BMC Cancer

RESEARCH Open Access

# Development and validation of a nomogram model of lung metastasis in breast cancer based on machine learning algorithm and cytokines

Zhaoyi Li[1†], Hao Miao[2†], Wei Bao[3†] and Lansheng Zhang[1*]

## Abstract

**Background** The relationship between cytokines and lung metastasis (LM) in breast cancer (BC) remains unclear and current clinical methods for identifying breast cancer lung metastasis (BCLM) lack precision, thus underscoring the need for an accurate risk prediction model. This study aimed to apply machine learning algorithms for identifying the key risk factors for BCLM before developing a reliable prediction model centered on cytokines.

**Methods** This population-based retrospective study included 326 BC patients admitted to the Second Affiliated Hospital of Xuzhou Medical University between September 2018 and September 2023. After randomly assigning the patients to a training cohort (70%; n = 228) or a validation cohort (30%; n = 98) the risk factors for BCLM were identified using Least Absolute Shrinkage and Selection Operator (LASSO), Extreme Gradient Boosting (XGBoost) and Random Forest (RF) models. Significant risk factors were visualized with a Venn diagram and incorporated into a nomogram model, the performance of which was then evaluated according to three criteria, namely discrimination, calibration and clinical utility using calibration plots, receiver operating characteristic (ROC) curves and decision curve analysis (DCA).

**Results** Among the cohort, 70 patients developed LM. A nomogram was then developed to predict the 5-year and 10-year BCLM risk by incorporating five key variables, namely endocrine therapy, hsCRP, IL6, IFN-α and TNF-α. For the 5-year prediction model, the training and validation cohorts had AUC values of 0.786 (95% CI: 0.691–0.881) and 0.627 (95% CI: 0.441–0.813), respectively, while for the 10-year prediction model, the corresponding AUC values were 0.687 (95% CI: 0.528–0.847) and 0.797 (95% CI: 0.605–0.988), respectively. ROC analysis further confirmed the model's strong discriminative ability, while calibration plots indicated that the predicted and observed outcomes were in good agreement in both cohorts. Finally, DCA demonstrated the model's effectiveness in clinical practice.

**Conclusion** Using machine learning algorithms, this study developed aa nomogram that could effectively identify BC patients who were at a higher risk of developing LM, thus providing a valuable tool for decision-making in clinical settings.

**Keywords** Breast cancer, Metastasis, Cytokine, Machine learning, Nomogram

[†]Zhaoyi Li, Hao Miao and Wei Bao share the first authorship.

*Correspondence:
Lansheng Zhang
lanshengyan@163.com
Full list of author information is available at the end of the article

## Introduction

Breast cancer (BC), the most frequently diagnosed malignant tumor in women worldwide, remains the major cause of cancer-related mortality among the female population [1]. It often begins as a localized disease but can subsequently spread to lymph nodes and distant organs, thereby posing significant challenges to effective treatment [2]. Approximately 10–15% of breast cancer patients develop invasive diseases, with distant metastases occurring within three years of the initial diagnosis. However, metastases may also emerge at least a decade after initial detection of the cancer. Furthermore, the heterogeneity of breast cancer metastasis complicates not only the determination of effective treatment strategies but also the assessment of metastasis risk factors [3]. Distant metastases of BC commonly affect organs, such as the brain, lungs, liver and bone, where they exhibit organ-specific patterns, and hence, each site is often associated with distinct symptoms, prognosis and treatment [4]. In nearly 25% of patients with metastatic BC, the lungs are the first and sometimes the only site of metastasis [5]. Additionally, the lungs represent the second most frequent site of BC metastasis, with a 5-year overall survival rate of only 16.8% [6]. Due to the typically asymptomatic nature of lung metastasis in BC, many patients remain undiagnosed until the disease becomes incurable, thus underscoring the need for timely intervention and proper care [7]. Early detection of lung metastases and accurate prognostic evaluation are crucial for improving outcomes of BC patients in clinical practice, especially to enable better clinical management and potentially achieve long-term survival. However, in Asian populations, the clinicopathological characteristics and risk factors associated with breast cancer lung metastases (BCLM) remain underexplored. This highlights the urgent need for predictive models to identify patients at the highest risk of developing lung metastases, thereby enabling physicians to tailor treatments according to patient needs.

In recent years, machine learning (ML) has emerged as a powerful tool for big data analysis, particularly for predicting the early stages of cancer [8–10]. ML enables the development of algorithms that can learn from data, predict outcomes and improve over time, thereby facilitating precise decision-making [11–13]. As such, its potential in exploring risk factors for disease progression and predict patient prognosis is significant [14]. Some predictive models, such as Extreme Gradient Boosting (XGBoost) [15], Least Absolute Shrinkage and Selection Operator (LASSO) [16] and Random Forest (RF) [17], have demonstrated superior generalization capabilities compared with traditional statistical models, especially since they excel at simulating and predicting complex relationships between variables and outcomes. However, despite these advances, few studies have explored the application of ML models for predicting the risk of lung metastases in BCLM.

Interleukins (ILs), a family of low-molecular-weight cytokines secreted by immune active cells, exhibit both pro-inflammatory and anti-inflammatory properties [18–21]. ILs are involved in cancer-related inflammation, influencing tumor progression through anti-tumor immune responses or by promoting a tumor-supporting microenvironment [22]. Previous research has identified specific ILs as risk factors in breast cancer subtypes. For instance, IL- 5, IL- 7 and IL- 16 were found to be associated with HER2-positive BC, while IL- 10 levels correlated with HER2-negative cases [23]. However, the role of various cytokines in BCLM patients remains unclear, especially with regards to those associated with enhanced risks of developing lung metastases in BC. Therefore, incorporating cytokines into predictive models of lung metastasis in BC is essential. In this context, the current study aimed to identify cytokine-based risk factors for lung metastases in breast cancer and establish a predictive risk model that could guide personalized treatment strategies and improve outcomes for BC patients.

## Methods and materials

### Study design and selection of patients

Approval for this study was obtained from the Ethics Committee of the Second Affiliated Hospital of Xuzhou Medical University (Ethics Approval Number: 120601). Using the inpatients' electronic medical record system of the Second Affiliated Hospital of Xuzhou Medical University, the current research retrospectively analyzed BC patients who were admitted to the hospital between September 2018 and September 2023. The following inclusion criteria was then applied: 1) a histologically-confirmed diagnosis of BC as the only primary malignant tumor; 2) Sufficient information about survival time and follow-up. In addition, BC patients were excluded if: 1) they were male; 2) the time interval from diagnosis to follow-up was less than one year; 3) rheumatic diseases and infections were present; 4) results for cytokine testing were incomplete; 5) they were unmarried; 6) multiple primary tumors were present. Overall, 326 patients met the above criteria, with 70 of them also presenting lung metastases which were diagnosed using radiological scans, biopsy or surgical resection specimens of metastatic lesions. The selected patients were then randomly assigned to two groups: a training group with 228 patients (accounting for 70%) and a validation one consisting of 98 patients (accounting for 30%). The above process of patient selection is visually represented in Fig. 1.

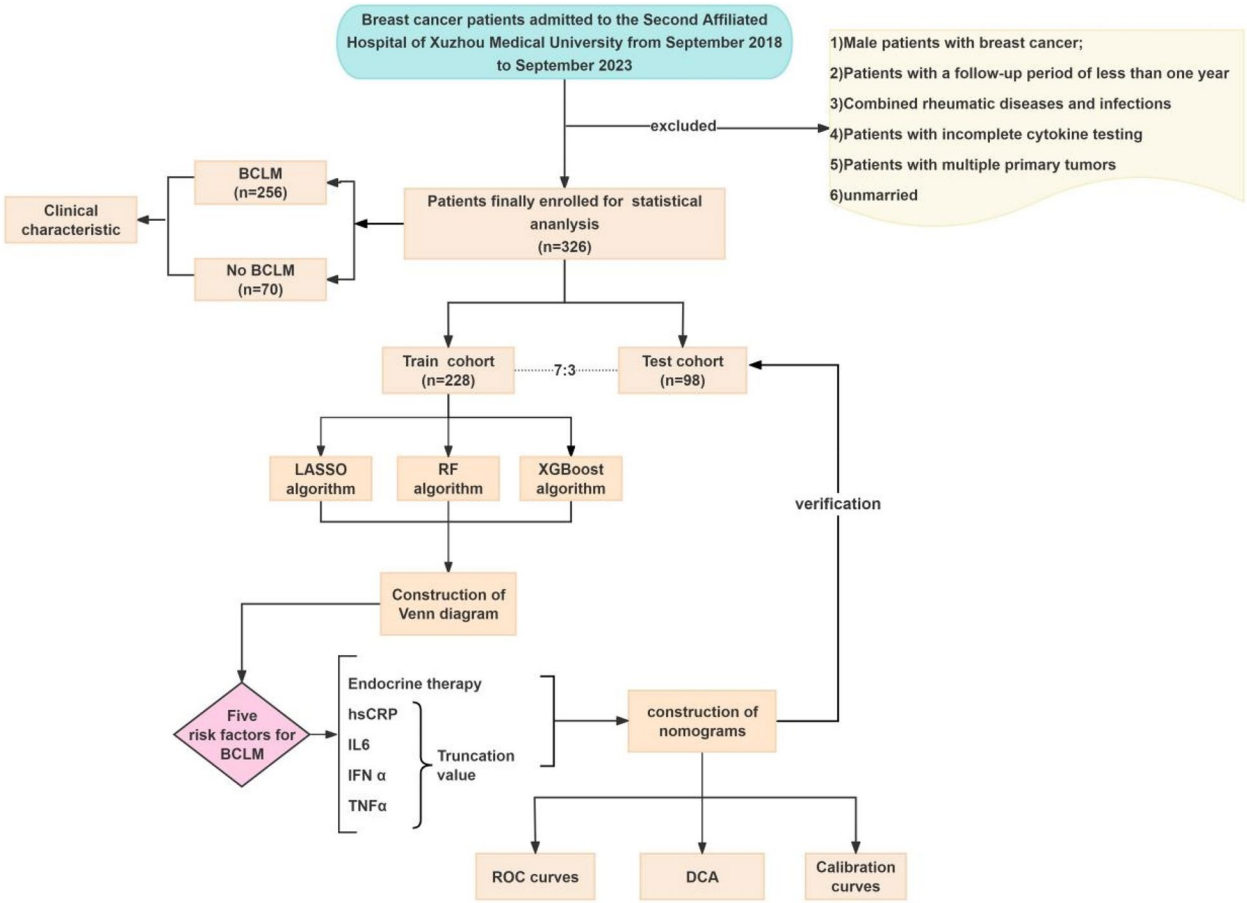


**Fig. 1** The fowchart described the process of conducting the study and statistical analysis. Note: LASSO = Least Absolute Shrinkage and Selection Operator,RF = Random Forest, BCLM = Breast cancer lung metastases,XGBoost = Extreme Gradient Boosting;DCA = Decision Curve Analysis,ROC = Receiver Operating Characteristic Curve

### Data collection and processing

The study collected comprehensive baseline demographic and clinicopathological data for each participant, including age at diagnosis, TNM staging (I–IV), and the number of extrathoracic metastatic sites prior to pulmonary involvement. Treatment history before lung metastasis was documented, encompassing radiotherapy, chemotherapy, endocrine therapy, targeted agents, and immunotherapy. Surgical approach (breast-conserving or radical resection), axillary lymph node involvement, maximal tumor dimension, and immunohistochemical profiles (ER, PR, HER2, Ki67) were recorded. Additional variables included histologic grade (I–III), molecular subtype classification (Luminal A/B, HER2-enriched, triple-negative), tumor laterality (unilateral/bilateral), and histopathological categorization (ductal, lobular, or other). Anthropometric (BMI), menopausal status, serum tumor markers, and novel biomarkers (adenosine kinase 1, high-sensitivity C-reactive protein) were analyzed alongside hematologic parameters, albumin-fibrinogen ratios, and a multiplex cytokine panel (IFN-α/γ, IL-1β/2/4/5/6/8/10/12p70/17 A, TNF-α).

### Endpoint of study

This study's primary endpoint was the occurrence of the first lung metastasis in BC patients. The follow-up deadline was defined as the time from the initial diagnosis of BC to the development of lung metastasis or the date of the last follow-up.

### Machine learning

Machine learning algorithms—including XGBoost, random forest (RF), and least absolute shrinkage and selection operator (LASSO)—were employed to systematically assess risk factors associated with lung metastasis in breast cancer (BC). LASSO, a regression-based method, facilitates feature selection and regularization by identifying the most predictive variables while minimizing overfitting [24, 25]. RF is an ensemble learning method that combines multiple predictions or classifications to

improve overall accuracy of prediction. It is highly versatile, capable of handling both categorical and continuous data, while also demonstrating strong noise resistance which effectively prevents overfitting, a key consideration when analyzing complex datasets [26]. Finally, XGBoost is an ML algorithm based on the gradient boosting framework and the CART decision tree algorithm. It offers high efficiency, flexibility and portability, resulting in superior prediction accuracy [27].

#### Nomogram model construction

Patients were randomly allocated to training and validation cohorts. The training cohort data were used to develop predictive models (RF, XGBoost, and LASSO), while internal validation was performed using the validation cohort [28, 29]. Key risk factors for breast cancer lung metastasis (BCLM) were identified through Venn diagram analysis and incorporated into nomogram models predicting 5- and 10-year BCLM risk. Lung metastasis-free survival was assessed via Kaplan–Meier analysis, with between-group differences evaluated using log-rank tests. Model performance was evaluated based on discrimination (AUC, equivalent to the C-index), calibration (calibration plots and Hosmer–Lemeshow test), and clinical utility (decision curve analysis, DCA).

#### Statistical analysis

Metric data with a normal distribution were presented as mean ± SD, while non-normally distributed ones were expressed as median (P25, P75). Additionally, categorical data were described as counts (percentages). Two-group comparisons were then performed using the Mann–Whitney U test and the independent sample t-test for non-parametric and parametric values, respectively. In the case of categorical variables, results were compared using the chi-square test. All statistical tests, performed using SPSS version 23.0 (SPSS Italy, Florence, Italy) and statistical software package R (version 4.0.0, R statistical calculation project), were two-tailed, with differences considered to be significant at $P < 0.05$.

## Results

#### Patients' baseline characteristics

This study included 326 breast cancer patients who visited the Second Affiliated Hospital of Xuzhou Medical University between September 2018 and September 2023, and among these, 70 patients also presented lung metastases. The cohort's median age was 52 years, and the majority had undergone modified radical surgery (85.28%). Postoperative histopathological analysis revealed that invasive ductal carcinoma was the predominant histological type (74.85%), with Luminal B being the most common molecular subtype (73.62%). Regarding tumor staging, 44.79% of patients were classified as T2 stage, and 34.97% had no lymph node metastases. Additionally, most patients (70.55%) did not experience metastases to other organs before developing lung metastases. In terms of treatment, chemotherapy (92.64%) was the most commonly administered one, followed by radiotherapy (64.42%), endocrine therapy (43.56%), targeted therapy (34.97%) and immunotherapy (3.07%). The non-BCLM and BCLM groups were also significantly different ($P < 0.05$) in terms of several parameters, including CA125, CA153, hsCRP, absolute monocyte count, TNF-α, IL- 8, IL- 6, IL- 2, IL- 1 β and IL- 12p70. The clinical characteristics of the 326 BC patients and baseline comparisons between the BCLM and non-BCLM groups are summarized in Table 1. Furthermore, demographic and clinicopathological characteristics did not differ significantly between patients of the training and validation groups (Table 2).

#### Identification of BCLM risk factors

To identify the risk factors for BCLM, the LASSO algorithm was employed, with Supplementary Figure S1-A showing the binomial deviation curve plotted against the logarithm of the tuning hyperparameter (λ). In this case, the solid vertical line indicates the binomial deviation ± standard error (SE), while the optimal λ value was determined using the minimum standard and 1-SE standard through tenfold cross validation. Furthermore, a coefficient profile was generated from the log (λ) sequence, with 49 clinical parameters integrated into the LASSO model to enable effective penalization of non-essential features. Following model training and the tenfold cross validation, 12 non-zero coefficients were identified as being significantly associated with lung metastasis (Supplementary S1B). According to the Lasso model's feature importance ranking (Fig. 2A), the relative importance of predictors from highest to lowest was as follows: other organ metastasis, endocrine therapy, PR status, absolute lymphocyte count, targeted therapy, IL- 2, INF-α, TNF-α, CEA, CA125, hsCRP, and IL- 6.

#### Identification of risk factors for BCLM

The RF machine learning algorithm was used to further refine the selection of risk factors. This algorithm works by randomly extracting subsets of features from the training data, with each subset subsequently utilized to construct a decision tree. For each node within these decision trees, the optimal feature was chosen from a random subset of features for node partitioning. The decision tree was then recursively built based on the selected features until a predefined stopping

**Table 1** Baseline characteristics of patients with lung metastases from breast cancer (BCLM)

| Characteristic | Overall, $N = 326$[1] | No BCLM, $N = 256$[1] | BCLM, $N = 70$[1] | $p$-value[2] |
|---|---|---|---|---|
| Age, years | 52 (45, 58) | 52 (45, 58) | 50 (42, 58) | 0.197 |
| BMI,kg/m$^2$ | 22.97 (21.48, 24.61) | 22.97 (21.48, 24.46) | 22.74 (19.81, 25.01) | 0.169 |
| β2MG, ug/ml | 1.80 (1.53, 2.24) | 1.79 (1.52, 2.23) | 1.93 (1.59, 2.44) | 0.171 |
| CEA2, ng/ml | 2 (2, 4) | 2 (2, 3) | 3 (2, 7) | 0.148 |
| TSGF, U/ml | 53 (46, 62) | 53 (46, 61) | 53 (45, 62) | 0.981 |
| SCC,ug/ml | 0.58 (0.47, 0.78) | 0.59 (0.48, 0.78) | 0.56 (0.39, 0.78) | 0.182 |
| CA125, U/ml | 14 (10, 22) | 14 (10, 20) | 22 (12, 41) | < 0.001 |
| CA153, U/ml | 10 (8, 20) | 10 (7, 19) | 18 (10, 44) | < 0.001 |
| CA50, IU/ml | 6 (4, 11) | 6 (4, 10) | 7 (4, 12) | 0.078 |
| SF,ng/ml | 110 (58, 219) | 104 (59, 191) | 175 (58, 330) | 0.053 |
| hsCRP,mg/L | 2 (1, 3) | 2 (1, 3) | 3 (1, 9) | 0.006 |
| Neutrophil count,10$^9$/L | 3.05 (2.34, 4.41) | 3.08 (2.30, 4.39) | 3.05 (2.44, 4.99) | 0.777 |
| Lymphocyte count,10$^9$/L | 1.36 (1.12, 1.72) | 1.36 (1.13, 1.76) | 1.41 (1.09, 1.63) | 0.725 |
| Hb,g/L | 125 (114, 132) | 126 (114, 132) | 123 (113, 131) | 0.357 |
| PLT,10$^9$/L | 236 (179, 290) | 239 (182, 292) | 231 (176, 283) | 0.454 |
| Monocyte count,10$^9$/L | 0.38 (0.31, 0.52) | 0.36 (0.31, 0.49) | 0.46 (0.35, 0.55) | 0.004 |
| Albumin,g/L | 43.6 (40.7, 46.5) | 43.5 (40.7, 46.5) | 44.9 (41.6, 46.6) | 0.215 |
| Fibrinogen,g/L | 3.86 (3.27, 4.21) | 3.81 (3.27, 4.20) | 3.96 (3.57, 4.26) | 0.155 |
| IFNα, pg/ml | 2.14 (1.46, 3.67) | 2.30 (1.46, 3.94) | 1.99 (1.46, 3.00) | 0.337 |
| IFNγ,pg/ml | 2.8 (1.9, 4.8) | 3.0 (2.0, 5.0) | 2.4 (1.7, 3.9) | 0.119 |
| IL12p70, pg/ml | 1.94 (1.14, 3.04) | 1.98 (1.21, 3.27) | 1.71 (0.95, 2.34) | 0.019 |
| IL17 A, pg/ml | 4 (2, 10) | 4 (2, 11) | 4 (2, 8) | 0.844 |
| IL1β, pg/ml | 1.75 (1.07, 2.86) | 1.92 (1.15, 3.21) | 1.62 (0.74, 2.72) | 0.042 |
| IL2, pg/ml | 1.75 (0.98, 3.17) | 1.84 (1.03, 3.37) | 1.38 (0.87, 2.26) | 0.005 |
| IL4, pg/ml | 2.06 (1.36, 3.35) | 2.08 (1.37, 3.40) | 1.94 (1.25, 2.99) | 0.281 |
| IL5, pg/ml | 1.11 (0.68, 1.54) | 1.15 (0.74, 1.59) | 1.05 (0.63, 1.31) | 0.178 |
| IL6, pg/ml | 5 (3, 9) | 5 (3, 8) | 8 (4, 21) | 0.001 |
| IL8, pg/ml | 9 (6, 14) | 9 (6, 13) | 12 (6, 22) | 0.005 |
| IL10, pg/ml | 3.35 (2.07, 4.48) | 3.53 (2.03, 4.63) | 3.10 (2.15, 4.15) | 0.452 |
| TNFα, pg/ml | 1.95 (1.23, 3.45) | 2.04 (1.28, 3.56) | 1.59 (1.17, 2.48) | 0.010 |
| Organ transfer | | | | 0.682 |
| No, n (%) | 230 (70.55%) | 182 (71.09%) | 48 (68.57%) | |
| Yes, n (%) | 96 (29.45%) | 74 (28.91%) | 22 (31.43%) | |
| Endocrine therapy | | | | 0.004 |
| No, n (%) | 70 (21.47%) | 22 (31.43%) | 48 (18.75%) | |
| Yes, n (%) | 256 (78.53%) | 48 (68.57%) | 208 (81.25%) | |
| Targeted therapy | | | | 0.067 |
| No, n (%) | 212 (65.03%) | 160 (62.50%) | 52 (74.29%) | |
| Yes, n (%) | 114 (34.97%) | 96 (37.50%) | 18 (25.71%) | |
| Immunotherapy | | | | 0.231 |
| No, n (%) | 316 (96.93%) | 250 (97.66%) | 66 (94.29%) | |
| Yes, n (%) | 10 (3.07%) | 6 (2.34%) | 4 (5.71%) | |
| Radiotherapy | | | | 0.046 |
| No, n (%) | 116 (35.58%) | 84 (32.81%) | 32 (45.71%) | |
| Yes, n (%) | 210 (64.42%) | 172 (67.19%) | 38 (54.29%) | |
| Chemotherapy | | | | 0.142 |
| No, n (%) | 24 (7.36%) | 16 (6.25%) | 8 (11.43%) | |
| Yes, n (%) | 302 (92.64%) | 240 (93.75%) | 62 (88.57%) | |
| AJCC-T | | | | 0.010 |

**Table 1** (continued)

| Characteristic | Overall, $N = 326$[1] | No BCLM, $N = 256$[1] | BCLM, $N = 70$[1] | $p$-value[2] |
|---|---|---|---|---|
| T1, n (%) | 90 (27.61%) | 78 (30.47%) | 12 (17.14%) | |
| T2, n (%) | 146 (44.79%) | 116 (45.31%) | 30 (42.86%) | |
| T3, n (%) | 24 (7.36%) | 20 (7.81%) | 4 (5.71%) | |
| T4, n (%) | 24 (7.36%) | 14 (5.47%) | 10 (14.29%) | |
| Unknown, n (%) | 42 (12.88%) | 28 (10.94%) | 14 (20.00%) | |
| AJCC-N | | | | 0.099 |
| N0, n (%) | 114 (34.97%) | 96 (37.50%) | 18 (25.71%) | |
| N1, n (%) | 82 (25.15%) | 68 (26.56%) | 14 (20.00%) | |
| N2, n (%) | 62 (19.02%) | 44 (17.19%) | 18 (25.71%) | |
| N3, n (%) | 42 (12.88%) | 30 (11.72%) | 12 (17.14%) | |
| Unknown, n (%) | 26 (7.98%) | 18 (7.03%) | 8 (11.43%) | |
| AJCC-M | | | | 0.682 |
| M0, n (%) | 318 (97.55%) | 250 (97.66%) | 68 (97.14%) | |
| M1, n (%) | 8 (2.45%) | 6 (2.34%) | 2 (2.86%) | |
| Surgery | | | | 0.038 |
| Modified radical mastectomy, n (%) | 278 (85.28%) | 212 (82.81%) | 66 (94.29%) | |
| Breast conserving surgery, n (%) | 32 (9.82%) | 30 (11.72%) | 2 (2.86%) | |
| No surgery, n (%) | 16 (4.91%) | 14 (5.47%) | 2 (2.86%) | |
| Pathological grading | | | | 0.191 |
| I, n (%) | 14 (4.29%) | 12 (4.69%) | 2 (2.86%) | |
| II, n (%) | 106 (32.52%) | 90 (35.16%) | 16 (22.86%) | |
| III, n (%) | 60 (18.40%) | 46 (17.97%) | 14 (20.00%) | |
| IV, n (%) | 4 (1.23%) | 4 (1.56%) | 0 (0.00%) | |
| Unknown, n (%) | 142 (43.56%) | 104 (40.63%) | 38 (54.29%) | |
| ER + | | | | 0.257 |
| Yes, n (%) | 192 (58.90%) | 156 (60.94%) | 36 (51.43%) | |
| No, n (%) | 104 (31.90%) | 76 (29.69%) | 28 (40.00%) | |
| Unknow, n (%) | 30 (9.20%) | 24 (9.38%) | 6 (8.57%) | |
| PR + | | | | < 0.001 |
| Yes, n (%) | 176 (53.99%) | 152 (59.38%) | 24 (34.29%) | |
| No, n (%) | 120 (36.81%) | 80 (31.25%) | 40 (57.14%) | |
| Unknow, n (%) | 30 (9.20%) | 24 (9.38%) | 6 (8.57%) | |
| HER2 + | | | | 0.970 |
| Yes, n (%) | 212 (65.03%) | 166 (64.84%) | 46 (65.71%) | |
| No, n (%) | 78 (23.93%) | 62 (24.22%) | 16 (22.86%) | |
| Unknow, n (%) | 36 (11.04%) | 28 (10.94%) | 8 (11.43%) | |
| ki67 > 14% | | | | 0.001 |
| Yes, n (%) | 202 (61.96%) | 160 (62.50%) | 42 (60.00%) | |
| No, n (%) | 54 (16.56%) | 50 (19.53%) | 4 (5.71%) | |
| Unknow, n (%) | 70 (21.47%) | 46 (17.97%) | 24 (34.29%) | |
| Subtype | | | | 0.109 |
| Luminal A, n (%) | 16 (4.91%) | 14 (5.47%) | 2 (2.86%) | |
| Luminal B, n (%) | 240 (73.62%) | 194 (75.78%) | 46 (65.71%) | |
| Triple-negative, n (%) | 42 (12.88%) | 30 (11.72%) | 12 (17.14%) | |
| HER2, n (%) | 28 (8.59%) | 18 (7.03%) | 10 (14.29%) | |
| Laterality | | | | 0.338 |
| Left, n (%) | 184 (56.44%) | 146 (57.03%) | 38 (54.29%) | |
| Right, n (%) | 138 (42.33%) | 108 (42.19%) | 30 (42.86%) | |
| Bilateral, n (%) | 4 (1.23%) | 2 (0.78%) | 2 (2.86%) | |

**Table 1** (continued)

| Characteristic | Overall, $N = 326$[1] | No BCLM, $N = 256$[1] | BCLM, $N = 70$[1] | *p*-value[2] |
|---|---|---|---|---|
| Pathological type | | | | 0.121 |
| Ductal carcinoma, n (%) | 244 (74.85%) | 188 (73.44%) | 56 (80.00%) | |
| Lobular carcinoma, n (%) | 6 (1.84%) | 6 (2.34%) | 0 (0.00%) | |
| Other types, n (%) | 36 (11.04%) | 26 (10.16%) | 10 (14.29%) | |
| Unknown, n (%) | 40 (12.27%) | 36 (14.06%) | 4 (5.71%) | |
| Menopausal | | | | 0.801 |
| No, n (%) | 154 (47.24%) | 120 (46.88%) | 34 (48.57%) | |
| Yes, n (%) | 172 (52.76%) | 136 (53.13%) | 36 (51.43%) | |

[1] Median (IQR); n (%)

[2] Wilcoxon rank sum test; Pearson's Chi-squared test; Fisher's exact test

BMI = Body Mass Index,SF = Ferritin, hs-CRP = Hypersensitive C-reactive protein,Hb = Haemoglobin,PLT = Platelet,ER = Estrogen receptors,PR = Progesterone receptors,Her2 = Human epidermal growth factor receptor- 2

condition was met. For classification problems, the final class was determined through a majority voting mechanism, while in the case of regression problems, the average of the predicted values from all trees served as the final prediction. The algorithm eventually combined the outputs from all constructed decision trees to calculate the average error rates separately for node-positive and node-negative groups. The importance of clinical features was subsequently assessed before visualizing their rankings (Fig. 2B). Overall, 24 clinically relevant features were identified as risk factors for BCLM, and they included IL- 6, PR, BMI, SCC, FIB, ER, endocrine therapy, IL- 8, ALB, IL- 17 A, Hb, IFN-γ, hsCRP, TNF-α, IL- 12p70, CA125, TSGF, PLT, INF-α, N, SF, B2M, CA50 and IL10.

The XGBoost model, based on a gradient boosting framework, is another ensemble method that uses decision trees to enhance predictive accuracy. Gradient boosting is a specific implementation of the Boosting technique which iteratively minimize the objective function by fitting each new tree with the negative gradient of the previous round's error. In this study, the XGBoost model identified 15 non-zero coefficients that were significantly correlated with lung metastasis. These features were then ranked by relative importance (Fig. 2C) as follows: endocrine therapy, IL- 6, hsCRP, IL- 17 A, PLT, IL- 8, SCC, SF, INF-α, CEA, ALB, Hb, IL2, M, TNF-α.

The RF, LASSO and XGBoost algorithms were used to independently identify BCLM-related risk factors, with overlapping variables among the three ML models subsequently selected as significant ones. The intersection of these factors was visualized using a Venn diagram (Fig. 2D), and the results highlighted five key variables for subsequent nomogram analysis: endocrine therapy, hsCRP, IL6, IFN-α and TNF-α. To enhance clinical applicability, the Maxstat method was then used to assess the optimal risk cut-off points for the five variables (Supplementary Figure S2). Using these cut-off values (hsCRP (16.8), IL6 (16.19), IFN-α (2.36) and TNF-α (1.35)), the biomarkers were reclassified into high- and low-risk groups prior to analysis using Kaplan–Meier curves (Supplementary Figure S3) to determine survival outcomes. Additionally, Supplementary Figure S4 shows the lung metastasis rates across different molecular subtypes. Compared with Luminal A patients, those with the Luminal B subtype demonstrated a lower risk of lung metastasis (unadjusted HR: 1.787; P = 0.435), while HER2 + (unadjusted HR: 3.571; $P = 0.094$) exhibited a higher risk. In particular, patients with TNBC faced the highest risk of lung metastasis (unadjusted HR: 6.487; $P = 0.018$).

### Establishment and validation of BCLM diagnostic nomogram

A nomogram model was constructed based on five key variables: endocrine therapy, hsCRP, IL6, IFN-α, and TNF-α. Each variable was assigned a point value ranging from 0 to 100 (Fig. 3). The cumulative score, obtained by summing these points, allowed estimation of the 5- and 10-year lung metastasis probability in breast cancer (BC) patients, aiding clinical decision-making. Risk prediction involved drawing a vertical line from the total score to the probability axis (ranging from 0.1 to 0.95), though not all probabilities aligned precisely with marked values. The model's performance was evaluated based on discrimination, calibration, and clinical utility in both training and validation cohorts, with results visualized using ROC curves, calibration plots, and decision curve analysis.

**Table 2** Comparison of baseline features between training group and validation group

| Characteristic | Overall, *N* = 3261 | Training group, *N* = 2281 | Validation groups, *N* = 981 | *p*-value[2] |
|---|---|---|---|---|
| Age, years | 52 (45, 58) | 52 (44, 57) | 52 (46, 59) | 0.685 |
| BMI,kg/m$^2$ | 22.97 (21.48, 24.61) | 22.89 (21.48, 24.61) | 23.44 (21.48, 24.61) | 0.973 |
| β2MG, ug/ml | 1.80 (1.53, 2.24) | 1.82 (1.50, 2.23) | 1.78 (1.54, 2.29) | 0.717 |
| CEA2, ng/ml | 2 (2, 4) | 2 (2, 4) | 2 (2, 3) | 0.631 |
| TSGF, U/ml | 53 (46, 62) | 52 (46, 61) | 53 (47, 62) | 0.217 |
| SCC,ug/ml | 0.58 (0.47, 0.78) | 0.59 (0.48, 0.79) | 0.56 (0.45, 0.73) | 0.146 |
| CA125, U/ml | 14 (10, 22) | 15 (10, 23) | 14 (10, 22) | 0.822 |
| CA153, U/ml | 10 (8, 20) | 11 (8, 20) | 10 (8, 19) | 0.934 |
| CA50, IU/ml | 6 (4, 11) | 6 (4, 11) | 7 (5, 10) | 0.904 |
| SF,ng/ml | 110 (58, 219) | 108 (58, 220) | 111 (60, 213) | 0.935 |
| hsCRP,mg/L | 2 (1, 3) | 2 (1, 3) | 1 (1, 3) | 0.132 |
| Neutrophil count,10$^9$/L | 3.05 (2.34, 4.41) | 3.05 (2.42, 4.39) | 3.17 (2.30, 4.45) | 0.411 |
| Lymphocyte count,10$^9$/L | 1.36 (1.12, 1.72) | 1.40 (1.12, 1.74) | 1.35 (1.12, 1.68) | 0.622 |
| Hb,g/L | 125 (114, 132) | 125 (115, 132) | 125 (111, 132) | 0.405 |
| PLT,10$^9$/L | 236 (179, 290) | 233 (180, 284) | 242 (181, 295) | 0.363 |
| Monocyte count,10$^9$/L | 0.38 (0.31, 0.52) | 0.38 (0.31, 0.50) | 0.40 (0.30, 0.54) | 0.904 |
| Albumin,g/L | 43.6 (40.7, 46.5) | 43.9 (40.9, 46.5) | 43.0 (40.4, 46.1) | 0.301 |
| Fibrinogen,g/L | 3.86 (3.27, 4.21) | 3.88 (3.27, 4.21) | 3.82 (3.33, 4.21) | 0.869 |
| IFNα, pg/ml | 2.14 (1.46, 3.67) | 2.15 (1.40, 3.64) | 2.14 (1.49, 4.14) | 0.866 |
| IFNγ,pg/ml | 2.8 (1.9, 4.8) | 2.8 (1.9, 4.9) | 2.7 (1.9, 4.0) | 0.507 |
| IL12p70, pg/ml | 1.94 (1.14, 3.04) | 1.96 (1.13, 3.05) | 1.93 (1.22, 3.01) | 0.809 |
| IL17 A, pg/ml | 4 (2, 10) | 4 (2, 11) | 3 (2, 8) | 0.059 |
| IL1β, pg/ml | 1.75 (1.07, 2.86) | 1.81 (1.16, 2.86) | 1.64 (0.95, 2.81) | 0.121 |
| IL2, pg/ml | 1.75 (0.98, 3.17) | 1.76 (1.01, 2.92) | 1.70 (0.96, 3.17) | 0.713 |
| IL4, pg/ml | 2.06 (1.36, 3.35) | 2.04 (1.39, 3.33) | 2.09 (1.28, 3.35) | 0.717 |
| IL5, pg/ml | 1.11 (0.68, 1.54) | 1.10 (0.67, 1.55) | 1.16 (0.75, 1.40) | 0.437 |
| IL6, pg/ml | 5 (3, 9) | 5 (3, 9) | 5 (3, 9) | 0.506 |
| IL8, pg/ml | 9 (6, 14) | 9 (6, 14) | 8 (4, 16) | 0.118 |
| IL10, pg/ml | 3.35 (2.07, 4.48) | 3.46 (2.23, 4.68) | 3.05 (1.83, 4.28) | 0.130 |
| TNFα, pg/ml | 1.95 (1.23, 3.45) | 1.95 (1.26, 3.52) | 1.92 (1.17, 3.22) | 0.739 |
| Organ transfer | | | | 0.015 |
| No, n (%) | 230 (70.55%) | 170 (74.56%) | 60 (61.22%) | |
| Yes, n (%) | 96 (29.45%) | 58 (25.44%) | 38 (38.78%) | |
| Endocrine therapy | | | | 0.867 |
| No, n (%) | 70 (21.47%) | 51 (22.37%) | 19 (19.39%) | |
| Yes, n (%) | 256 (78.53%) | 177 (77.63%) | 79 (80.61%) | |
| Targeted therapy | | | | 0.014 |
| No, n (%) | 212 (65.03%) | 158 (69.30%) | 54 (55.10%) | |
| Yes, n (%) | 114 (34.97%) | 70 (30.70%) | 44 (44.90%) | |
| Immunotherapy | | | | 0.292 |
| No, n (%) | 316 (96.93%) | 219 (96.05%) | 97 (98.98%) | |
| Yes, n (%) | 10 (3.07%) | 9 (3.95%) | 1 (1.02%) | |
| Radiotherapy | | | | 0.826 |
| No, n (%) | 116 (35.58%) | 82 (35.96%) | 34 (34.69%) | |
| Yes, n (%) | 210 (64.42%) | 146 (64.04%) | 64 (65.31%) | |
| Chemotherapy | | | | 0.716 |
| No, n (%) | 24 (7.36%) | 16 (7.02%) | 8 (8.16%) | |
| Yes, n (%) | 302 (92.64%) | 212 (92.98%) | 90 (91.84%) | |
| AJCC-T | | | | 0.608 |

**Table 2** (continued)

| Characteristic | Overall, *N* = 3261 | Training group, *N* = 2281 | Validation groups, *N* = 981 | *p*-value[2] |
|---|---|---|---|---|
| T1, n (%) | 90 (27.61%) | 63 (27.63%) | 27 (27.55%) | |
| T2, n (%) | 146 (44.79%) | 107 (46.93%) | 39 (39.80%) | |
| T3, n (%) | 24 (7.36%) | 15 (6.58%) | 9 (9.18%) | |
| T4, n (%) | 42 (12.88%) | 26 (11.40%) | 16 (16.33%) | |
| Unknown, n (%) | 24 (7.36%) | 17 (7.46%) | 7 (7.14%) | |
| AJCC-N | | | | 0.128 |
| N0, n (%) | 114 (34.97%) | 82 (35.96%) | 32 (32.65%) | |
| N1, n (%) | 82 (25.15%) | 53 (23.25%) | 29 (29.59%) | |
| N2, n (%) | 62 (19.02%) | 45 (19.74%) | 17 (17.35%) | |
| N3, n (%) | 42 (12.88%) | 34 (14.91%) | 8 (8.16%) | |
| Unknown, n (%) | 26 (7.98%) | 14 (6.14%) | 12 (12.24%) | |
| AJCC-M | | | | > 0.999 |
| M0, n (%) | 318 (97.55%) | 222 (97.37%) | 96 (97.96%) | |
| M1, n (%) | 8 (2.45%) | 6 (2.63%) | 2 (2.04%) | |
| Surgery | | | | 0.306 |
| Modified radical mastectomy, n (%) | 278 (85.28%) | 194 (85.09%) | 84 (85.71%) | |
| Breast conserving surgery, n (%) | 32 (9.82%) | 25 (10.96%) | 7 (7.14%) | |
| No surgery, n (%) | 16 (4.91%) | 9 (3.95%) | 7 (7.14%) | |
| Pathological grading | | | | 0.417 |
| I, n (%) | 14 (4.29%) | 9 (3.95%) | 5 (5.10%) | |
| II, n (%) | 106 (32.52%) | 76 (33.33%) | 30 (30.61%) | |
| III, n (%) | 60 (18.40%) | 47 (20.61%) | 13 (13.27%) | |
| IV, n (%) | 4 (1.23%) | 3 (1.32%) | 1 (1.02%) | |
| Unknown, n (%) | 142 (43.56%) | 93 (40.79%) | 49 (50.00%) | |
| ER + | | | | 0.022 |
| Yes, n (%) | 176 (53.99%) | 134 (58.77%) | 42 (42.86%) | |
| No, n (%) | 120 (36.81%) | 77 (33.77%) | 43 (43.88%) | |
| Unknow, n (%) | 30 (9.20%) | 17 (7.46%) | 13 (13.27%) | |
| PR + | | | | 0.013 |
| Yes, n (%) | 192 (58.90%) | 146 (64.04%) | 46 (46.94%) | |
| No, n (%) | 104 (31.90%) | 65 (28.51%) | 39 (39.80%) | |
| Unknow, n (%) | 30 (9.20%) | 17 (7.46%) | 13 (13.27%) | |
| HER2 + | | | | 0.043 |
| Yes, n (%) | 212 (65.03%) | 143 (62.72%) | 69 (70.41%) | |
| No, n (%) | 78 (23.93%) | 63 (27.63%) | 15 (15.31%) | |
| Unknow, n (%) | 36 (11.04%) | 22 (9.65%) | 14 (14.29%) | |
| ki67 > 14% | | | | 0.264 |
| Yes, n (%) | 202 (61.96%) | 143 (62.72%) | 59 (60.20%) | |
| No, n (%) | 54 (16.56%) | 41 (17.98%) | 13 (13.27%) | |
| Unknow, n (%) | 70 (21.47%) | 44 (19.30%) | 26 (26.53%) | |
| Subtype | | | | 0.158 |
| Luminal A, n (%) | 16 (4.91%) | 13 (5.70%) | 3 (3.06%) | |
| Luminal B, n (%) | 240 (73.62%) | 164 (71.93%) | 76 (77.55%) | |
| Triple-negative, n (%) | 42 (12.88%) | 27 (11.84%) | 15 (15.31%) | |
| HER2, n (%) | 28 (8.59%) | 24 (10.53%) | 4 (4.08%) | |
| Laterality | | | | 0.252 |
| Left, n (%) | 184 (56.44%) | 135 (59.21%) | 49 (50.00%) | |
| Right, n (%) | 138 (42.33%) | 90 (39.47%) | 48 (48.98%) | |
| Bilateral, n (%) | 4 (1.23%) | 3 (1.32%) | 1 (1.02%) | |

**Table 2** (continued)

| Characteristic | Overall, *N* = 3261 | Training group, *N* = 2281 | Validation groups, *N* = 981 | *p*-value[2] |
|---|---|---|---|---|
| Pathological type | | | | 0.655 |
| Ductal carcinoma, n (%) | 244 (74.85%) | 170 (74.56%) | 74 (75.51%) | |
| Lobular carcinoma, n (%) | 6 (1.84%) | 4 (1.75%) | 2 (2.04%) | |
| Other types, n (%) | 36 (11.04%) | 28 (12.28%) | 8 (8.16%) | |
| Unknown, n (%) | 40 (12.27%) | 26 (11.40%) | 14 (14.29%) | |
| Menopausal | | | | 0.299 |
| No, n (%) | 154 (47.24%) | 112 (49.12%) | 42 (42.86%) | |
| Yes, n (%) | 172 (52.76%) | 116 (50.88%) | 56 (57.14%) | |
| Lung Metastasis | | | | 0.234 |
| No BCLM, n (%) | 256 (78.53%) | 175 (76.75%) | 81 (82.65%) | |
| BCLM, n (%) | 70 (21.47%) | 53 (23.25%) | 17 (17.35%) | |

[1] Median (IQR); n (%)

2 Wilcoxon rank sum test; Pearson's Chi-squared test; Fisher's exact test

BMI = Body Mass Index,SF = Ferritin, hs-CRP = Hypersensitive C-reactive protein,Hb = Haemoglobin,PLT = Platelet,ER = Estrogen receptors,PR = Progesterone receptors,Her2 = Human epidermal growth factor receptor- 2

In the training set, the AUC values for 5- and 10-year metastasis prediction were 0.786 (95% CI: 0.691–0.881) and 0.787 (95% CI: 0.749–0.824), respectively. The validation set yielded AUCs of 0.627 (95% CI: 0.441–0.813) for 5-year and 0.797 (95% CI: 0.605–0.988) for 10-year prediction. These findings indicate robust predictive accuracy across both datasets (Fig. 4A, B).

### Calibration curve and DCA analysis

Calibration curves were generated for evaluating the nomogram's performance. Following internal validation with 1000 bootstrap iterations, the calibration curves for both the training and validation sets (Fig. 5A, B) closely aligned with the diagonal line, indicating that the predicted and actual probabilities of lung metastasis were in strong agreement. The nomogram's clinical utility was assessed using DCA (Fig. 6A, B), in which the horizontal line represented the assumption of no lung metastasis where the net benefit was zero and the diagonal line represented the scenario where all patients were assumed to have BCLM. Overall, the decision curves demonstrated that the range of high threshold probabilities was broad and applicable to both the training and validation sets. Compared with individual variables, the nomogram exhibited a higher net benefit for both datasets, thus underscoring its superior predictive ability. This indicates that the nomogram can effectively predict the 5-year and 10-year risk of lung metastasis in BC patients.

## Discussion

In this study, multiple ML algorithms were applied to determine the risk factors for BCLM, with the following five significant predictors subsequently identified: endocrine therapy, hsCRP, IL6, IFN-α, and TNF-α. These variables were then integrated into a nomogram model. The findings provided a framework for identifying BC patients who were at a higher risk of lung metastasis, thereby improving prognostic evaluation and clinical management while offering new insights for developing more effective treatments. Additionally, this study was also the first one to construct a lung metastasis prediction model for BC patients based on cytokines. The model demonstrated high accuracy in predicting survival outcomes for BCLM patients, and in practice, the nomogram, which integrated predictions from RF, LASSO and XGBoost algorithms, exhibited robust performance across both the training and validation groups.

A Mendelian randomization analysis involving 420,964 cancer-free patients from the UK Biobank cohort showed that elevated serum c-reactive protein (CRP) levels were linked to higher risks of breast cancer, colorectal cancer, head and neck as well as other malignancies over a 7.1-year follow-up period [30]. Similarly, a meta-analysis examining 119 inflammatory markers (with CRP as the primary focus) across 26 cancer types reached comparable conclusions [31]. These pan-cancer studies identified individuals

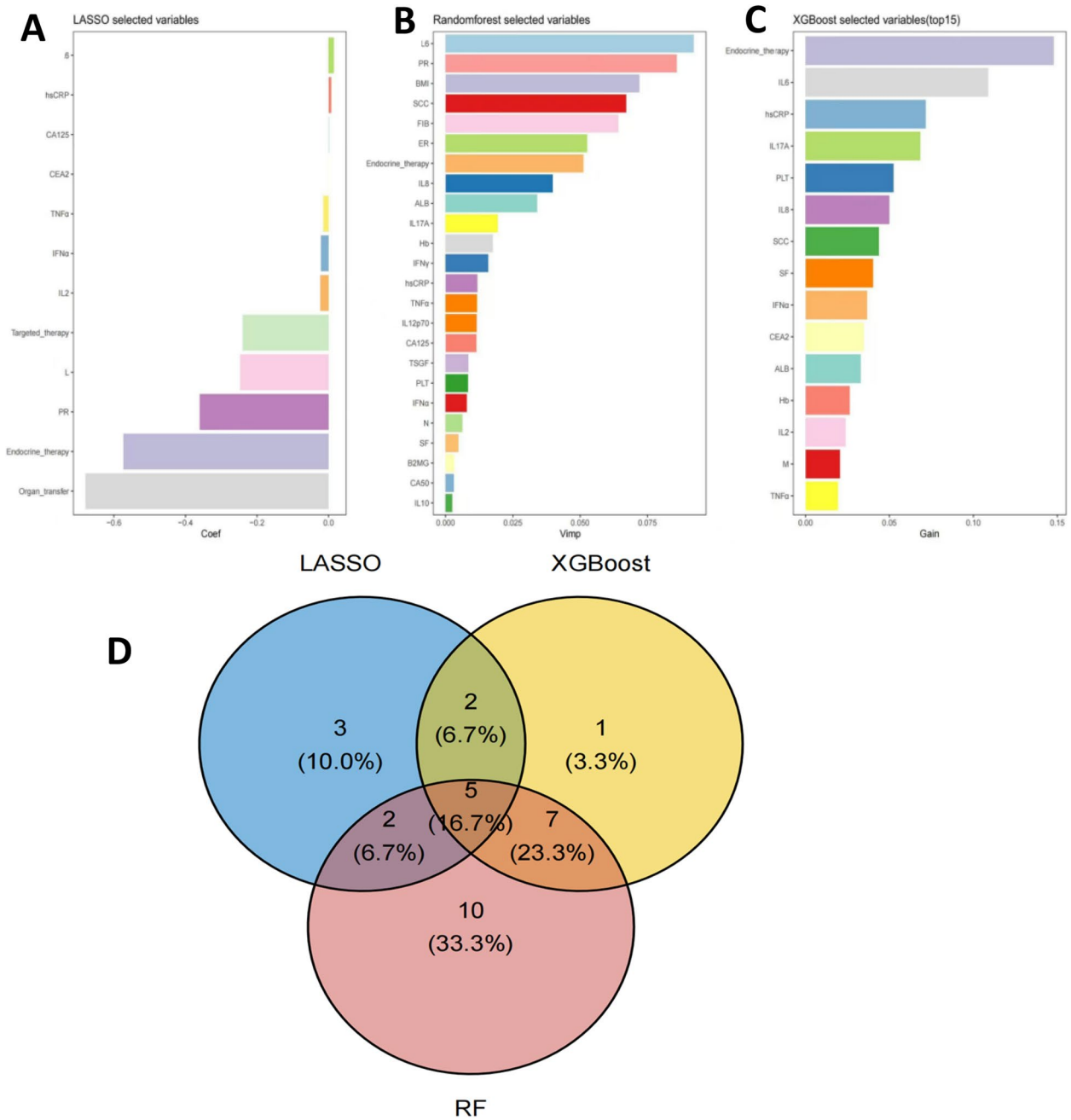


**Fig. 2** Features selection using Lasso algorithm. **A** The importance of 12 features was ranked using Lasso algorithm. **B** Identification of risk factors for BCLM using RF. **C** Ranking of relative importance of features of XGBoost model. **D** Five common risk factors for BCLM were visualized using a Venn diagram

with CRP levels above 3 mg/L as having a high risk of inflammation but this threshold may not apply specifically to BC [30, 32]. The findings further corroborated the link between elevated hypersensitive CRP and an increased risk of BCLM. This suggests that inflammation in BC patients may contribute to tumor proliferation and metastasis, including lung metastasis. The above analyses also determined the optimal hsCRP cut-off value for predicting BCLM to be 16.8 mg/L, hence providing a potential reference point for individualized breast cancer treatment. Interestingly, CRP was consistently identified as a key risk factor for BCLM across all three ML models used in this study. While prior meta-analyses have highlighted

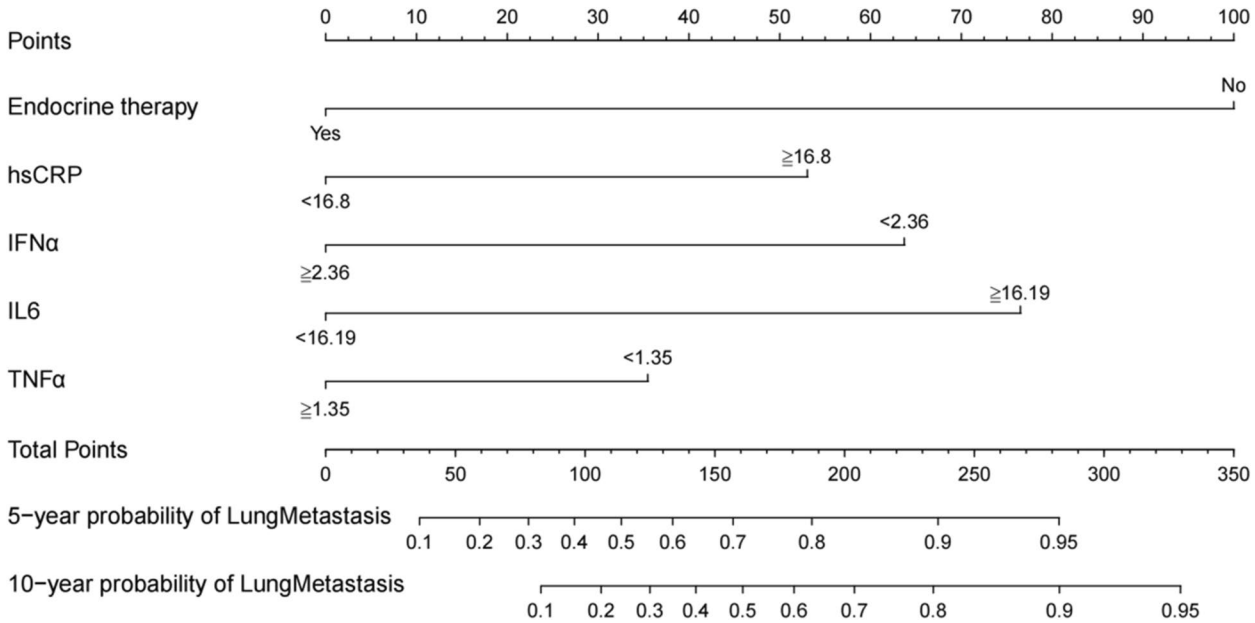


**Fig. 3** Nomogram for the prediction of LM in breast cancer

the limited predictive value of CRP in non-metastatic BC, the association between elevated CRP levels and poor prognosis is well documented in metastatic cases [33, 34]. For instance, in vitro studies revealed that CRP could promote the adhesion of MCF10 A human breast epithelial cells through activation of the integrin α 2 signaling pathway and Fcγ receptor I (FcγRI), with the process subsequently activating paxillin, FAK and ERKs to drive autocrine effects [35]. Furthermore, using an invasion model of MDA-MB- 231 TNBC cells and mouse tumor models, CRP was shown to be involved in tumor growth. Additional animal experiments further demonstrated that CRP impaired immune surveillance by inhibiting the activation of pulmonary macrophages, induced by symbiotic bacteria through an FcγR2B dependent mechanism, thereby fostering the formation of pre-metastatic niches in the lungs of tumor-bearing mice [36]. Altogether, these findings highlight the significant role of CRP in lung metastasis, thus supporting this study's results.

This study underscores the potential of endocrine therapy to reduce the risk of BCLM, with this lower risk being particularly evident among hormone receptor-positive patients who constituted over half of the total study population. Of these patients, 80% received endocrine therapy, including options such as tamoxifen and steroidal (exemestane) or nonsteroidal (letrozole or anastrozole) aromatase inhibitors. Tamoxifen is known to improve disease-free and overall survival in postmenopausal women with ER-positive tumors [37, 38]. However, the BIG trial demonstrated that first-line treatment with aromatase inhibitors lowered the absolute risk of 10-year recurrence by 3.6%, increased overall survival by 2.1% and outperformed tamoxifen monotherapy [39]. Furthermore, post hoc analyses of the SOFT and TEXT trials revealed that combining ovarian suppression with tamoxifen significantly improved 8-year disease-free and overall survival rates in comparison with tamoxifen alone [40]. Despite the success of endocrine therapy in reducing BC recurrence and mortality, both intrinsic and acquired drug resistance remain a challenge. In this context, recent advances in understanding the drivers and mechanisms underlying endocrine therapy resistance in estrogen receptor-positive BC has led to the development of targeted drugs, such as mTOR inhibitors and cyclin dependent kinase 4/6 inhibitors can markedly extend progression-free survival [41, 42]. When lung metastasis rates were further analyzed by molecular subtypes, it was found that the risk of lung metastasis was significantly lower in hormone receptor-positive patients compared with HER2 +ones, with TNBC patients exhibiting the highest risk. These results underscore the importance of endocrine therapy in mitigating the risk of BCLM.

The TME comprises both cellular elements, including adipocytes, immune cells [43, 44], endothelial cells and cancer-associated fibroblasts, as well as non-cellular

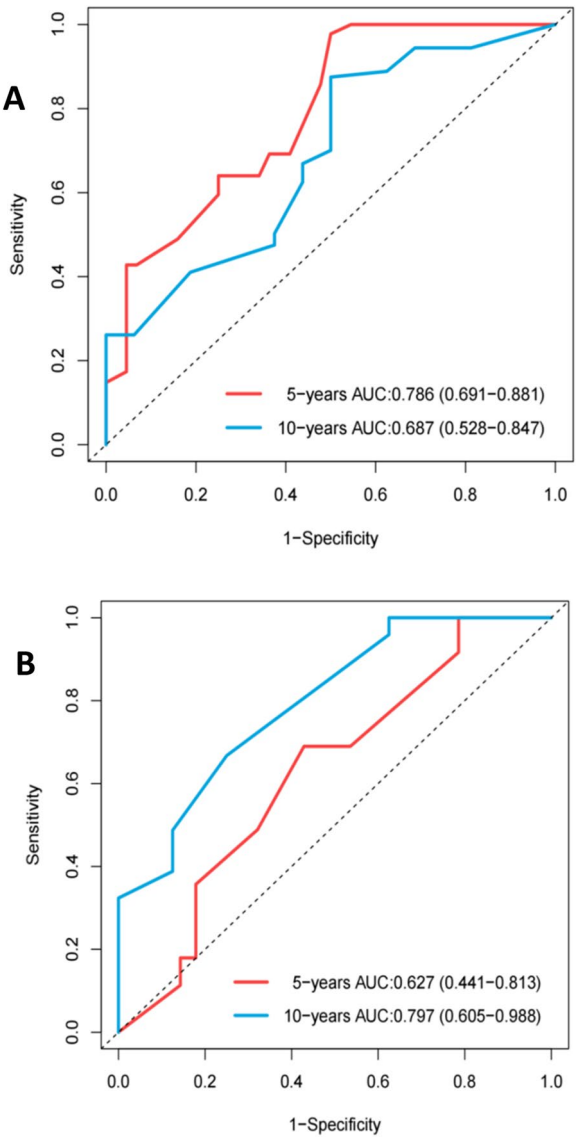


**Fig. 4** Validation of the ability of nomogram to predict the risk of lung metastasis in breast cancer patients within 5 years and 10 years. (A) ROC curve in the training cohort; (**B**) ROC curve in the validation cohort

components [45–48], such as cytokines and the ECM. It promotes tumor progression and invasion through the secretion of growth factors and pro-inflammatory mediators as well as through intercellular interactions and metabolic crosstalk with tumor cells [49, 50]. In this study, cytokines were innovatively incorporated into a lung metastasis model, and three key inflammatory factors (IL6, IFN-α and TNF-α) associated with lung metastasis were then identified using RF, LASSO and XGBoost ML algorithms. These cytokines have been extensively studied in the context of BC metastasis mechanisms. For instance, early research has demonstrated that IL- 6–174 promoter polymorphism was linked to clinical outcomes in a group of lymph node-positive BC patients undergoing high-dose adjuvant therapy [51]. Additionally, Adam et al. reported that fibroblasts isolated from common sites of breast cancer metastasis enhanced the growth and invasiveness of cancer cells in an IL- 6-dependent manner [52]. Similarly, Laura et al. found that p53 inactivation triggered a methylation-dependent autocrine IL- 6 loop that led to epigenetic reprogramming and the development of basal/stem cell-like gene expression profiles in BC cells [53]. HER2 overexpression has also been shown to induce IL- 6 secretion, activate STAT3, alter gene expression and reinforce the autocrine IL- 6/STAT3 loop [54]. In one study, Luca et al. reported that the combination of VEGF and IL- 6 synergistically and durably activated intracellular signaling pathways, such as MAPK, AKT and p38MAPK, in BC cells [55], while Rasmus et al. demonstrated that, in ER + breast cancer, the IL6/STAT3 signaling pathway could drive metastasis independently of the estrogen receptor. Although STAT3 and ER share enhancers, the former can hijack a subset of ER enhancers to induce unique transcriptional programs. This decoupling of ER and IL6/STAT3 oncogenic pathways underscores the therapeutic potential of targeting IL6/STAT3 in ER + breast cancer [56]. In contrast to IL- 6, IFN-α and TNF-α inhibit breast cancer growth and invasion through distinct mechanisms. Specifically, IFN-α is involved in tumor immune surveillance by activating CD8 α+ dendritic cells (DCs) and enhancing CD8 + T cell recognition of tumor antigens [57]. Thus, a deficiency in IFN-α can disrupt this process, leading to the expansion of regulatory T cells (Tregs) which suppresses plasma cell-like DCs and facilitate BC metastasis [58]. On the other hand, TNF-α can restrict the migration of triple negative, mesenchymal-like BC cells with high TNFR1 expression, while inhibiting the migration of epithelioid cells with low TNFR1 expression [59].

## Conclusions

This study applied three ML methods to systematically analyze clinical information and surgical pathology results, integrated treatment exposure and inflammatory markers, and established a predictive model for BCLM. This model exhibits strong discriminative ability in both training and validation queues. In fact, through this nomogram, doctors can estimate the likelihood of lung metastasis in BC patients based on the cumulative score of each risk factor. Therefore, this tool can achieve personalized risk assessment by regularly reviewing inflammation

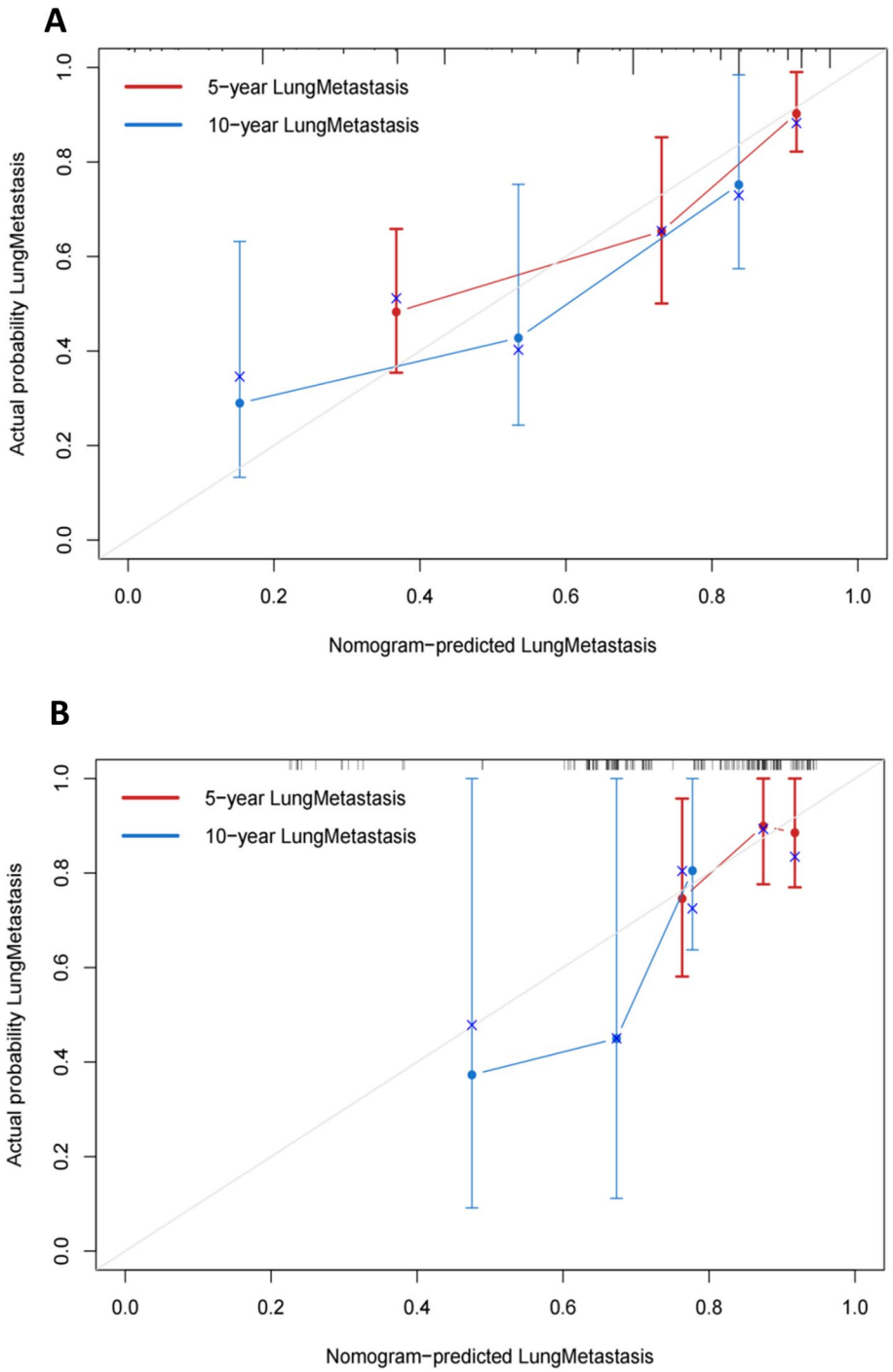


**Fig. 5** Calibration curves in the training set (**A**) and validation set (**B**). The x-axis represents the predicted probability of the nomogram plot, and the y-axis represents the actual probability of lung metastasis in breast cancer patients

indicators for high-risk patients, and immediately initiating imaging screening for patients with improved scores. In addition, the findings highlighted the contrasting roles of cytokines in BC, with IL- 6 promoting BCLM, while IFN-α and TNF-α inhibited tumor metastasis. These insights deepen current understanding of the interplay

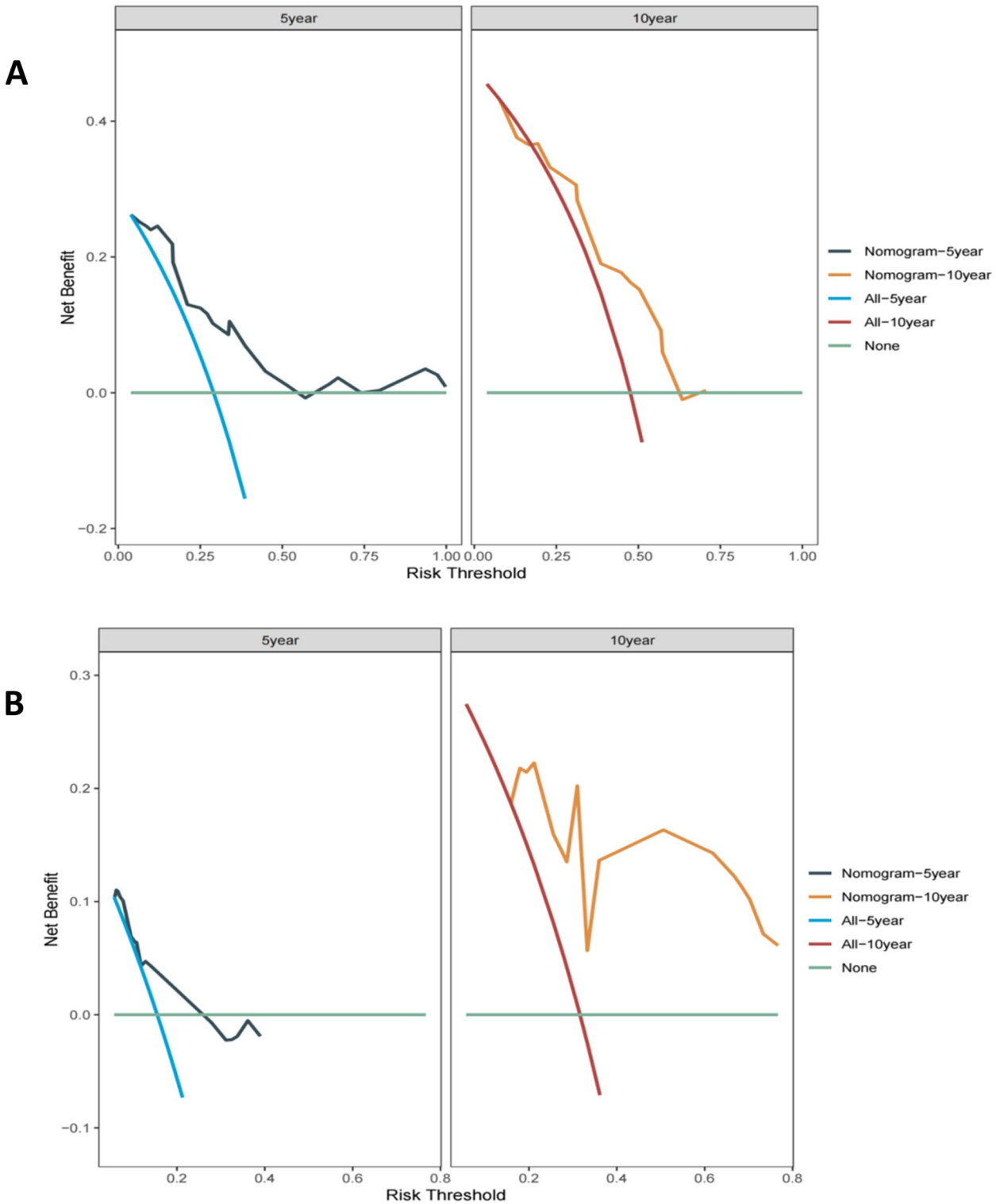


**Fig. 6** Decision curve analysis for the training set (**A**) and validation set (**B**). The horizontal line indicates that all samples are negative and untreated, with zero net benefit. A slash indicates that all samples are positive. Net income has a negative slope

between cytokines and BCLM, thus underscoring the importance of detecting and managing inflammation associated with BC. Future works should validate the current findings through large, prospective, multi-center trials.

**Abbreviations**

| | |
|---|---|
| BMI | Body Mass Index |
| SF | Serum Ferritin |
| hs-CRP | Hypersensitive C-reactive protein |
| Hb | Haemoglobin |
| PLT | Platelet |
| ER | Estrogen receptors |
| PR | Progesterone receptors |
| Her2 | Human epidermal growth factor receptor-2 |

## Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12885-025-14101-3.

Supplementary Material 1.

**Acknowledgements**
Not applicable.

**Authors' contributions**
ZL conceived of the study. ZL and HM designed the study and collected data. WB contributed to data analysis. ZL, HM, LZ contributed to the writing and revision of the paper.All authors read and approved the fnal manuscript.

**Funding**
This work was supported by the Project of Jiangsu Health Commission(M2020072).

**Data availability**
The datasets generated and analysed during the current study are not publicly available due to privacy or ethical restrictions but are available from the corresponding author on reasonable request.

## Declarations

**Ethics approval and consent to participate**
This study was approved by the Human Research Ethics Committee, the Second Affiliated Hospital of Xuzhou Medical University (120601). The Ethics Committee of the Second Affiliated Hospital of Xuzhou Medical University agreed to waive informed consent because this retrospective observational study uses medical recordsfrom previous clinical diagnosis, the risk of which was not greater than the minimum risk.And the research process was in accordance with the content of the Declaration of Helsinki.

**Consent for publication**
Not applicable.

**Competing Interests**
The authors declare no competing interests.

**Author details**
[1]Department of Radiotherapy, The Second Affiliated Hospital of Xuzhou Medical University, Meijian Road 32, Xuzhou 221000, China. [2]Xuzhou Medical University, Tongshan Road 209, Xuzhou 221000, China. [3]Tongji University, Yangpu District, Siping Road 1239, Shanghai 310000, China.

Received: 16 January 2025 Accepted: 7 April 2025
Published online: 14 April 2025

**References**

1. Li Q, Xia C, Li H, Yan X, Yang F, Cao M, Zhang S, Teng Y, He S, Cao M, Chen W. Disparities in 36 cancers across 185 countries: secondary analysis of global. Front Med. 2024;18(5):911–20. https://doi.org/10.1007/s11684-024-1058-6.
2. Park M, Kim D, Ko S, Kim A, Mo K, Yoon H: Breast Cancer Metastasis: Mechanisms and Therapeutic Implications. Int J Mol Sci 2022, 23.
3. Weigelt B, Peterse JL. van 't Veer LJ: Breast cancer metastasis: markers and models. Nat Rev Cancer. 2005;5:591–602.
4. Wang C, Xu K, Wang R, Han X, Tang J, Guan X. Heterogeneity of BCSCs contributes to the metastatic organotropism of breast cancer. J Exp Clin Cancer Res. 2021;40:370.
5. Pillar N, Polsky AL, Weissglas-Volkov D, Shomron N. Comparison of breast cancer metastasis models reveals a possible mechanism of tumor aggressiveness. Cell Death Dis. 2018;9:1040.
6. Liang Y, Zhang H, Song X, Yang Q. Metastatic heterogeneity of breast cancer: Molecular mechanism and potential therapeutic targets. Semin Cancer Biol. 2020;60:14–27.
7. Lin S, Mo H, Li Y, Guan X, Chen Y, Wang Z, Xu B. Clinicopathological characteristics and survival outcomes in patients with synchronous lung metastases upon initial metastatic breast cancer diagnosis in Han population. BMC Cancer. 2021;21:1330.
8. Jin W, Yang Q, Chi H, Wei K, Zhang P, Zhao G, Chen S, Xia Z, Li X. Ensemble deep learning enhanced with self-attention for predicting immunotherapeutic responses to cancers. Front Immunol. 2022;13:1025330.
9. Ren Q, Zhang P, Lin H, Feng Y, Chi H, Zhang X, Xia Z, Cai H, Yu Y. A novel signature predicts prognosis and immunotherapy in lung adenocarcinoma based on cancer-associated fibroblasts. Front Immunol. 2023;14:1201573.
10. Zhang S, Jiang C, Jiang L, Chen H, Huang J, Gao X, Xia Z, Tran LJ, Zhang J, Chi H, et al. Construction of a diagnostic model for hepatitis B-related hepatocellular carcinoma using machine learning and artificial neural networks and revealing the correlation by immunoassay. Tumour Virus Res. 2023;16: 200271.
11. Zhao S, Zhang X, Gao F, Chi H, Zhang J, Xia Z, Cheng C, Liu J. Identification of copper metabolism-related subtypes and establishment of the prognostic model in ovarian cancer. Front Endocrinol (Lausanne). 2023;14:1145797.
12. Chi H, Yang J, Peng G, Zhang J, Song G, Xie X, Xia Z, Liu J, Tian G. Circadian rhythm-related genes index: A predictor for HNSCC prognosis, immunotherapy efficacy, and chemosensitivity. Front Immunol. 2023;14:1091218.
13. Zhao S, Chi H, Yang Q, Chen S, Wu C, Lai G, Xu K, Su K, Luo H, Peng G, et al. Identification and validation of neurotrophic factor-related gene signatures in glioblastoma and Parkinson's disease. Front Immunol. 2023;14:1090040.
14. Chi H, Jiang P, Xu K, Zhao Y, Song B, Peng G, He B, Liu X, Xia Z, Tian G. A novel anoikis-related gene signature predicts prognosis in patients with head and neck squamous cell carcinoma and reveals immune infiltration. Front Genet. 2022;13: 984273.
15. Kang J, Choi YJ, Kim IK, Lee HS, Kim H, Baik SH, Kim NK, Lee KY. LASSO-Based Machine Learning Algorithm for Prediction of Lymph Node Metastasis in T1 Colorectal Cancer. Cancer Res Treat. 2021;53:773–83.
16. Chi H, Xie X, Yan Y, Peng G, Strohmer DF, Lai G, Zhao S, Xia Z, Tian G. Natural killer cell-related prognosis signature characterizes immune landscape and predicts prognosis of HNSCC. Front Immunol. 2022;13:1018685.
17. Li J, Shi Z, Liu F, Fang X, Cao K, Meng Y, Zhang H, Yu J, Feng X, Li Q, et al. XGBoost Classifier Based on Computed Tomography Radiomics for Prediction of Tumor-Infiltrating CD8(+) T-Cells in Patients With Pancreatic Ductal Adenocarcinoma. Front Oncol. 2021;11: 671333.
18. Zhai X, Zhang H, Xia Z, Liu M, Du G, Jiang Z, Zhou H, Luo D, Dou D, Li J, et al. Oxytocin alleviates liver fibrosis via hepatic macrophages. JHEP Rep. 2024;6: 101032.
19. Xiao J, Lin H, Liu B, Xia Z, Zhang J, Jin J. Decreased S1P and SPHK2 are involved in pancreatic acinar cell injury. Biomark Med. 2019;13:627–37.
20. Xiao J, Huang K, Lin H, Xia Z, Zhang J, Li D, Jin J. Mogroside II(E) Inhibits Digestive Enzymes via Suppression of Interleukin 9/Interleukin 9 Receptor Signalling in Acute Pancreatitis. Front Pharmacol. 2020;11:859.
21. Zhang H, Xia T, Xia Z, Zhou H, Li Z, Wang W, Zhai X, Jin B. KIF18A inactivates hepatic stellate cells and alleviates liver fibrosis through the TTC3/Akt/mTOR pathway. Cell Mol Life Sci. 2024;81:96.

22. Habanjar O, Bingula R, Decombat C, Diab-Assaf M, Caldefie-Chezet F, Delort L: Crosstalk of Inflammatory Cytokines within the Breast Tumor Microenvironment. Int J Mol Sci 2023, 24.
23. Zhou H, Cai Z, Yang Q, Yang X, Chen J, Huang T. Inflammatory cytokines and two subtypes of breast cancer: A two-sample mendelian randomization study. PLoS ONE. 2023;18: e0293230.
24. Zhang X, Zhuge J, Liu J, Xia Z, Wang H, Gao Q, Jiang H, Qu Y, Fan L, Ma J, et al. Prognostic signatures of sphingolipids: Understanding the immune landscape and predictive role in immunotherapy response and outcomes of hepatocellular carcinoma. Front Immunol. 2023;14:1153423.
25. Wang X, Zhao Y, Strohmer DF, Yang W, Xia Z, Yu C. The prognostic value of MicroRNAs associated with fatty acid metabolism in head and neck squamous cell carcinoma. Front Genet. 2022;13: 983672.
26. Yu Y, He Z, Ouyang J, Tan Y, Chen Y, Gu Y, Mao L, Ren W, Wang J, Lin L, et al. Magnetic resonance imaging radiomics predicts preoperative axillary lymph node metastasis to support surgical decisions and is associated with tumor microenvironment in invasive breast cancer: A machine learning, multicenter study. EBioMedicine. 2021;69: 103460.
27. Zhang H, Lin F, Zheng T, Gao J, Wang Z, Zhang K, Zhang X, Xu C, Zhao F, Xie H, et al. Artificial intelligence-based classification of breast lesion from contrast enhanced mammography: a multicenter study. Int J Surg. 2024;110:2593–603.
28. Liu J, Zhang P, Yang F, Jiang K, Sun S, Xia Z, Yao G, Tang J. Integrating single-cell analysis and machine learning to create glycosylation-based gene signature for prognostic prediction of uveal melanoma. Front Endocrinol (Lausanne). 2023;14:1163046.
29. Chi H, Gao X, Xia Z, Yu W, Yin X, Pan Y, Peng G, Mao X, Teichmann AT, Zhang J, et al. FAM family gene prediction model reveals heterogeneity, stemness and immune microenvironment of UCEC. Front Mol Biosci. 2023;10:1200335.
30. Zhu M, Ma Z, Zhang X, Hang D, Yin R, Feng J, Xu L, Shen H. C-reactive protein and cancer risk: a pan-cancer study of prospective cohort and Mendelian randomization analysis. BMC Med. 2022;20:301.
31. Michels N, van Aart C, Morisse J, Mullee A, Huybrechts I. Chronic inflammation towards cancer incidence: A systematic review and meta-analysis of epidemiological studies. Crit Rev Oncol Hematol. 2021;157: 103177.
32. Siemes C, Visser LE, Coebergh JW, Splinter TA, Witteman JC, Uitterlinden AG, Hofman A, Pols HA, Stricker BH. C-reactive protein levels, variation in the C-reactive protein gene, and cancer risk: the Rotterdam Study. J Clin Oncol. 2006;24:5216–22.
33. Han Y, Mao F, Wu Y, Fu X, Zhu X, Zhou S, Zhang W, Sun Q, Zhao Y. Prognostic role of C-reactive protein in breast cancer: a systematic review and meta-analysis. Int J Biol Markers. 2011;26:209–15.
34. Mikkelsen MK, Lindblom NAF, Dyhl-Polk A, Juhl CB, Johansen JS, Nielsen D. Systematic review and meta-analysis of C-reactive protein as a biomarker in breast cancer. Crit Rev Clin Lab Sci. 2022;59:480–500.
35. Kim ES, Kim SY, Koh M, Lee HM, Kim K, Jung J, Kim HS, Moon WK, Hwang S, Moon A. C-reactive protein binds to integrin α2 and Fcγ receptor I, leading to breast cell adhesion and breast cancer progression. Oncogene. 2018;37:28–38.
36. Feng JR, Li X, Han C, Chang Y, Fu Y, Feng GC, Lei Y, Li HY, Tang PM, Ji SR, et al. C-Reactive Protein Induces Immunosuppression by Activating FcγR2B in Pulmonary Macrophages to Promote Lung Metastasis. Cancer Res. 2024;84:4184–98.
37. Jaiyesimi IA, Buzdar AU, Decker DA, Hortobagyi GN. Use of tamoxifen for breast cancer: twenty-eight years later. J Clin Oncol. 1995;13:513–29.
38. Legha SS. Tamoxifen in the treatment of breast cancer. Ann Intern Med. 1988;109:219–28.
39. Ruhstaller T, Giobbie-Hurder A, Colleoni M, Jensen MB, Ejlertsen B, de Azambuja E, Neven P, Láng I, Jakobsen EH, Gladieff L, et al. Adjuvant Letrozole and Tamoxifen Alone or Sequentially for Postmenopausal Women With Hormone Receptor-Positive Breast Cancer: Long-Term Follow-Up of the BIG 1–98 Trial. J Clin Oncol. 2019;37:105–14.
40. Francis PA, Pagani O, Fleming GF, Walley BA, Colleoni M, Láng I, Gómez HL, Tondini C, Ciruelos E, Burstein HJ, et al. Tailoring Adjuvant Endocrine Therapy for Premenopausal Breast Cancer. N Engl J Med. 2018;379:122–37.
41. Hanker AB, Sudhan DR, Arteaga CL. Overcoming Endocrine Resistance in Breast Cancer. Cancer Cell. 2020;37:496–513.
42. Turner NC, Neven P, Loibl S, Andre F. Advances in the treatment of advanced oestrogen-receptor-positive breast cancer. Lancet. 2017;389:2403–14.
43. Xie H, Xi X, Lei T, Liu H, Xia Z. CD8(+) T cell exhaustion in the tumor microenvironment of breast cancer. Front Immunol. 2024;15:1507283.
44. Deng Y, Shi M, Yi L, Naveed Khan M, Xia Z, Li X. Eliminating a barrier: Aiming at VISTA, reversing MDSC-mediated T cell suppression in the tumor microenvironment. Heliyon. 2024;10: e37060.
45. Zhang X, Zhang P, Cong A, Feng Y, Chi H, Xia Z, Tang H. Unraveling molecular networks in thymic epithelial tumors: deciphering the unique signatures. Front Immunol. 2023;14:1264325.
46. Xiong J, Chi H, Yang G, Zhao S, Zhang J, Tran LJ, Xia Z, Yang F, Tian G. Revolutionizing anti-tumor therapy: unleashing the potential of B cell-derived exosomes. Front Immunol. 2023;14:1188760.
47. Zhao Y, Wei K, Chi H, Xia Z, Li X. IL-7: A promising adjuvant ensuring effective T cell responses and memory in combination with cancer vaccines? Front Immunol. 2022;13:1022808.
48. Gong X, Chi H, Strohmer DF, Teichmann AT, Xia Z, Wang Q. Exosomes: A potential tool for immunotherapy of ovarian cancer. Front Immunol. 2022;13:1089410.
49. Xia Z, Chen S, He M, Li B, Deng Y, Yi L, Li X. Editorial: Targeting metabolism to activate T cells and enhance the efficacy of checkpoint blockade immunotherapy in solid tumors. Front Immunol. 2023;14:1247178.
50. Zhang P, Pei S, Wu L, Xia Z, Wang Q, Huang X, Li Z, Xie J, Du M, Lin H. Integrating multiple machine learning methods to construct glutamine metabolism-related signatures in lung adenocarcinoma. Front Endocrinol (Lausanne). 2023;14:1196372.
51. DeMichele A, Martin AM, Mick R, Gor P, Wray L, Klein-Cabral M, Athanasiadis G, Colligan T, Stadtmauer E, Weber B. Interleukin-6 -174G–>C polymorphism is associated with improved outcome in high-risk breast cancer. Cancer Res. 2003;63:8051–6.
52. Studebaker AW, Storci G, Werbeck JL, Sansone P, Sasser AK, Tavolari S, Huang T, Chan MW, Marini FC, Rosol TJ, et al. Fibroblasts isolated from common sites of breast cancer metastasis enhance cancer cell growth rates and invasiveness in an interleukin-6-dependent manner. Cancer Res. 2008;68:9087–95.
53. D'Anello L, Sansone P, Storci G, Mitrugno V, D'Uva G, Chieco P, Bonafé M. Epigenetic control of the basal-like gene expression profile via Interleukin-6 in breast cancer cells. Mol Cancer. 2010;9:300.
54. Hartman ZC, Yang XY, Glass O, Lei G, Osada T, Dave SS, Morse MA, Clay TM, Lyerly HK. HER2 overexpression elicits a proinflammatory IL-6 autocrine signaling loop that is critical for tumorigenesis. Cancer Res. 2011;71:4380–91.
55. De Luca A, Lamura L, Gallo M, Maffia V, Normanno N. Mesenchymal stem cell-derived interleukin-6 and vascular endothelial growth factor promote breast cancer cell migration. J Cell Biochem. 2012;113:3363–70.
56. Siersbæk R, Scabia V, Nagarajan S, Chernukhin I, Papachristou EK, Broome R, Johnston SJ, Joosten SEP, Green AR, Kumar S, et al. IL6/STAT3 Signaling Hijacks Estrogen Receptor α Enhancers to Drive Breast Cancer Metastasis. Cancer Cell. 2020;38:412–423.e419.
57. Zhang J, Peng G, Chi H, Yang J, Xie X, Song G, Tran LJ, Xia Z, Tian G. CD8 + T-cell marker genes reveal different immune subtypes of oral lichen planus by integrating single-cell RNA-seq and bulk RNA-sequencing. BMC Oral Health. 2023;23:464.
58. Sisirak V, Vey N, Goutagny N, Renaudineau S, Malfroy M, Thys S, Treilleux I, Labidi-Galy SI, Bachelot T, Dezutter-Dambuyant C, et al. Breast cancer-derived transforming growth factor-β and tumor necrosis factor-α compromise interferon-α production by tumor-associated plasmacytoid dendritic cells. Int J Cancer. 2013;133:771–8.
59. Cruceriu D, Balacescu L, Baldasici O, Gaal OI, Balacescu O, Russom A, Irimia D, Tudoran O. Gene expression-phenotype association study reveals the dual role of TNF-α/TNFR1 signaling axis in confined breast cancer cell migration. Life Sci. 2024;354: 122982.


## Publisher's Note