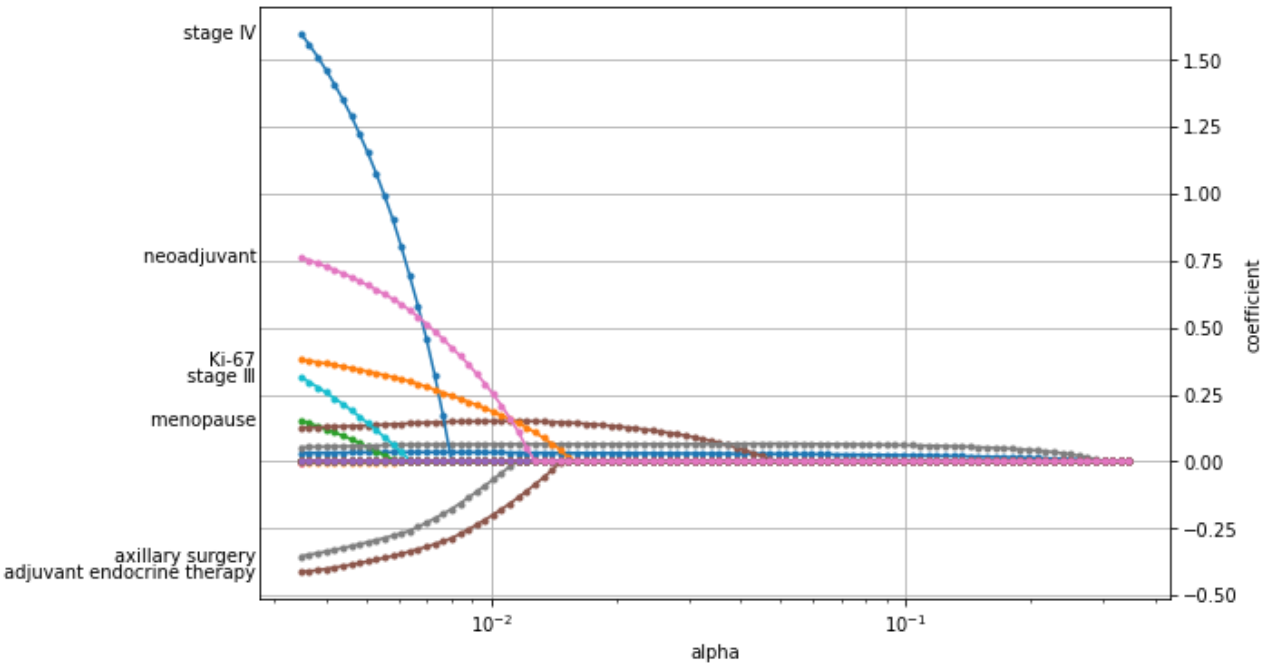


Figure 2. The important coefficient of each feature corresponding to the optimal α by elastic net. Ln: lymph node; PR: progesterone receptors.

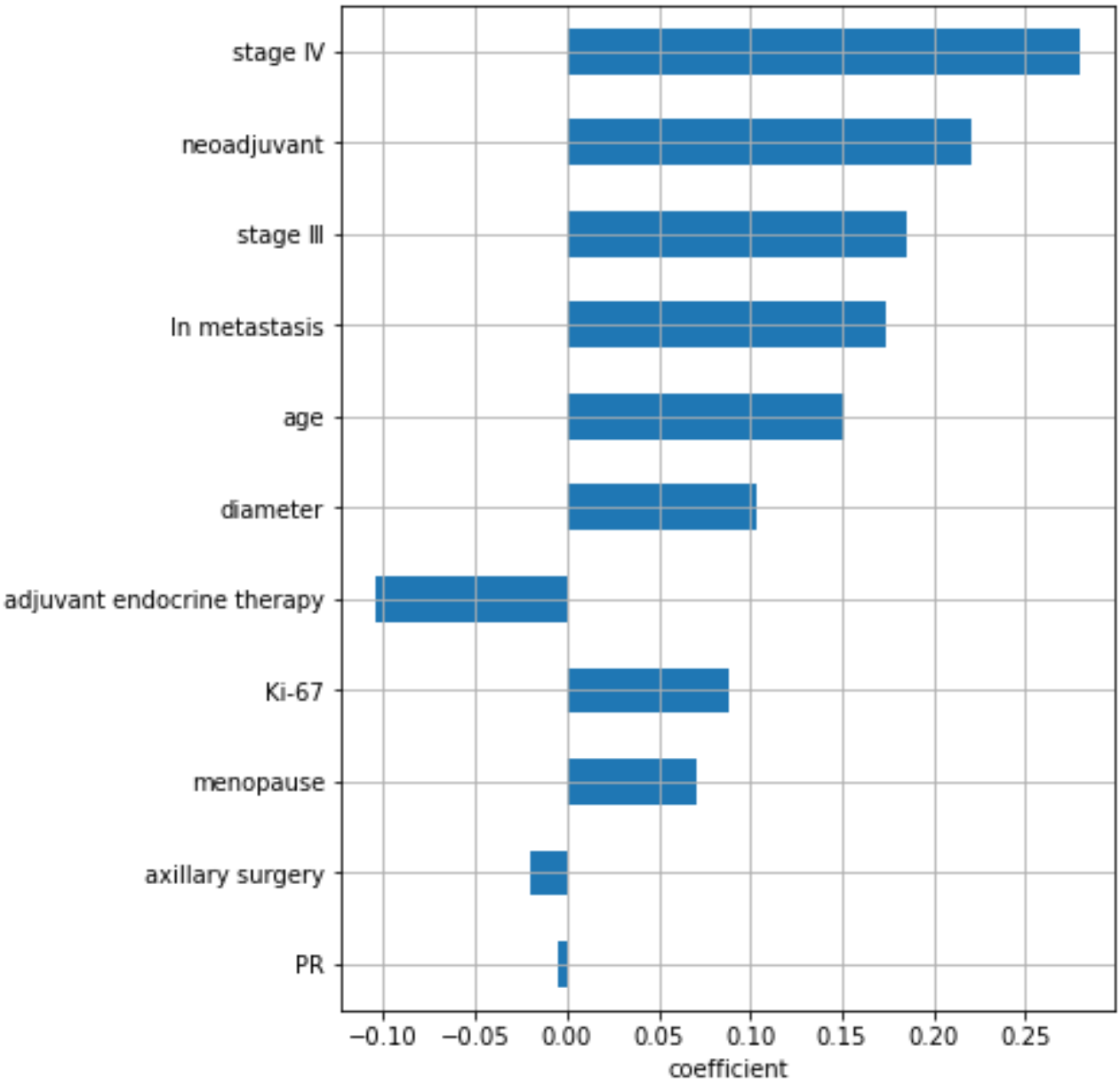
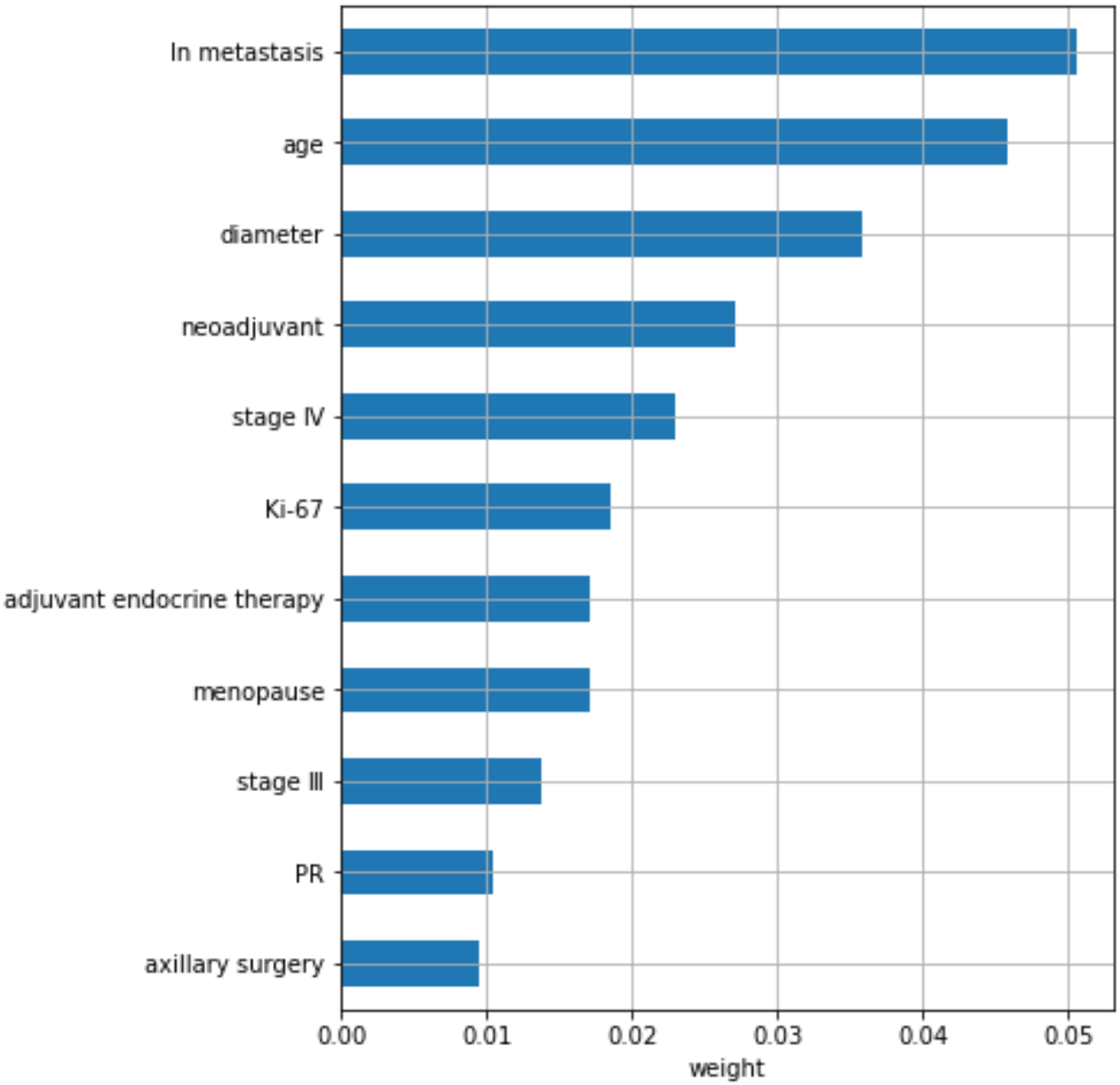


Figure 3. The important coefficient of each feature by random survival forest. Ln: lymph node; PR: progesterone receptors.



Methods Performance

Evaluation results of the 4 models are shown in Table 1. From the point of view of the C-index, the RSF model slightly and significantly outperformed the Cox-EN model (P=.01), the Cox model (P=.003), and the SVM model (P<.001) on discriminative ability, and no significant difference was found between the discriminative ability of other models. Time-dependent receiver operating characteristic curves of each model at 3 years, 5 years, and 10 years are shown in Figure 4. The time-dependent AUC of each model over time is shown in Figure 5. As shown in Figure 5, the time-dependent AUC of RSF was the highest at most times. Survival curves of the high-risk and low-risk groups divided according to the risk score are shown in Figure 6. The D-index of 7.643 from the RSF model was also the highest, and it can be interpreted as the risk of death in the high-risk group, which is 7.643 times the risk of death in the low-risk group. The 4 models' 3-year, 5-year, and 10-year Brier scores were all <0.1, suggesting that all models had good calibration.

Table 1. Performance of different methods.

Indexes	Cox	Cox-EN[a]	SVM[b]	RSF[c]
C-index[d] (95% CI)	0.814 (0.794,0.835)	0.816 (0.796,0.836)	0.812 (0.793,0.832)	0.827 (0.809,0.845)
AUC[e] (3 years)	0.850	0.857	0.847	0.857
AUC (5 years)	0.821	0.822	0.823	0.838
AUC (10 years)	0.770	0.769	0.760	0.781
D-index (95% CI)	7.210 (6.172,8.424)	7.466 (6.383,8.733)	6.522 (5.606,7.583)	7.643 (6.542,8.930)
Brier score (3 years)	0.027	0.027	—[f]	0.027
Brier score (5 years)	0.044	0.045	—	0.045
Brier score (10 years)	0.094	0.093	—	0.093

[a]EN: elastic net.

[b]SVM: support vector machine.

[c]RSF: random survival forest.

[d]C-index: concordance index.

[e]AUC: area under the curve.

[f]Not available. Survival support vector machine can only predict a risk score and not a probability. Therefore, Brier score is not available for survival support vector machine.

Figure 4. Time-dependent receiver operating characteristic curves of models at 3 years, 5 years, and 10 years. EN: elastic net; RSF: random survival forest; SVM: support vector machine.

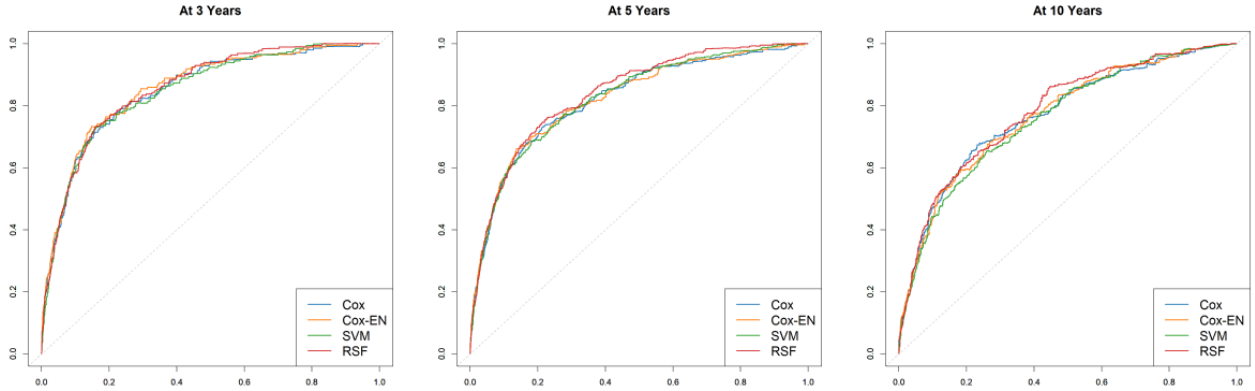


Figure 5. Time-dependent AUC of models over time. AUC: area under the curve; EN: elastic net; RSF: random survival forest; SVM: support vector machine.

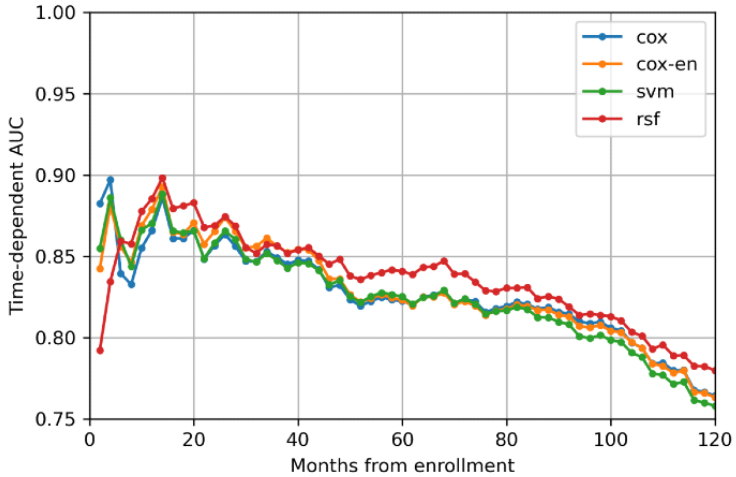
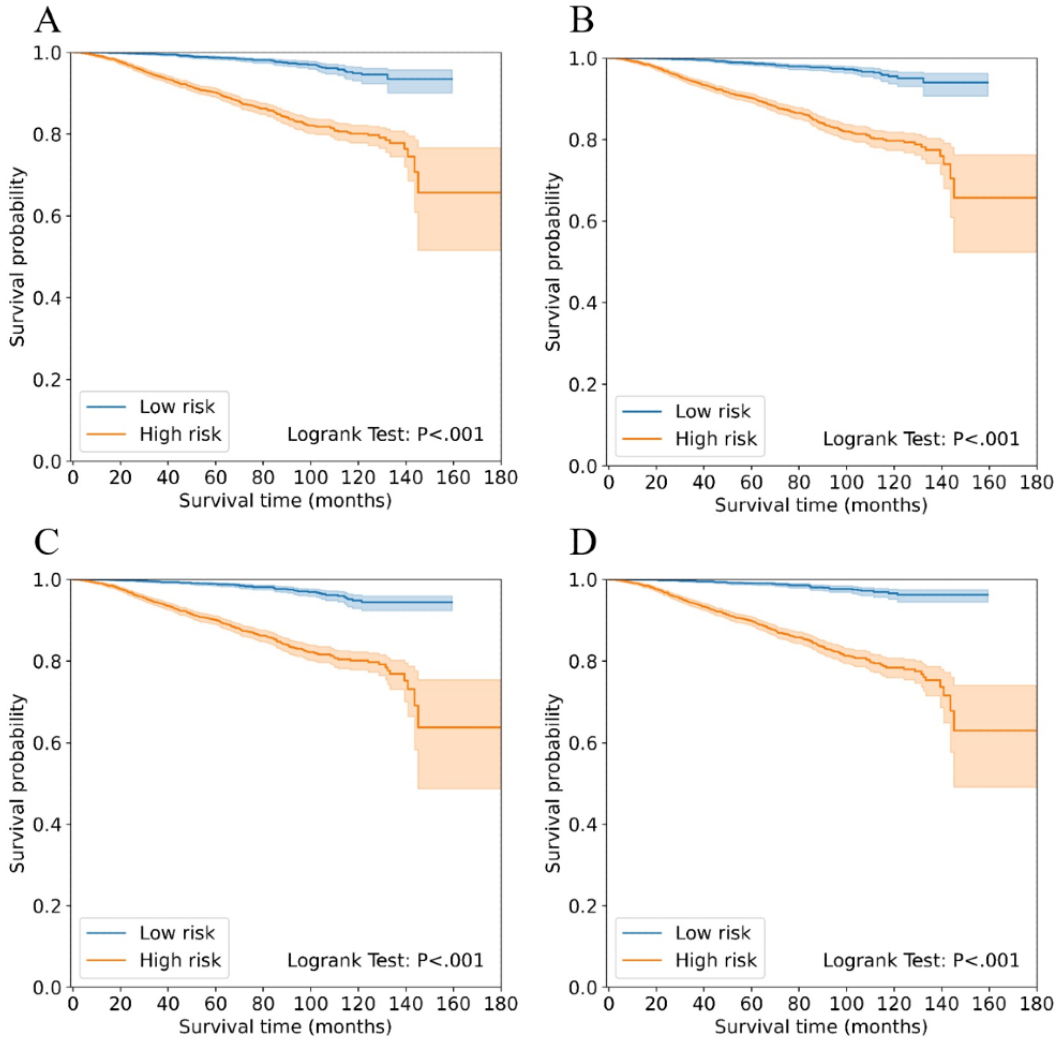


Figure 6. Survival curves of high-risk and low-risk groups divided according to the risk score from (A) Cox, (B) Cox-EN (elastic net), (C) SVM (support vector machine), and (D) RSF (random survival forest).



Online Prognostic Prediction Tool

Although the RSF model achieved the best performance among these models, the interpretability and computational efficiency of the RSF model had to be considered at the same time in the deployment of the online prognostic prediction tool. The memory usage of the RSF model was too large for the model to be deployed on a website and have good computational efficiency. The Cox-EN model achieved suboptimal performance in the study and had better interpretability and computational efficiency compared with the RSF model. Therefore, it was selected as a backend for the online prognostic prediction tool [34].

Discussion

In this paper, we compared the performance of traditional (Cox) and machine learning-based (SVM and RSF) prognostic prediction models for patients diagnosed with breast cancer and found out the RSF model slightly and significantly outperformed the Cox-EN model, the Cox model, and the SVM model on discriminative ability. Compared with Cox, Cox-EN, and SVM, the RSF model had a slightly better performance with a C-index of 0.827 (95% CI 0.809, 0.845) and 3-year, 5-year, and 10-year time-dependent AUC of 0.857, 0.838, and 0.781, respectively. The results in this study were similar to those reported by some previous studies. For example, Liu et al [10] used several methods, including RSF and Cox, to predict breast cancer progression with a sample size of 4575 patients. The results showed that the RSF model achieved better performance with a C-index of 0.814 compared with the Cox model with a C-index of 0.759. Rahman et al [35] showed that RSF (5-year time-dependent AUC 0.839, 95% CI 0.826, 0.849) outperformed Cox (5-year time-dependent AUC 0.823, 95% CI 0.811, 0.833) in the survival prediction of patients with esophageal cancer.

The possible reason for RSF achieving better performance may be that RSF is able to detect and account for higher-order interactions and nonlinear relationships. However, despite the great predictive performance of RSF, there are several shortcomings that limit the wide adoption of RSF. Firstly, the theoretical properties and the inferential procedures of RSF are not well understood. Secondly, RSF creates a “black-box” model that is hard to interpret or visualize [8]. Nonetheless, RSF still has the potential to be used as an effective approach to build prognostic prediction models in the context of survival analysis.

A major advantage of this paper was the large-scale prospective cohort design with a long follow-up time. To the best of our knowledge, this is the study with the largest sample size for breast cancer prognostic prediction modeling based on machine learning in the Chinese population thus far. Even though the study is based on a single institution, the large-scale prospective cohort and long follow-up time make the results valuable and credible.

There are some limitations in this study that should be acknowledged. The main limitation is that this study was performed in a single center in China with no external validation. Therefore, the current results need further multi-institutional validation with larger samples before the prediction models could be used in clinical practice. Another limitation relates to missing data that were imputed, and we could not ascertain the effect of the imputation of missing data on the overall results and subsequent conclusions. Moreover, we chose the randomized search method with 50 parameter settings sampled instead of grid search in the process of tuning the hyperparameters of the RSF due to the limitation of the computational efficiency. This may cause an underestimate of the performance of the RSF model.

In summary, the RSF model slightly outperformed the other models on discriminative ability, revealing the potential of the RSF method to be used as an effective approach to build prognostic prediction models in the context of survival analysis. Our future work will focus on additional external validation of the model using data from multiple centers to verify the extrapolation of our results.

Conflicts of Interest

None declared.

Multimedia Appendix 1

The statistical description of features and the survival curves of patients in the training and test set.
[DOCX File , 204 KB-Multimedia Appendix 1]

Multimedia Appendix 2

Statistics for missing fields and missing features before and after processing.
[DOCX File , 16 KB-Multimedia Appendix 2]

Multimedia Appendix 3

Technical implementation details.
[DOCX File , 16 KB-Multimedia Appendix 3]

Multimedia Appendix 4

Results of univariate survival analysis and multivariate survival analysis.
[DOCX File , 18 KB-Multimedia Appendix 4]

References

1. Candido Dos Reis FJ, Wishart GC, Dicks EM, Greenberg D, Rashbass J, Schmidt MK, et al. An updated PREDICT breast cancer prognostication and treatment benefit prediction model with independent validation. Breast Cancer Res 2017 May 22;19(1):58 [FREE Full text] [doi: 10.1186/s13058-017-0852-3] [Medline: 28532503]
2. Ravdin PM, Siminoff LA, Davis GJ, Mercer MB, Hewlett J, Gerson N, et al. Computer program to assist in making decisions about adjuvant therapy for women with early breast cancer. J Clin Oncol 2001 Feb 15;19(4):980-991. [doi: 10.1200/JCO.2001.19.4.980] [Medline: 11181660]
3. Bhoo-Pathy N, Yip C, Hartman M, Saxena N, Taib NA, Ho G, et al. Adjuvant! Online is overoptimistic in predicting survival of Asian breast cancer patients. Eur J Cancer 2012 May;48(7):982-989 [FREE Full text] [doi: 10.1016/j.ejca.2012.01.034] [Medline: 22366561]
4. Wong H, Subramaniam S, Alias Z, Taib NA, Ho G, Ng C, et al. The predictive accuracy of PREDICT: a personalized decision-making tool for Southeast Asian women with breast cancer. Medicine (Baltimore) 2015 Feb;94(8):e593 [FREE Full text] [doi: 10.1097/MD.0000000000000593] [Medline: 25715267]
5. Zaguirre K, Kai M, Kubo M, Yamada M, Kurata K, Kawaji H, et al. Validity of the prognostication tool PREDICT version 2.2 in Japanese breast cancer patients. Cancer Med 2021 Mar;10(5):1605-1613 [FREE Full text] [doi: 10.1002/cam4.3713] [Medline: 33452761]
6. Zhong X, Luo T, Deng L, Liu P, Hu K, Lu D, et al. Multidimensional Machine Learning Personalized Prognostic Model in an Early Invasive Breast Cancer Population-Based Cohort in China: Algorithm Validation Study. JMIR Med Inform 2020 Nov 09;8(11):e19069 [FREE Full text] [doi: 10.2196/19069] [Medline: 33164899]

7. Goerdten J, Carrière I, Muniz-Terrera G. Comparison of Cox proportional hazards regression and generalized Cox regression models applied in dementia risk prediction. Alzheimers Dement (N Y) 2020;6(1):e12041 [FREE Full text] [doi: 10.1002/trc2.12041] [Medline: 32548239]
8. Hu C, Steingrimsson JA. Personalized Risk Prediction in Clinical Oncology Research: Applications and Practical Issues Using Survival Trees and Random Forests. J Biopharm Stat 2018;28(2):333-349 [FREE Full text] [doi: 10.1080/10543406.2017.1377730] [Medline: 29048993]
9. Senders JT, Staples P, Mehrtash A, Cote DJ, Taphoorn MJB, Reardon DA, et al. An Online Calculator for the Prediction of Survival in Glioblastoma Patients Using Classical Statistics and Machine Learning. Neurosurgery 2020 Feb 01;86(2):E184-E192 [FREE Full text] [doi: 10.1093/neuros/nyz403] [Medline: 31586211]
10. Liu P, Fu B, Yang SX, Deng L, Zhong X, Zheng H. Optimizing Survival Analysis of XGBoost for Ties to Predict Disease Progression of Breast Cancer. IEEE Trans Biomed Eng 2021 Jan;68(1):148-160. [doi: 10.1109/TBME.2020.2993278] [Medline: 32406821]
11. Qiu X, Gao J, Yang J, Hu J, Hu W, Kong L, et al. A Comparison Study of Machine Learning (Random Survival Forest) and Classic Statistic (Cox Proportional Hazards) for Predicting Progression in High-Grade Glioma after Proton and Carbon Ion Radiotherapy. Front Oncol 2020;10:551420 [FREE Full text] [doi: 10.3389/fonc.2020.551420] [Medline: 33194609]
12. Tran BX, Latkin CA, Sharafeldin N, Nguyen K, Vu GT, Tam WWS, et al. Characterizing Artificial Intelligence Applications in Cancer Research: A Latent Dirichlet Allocation Analysis. JMIR Med Inform 2019 Sep 15;7(4):e14401 [FREE Full text] [doi: 10.2196/14401] [Medline: 31573929]
13. Cos H, Li D, Williams G, Chininis J, Dai R, Zhang J, et al. Predicting Outcomes in Patients Undergoing Pancreatectomy Using Wearable Technology and Machine Learning: Prospective Cohort Study. J Med Internet Res 2021 Mar 18;23(3):e23595 [FREE Full text] [doi: 10.2196/23595] [Medline: 33734096]
14. Abdikenov B, Iklassov Z, Sharipov A, Hussain S, Jamwal PK. Analytics of Heterogeneous Breast Cancer Data Using Neuroevolution. IEEE Access 2019;7:18050-18060. [doi: 10.1109/access.2019.2897078]
15. Zhao M, Tang Y, Kim H, Hasegawa K. Machine Learning With K-Means Dimensional Reduction for Predicting Survival Outcomes in Patients With Breast Cancer. Cancer Inform 2018;17:1176935118810215 [FREE Full text] [doi: 10.1177/1176935118810215] [Medline: 30455569]
16. García-Laencina PJ, Abreu PH, Abreu MH, Afonoso N. Missing data imputation on the 5-year survival prediction of breast cancer patients with unknown discrete values. Comput Biol Med 2015 Apr;59:125-133. [doi: 10.1016/j.compbiomed.2015.02.006] [Medline: 25725446]
17. Lotfnezhad Afshar H, Ahmadi M, Roudbari M, Sadoughi F. Prediction of breast cancer survival through knowledge discovery in databases. Glob J Health Sci 2015 Jan 26;7(4):392-398 [FREE Full text] [doi: 10.5539/gjhs.v7n4p392] [Medline: 25946945]
18. Chao C, Yu Y, Cheng B, Kuo Y. Construction the model on the breast cancer survival analysis use support vector machine, logistic regression and decision tree. J Med Syst 2014 Oct;38(10):106. [doi: 10.1007/s10916-014-0106-1] [Medline: 25119239]
19. Ganggayah MD, Taib NA, Har YC, Lio P, Dhillon SK. Predicting factors for survival of breast cancer patients using machine learning techniques. BMC Med Inform Decis Mak 2019 Mar 22;19(1):48 [FREE Full text] [doi: 10.1186/s12911-019-0801-4] [Medline: 30902088]
20. Pölsterl S, Amin Katouzian NN. Fast Training of Support Vector Machines for Survival Analysis. 2015 Presented at: ECML PKDD: Joint European Conference on Machine Learning and Knowledge Discovery in Databases; September 7-11, 2015; Porto, Portugal p. 243-259 URL: https://link.springer.com/chapter/10.1007/978-3-319-23525-7_15 [doi: 10.1007/978-3-319-23525-7_15]
21. Ishwaran H, Kogalur UB, Blackstone EH, Lauer MS. Random survival forests. Ann. Appl. Stat 2008 Sep 1;2(3):841-860. [doi: 10.1214/08-AOAS169]
22. Stekhoven DJ, Bühlmann P. MissForest--non-parametric missing value imputation for mixed-type data. Bioinformatics 2012 Jan 01;28(1):112-118. [doi: 10.1093/bioinformatics/btr597] [Medline: 22039212]
23. Holford TR. Life tables with concomitant information. Biometrics 1976 Sep;32(3):587-597. [Medline: 963172]
24. Van Belle V, Kristiaan P, Suykens J, Van Huffel S. Support vector machines for survival analysis. 2007 Jan 01 Presented at: Third International Conference on Computational Intelligence in Medicine and Healthcare (CIMED2007); July 1-7, 2007; Plymouth, England URL: https://limo.libis.be/primo-explore/fulldisplay?docid=LIRIAS1690710&context=L&vid=Lirias&search_scope=Lirias&tab=default_tab&lang=en_US&fromSitemap=1
25. ml-for-survival. GitHub. URL: https://github.com/xiaojialong0518/ml-for-survival [accessed 2021-12-12]
26. Stensrud MJ, Hernán MA. Why Test for Proportional Hazards? JAMA 2020 Apr 14;323(14):1401-1402. [doi: 10.1001/jama.2020.1267] [Medline: 32167523]
27. Shivaswamy PK, Chu W, Jansche M. A support vector approach to censored targets. 2007 Oct 28 Presented at: Seventh IEEE International Conference on Data Mining; October 28-31, 2007; Omaha, NE, USA. [doi: 10.1109/icdm.2007.93]
28. Van Belle V, Pelckmans K, Van Huffel S, Suykens JAK. Support vector methods for survival analysis: a comparison between ranking and regression approaches. Artif Intell Med 2011 Oct;53(2):107-118. [doi: 10.1016/j.artmed.2011.06.006] [Medline: 21821401]

29. Zou H, Hastie T. Regularization and variable selection via the elastic net. J Royal Statistical Soc B 2005 Apr;67(2):301-320. [doi: 10.1111/j.1467-9868.2005.00503.x]
30. Uno H, Cai T, Pencina MJ, D'Agostino RB, Wei LJ. On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. Stat Med 2011 May 10;30(10):1105-1117 [FREE Full text] [doi: 10.1002/sim.4154] [Medline: 21484848]
31. Heagerty PJ, Lumley T, Pepe MS. Time-dependent ROC curves for censored survival data and a diagnostic marker. Biometrics 2000 Jun;56(2):337-344. [doi: 10.1111/j.0006-341x.2000.00337.x] [Medline: 10877287]
32. Royston P, Sauerbrei W. A new measure of prognostic separation in survival data. Stat Med 2004 Mar 15;23(5):723-748. [doi: 10.1002/sim.1621] [Medline: 14981672]
33. Brier GW. Verification of forecasts expressed in terms of probability. Mon. Wea. Rev 1950 Jan;78(1):1-3. [doi: 10.1175/1520-0493(1950)078<0001:vofeit>2.0.co;2]
34. Online Breast Cancer Prognosis Tool. Fudan Breast Cancer: shinyapps.io. URL: https://fudanbreastcancer.shinyapps.io/DynNomapp [accessed 2021-12-15]
35. Rahman SA, Walker RC, Maynard N, Trudgill N, Crosby T, Cromwell DA, NOGCA project team AUGIS. The AUGIS Survival Predictor: Prediction of Long-term and Conditional Survival after Esophagectomy Using Random Survival Forests. Ann Surg 2021 Feb 17:online ahead of print. [doi: 10.1097/SLA.0000000000004794] [Medline: 33630434]

Abbreviations

AUC: area under the curve
C-index: concordance index
EN: elastic net
LASSO: least absolute shrinkage and selection operator
RSF: random survival forest
SVM: support vector machine

Edited by C Lovis; submitted 08.09.21; peer-reviewed by X Dong, K Rathi, JA Benítez-Andrades; comments to author 14.11.21; revised version received 15.12.21; accepted 02.01.22; published 18.02.22

Please cite as:
Xiao J, Mo M, Wang Z, Zhou C, Shen J, Yuan J, He Y, Zheng Y
The Application and Comparison of Machine Learning Models for the Prediction of Breast Cancer Prognosis: Retrospective Cohort Study
JMIR Med Inform 2022;10(2):e33440
URL: https://medinform.jmir.org/2022/2/e33440
doi: 10.2196/33440
PMID: