BMC Cancer

RESEARCH Open Access

Clinical prediction of pathological complete response in breast cancer: a machine learning study

Chongwu He[1†], Tenghua Yu[1†], Liu Yang[2†], Longbo He[1], Jin Zhu[1] and Jing Chen[3*]

Abstract

Background This study aimed to develop and validate machine learning models to predict pathological complete response (pCR) after neoadjuvant therapy in patients with breast cancer patients.

Methods Clinical and pathological data from 1143 patients were analyzed, encompassing variables such as age, gender, marital status, histologic grade, T stage, N stage, months from diagnosis to treatment, molecular subtype, and response to neoadjuvant therapy. Seven machine learning models were trained and validated using both internal and external datasets. Model performance was evaluated using multiple metrics, and interpretability analysis was conducted to assess feature importance.

Results Key variables influencing pCR included grade, N stage, months from diagnosis to treatment, and molecular subtype. The Naive Bayes model emerged as the most effective, with accuracy (0.746), sensitivity (0.699), specificity (0.808), and F1 score (0.759) surpassing other models. Both internal and external validation confirmed the model's robust predictive power. A web tool was developed for clinical use, aiding in personalized treatment planning. Interpretability analysis further elucidated the contribution of features to pCR prediction, enhancing clinical applicability.

Conclusion The Naive Bayes model provides a robust tool for personalized treatment decisions in patients with breast cancer undergoing neoadjuvant therapy. By accurately predicting pCR rates, it enables clinicians to tailor treatment strategies, potentially improving outcomes.

Keywords Breast cancer, Machine learning, Pathological complete response, PCR, Prediction

Introduction

In 2022, breast cancer accounted for 287,000 new cases in the United States, representing 31% of all new malignancies in women [1]. In early invasive breast cancer, treatment strategies such as neoadjuvant or adjuvant therapy may be employed, depending on the timing relative to surgery. Neoadjuvant therapy has been shown to be as effective as adjuvant therapy in reducing the risk of cancer recurrence and mortality. The primary objectives of neoadjuvant therapy include: 1) downstaging tumors to enhance the effect of surgical resection [2]; 2) eliminating or reducing metastatic cancer cells and micrometastatic foci, thus lowering recurrence rates [3]; 3) reducing tumor size to increase the likelihood of breast conserving

[†]Chongwu He, Tenghua Yu, and Liu Yang contributed equally to this work.

*Correspondence:
Jing Chen
chenjing1247@126.com
[1] Department of Breast Surgery, The Second Affiliated Hospital of Nanchang Medical College, Jiangxi Cancer Hospital, Nanchang, Jiangxi Province, China
[2] Department of Pathology, Nanchang People's Hospital, Nanchang, Jiangxi Province, China
[3] Department of Nursing, Nanchang Medical College, No. 689, Huiren Avenue, Xiaolan Economic Development Zone, Nanchang, Jiangxi Province, China

surgery [4];4) guiding adjuvant treatment decisions based on tumor's response to the initial therapy [5].

Neoadjuvant chemotherapy, a critical component of this approach, plays a pivotal role in treatment outcomes. Esserman et al. demonstrated a significant correlation between residual cancer burden (RCB) after neoadjuvant therapy and event-free survival (EFS), with lower RCB class predicting a reduced likelihood of metastatic recurrence [6]. Although there are variations in the criteria used to assess pathological response post-neoadjuvant therapy, numerous studies have confirmed a strong link between the degree of pathological response and patient prognosis [7, 8]. Pathological complete response (pCR) or non-pCR serves as a robust prognostic marker following neoadjuvant chemotherapy [6]. Consequently, identifying reliable methods to predict pCR is crucial for clinical practice.

Artificial intelligence (AI), a rapidly advancing field within computer science, has seen widespread applications across diverse sectors [9]. In healthcare, machine learning (ML), a subset of AI focused on automating complex tasks and analyzing large datasets, has garnered attention. ML models, trained on vast datasets, are adept at delivering precise classification and predictive outputs, effectively capturing complex relationships within the data [10]. ML encompasses supervised, unsupervised, and semi-supervised learning techniques, enabling the analysis of labeled, unlabeled, and partially labeled data, respectively, to optimize model efficiency and reduce costs. Supervised learning, which involves classification and regression tasks, is commonly used for assigning data points to predefined categories, such as'cancer'or'non-cancer,'with classification features typically being binary(true/false, 1/0) [11]. This study employed seven classifiers to train models aimed at predicting treatment outcomes.

Clinical and pathological data from 909 patients, including pre-neoadjuvant therapy clinical details and post-therapy pathological information, were analyzed. Seven distinct machine learning models were developed to predict the efficacy of neoadjuvant therapy based on this dataset. The models were then validated using an independent external dataset comprising 234 patients to evaluate their predictive accuracy. The goal of this study is to establish a reliable framework for identifying patients who are likely to respond favorably to neoadjuvant therapy in clinical settings.

Methods

Study population

A retrospective analysis was conducted on data from patients treated between September 2018 to April 2022 at Nanchang People's Hospital (NPH cohort) and Jiangxi Cancer Hospital (JCH cohort). Eligible patients were those who met the following criteria: (1) untreated unilateral primary early or locally advanced breast cancer, (2) pathologically confirmed diagnosis via core biopsy, (3) available immunohistochemical (IHC) data, including estrogen receptor (ER), progesterone receptor (PR), human epidermal growth factor receptor 2 (HER2) status, with a cutoff of 1% for ER and PR positivity. In cases of HER2 IHC score of 2+, fluorescence in situ hybridization (FISH) was performed to assess HER2 gene amplification, (4) complete clinical information, including age, gender, marital status, laterality, histology grade, T stage, N stage, M stage, months from diagnosis to treatment, molecular subtype and response to neoadjuvant therapy; and (5) completion of neoadjuvant therapy followed by surgery. Hormone receptor (HR) positive was defined as ER and/or PR positive. Exclusion criteria included (1) presence of other primary neoplastic disease and (2) unknown metastatic status. All patients underwent neoadjuvant therapy and completed treatment. The patients were subsequently categorized into four subgroups: HR +HER2 +, HR +HER2-, HR-HER2 +, and HR-HER2-. For multicenter data, the study received approved from the ethics review committees of Jiangxi Cancer Hospital (Grant Number: 2023ky170) and Nanchang People's Hospital (Grant Number: K-ky2022087) and was conducted in accordance with the principles outlined in the Helsinki Declaration.

Pathologic evaluation after neoadjuvant therapy

The response to neoadjuvant therapy was assessed using the Miller-Payne grading system [12], as follows: Grade 1: No change or minor alteration in individual malignant cells without a reduction in overall cellularity; Grade 2: Minor loss of tumor cells, with up to 30% reduction in overall cellularity; Grade 3: An estimated 30% to 90% reduction in tumor cells; Grade 4: Marked disappearance of tumor cells, with only small clusters or widely dispersed individual cells remaining, more than 90% loss of tumor cells; Grade 5: No malignant cells identifiable, with only vascular fibroelastotic stroma remaining, and ductal carcinoma in situ (DCIS) may be present. pCR was defined as Grade 5 and negative axillary lymph nodes negative, while non-pCR was defined as Grades 1–4 or positive axillary lymph nodes.

Predictor characteristics

Ten variables were selected for analysis: age, gender, marital status, laterality, histology grade, T stage, N stage, months from diagnosis to treatment, molecular subtype, and response to neoadjuvant therapy. Age and months from diagnosis to treatment were treated as continuous variables, while the other variables were categorized as categorical. Descriptive statistics were presented as mean and standard deviation (SD) for continuous variables and counts with percentages for categorical variables. Missing data were handled by imputing a separate category for predictor variables with missing values [13]. Differences between the pCR and non-pCR groups were assessed using the Fisher exact test for categorical variables and the Kolmogorov–Smirnov test for continuous variables. Statistical significance was defined $p < 0.05$. Relevant features for pCR prediction were selected using the least absolute shrinkage and selection operator (LASSO) [14].

Model building and validation

Seven machine learning models, including logistic regression [15], k-nearest neighbor (KNN) [16], support vector machine (SVM) [17], Naive Bayes(Bayes) [18], random forest(RF) [19], extreme gradient boosting (XGBoost) model [20], and neural network (Nnet) [21], were utilized for pCR prediction. Hyperparameter tuning was conducted using nested resampling, incorporating a k-fold cross-validation procedure for both model selection and hyperparameter optimization [22]. For each model, hyperparameter optimization was performed with 1000 iterations of random search with fivefold cross-validation. Model discrimination was evaluated using multiple metrics, including the area under the receiver operator characteristic curve (AUROC), calibration curves, decision curve analysis (DCA) curves, sensitivity, specificity, F1 scores, Youden's index, and accuracy. The fivefold cross-validation method was applied for both internal and external validation, with model performance averaged from five evaluations during external cross-validation. Data analysis was conducted using R software version 4.0.3 (R Foundation for Statistical Computing, Vienna, Austria) and were trained using the mlr3 package [23].

Result

Patient characteristics

A total of 1143 patients were included in this study, with 909 in the NPH cohort and 234 in the JCH cohort. Detailed demographic information and a comparison of nine potential features between the pCR and non-pCR groups are provided in Supplementary Table 1 and Supplementary Table 2. The pCR rate was 42.9% in the NPH cohort and 42.3% in the JCH cohort. The study variables, including age, sex, marital status, laterality, histology, grade, T stage, N stage, months from diagnosis to treatment, molecular subtype and response to neoadjuvant treatment, were thoroughly reviewed, with only a few missing values. The NPH cohort ($n = 909$) was used for both training and internal validation of the predictive models, while the JCH cohort served as the testing set for external validation. The flowchart illustrating the machine learning process in this study is shown in Fig. 1.

Key variables

Lasso regression selects the optimal regularization parameter (λ, lambda) through cross-validation to balance the model fit and sparsity. This method forces some coefficients to zero, effectively filtering important variables based on the magnitude of the coefficients. Four features with nonzero coefficients were ultimately filtered out (Fig. 2A, B). Four variables, including grade, N stage, months from diagnosis to treatment and molecular subtype, were most strongly to be associated with pCR. Subsequently, the model was constructed using these four variables.

Performance of machine learning models

After identifying these four variables, machine learning models were employed to predict pCR following neoadjuvant therapy. Seven learners from the mlr3 package, including Logistic, KNN, Bayes, SVM, RF, XGBoost,

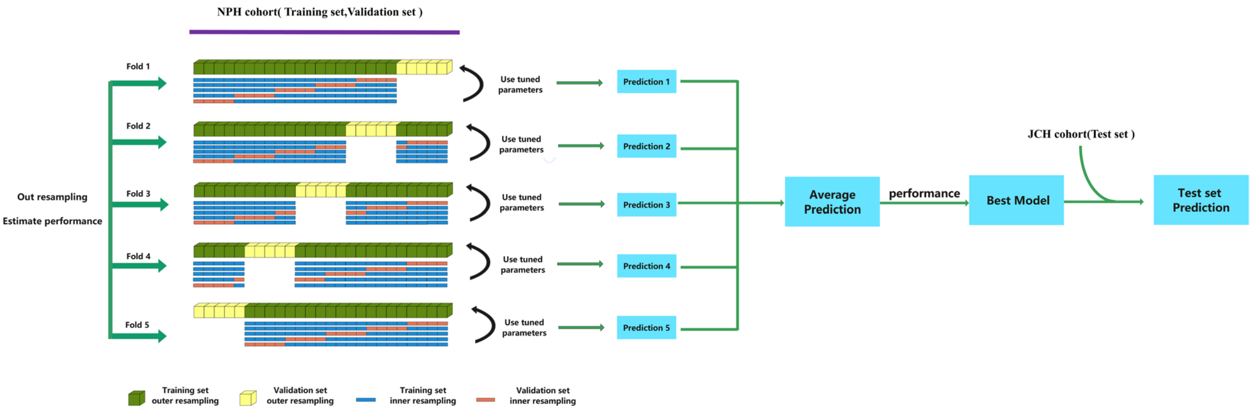


Fig. 1 Flowchart of machine learning in this study

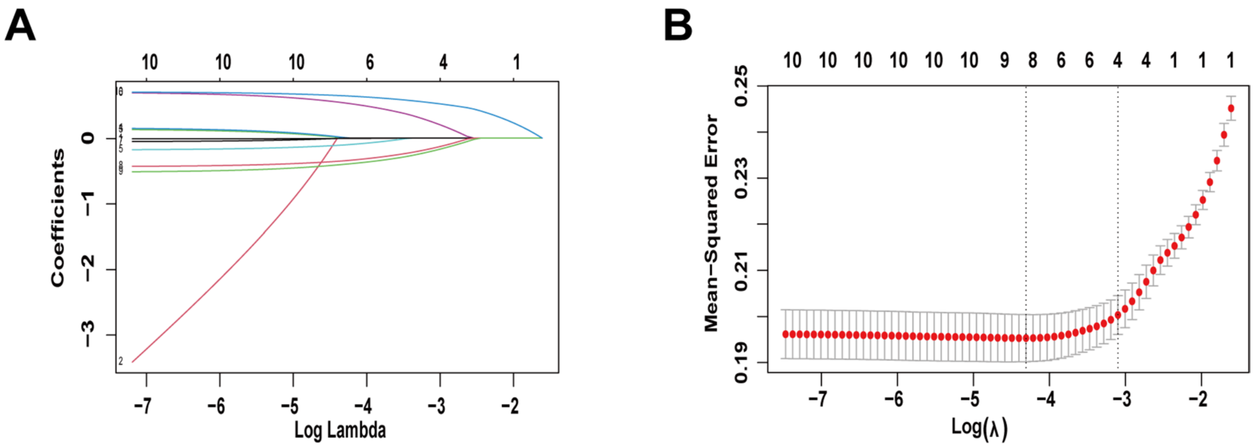


Fig. 2 Screening for key variables using LASSO regression. **A** LASSO coefficient profiles for the eight variables. **B** Four risk factors selected using LASSO Cox regression analysis. The dotted vertical lines correspond to the optimal scores by minimum criteria and1-standard error (1-s.e.) criteria. At the minimum criteria, the variables include age, marital status, laterality, histology grade, T stage, N stage, months from diagnosis to treatment and molecular subtype. At 1-s.e. criteria, the selected variables are grade, N stage and months from diagnosis to treatment and molecular subtype

Table 1 Performance metrics for seven models in validation dataset

	Sensitivity	Specificity	F1 score	Youden index	Accuracy
Logistic	0.887	0.601	0.734	0.488	0.724
KNN	0.746	0.703	0.697	0.449	0.722
Bayes	0.699	0.808	0.759	0.507	0.746
SVM	0.557	0.933	0.693	0.490	0.718
RF	0.651	0.849	0.738	0.500	0.736
XGBoost	0.628	0.872	0.728	0.500	0.733
Nnet	0.640	0.846	0.729	0.486	0.728

Nnet, were selected for evaluation. Among these, the Naive Bayes model demonstrated relatively high AUC, F1 score, and Youden index, coupled with low Brier scores (Table 1, Fig. 3). Calibration plots for the seven models are shown in Fig. 3. DCA further highlighted that the Naive Bayes model was more accurate in predicting pCR outcomes. Based on these results, the Naive Bayes model outperformed the other six models in predicting pCR. Performance metrics for each model in the validation set are presented in Table 1.

The SHAP package was used to analyze the Naive Bayes model. Shapley Additive explanation (SHAP) generated Shap values for each sample in the model. Visual analysis revealed that molecular subtype was a highly influential feature, generally correlating positively with response. Patients with the HR-, HER2 +or HR-, HER2- molecular subtypes had a higher likelihood of achieving pCR after neoadjuvant therapy. The mean absolute SHAP value for each feature reflects its importance, with larger values indicating greater importance. The molecular subtype feature exhibited the highest mean SHAP value, indicating its primary influence on the model's predictions. Other important features, ranked by descending SHAP mean values, included months from diagnosis to treatment, grade, and N stage (Fig. 4B).

Performance of testing set

The testing dataset, consisting of 234 patients from the JCH cohort, was used to validate the performance of the Naive Bayes model and the other six models. The following performance metrics were observed for the Naive Bayes model in the testing group: AUC =75.1%, 95% CI: 68.8%–81.4%, brier score =0.211, Sensitivity =0.768, Specificity =0.644, F1 score =0.682, Youden index =0.412, accuracy =0.697(Fig. 5 and Table 2). Based on these evaluation results, the Naive Bayes model demonstrated the best generalization ability on the JCH cohort.

Application of the model

To further enable the use of this prediction tool in clinical settings, a web tool was created based on the Naive Bayes model from the NPH cohort's model. This tool, developed, using the Shiny package, is accessible at https://pcrpredictior.shinyapps.io/pCRpredictior/ (Fig. 6).

Discussion

Machine learning techniques, particularly those based on large-scale data analysis, have significantly transformed how tumor diagnosis and prognosis are predicted. As a key branch of AI, machine learning enables more sophisticated data interpretation and analysis, often surpassing traditional statistical methods in terms of predictive accuracy and flexibility [24]. This study explored the

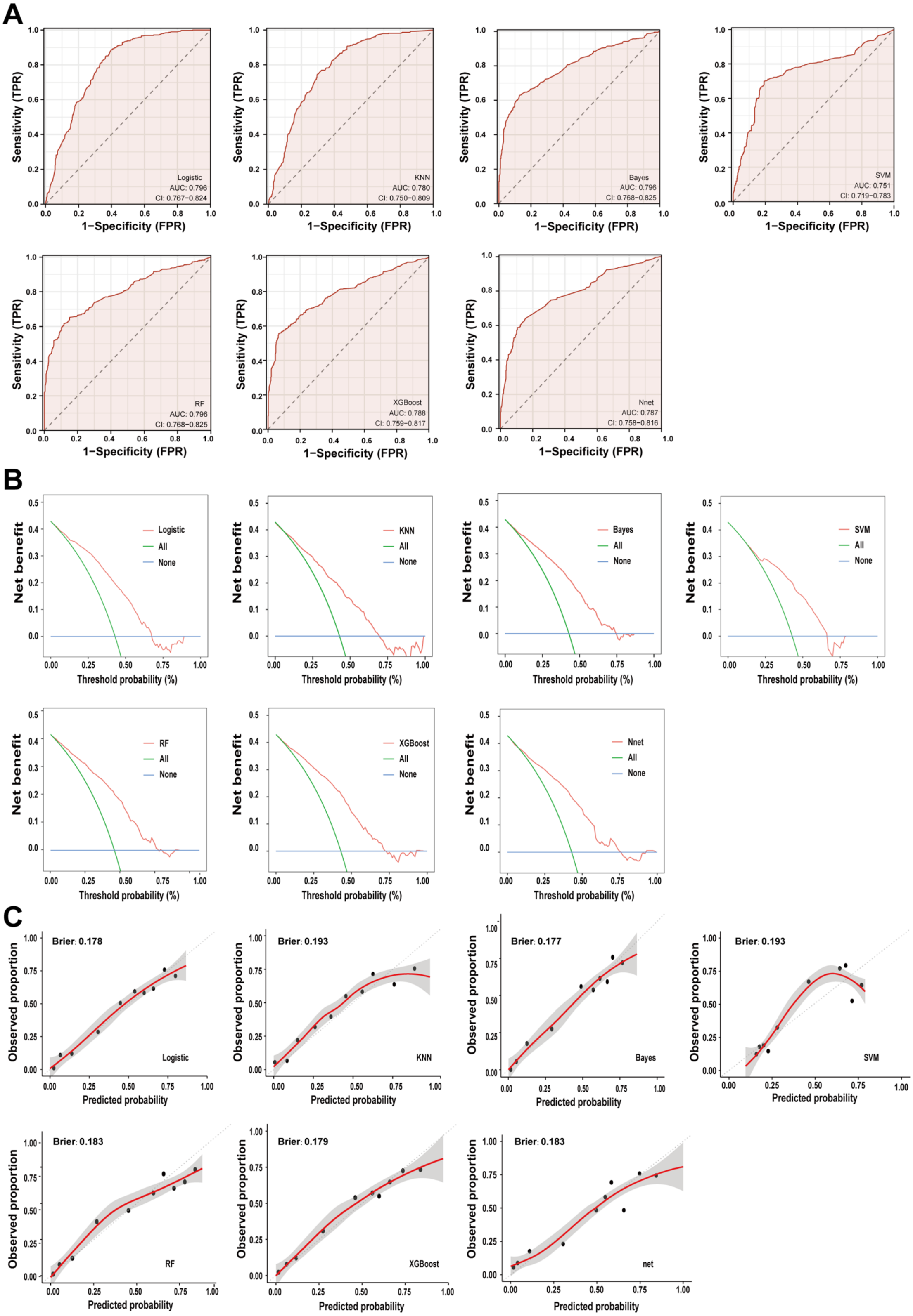


Fig. 3 Performance of seven machine learning algorithms in NPH cohort. **A** ROC curves, (**B**) DCA (Decision Curve Analysis) curves, (**C**) calibration curves

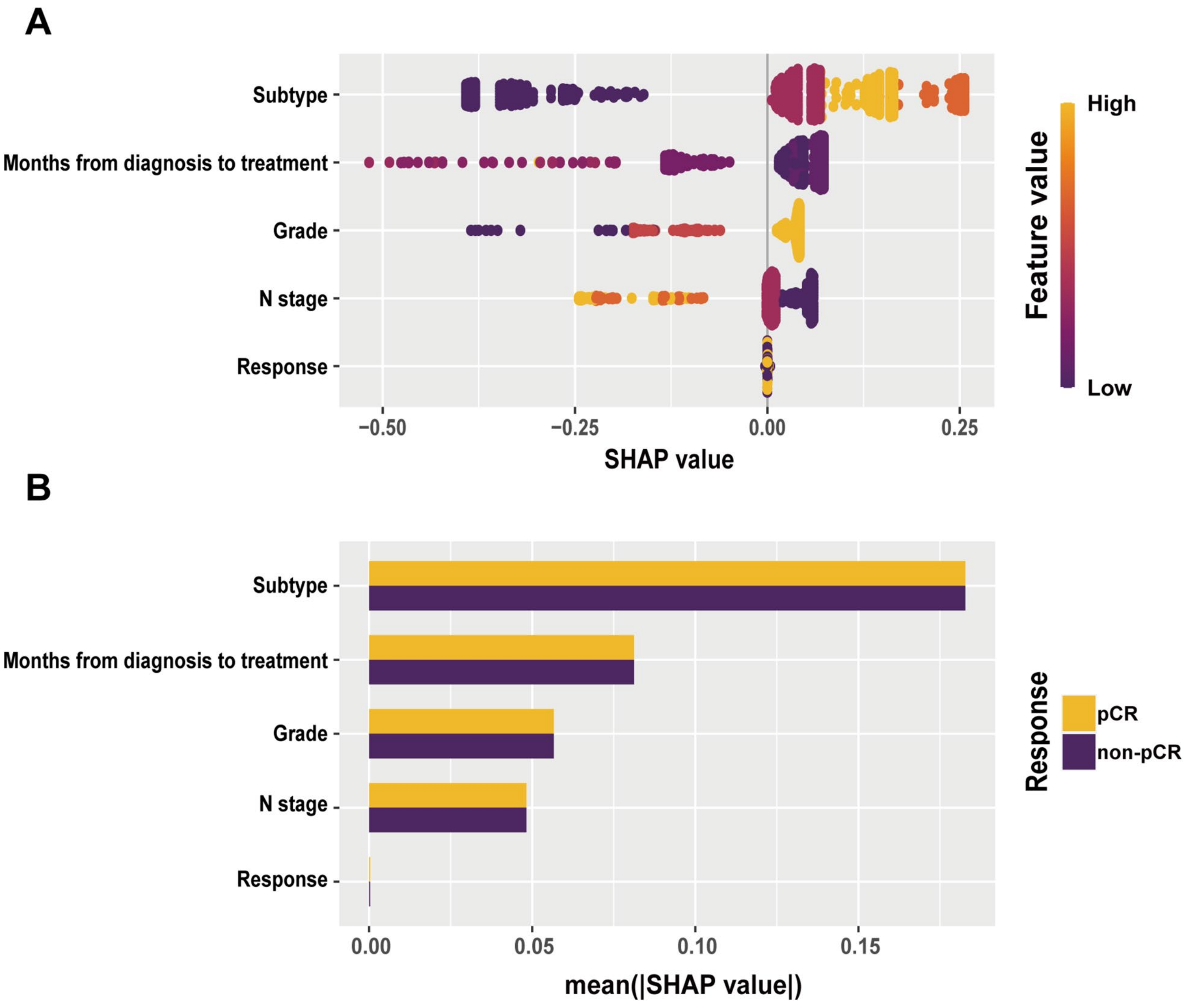


Fig. 4 SHAP analysis of the Naive Baye model. **A** SHAP values for each variable in each sample; (**B**) ranking of variable importance

prediction of pCR after neoadjuvant therapy in patients with breast cancer using seven distinct machine learning algorithms, based on a comprehensive set of clinicopathologic data. Our findings demonstrate that the Naive Bayes model, in particular, shows great promise in predicting pCR. In the validation cohort, the model achieved an accuracy of 74.6%, sensitivity of 69.9%, specificity of 80.8%, F1 score of 75.9%, and a Youden index of 0.507. In the test cohort, the model's performance remained strong, with an accuracy of 69.7%, sensitivity of 76.8%, specificity of 64.4%, F1 score of 68.2%, and Youden index of 41.2%. These metrics highlight the model's robust predictive capability and generalizability across different datasets.

In 2020, the Food and Drug Administration (FDA) approved pCR as a surrogate endpoint for research on neoadjuvant therapy in high-risk breast cancer. A meta-analysis conducted by Laura J. Esserman's research team, based on data from the I-SPY2 clinical trial, revealed a significant correlation between RCB and EFS in patients with breast cancer undergoing neoadjuvant therapy. The study found that a lower the RCB following neoadjuvant therapy was associated with a reduced probability of recurrence and metastasis [6]. The pathological status after neoadjuvant therapy significantly affects the prognosis of patients with breast cancer.

Deep learning has been applied to predict the outcome of neoadjuvant therapy using pre- and post-neoadjuvant therapy magnetic resonance imaging(MRI) [25, 26], color Doppler ultrasound [27] or pathological images [28]. In this study, patient outcomes were predicted by analyzing MRI changes after two cycles of neoadjuvant therapy in the early stage [29]. Fanizzi utilized a study, random forest algorithm to develop a

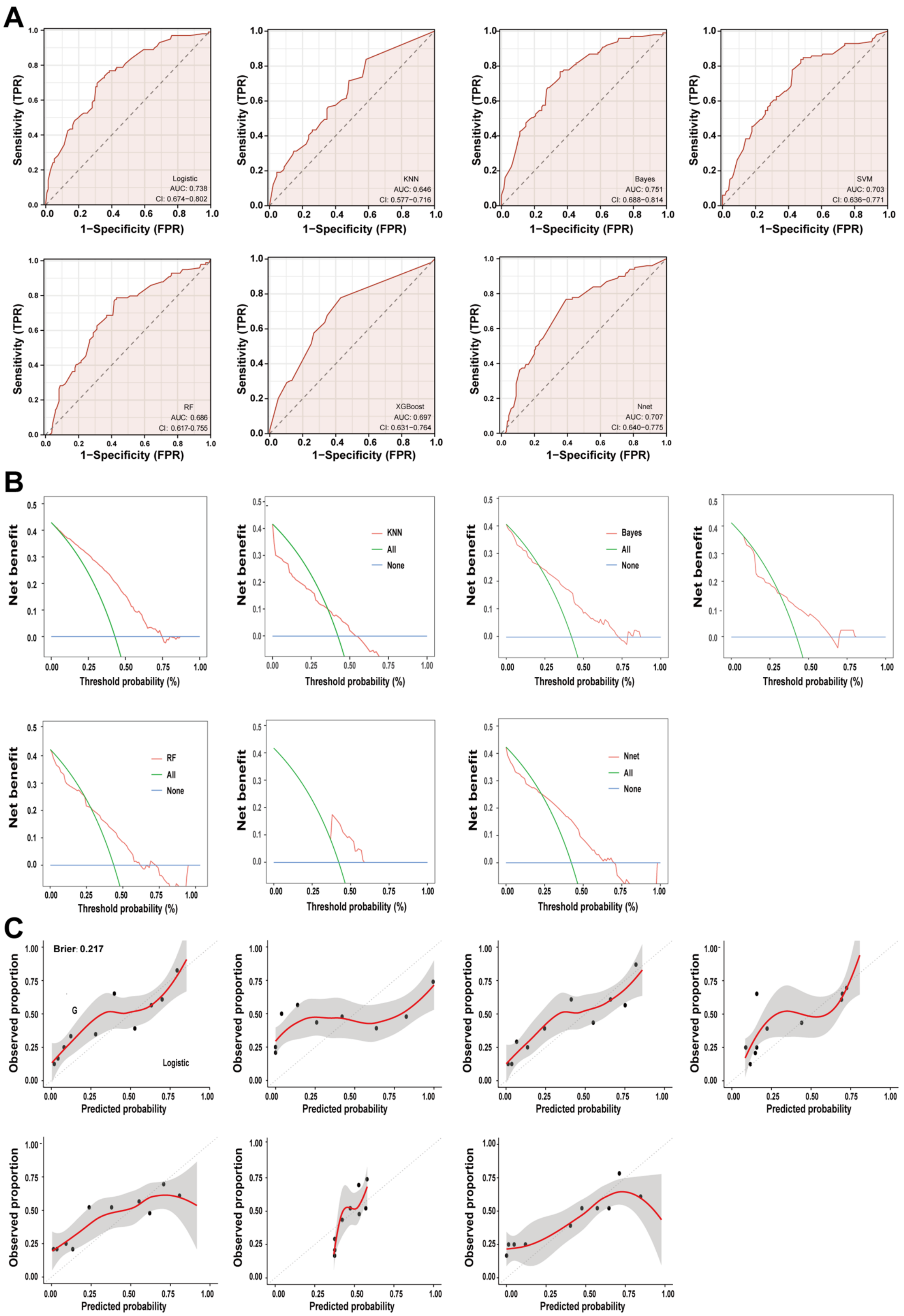


Fig. 5 Performance of seven machine learning algorithms in JCH cohort. **A** ROC curves, (**B**) DCA (Decision Curve Analysis) curves, (**C**) calibration curves

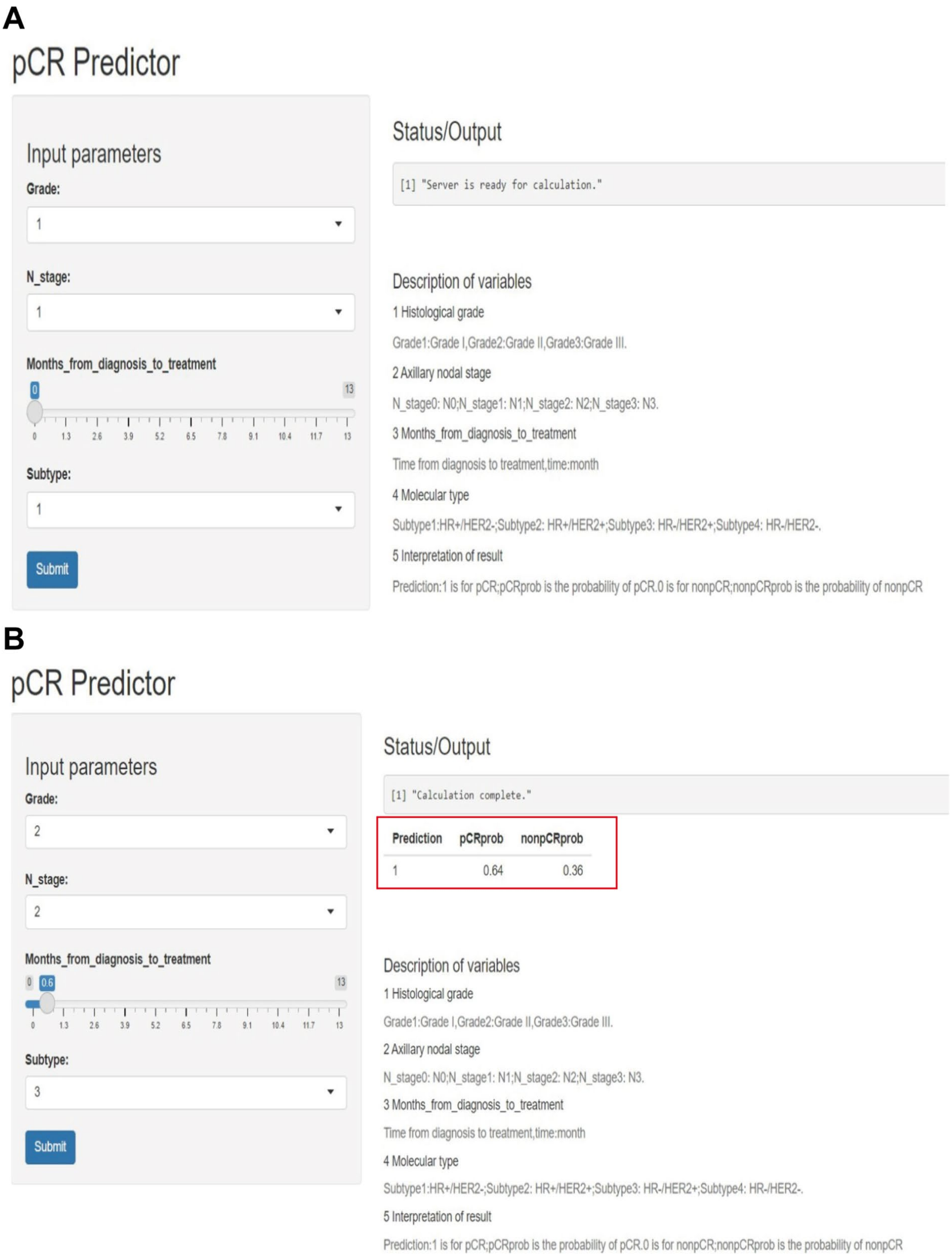
Table 2 Performance metrics for seven models in test dataset

	Sensitivity	Specificity	F1 score	Youden index	Accuracy
Logistic	0.697	0.689	0.657	0.386	0.692
KNN	0.838	0.415	0.636	0.253	0.594
Bayes	0.768	0.644	0.682	0.412	0.697
SVM	0.838	0.526	0.675	0.364	0.658
RF	0.788	0.570	0.664	0.358	0.662
XGBoost	0.778	0.570	0.658	0.348	0.658
Nnet	0.768	0.607	0.667	0.375	0.675

model for predicting neoadjuvant therapy outcome in patients with HER2-positive breast cancer based on clinical and pathological data. In contrast, this study encompasses all subtypes of breast cancer patients. Moreover, seven different learners were employed to identify the optimal model for distinguishing between pCR and non-pCR patients. The sample size in this study exceeds that of Fanizzi's study, which only included training and validation sets without an external test set. Aswolinskiy et al. developed interpretable biomarkers for predicting pCR to neoadjuvant chemotherapy using deep learning, based solely on digital pathology H&E images of pre-treatment breast biopsies. These biomarkers exhibited AUROC values ranging from 0.66 to 0.88 across various cohorts [28]. Chen et al. applied machine learning algorithms and a model stacking approach to develop two models: the immunological gene-based Ipredictor model and the immunological gene and receptor status-based ICpredictor model, utilizing RNA-seq data. The AUROCs for these models in an independent external test set were 0.716 and 0.752, respectively, based on a microarray platform [30]. Predicting pCR based on imaging requires patients to undergo treatment first, resulting in a delayed prediction process. Moreover, changes in neoadjuvant chemotherapy regimens may influence prediction accuracy. In contrast, while high-throughput sequencing can predict therapy outcomes based on tumor gene expression, its implementation may increase treatment costs and patient burden in clinical practice. The predictions from the models discussed above are consistent with those generated by the Naive Bayes algorithm in this study. Additionally, the complexity of variable collection and model training methods in those studies presents greater challenges for clinical application compared to this study. This study primarily focuses on patients'clinical and pathological data, with only four variables selected through LASSO regression. The objective is to enable the prediction of neoadjuvant therapy outcomes at the time of diagnosis, thereby ensuring practical application in clinical practice. Undoubtedly, future multi-omics or multi-modal analyses offer promising tools for predicting pCR. For instance, combining genomic or transcriptomic data with clinical features could yield a more comprehensive framework, improving prediction accuracy and patient stratification. Additionally, incorporating imaging data alongside clinical and molecular features could further enhance prediction models, supporting personalized treatment strategies.

For model interpretability, a comprehensive analysis was conducted to assess the impact of various clinical and pathological features on pCR rates. This analysis enhances clinicians'understanding of the model's decision-making process, fostering greater confidence in the predicted outcomes. Machine learning's ability to process large volumes of multidimensional data facilitates the identification of novel associations between clinicopathologic features and pCR, highlighting its potential in medical oncology. Several studies have demonstrated the advantages of machine learning algorithms in predictive tasks [31–33]. In this study, four key variables influencing pCR were identified using LASSO regression, based on patient clinicopathologic characteristics, including age, gender, marital, laterality, histology grade, T stage, N stage, M stage, months from diagnosis to treatment, molecular subtype and response to neoadjuvant therapy. SHAP analysis was applied to interpret the model output, providing insight into the contribution of each feature to the model predictions [34]. The SHAP results for the Naive Bayes model revealed the relative importance of the four variables in predicting pCR. Molecular subtype emerged as the most influential feature, consistent with previous studies showing that TNBC and HER2-positive breast cancers are more responsive to neoadjuvant chemotherapy and targeted therapies [35, 36], whereas luminal breast cancers tend to exhibit a higher RCB after neoadjuvant therapy [6]. The time from diagnosis to treatment refers to the interval between diagnosis and the initiation of neoadjuvant therapy, which may include chemotherapy or targeted molecular therapy in the study. Treatment delays are common in cancer care, often resulting from factors such as treatment transfers, histological biopsies, treatment planning, and other external influences [37]. Several studies have indicated that treatment delays can increase all-cause mortality in cancer individuals [38–40]. Specifically, a breast cancer study identified the time from diagnosis to surgery as an independent risk factor for overall survival (OS) [41]. In this study, the"months from diagnosis to treatment"variable was ranked second in importance for pCR, after molecular subtype (Fig. 4).

While numerous biomarkers have been proposed to predict the response to neoadjuvant therapy [42–44],



A

pCR Predictor

Input parameters

Grade: 1

N_stage: 1

Months_from_diagnosis_to_treatment: 0

Subtype: 1

Submit

Status/Output

```
[1] "Server is ready for calculation."
```

Description of variables

1 Histological grade

Grade1:Grade I,Grade2:Grade II,Grade3:Grade III.

2 Axillary nodal stage

N_stage0: N0;N_stage1: N1;N_stage2: N2;N_stage3: N3.

3 Months_from_diagnosis_to_treatment

Time from diagnosis to treatment,time:month

4 Molecular type

Subtype1:HR+/HER2-;Subtype2: HR+/HER2+;Subtype3: HR-/HER2+;Subtype4: HR-/HER2-.

5 Interpretation of result

Prediction:1 is for pCR;pCRprob is the probability of pCR.0 is for nonpCR;nonpCRprob is the probability of nonpCR

B

pCR Predictor

Input parameters

Grade: 2

N_stage: 2

Months_from_diagnosis_to_treatment: 0.6

Subtype: 3

Submit

Status/Output

```
[1] "Calculation complete."
```

Prediction	pCRprob	nonpCRprob
1	0.64	0.36

Description of variables

1 Histological grade

Grade1:Grade I,Grade2:Grade II,Grade3:Grade III.

2 Axillary nodal stage

N_stage0: N0;N_stage1: N1;N_stage2: N2;N_stage3: N3.

3 Months_from_diagnosis_to_treatment

Time from diagnosis to treatment,time:month

4 Molecular type

Subtype1:HR+/HER2-;Subtype2: HR+/HER2+;Subtype3: HR-/HER2+;Subtype4: HR-/HER2-.

5 Interpretation of result

Prediction:1 is for pCR;pCRprob is the probability of pCR.0 is for nonpCR;nonpCRprob is the probability of nonpCR

Fig. 6 A web tool based on the Naive Baye model constructed in the NPH cohort. **A** Default tool interface; (**B**) Probability of pCR calculated using grading:2, N staging:2, months from diagnosis to treatment:0.6, subtype:3

their real-world applicability remains limited, with suboptimal predictive efficacy for individual markers. The Naive Bayes model developed in this study offers several advantages. By combination multiple clinicopathological factors, it improves predictive accuracy. Additionally, the four variables are readily obtainable, enhancing the model's practicality. Patients can be made at the time of diagnosis, eliminating the need for post-treatment evaluation and enabling early patient screening.

This study still has several limitations. First, small sample size and retrospective nature may have impacted the model's generalization. Moreover, the study only incorporated clinicopathologic data, excluding other factors that could influence pCR, such as the specific therapeutic agents used. Medications are critical in determining a patient's likelihood of achieving pCR with neoadjuvant therapy. Previous studies have shown that patients treated with a combination of anthracyclines and paclitaxel had higher pCR rates and better survival outcomes than those treated with anthracyclines alone [45]. Furthermore, the advent of targeted therapies has significantly improved pCR rates in patients with HER2-positive breast cancer [46]. However, the study did not account for differences in chemotherapeutic and targeted drug regimens used due to the long period between patient inclusion and the variations in treatment protocols. Future studies should aim to expand the sample size and incorporate additional factors, including therapeutic agents. Second, the data for this study were derived from a single cohort of patients across two hospitals, which may limit the model's generalizability to other populations. The lack of diversity in ethnic groups and geographical regions in the sample could impact the model's applicability to a broader patient population. Future studies should aim to include a more diverse cohort to enhance the external validity of the model. Third, LASSO was used to identify predictors of pCR. Although LASSO is a widely applied and effective technique for variable selection, it may fail to capture important interactions or non-linear relationships between features that could be vital for pCR prediction. Furthermore, LASSO assumes the inclusion of all relevant features in the initial set. However, other potentially significant features, such as novel biomarkers or genetic data, were not considered in the analysis. The exclusion of these variables may limit the model's predictive power and its ability to fully reflect the complexity of pCR. Future studies could explore alternative feature selection techniques or incorporate more comprehensive feature sets to improve model accuracy and clinical applicability.

Conclusions

In this study, seven pCR prediction models were developed and evaluated using various metrics such as AUC, Brier score, and DCA, to identify the most effective model. The Naive Bayes algorithm, which demonstrated the best performance and practical applicability, was selected. The model enables personalized prediction of pCR following neoadjuvant therapy in patients with breast cancer, aiding in the selection of appropriate treatment regimens. Ultimately, it may serve as a stratification tool for clinical trials and, if validated in prospective studies, could become a useful approach for identifying patients with the potential for pCR.

Abbreviations

pCR	Pathological complete response
RCB	Residual cancer burden
EFS	Event-free survival
AI	Artificial intelligence
ML	Machine learning
IHC	Immunohistochemical
ER	Estrogen receptor
PR	Progesterone receptor
HER2	Human epidermal growth factor receptor 2
DCIS	Ductal carcinoma in situ
SD	Standard deviation
LASSO	Least absolute shrinkage and selection operator
KNN	K-nearest neighbor
SVM	Support vector machine
Bayes	Naive Bayes
RF	Random forest
XGBoost	Extreme gradient boosting
Nnet	Neural network
AUROC	Area under the receiver operator characteristic curve
DCA	Decision curve analysis

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12885-025-14335-1.

Supplementary Material 1.

Supplementary Material 2.

Acknowledgements

Not applicable.

Authors' contributions

JC and CH contributed to study conceptualization and project administration. CH, LY and JC contributed to data curation and formal analysis. LH, JZ and CH contributed validation and visualization; CH, JC, and TY contributed to writing-original draft and manuscript revision. LH and JC contributed to investigation and methodology. All authors read and approved the final manuscript.

Funding

This work was supported by Science and Technology Research Project of Jiangxi Provincial Education Department (Grant Number: GJJ2208202). Science and Technology Program Project of Health Commission of Jiangxi Province (Grant Number: 202510069). General Project of Jiangxi Provincial Administration of Traditional Chinese Medicine(Grant Number: 2024B0854). Jiangxi Cancer Hospital Doctoral Start-up Fund (Grant Number: BSQDJ202309). Jiangxi Province Gan Po Talent Support Program (Grant Number: 20232BCJ23035). National Natural Science Foundation of China (Grant Number: 82160565).

Data availability

The original contributions presented in the study are included in the article. Further inquiries can be directed to the corresponding author.

Declarations

Ethics approval and consent to participate

The study was performed in accordance with the principles of the Declaration of Helsinki. The study was approved by the Ethics Committee of the Jiangxi Cancer Hospital (Grant Number: 2023ky170) and the Ethics Committee of the Nanchang People's Hospital (Grant Number: K-ky2022087) and all participants provided written in-formed consent to participate and for publication.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Received: 12 November 2024 Accepted: 14 May 2025
Published online: 23 May 2025

References

1. Siegel RL, Miller KD, Fuchs HE, Jemal A. Cancer statistics, 2022. CA Cancer J Clin. 2022;72(1):7–33.
2. Zaborowski AM, Wong SM. Neoadjuvant systemic therapy for breast cancer. Br J Surg. 2023;110(7):765–72.
3. Roth MT, Eng C. Neoadjuvant chemotherapy for colon cancer. Cancers. 2020;12(9):2368.
4. Golshan M, Cirrincione CT, Sikov WM, Carey LA, Berry DA, Overmoyer B, Henry NL, Somlo G, Port E, Burstein HJ, et al. Impact of neoadjuvant therapy on eligibility for and frequency of breast conservation in stage II-III HER2-positive breast cancer: surgical results of CALGB 40601 (Alliance). Breast Cancer Res Treat. 2016;160(2):297–304.
5. Cortina CS, Lloren JI, Rogers C, Johnson MK, Cobb AN, Huang CC, Kong AL, Singh P, Teshome M. Does neoadjuvant chemotherapy in clinical T1–T2 N0 triple-negative breast cancer increase the extent of axillary surgery? Ann Surg Oncol. 2024;31(5):3128–40.
6. Symmans WF, Yau C, Chen YY, Balassanian R, Klein ME, Pusztai L, Nanda R, Parker BA, Datnow B, Krings G, et al. Assessment of residual cancer burden and event-free survival in neoadjuvant treatment for high-risk breast cancer: an analysis of data from the I-SPY2 randomized clinical trial. JAMA Oncol. 2021;7(11):1654–63.
7. Cortazar P, Zhang L, Untch M, Mehta K, Costantino JP, Wolmark N, Bonnefoi H, Cameron D, Gianni L, Valagussa P, et al. Pathological complete response and long-term clinical benefit in breast cancer: the CTNeoBC pooled analysis. Lancet. 2014;384(9938):164–72.
8. Davey MG, Ryan ÉJ, Davey MS, Lowery AJ, Miller N, Kerin MJ. Clinicopathological and prognostic significance of programmed cell death ligand 1 expression in patients diagnosed with breast cancer: meta-analysis. Br J Surg. 2021;108(6):622–31.
9. Bhat M, Rabindranath M, Chara BS, Simonetto DA. Artificial intelligence, machine learning, and deep learning in liver transplantation. J Hepatol. 2023;78(6):1216–33.
10. Rattan P, Penrice DD, Simonetto DA. Artificial intelligence and machine learning: what you always wanted to know but were afraid to ask. Gastro Hep Adv. 2022;1(1):70–8.
11. Greener JG, Kandathil SM, Moffat L, Jones DT. A guide to machine learning for biologists. Nat Rev Mol Cell Biol. 2022;23(1):40–55.
12. Ogston KN, Miller ID, Payne S, Hutcheon AW, Sarkar TK, Smith I, Schofield A, Heys SD. A new histological grading system to assess response of breast cancers to primary chemotherapy: prognostic significance and survival. Breast. 2003;12(5):320–7.
13. Ding Y, Simonoff JS. An investigation of missing data methods for classification trees applied to binary response data. J Machine Learn Res. 2010;11(6):131–70.
14. Mullah MAS, Hanley JA, Benedetti A. LASSO type penalized spline regression for binary data. BMC Med Res Methodol. 2021;21(1):83.
15. Zhou Z, Huang H, Liang Y. Cancer classification and biomarker selection via a penalized logsum network-based logistic regression model. Technol Health Care. 2021;29(S1):287–95.
16. Tamaru A, Hara J, Higashi H, Tanaka Y, Ortega A. Optimizing k in kNN Graphs with Graph Learning Perspective. In: ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP): 14–19 April 2024. 2024;2024:9441–5.
17. Amarappa S, Sathyanarayana S. Data classification using Support vector Machine (SVM), a simplified approach. Int J Electron Comput Sci Eng. 2014;3:435–45.
18. Webb GI, Keogh E, Miikkulainen R. Naïve Bayes. Encyclopedia Mach Learn. 2010;15(1):713–4.
19. Ziegler A, König IR. Mining data with random forests: current options for real-world applications. Wiley Interdisc Rev Data Mining Knowl Discov. 2014;4(1):55–63.
20. Perez BC, Bink M, Svenson KL, Churchill GA, Calus MPL. Prediction performance of linear models and gradient boosting machine on complex phenotypes in outbred mice. G3 (Bethesda, Md). 2022;12(4):jkac039.
21. Günther F, Fritsch S. Neuralnet: training of neural networks. R J. 2010;2(1):30.
22. Cawley GC, Talbot NL. On over-fitting in model selection and subsequent selection bias in performance evaluation. J Mach Learn Res. 2010;11:2079–107.
23. Lang M, Binder M, Richter J, Schratz P, Pfisterer F, Coors S, Au Q, Casalicchio G, Kotthoff L, Bischl B. mlr3: A modern object-oriented machine learning framework in R. J Open Source Softw. 2019;4(44):1903.
24. Visco V, Ferruzzi GJ, Nicastro F, Virtuoso N, Carrizzo A, Galasso G, Vecchione C, Ciccarelli M. Artificial intelligence as a business partner in cardiovascular precision medicine: an emerging approach for disease detection and treatment optimization. Curr Med Chem. 2021;28(32):6569–90.
25. Sutton EJ, Onishi N, Fehr DA, Dashevsky BZ, Sadinski M, Pinker K, Martinez DF, Brogi E, Braunstein L, Razavi P, et al. A machine learning model that classifies breast cancer pathologic complete response on MRI post-neoadjuvant chemotherapy. Breast Cancer Res. 2020;22(1):57.
26. Zhou Z, Adrada BE, Candelaria RP, Elshafeey NA, Boge M, Mohamed RM, Pashapoor S, Sun J, Xu Z, Panthi B, et al. Prediction of pathologic complete response to neoadjuvant systemic therapy in triple negative breast cancer using deep learning on multiparametric MRI. Sci Rep. 2023;13(1):1171.
27. Gu J, Tong T, Xu D, Cheng F, Fang C, He C, Wang J, Wang B, Yang X, Wang K, et al. Deep learning radiomics of ultrasonography for comprehensively predicting tumor and axillary lymph node status after neoadjuvant chemotherapy in breast cancer patients: a multicenter study. Cancer. 2023;129(3):356–66.
28. Aswolinskiy W, Munari E, Horlings HM, Mulder L, Bogina G, Sanders J, Liu YH, van den Belt-Dusebout AW, Tessier L, Balkenhol M, et al. PROACTING: predicting pathological complete response to neoadjuvant chemotherapy in breast cancer from routine diagnostic histopathology biopsies with deep learning. Breast Cancer Res. 2023;25(1):142.
29. Zeng Q, Ke M, Zhong L, Zhou Y, Zhu X, He C, Liu L. Radiomics based on dynamic contrast-enhanced MRI to early predict pathologic complete response in breast cancer patients treated with neoadjuvant therapy. Acad Radiol. 2023;30(8):1638–47.
30. Chen J, Hao L, Qian X, Lin L, Pan Y, Han X. Machine learning models based on immunological genes to predict the response to neoadjuvant therapy in breast cancer patients. Front Immunol. 2022;13:948601.
31. Moon I, LoPiccolo J, Baca SC, Sholl LM, Kehl KL, Hassett MJ, Liu D, Schrag D, Gusev A. Machine learning for genetics-based classification and treatment response prediction in cancer of unknown primary. Nat Med. 2023;29(8):2057–67.
32. Swanson K, Wu E, Zhang A, Alizadeh AA, Zou J. From patterns to patients: Advances in clinical machine learning for cancer diagnosis, prognosis, and treatment. Cell. 2023;186(8):1772–91.
33. Zuo D, Yang L, Jin Y, Qi H, Liu Y, Ren L. Machine learning-based models for the prediction of breast cancer recurrence risk. BMC Med Inform Decis Mak. 2023;23(1):276.
34. Smith M, Alvarez F. Identifying mortality factors from Machine Learning using Shapley values - a case of COVID19. Expert Syst Appl. 2021;176:114832.

35. Houvenaeghel G, Lambaudie E, Classe JM, Mazouni C, Giard S, Cohen M, Faure C, Charitansky H, Rouzier R, Daraï E, et al. Lymph node positivity in different early breast carcinoma phenotypes: a predictive model. BMC Cancer. 2019;19(1):45.
36. Noske A, Anders SI, Ettl J, Hapfelmeier A, Steiger K, Specht K, Weichert W, Kiechle M, Klein E. Risk stratification in luminal-type breast cancer: comparison of Ki-67 with EndoPredict test results. Breast. 2020;49:101–7.
37. Hanna TP, King WD, Thibodeau S, Jalink M, Paulin GA, Harvey-Jones E, O'Sullivan DE, Booth CM, Sullivan R, Aggarwal A. Mortality due to cancer treatment delay: systematic review and meta-analysis. BMJ (Clinical research ed). 2020;371:m4087.
38. Cone EB, Marchese M, Paciotti M, Nguyen DD, Nabi J, Cole AP, Molina G, Molina RL, Minami CA, Mucci LA, et al. Assessment of time-to-treatment initiation and survival in a cohort of patients with common cancers. JAMA Netw Open. 2020;3(12):e2030072.
39. Nguyen DD, Haeuser L, Paciotti M, Reitblat C, Cellini J, Lipsitz SR, Kibel AS, Choudhury AD, Cone EB, Trinh QD. Systematic review of time to definitive treatment for intermediate risk and high risk prostate cancer: are delays associated with worse outcomes? J Urol. 2021;205(5):1263–74.
40. Yabroff KR, Wu XC, Negoita S, Stevens J, Coyle L, Zhao J, Mumphrey BJ, Jemal A, Ward KC. Association of the COVID-19 pandemic with patterns of statewide cancer services. J Natl Cancer Inst. 2022;114(6):907–9.
41. Polverini AC, Nelson RA, Marcinkowski E, Jones VC, Lai L, Mortimer JE, Taylor L, Vito C, Yim J, Kruper L. Time to treatment: measuring quality breast cancer care. Ann Surg Oncol. 2016;23(10):3392–402.
42. Bownes RJ, Turnbull AK, Martinez-Perez C, Cameron DA, Sims AH, Oikonomidou O. On-treatment biomarkers can improve prediction of response to neoadjuvant chemotherapy in breast cancer. Breast Cancer Res. 2019;21(1):73.
43. Magbanua MJM, Swigart LB, Wu HT, Hirst GL, Yau C, Wolf DM, Tin A, Salari R, Shchegrova S, Pawar H, et al. Circulating tumor DNA in neoadjuvant-treated breast cancer reflects response and survival. Ann Oncol. 2021;32(2):229–39.
44. Papakonstantinou A, Gonzalez NS, Pimentel I, Suñol A, Zamora E, Ortiz C, Espinosa-Bravo M, Peg V, Vivancos A, Saura C, et al. Prognostic value of ctDNA detection in patients with early breast cancer undergoing neoadjuvant therapy: a systematic review and meta-analysis. Cancer Treat Rev. 2022;104:102362.
45. Gianni L, Baselga J, Eiermann W, Guillem Porta V, Semiglazov V, Lluch A, Zambetti M, Sabadell D, Raab G, Llombart Cussac A, et al. Feasibility and tolerability of sequential doxorubicin/paclitaxel followed by cyclophosphamide, methotrexate, and fluorouracil and its effects on tumor response as preoperative therapy. Clin Cancer Res. 2005;11(24 Pt 1):8715–21.
46. Gianni L, Pienkowski T, Im YH, Tseng LM, Liu MC, Lluch A, Starosławska E, de la Haba-Rodriguez J, Im SA, Pedrini JL, et al. 5-year analysis of neoadjuvant pertuzumab and trastuzumab in patients with locally advanced, inflammatory, or early-stage HER2-positive breast cancer (NeoSphere): a multicentre, open-label, phase 2 randomised trial. Lancet Oncol. 2016;17(6):791–800.

Publisher's Note