


Prediction of Distant Metastasis of Lymph-Node-Negative Primary Breast Cancer From Gene Expression Profiling Using Cox-Boost Regression Model

Cancer Informatics
Volume 23: 1–13
© The Author(s) 2024
DOI: 10.1177/11769351241297493


Omid Hamidi¹, Payam Amini², Leili Tapak^{3,4},
Yasaman Zohrab Beigi⁵, Saeid Afshar^{6,7} and Irina Dinu⁸

¹Department of Science, Hamedan University of Technology, Hamedan, Iran. ²School of Medicine, Keele University, Keele, Staffordshire, UK. ³Department of Biostatistics, School of Public Health, Hamadan University of Medical Sciences, Hamadan, Iran. ⁴Modeling of Noncommunicable Diseases Research Center, Hamadan University of Medical Sciences, Hamadan, Iran. ⁵Department of Biology, Yazd University, Yazd, Iran. ⁶Cancer Research Center, Institute of cancer, Avicenna Health Research Institute, Hamadan University of Medical Sciences, Hamadan, Iran. ⁷Department of Medical Biotechnology, School of Advanced Medical Sciences and Technologies, Hamadan University of Medical Sciences, Hamadan, Iran. ⁸School of Public Health, Health Academy, University of Alberta, Edmonton, AB, Canada.

ABSTRACT

BACKGROUNDS: Distant metastasis in breast cancer patients contributes to increased breast cancer mortality, highlighting the urgent need for effective predictive strategies. Understanding metastasis mechanisms and identifying relevant biomarkers are crucial for improving patient outcomes and informing targeted therapies. This study employed a high-dimensional regression model to identify biomarkers linked to distant metastasis-free survival in breast cancer patients, with the goal of enhancing prognostic accuracy and guiding clinical decisions.

METHODS: We utilized the publicly available breast cancer dataset (GSE2034), which includes gene expression profiles for 22,283 genes across 286 samples. To identify relevant genes, we applied Cox-Boost regression and a random forest (RF) model. We then explored the association between the selected genes and metastasis-free survival outcomes using quantile regression, chosen for its ability to assess the impact of these genes across different survival quantiles ($P < .05$). This approach complements the Cox-Boost model by providing a more detailed understanding of gene-survival relationships at various points in the survival distribution, thereby strengthening the robustness of our findings.

RESULTS: We identified 222 significant transcripts using univariate Cox regression models. By applying Cox-Boost, both with and without adjustment for ER+/- status, we identified 7 genes associated with time-to-relapse/metastasis in breast cancer patients: SNU13, CLINT1, ACBD3, NEK2, COL2A1, WFDC1, and RACGAP1. A similar approach was used for ER-positive patients. Patients were classified as high or low risk for metastasis based on the median prognostic index calculated from the identified genes ($P < .001$). The top-ranked genes associated with high/low risk groups using RF were RACGAP1, NEK2, CCNA2, DTL, ACBD3, ARL6IP5, WFDC1, and PDCD4.

CONCLUSIONS: We identified eleven key genes, including SNU13, CLINT1, ACBD3, NEK2, COL2A1, WFDC1, and RACGAP1, as well as CCNA2, DTL, ARL6IP5, and PDCD4, that are related to the risk of distant metastasis and may be used as biomarkers to predict distant metastasis of breast cancer.

KEYWORDS: Gene expression, Bioinformatics, biomarker, prognosis, machine learning

RECEIVED: July 3, 2024. **ACCEPTED:** October 7, 2024.

TYPE: Original Research

FUNDING: The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This study was partially supported by Hamadan University of Medical Sciences (Grant NO. 140308016662).

DECLARATION OF CONFLICTING INTERESTS: The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

CORRESPONDING AUTHORS: Payam Amini, School of medicine, Keele University, Stoke On Trent, Keele, Staffordshire ST5 5BG, UK. Email: p.amini@keele.ac.uk.

Leili Tapak, Department of Biostatistics, School of Public Health, and Modeling of Noncommunicable Diseases Research Center, Hamadan University of Medical Sciences, Hamadan, 6517838636, Iran. Email: ltapak@yahoo.com

Introduction

Breast cancer (BC) is the most prevalent cancer among women, representing 10.4% of all cancer cases.¹ Its incidence and mortality rates are approximately 30% and 15%, respectively, among females worldwide.² Risk factors include lifestyle choices (such as smoking and diet), genetics, and certain health conditions. Breast cancer often develops due to genetic mutations or DNA damage, which can be associated with inherited genetic defects.³ Although many causes remain unknown, the risk of developing breast cancer increases with age, a family history of the disease, a previous diagnosis of breast cancer or benign

lumps, dense breast tissue, exposure to estrogen, hormone replacement therapy, obesity, alcohol consumption, and radiation.⁴ Early detection is crucial for effective treatment, which typically includes surgery, radiation therapy, hormone therapy, biological treatments, and chemotherapy. However, despite treatment, some cases have poor prognoses due to unknown molecular factors, and local, regional, or distant recurrence can occur months or even years later as a result of residual cancer cells.⁵ In 2020, approximately 2.3 million women worldwide were diagnosed with breast cancer, and by the end of the year, 7.8 million women were living as breast cancer survivors within



Creative Commons CC BY: This article is distributed under the terms of the Creative Commons Attribution 4.0 License (<https://creativecommons.org/licenses/by/4.0/>) which permits any use, reproduction and distribution of the work without further permission provided the original work is attributed as specified on the SAGE and Open Access pages (<https://us.sagepub.com/en-us/nam/open-access-at-sage>).

the past 5 years.⁶ Also, a 40% increase from the estimated 1.7 million cases in 2012.⁷ The 5-year survival rates for breast cancer after diagnosis vary widely: 85% to 90% in high-income countries such as the United States and the United Kingdom, 66% in India, and 40% in South Africa.⁸ Poorer survival rates in developing countries can be attributed to factors such as limited awareness, lack of early detection programs, and insufficient diagnosis and treatment facilities.⁹ Furthermore, a longer interval between diagnosis and BC surgery is associated with decreased overall and disease-specific survival.¹⁰

Altered genes, whether inherited or acquired, are significant risk factors for both breast cancer incidence and survival. In breast cancer patients, various cellular activities and signaling pathways are involved, underscoring the importance of understanding the underlying molecular mechanisms. This knowledge could enable the identification of biomarkers that can predict clinical outcomes and treatment responses.¹¹ Breast cancer is a heterogeneous disease with diverse molecular subtypes and varying clinical behaviors. Understanding the molecular mechanisms that drive breast cancer progression is essential for developing more effective diagnostic, prognostic, and therapeutic strategies. By unraveling the complex signaling pathways and genetic alterations that regulate tumor initiation, growth, and metastasis, researchers can identify novel molecular targets for personalized treatment approaches.^{12,13} The most well-known genetic risk factors for BC are mutated BRCA1 and BRCA2 which can result in abnormal cell growth and cancer.¹⁴ Other genes, such as ATM, PALB2, TP53, CHEK2, PTEN, and CDH1 influence BC risk, albeit to a lesser extent.¹¹

Numerous studies have explored prognostic markers, including gene expression signatures that differentiate between tumor and normal tissues, focusing on survival outcomes.¹⁵⁻¹⁷ However, progress in developing diagnostic tools to detect recurrence in breast cancer patients has been limited. Such tools are essential for ensuring that high-risk patients receive appropriate therapy. Although gene expression profiles have been used to classify different breast tumor subtypes, the influence of genetic alterations on breast cancer progression and survival is still not fully understood. Therefore, it is crucial to explore the association between a broad set of gene expression variables and breast cancer incidence and survival, particularly in relation to disease recurrence.

Boosting techniques were originally developed as powerful tools for classification, aimed at enhancing the predictive performance of weak learners by combining them into a single strong learner. This iterative ensemble method works by sequentially applying a learning algorithm, re-weighting data points to focus on those misclassified in earlier iterations. Over time, boosting has expanded beyond classification to address various statistical challenges, including survival analysis. One notable adaptation is the Cox-Boost method, which combines the principles of boosting with the Cox proportional hazards model. This likelihood-based approach is particularly effective for analyzing high-dimensional data, as it facilitates variable selection and

promotes sparsity through its iterative process. By identifying and prioritizing the most relevant predictors, Cox-Boost enables accurate prediction of patient survival probabilities while assessing the influence of multiple covariates.¹⁸ This method efficiently handles high-dimensional data analysis by promoting sparsity and variable selection through its iterative process, allowing for accurate prediction of patient survival probabilities and evaluating the impact of multiple predictors.¹⁹⁻²²

To address the need for assessing the impact of a large number of genes on disease recurrence in breast cancer patients, this study aims to utilize the Cox-Boost approach to identify genes highly associated with this cancer. A comprehensive analysis of gene expression was performed to establish a reliable set of prognostic markers and offer quantitative predictions on recurrence for patients with lymph-node-negative breast cancer.

Methods

Data

A publicly available dataset of BC is available in GEO repository with the ID: GSE2034, which was generated using the Affymetrix Human Genome U133A Array, was used. This series represents 180 lymph-node negative relapse free patients and 106 lymph-node negative patients that developed a distant metastasis.²³ All cases had sufficient tumor and uniform involvement of tumor in 5 μ m frozen sections stained with hematoxylin and eosin. Immunohistochemistry or ligand binding assay were utilized to measure estrogen receptor (ER) and a cut-off value of 10 fmol/mg or 10% positive tumor cells was assumed to classify cases in positive or negative ER.²³ In this study, Cox-Boost analysis was performed to assess and determine the most influencing genes on the BC relapse free survival.

Statistical analysis

We used Bioconductor packages in the R language for data analysis. The raw data was processed using Log2 and normalization. Figure 1 presents the flowchart of the modeling workflow. To evaluate the unadjusted association of transcripts with survival probability, simple Cox proportional hazards regression models were applied. Multiple versions of Cox regression were conducted using component-wise likelihood-based boosting, a method particularly well-suited for handling numerous predictors. This approach also allows for the inclusion of mandatory covariates with unpenalized parameter estimates. A Cox proportional hazards model is formulated as follows, where t is the n -dimensional vector of the observed survival times, X the $n \times p$ matrix of the data, β is the vector of coefficients, and δ an n -dimensional vector reporting whether the i th observed survival time is censored or not.

$$\lambda(t|X) = \lambda_0(t)e^{X\beta}$$

For data filtering, separate Cox regression models were applied using each gene expression as the only covariate and those with

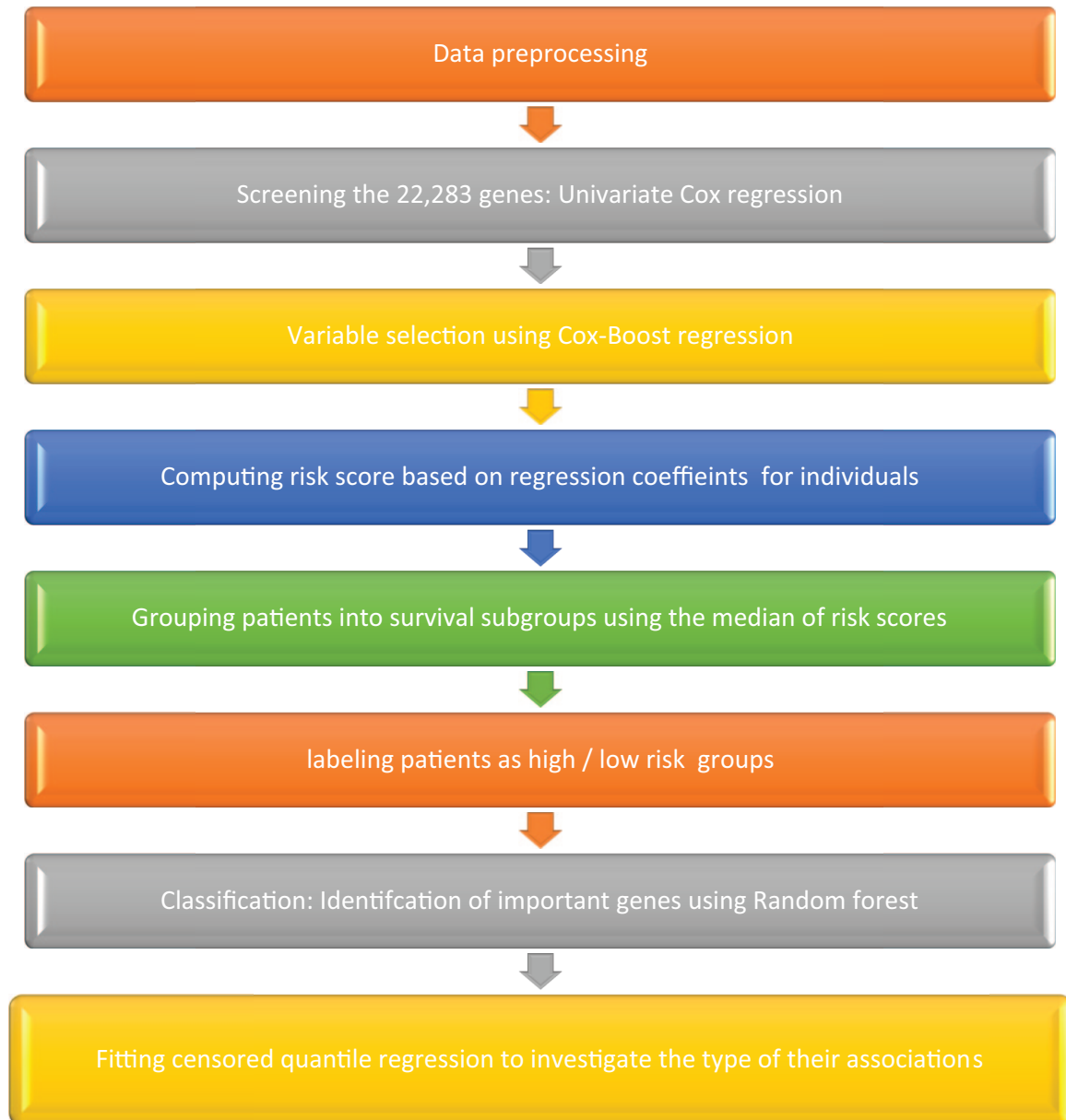


Figure 1. The flowchart of the modeling workflow.

P-value less than .05 were evaluated in the Cox Boost step of data analysis. In boosting procedure, one follows these steps:

1. Set the vector of regression coefficients as zero.
2. Compute the negative gradient vector in which $L(y, F(X, \beta))$ is a loss function: $u = \frac{\delta L(y, F(X, \beta))}{\delta F(X, \beta)}$
3. Computing the updates:
 - 3.1 fitting the base learner to the negative gradient vector, $h(u, X_j)$
 - 3.2 penalizing it, $\hat{b}_j = v \hat{h}(u, X_j)$
4. Select the best update j^* (usually by minimizing the loss function).
5. Updating the estimations $\hat{\beta}_j = \hat{\beta}_{j^*} + \hat{b}_{j^*}$ where \hat{b} is called a weak estimator.

The above steps between 2 and 5 are repeated multiple times. The approach for estimation is based on a likelihood function. The loss function, L_2 norm penalized partial log-likelihood is $pl_{pen}(\beta) = pl(\beta) - 0.5\lambda\beta^T P \beta$, where P is a $p \times p$ matrix usually corresponding to the identity matrix and λ is the penalty term. In each iteration of the component-wise boosting, the restricted partial log-likelihoods are “shifted” toward $\hat{\beta}_j$, obtaining the restricted penalized partial log-likelihoods. More details about this approach is well described by Riccardo De Bin.²⁰ The number of boosting steps and shrinkage parameter (ν) are hyper-parameters that should be tuned to prevent overfitting problem and to find the best performance of the model. In this study, the number of steps was considered as 500 which is assumed as a sufficient number for convergence in high-dimensional survival data.²⁴ The shrinkage parameter

was obtained as 0.1, which is a commonly used value that ensures a gradual update of the coefficient estimates, preventing overfitting, and improving the model's generalization ability.

Using the estimated coefficients from the Cox Boost approach, prognostic scores were calculated for each case. The median score of prognostic was used as a cut-off point to distinguish the high/low risk individuals. To assess the importance of each gene on the categorized prognostic scores, Random Forest approach was utilized.

Random Forest employs an algorithm for regression, classification, and building trees.²⁵ For each tree, a bootstrap sample from the training data is taken, and a random-forest tree for each group is grown until reaching the minimum number of nodes. The method randomly selects a subset of variables, chooses the best variable for splitting, and divides the node into 2 daughter nodes. Once the important variables are identified in the classification, partial dependence and marginal effects are used to assess the impact of different variable values on classification. This technique also utilizes the average increase in the Gini index to measure variable importance. Two indices are used to determine the significance of genes: the increase in MSE of predictions, assessed via a permutation-based approach, and the increase in node purity, which relates to the loss function. Genes with higher indices are considered more important. Hyperparameter tuning was conducted using cross-validation and grid search to optimize the number of trees and the number of variables considered at each split, ensuring the best predictive performance. To do so, a range of hyperparameter values will be defined, including different numbers of trees (eg, 100, 200, 500) and the number of variables to be selected at each split. The grid search method systematically tests all combinations of these hyperparameter values by fitting the model with each combination and evaluating its performance using cross-validation.

Moreover, we used censored quantile regression to find the overall survival of the patients in different quantiles of time.²⁶ This approach allows us to estimate the p^{th} quantile of survival time (Q_p) using the following model in which X 's are the covariates and factors, β_p is the coefficient for the p^{th} quantile.

$$Q(p | X) = X'\beta_p$$

For example, $Q_{.50} = a(\text{days})$ means that a randomly selected individual from the sample has a 0.5 probability of experiencing the event within "a" given number of days. Censored quantile regression was chosen because it offers a more detailed perspective on survival outcomes by examining different quantiles, which is especially valuable in clinical settings where survival times can vary significantly among patients. This method complements the Cox-Boost regression, which focuses on the overall hazard function, by allowing us to assess the influence of covariates at specific points in

the survival distribution. Additionally, the model employs a bootstrap resampling method to estimate the standard errors of the coefficients, thereby increasing the robustness of our estimates.

Software. To analyze the data, we used R software version 4.1.1 and "CoxBoost," "survival" and "quantreg" packages.

Enrichment analysis

For further validation the expression levels of the selected genes were evaluated through the Human Protein Atlas (<https://www.proteinatlas.org>). Moreover, in this study the GEPIA (<http://gepia.cancer-pku.cn/>) was used for analyzing the RNA sequencing expression data of breast cancer tissues and normal tissues from the TCGA.

Result

We analyzed gene expression data from 22283 genes across 286 samples. Initially, we conducted separate simple Cox regression models, which identified 222 significant transcripts with P -values less than .05. The results of the CoxBoost analysis are summarized in Table 1, detailing the hazard ratios (HRs), 95% confidence intervals, and corresponding P -values for each transcript after applying Cox regression.

This analysis was performed on the entire dataset, both with and without adjustments for estrogen receptor (ER) status, as well as within 2 distinct subsets: ER-positive and ER-negative samples. The CoxBoost model was run 100 times, allowing for variation in the selection of significant predictors across iterations. The "Frequency" column in Table 1 indicates the number of iterations in which a transcript was identified as a significant predictor. Only transcripts with a frequency greater than 80% are included in Table 1. A similar analysis was conducted for the ER-positive dataset; however, no significant transcripts were identified in the ER-negative dataset.

The results from both the adjusted and unadjusted analyses for estrogen receptor (ER) status indicate a slightly stronger association after adjustment in the Cox regression models. In the adjusted analysis, the expression of Small Nuclear Ribonucleoprotein 13 (SNU13) was associated with a lower hazard ratio (HR: 0.646; 95%CI: 0.533-0.784). In contrast, several other transcripts, including Clathrin Interactor 1 (CLINT1), Acyl-CoA Binding Domain Containing 3 (ACBD3), Never in Mitosis-Related Kinase 2 (NEK2), Collagen Type II Alpha-1 (COL2A1), WAP Four-Disulfide Core Domain 1 (WFDC1), and Rac GTPase Activating Protein 1 (RACGAP1), were associated with higher hazard ratios.

For the ER-positive subset, CD44 and SNU13 expression were associated with lower hazard ratios, while Actin-Like Protein 6A (ACTL6A), Never in Mitosis-Related Kinase 2 (NEK2), SHC Adaptor Protein 1 (SHC1), Zinc Finger CCHC-Type Containing 8 (ZCCHC8), and WAP

Table 1. The hazard ratio (95% confidence interval) resulted by the CoxBoost approach using adjusted and unadjusted for ER and in the sub-population of ER+ for 100 iterations.

PROBE-SET ID	GENE SYMBOL	COX BOOST COEFFICIENT	FREQUENCY	SIMPLE COX HR (95%CI)	MULTIPLE COX HR (95%CI)
<i>Unadjusted for ER</i>					
201076_at	SNU13	−0.041	100	0.647(0.533-0.784)	0.664(0.541-0.815)
201769_at	CLINT1	0.045	100	1.695(1.365-2.104)	1.685(1.316-2.157)
202324_s_at	ACBD3	0.040	100	1.557(1.295-1.872)	1.144(0.926-1.413)
204641_at	NEK2	0.043	100	1.621(1.337-1.967)	1.299(0.949-1.777)
217404_s_at	COL2A1	0.039	90	1.475(1.219-1.785)	1.473(1.214-1.788)
219478_at	WFDC1	0.131	100	1.661(1.362-2.024)	1.832(1.471-2.283)
222077_s_at	RACGAP1	0.047	100	1.622(1.337-1.967)	1.214(0.885-1.666)
<i>Adjusted for ER</i>					
201076_at	SNU13	−0.041	100	0.646(0.533-0.784)	0.663(0.541-0.815)
201769_at	CLINT1	0.045	100	1.696(1.365-2.106)	1.686(1.316-2.159)
202324_s_at	ACBD3	0.040	100	1.579(1.31-1.904)	1.135(0.911-1.413)
204641_at	NEK2	0.043	100	1.658(1.358-2.023)	1.306(0.951-1.794)
217404_s_at	COL2A1	0.039	91	1.481(1.224-1.793)	1.471(1.211-1.786)
219478_at	WFDC1	0.130	100	1.673(1.367-2.048)	1.839(1.473-2.295)
222077_s_at	RACGAP1	0.047	100	1.662(1.362-2.028)	1.216(0.885-1.672)
<i>ER+ subset</i>					
201076_at	SNU13	−0.046	100	0.599(0.479-0.751)	0.691(0.545-0.876)
202666_s_at	ACTL6A	0.051	100	1.828(1.443-2.316)	1.171(0.877-1.563)
204641_at	NEK2	0.099	100	1.875(1.472-2.388)	1.2507(1.135-2.001)
209835_x_at	CD44	−0.106	100	0.509(0.397-0.653)	0.686(0.525-0.898)
214853_s_at	SHC1	0.042	88	1.647(1.311-2.068)	1.303(1.013-1.676)
218478_s_at	ZCCHC8	0.047	100	1.808(1.417-2.308)	1.245(0.948-1.634)
219478_at	WFDC1	0.044	98	1.742(1.344-2.258)	1.625(1.248-2.117)

Abbreviations: ER, estrogen receptor; HR (95%CI), hazard ratio (95% confidence interval).

Four-Disulfide Core Domain 1 (WFDC1) were linked to higher hazard ratios. Multiple Cox regression analyses showed non-significant hazard ratios for Acyl-CoA Binding Domain Containing 3 (ACBD3), NEK2, and Rac GTPase Activating Protein 1 (RACGAP1) after adjusting for the effects of the identified genes, both with and without adjustment for ER status. Additionally, the multiple Cox regression analysis of the ER-positive subset indicated that ACTL6A and ZCCHC8 were not significantly associated with survival probabilities after adjusting for the influence of the other genes in the model.

To evaluate the influence of genes on various percentiles of survival probabilities, censored quantile regression was applied, with results presented in Tables 2 and 3 for the entire dataset

and the ER-positive subset, respectively. Table 2 indicates that the selected genes (identified through CoxBoost analysis) demonstrate near-significance across most survival probability deciles. The deciles of survival time can be calculated by multiplying each gene's expression by its estimated coefficient. For instance, the 10th percentile of survival time for a case with a normalized expression of SNU13 of −0.347 is calculated as $Q(0.10|SNU13) = 20.36 + 6.42 \times (-0.347) = 22.58$. This means that the probability of survival for a case with an SNU13 expression of −0.347 at 22.58 months is 10%. Similarly, other coefficients can be interpreted in the same manner. For example, the median survival time for a case with a normalized expression of CLINT1 equal to 1.03 is calculated as

Table 2. The results of censored quantile regression in the entire population.

QUANTILES	SNU13	CLINT1	ACBD3	NEK2	COL2A1	WFDC1	RACGAP1
10							
Intercept	20.36(1.95)	21.15(1.82)	16.37(0.77)	20.82(0.78)	17.28(0.89)	19.16(1.21)	20.93(0.99)
$\hat{\beta}$	6.42(1.33)*	-2.37(0.55)*	-8.81(0.75)*	-3.72(0.92)*	-6.64(1.24)*	-9.42(1.02)*	-10.38(2.39)*
20							
Intercept	38.72(1.99)	46.36(3.45)	37.12(4.17)	41.11(2.16)	35.39(2.54)	38.67(2.32)	39.52(2.45)
$\hat{\beta}$	16.30(1.59)*	-12.40(3.68)*	-19.31(2.28)*	-12.29(2.05)*	-16.32(2.42)*	-16.78(2.30)*	-29.69(4.11)*
30							
Intercept	76.74(14.66)	78.92(10.85)	118.53(42.34)	109.23(66.36)	80.73(11.73)	88.38(12.81)	127.38(11.85)
$\hat{\beta}$	40.77(17.33)*	-88.64(17.93)*	-67.12(41.64)	-41.80(10.89)*	-51.42(21.78)*	-64.80(27.26)*	-55.55(8.29)*
40							
Intercept	132.01(15.76)	177.32(15.37)	126.43(53.24)	144.44(25.19)	152.91(63.54)	198.70(81.54)	137.38(2.87)
$\hat{\beta}$	62.74(8.19)*	-61.36(11.27)*	-78.93(26.95)*	-83.89(66.29)	-105.07(41.81)*	-74.65(18.27)*	-151.73(36.13)
50							
Intercept	230.74(163.07)	264.92(23.16)	140.96(13.23)	167.62(39.35)	157.21(17.82)	192.32(23.01)	172.25(4.91)
$\hat{\beta}$	73.35(28.07)*	-63.24(10.18)*	-81.31(22.07)*	-63.85(10.11)*	-98.98(20.64)*	-88.39(49.84)	-131.23(52.98)*
60							
Intercept	227.52(49.73)	223.64(25.17)	152.89(21.52)	179.85(40.53)	174.35(33.36)	189.78(21.57)	170.18(2.71)
$\hat{\beta}$	86.15(23.34)*	-44.33(15.18)*	-61.61(20.77)*	-69.06(39.93)	-69.34(8.66)*	-93.57(66.56)	-112.40(14.56)*
70							
Intercept	248.76(61.47)	179.31(44.54)	173.17(23.45)	188.79(58.05)	258.56(18.75)	198.82(13.71)	169.33(2.17)
$\hat{\beta}$	104.98(7.26)*	-30.35(27.29)	-10.41(27.38)	-58.34(96.41)	-43.32(28.01)*	-53.57(38.38)	-26.70(41.56)
80							
Intercept	-----	167.53(4.04)	186.49(31.33)	166.41(4.02)	241.62(17.39)	166.37(13.29)	163.83(4.03)
$\hat{\beta}$	-----	-11.74(19.61)	4.97(6.13)	4.29(4.85)	0.74(4.41)	4.09(5.78)	-9.52(11.69)
90							
Intercept	-----	168.86(3.39)	175.82(11.02)	167.59(2.65)	245.81(19.22)	157.35(16.26)	167.62(12.71)
$\hat{\beta}$	-----	-6.81(10.55)	7.48(5.04)	4.15(5.13)	0.91(5.04)	6.84(3.77)	-11.14(12.83)

*: Significant P -value ($P < .05$); $\hat{\beta}$: The estimated coefficient of the gene.

$Q(0.50|CLINT1) = 264.92 - 63.24 \times (1.03) = 199.78$, indicating that the probability of survival for a case with a CLINT1 expression of 1.03 at 199.78 months is 50%.

The relapse score was calculated using the estimated coefficients from the final Cox model and was categorized into 2 groups based on values below and above the median. The mean estimated survival time, standard error, and 95% confidence interval for both the entire dataset and the ER-positive subsets are presented in Table 4. The log-rank test indicates a significant difference in the survival distributions between the 2 groups ($P < .001$). Additionally, the

Kaplan-Meier survival curves for both datasets are shown in Figure 2.

The results of the random forest analysis assessing the impact of various genes on the categorized relapse scores are presented in Table 5. This table highlights the increase in mean squared error of predictions and node purities, which the random forest method uses to evaluate the importance of genes in relation to the categorized prognostic scores (low/high risk patients). The 8 most important genes identified in this analysis are RACGAP1, NEK2, CCNA2, DTL, ACBD3, ARL6IP5, WFDC1, and Programmed Cell Death 4 (PDCD4).

Table 3. The results of censored quantile regression on the ER+ subset .

QUANTILES	SNU13	ACTL6A	NEK2	CD44	SHC1	ZCCHC8	WFDC1
10							
Intercept	24.88(2.39)	24.36(1.16)	23.01(1.22)	26.02(2.26)	19.13(1.81)	26.95(3.74)	20.85(1.57)
$\hat{\beta}$	9.03(1.82)*	-17.05(1.73)*	-17.57(2.12)*	8.63(1.94)*	-1.25(2.17)	-11.04(3.66)*	-6.61(1.53)*
20							
Intercept	45.01(2.41)	41.76(2.45)	44.91(1.78)	44.71(3.15)	45.05(3.77)	47.22(1.63)	40.71(3.43)
$\hat{\beta}$	19.92(1.95)*	-26.16(2.74)*	-26.85(3.60)*	14.25(2.02)*	-17.14(5.05)*	-20.72(1.70)*	-16.34(2.97)*
30							
Intercept	86.09(26.09)	84.27(32.26)	111.06(74.92)	112.32(2.24)	95.92(27.42)	90.86(22.82)	82.61(19.51)
$\hat{\beta}$	48.78(18.64)*	-78.37(66.20)	-66.58(28.66)*	70.69(74.85)	-51.94(13.49)*	-53.26(22.39)*	-43.01(9.19)*
40							
Intercept	124.35(11.22)	121.03(14.09)	158.24(99.42)	155.48(7.18)	126.95(50.77)	119.47(24.18)	166.03(80.23)
$\hat{\beta}$	64.42(21.96)*	-79.91(17.63)*	-91.37(62.90)	69.64(22.73)*	-60.34(25.62)*	-57.09(15.66)*	-86.93(21.51)*
50							
Intercept	182.89(86.51)	164.35(84.91)	157.81(18.46)	186.74(7.68)	147.24(17.44)	164.61(19.94)	186.52(19.66)
$\hat{\beta}$	82.55(49.06)	-73.99(19.91)*	-75.48(20.05)*	84.20(45.73)*	-79.75(52.10)	-73.80(13.65)*	-103.91(16.83)*
60							
Intercept	230.75(12.32)	150.03(12.63)	189.82(52.71)	187.47(2.37)	162.62(9.62)	182.81(27.66)	205.03(34.79)
$\hat{\beta}$	88.53(62.33)	-46.12(8.19)*	-48.95(11.81)*	68.51(11.89)*	-69.08(14.44)*	-64.56(12.39)*	-102.21(31.28)*
70							
Intercept	212.61(25.87)	169.55(13.22)	175.42(34.71)	218.33(3.61)	179.41(13.25)	-----	191.41(16.80)
$\hat{\beta}$	67.33(37.80)	-47.755(90)*	-47.07(4.70)*	77.77(16.88)*	-52.67(18.24)*	-----	-57.55(44.24)
80							
Intercept	188.39(33.92)	183.61(20.25)	162.96(3.96)	253.77(5.05)	192.92(18.84)	-----	201.53(44.59)
$\hat{\beta}$	48.27(50.70)	-49.20(11.74)*	-23.54(28.22)	74.51(30.53)*	-43.86(17.30)*	-----	-3.36(13.52)
90							
Intercept	173.43(10.33)	179.34(1.74)	164.03(1.62)	248.34(3.59)	-----	-----	170.01(19.22)
$\hat{\beta}$	-14.83(8.93)	-35.15(27.01)	-13.63(38.63)	51.85(45.08)	-----	-----	-1.40(9.41)

*: Significant P -value ($P < .05$); $\hat{\beta}$: The estimated coefficient of the genes.

Table 6 presents the expression levels of proteins encoded by SNU13, CLINT1, ACBD3, NEK2, COL2A1, WFDC1, RACGAP1, CCNA2, DTL, ARL6IP5, and Programed Cell Death 4 (PDCD4) in breast tumors. The analysis of RNA sequencing expression data from breast cancer tissues and normal tissues in The Cancer Genome Atlas (TCGA) is illustrated in Figure S1 (Supplemental File), showing a significant difference in expression between normal and cancerous tissues. Additionally, the protein-protein interaction

network involving RACGAP1, NEK2, and CCNA2 in breast cancer is displayed in Figure 3.

External validation

The publicly available dataset (GEO repository ID: GSE26971), generated using the Affymetrix Human Genome U133A Array, was utilized for validation purposes. The results of the concordance index are presented in Figure 4, indicating that the model based on the defined genes is significantly predictive.

Table 4. The comparison of survival probability distribution between the 2 levels using log-rank test.

PROGNOSTIC SCORE LEVELS	MEAN	STANDARD ERROR	95% CONFIDENCE INTERVAL		LOG RANK	P-VALUE
			LOWER BOUND	UPPER BOUND		
All data					51.021	<.001
Low risk	138.781	4.067	130.810	146.752		
High risk	89.760	5.963	78.073	101.447		
ER +					38.599	<.001
Low risk	138.862	4.641	129.766	147.958		
High risk	90.086	6.802	76.754	103.418		
Validation data					51.633	<.001
Low risk	138.938	4.041	131.018	146.858		
High risk	89.760	5.963	78.073	101.447		

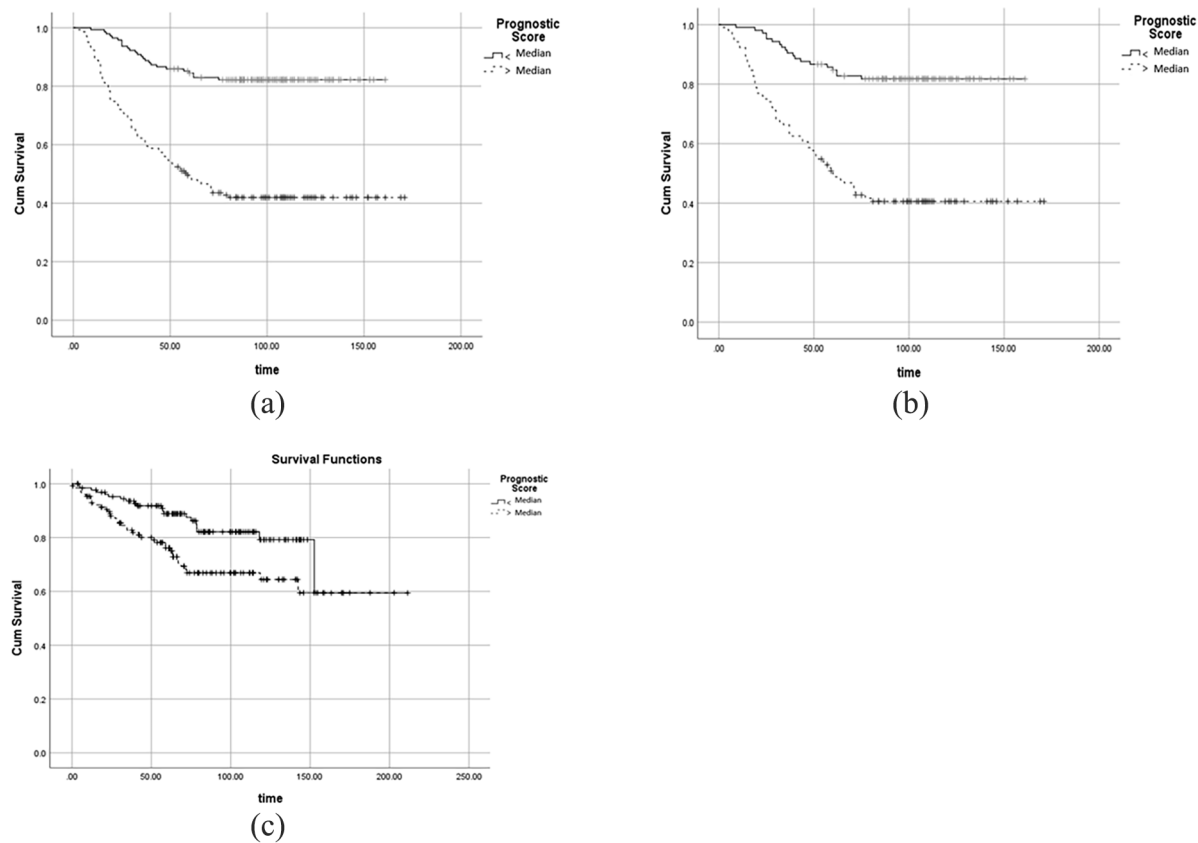


Figure 2. Kaplan-Meier survival function (a)comparing the survival probabilities of high (prognostic score> median) and low (prognostic score< median) risk groups of patients in the entire population, (b) comparing the survival probabilities of high (prognostic score> median) and low (prognostic score< median) risk groups of patients in the ER positive subset, (c) comparing the survival probabilities of high (prognostic score> median) and low (prognostic score< median) risk groups of patients in the validation data.

Enrichment analysis

The STRING database (<https://string-db.org/>) serves as a valuable resource for investigating potential gene interactions and their relevance to breast cancer. In this study, we examined gene interactions focused on a set of eleven genes:

SNU13, CLINT1, ACBD3, NEK2, COL2A1, WFDC1, RACGAP1, CCNA2, DTL, ARL6IP5, and Programed Cell Death 4 (PDCD4). By leveraging insights gained from the interactions among RACGAP1, NEK2, and CCNA2, we can enhance our understanding of their significance in various aspects of breast cancer, including time to relapse/

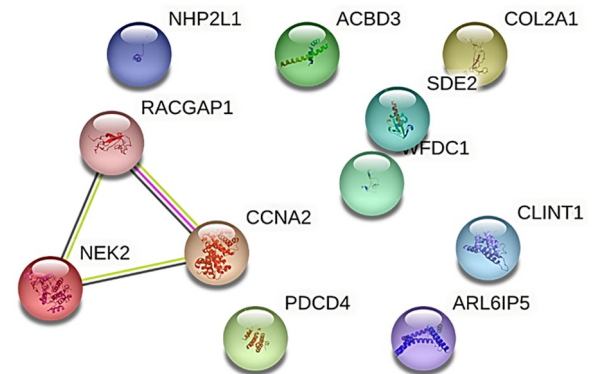
Table 5. The increase in mean squared error of predictions and node purities.

GENE	INCREASE IN MSE OF PREDICTIONS	INCREASES IN NODE PURITIES
RACGAP1	0.024956837	5.456876583
NEK2	0.011841553	3.005430556
CCNA2	0.009121284	2.140263706
DTL	0.008553764	2.058999355
ACBD3	0.007398398	1.721430325
ARL6IP5	0.007345479	2.192671492
WFDC1	0.004281475	1.220224785
PDCD4	0.004089229	1.110835191
PRC1	0.002938767	0.745111081
PRKCH	0.002884251	0.684918325
CENPF	0.002870078	0.686850513
SLC25A28	0.002830814	0.728713811
SMC4	0.001932006	0.488324668
CENPU	0.001322281	0.212405994
UBXN6	0.001120918	0.375613913
PDCD4	0.001014161	0.170623622
ELOC	0.001011655	0.281563336
HLA-DMA	0.000864591	0.124558936

Abbreviation: MSE, mean squared error.

Table 6. The results of expression level of proteins encoded by the selected genes in breast tumors.

GENE NAME	HIGH	MEDIUM	LOW	NOT DETECTED
SNU13	0	0	4	8
CLINT1	1	9	0	0
ACBD3	0	11	0	0
NEK2	0	0	0	10
COL2A1	0	1	0	9
WFDC13				
RACGAP1	0	6	4	0
CCNA2	0	5	5	1
DTL	0	0	1	10
ARL6IP5	2	7	1	0
PDCD4	8	2	1	0

**Figure 3.** Protein-Protein Interaction Network Involving RACGAP1, NEK2, and CCNA2 Proteins in Breast Cancer.

metastasis, identification of high-risk groups, and the risk of distant metastasis.

Discussion

BC is considered as one of the most prevalent life-threatening cancers among women worldwide. Although BC without metastasis can be controlled using advanced therapies, strategies to prevent the recurrence/metastasis of BC are rare. As the whole body can be involved in advanced stages of BC that can lead to cancer-related death of the patient, it is of great importance to identify biomarkers associated with metastasis of BC.²⁷ In this study, advanced statistical and machine learning methods were employed to achieve more reliable results. Gene expression data related to breast cancer (GSE2034) was reanalyzed using the component-wise likelihood-based boosting method for survival data. A small set of genes, including SNU13, CLINT1, ACBD3, NEK2, COL2A1, WFDC1, and RACGAP1, were identified as being related to time-to-relapse/metastasis in breast cancer patients. The relationship between these identified genes and metastasis time was further examined using a quantile regression model. A relapse index was created based on the 7 selected genes to categorize patients into low- and high-risk survival groups. Our findings, supported by the log-rank test, demonstrated the relapse index's strong discrimination ability. A supervised random forest algorithm was then applied to classify the survival groups and rank the genes most critical to survival outcomes. The top 8 genes associated with the identified survival groups were RACGAP1, NEK2, CCNA2, DTL, ACBD3, ARL6IP5, WFDC1, and PDCD4. Four of these genes overlapped with the Cox-Boost results, while the other 4 were newly identified in this step.

In analyzing gene expression profiles and corresponding time-to-metastasis of patients with BC, our finding showed that CCNA2 is highly expressed in identified high-risk group, which agreed with the results of other studies. Cyclin-A2, which belongs to the cyclin family, is a protein that is encoded by CCNA2 gene. The function of Cyclin-A2 is to regulate the

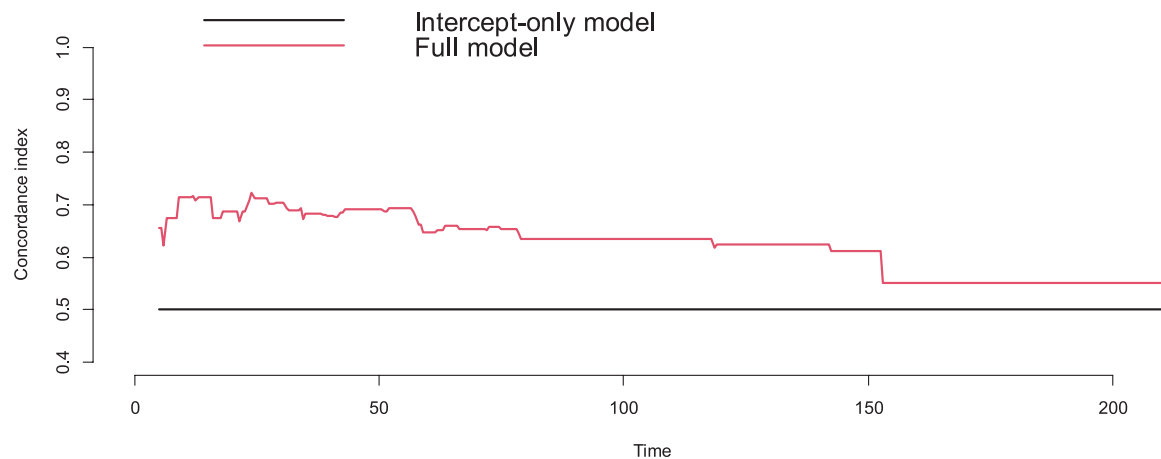


Figure 4. Plotting the concordance index for intercept-only versus full model (7 genes) during the study period in the validation data comparing the probability that a randomly selected subject who experienced the outcome will have a higher predicted probability of having the outcome occur than a randomly selected subject who did not experience the outcome.

cyclin-dependent protein CDK kinases. The findings of the present study showed that PDC4 was one of the top rank genes associated with identified survival groups. PDC4 has been shown to be a tumor suppressor and acts as an inhibitor of protein translation by interacting with the eukaryotic translation initiation factor 4A1 (eIF4A1) through binding to some mRNAs.²⁸ Under-expression of PDCD4 is reported to be correlated with poor prognosis in BC tumors and correlated with over-expression of oncogenic microRNA-21 (miR-21), targeting PDCD4 mRNA, which agreed with our study.²⁹

Our finding showed that DTL was over-expressed in the high-risk group (patients with a lower survival), which was in agreement with other studies.³⁰ Studies have shown that cancer patients, including those with breast cancer, who exhibit over-expression of DTL tend to have lower survival rates. DTL is upregulated in cancerous tissues compared to normal tissues and is associated with poor outcomes.³⁰ Additionally, it has been demonstrated that overexpression of DTL decreases the protein levels and accelerates the degradation of programmed cell death 4 (PDCD4), a potential substrate of DTL. Furthermore, DTL enhances the proliferation rate and migration capabilities of cancerous cells.

The findings of this study also revealed that ADP ribosylation factor-like GTPase 6 interacting protein 5 (ARL6IP5) was associated with high- and low-risk groups. This gene has been shown to be a microtubule-binding protein involved in differentiation, apoptosis, and response to stress stimuli. ARL6IP5 has been reported to regulate cancer cell migration through its involvement in the mitogen-activated protein kinase (MAPK) signaling pathway.³¹ The expression of ARL6IP5 has been shown to be mediated by ER and its increased level of expression caused by increased H₂O₂ levels indicates its role in oxidative stress.³¹ Studies have confirmed that knocking down the ARL6IP5 in BC enhances invasion and migration of MDA-MB-231 cells and decreases inducing cell apoptosis.³²

SNU13 as a component of U4 snRNP has an essential role in mRNA splicing pathways.^{33,34} Zhang et al. in their study showed that the expression level of *SNU13* is increased in Her-2 positive breast tumors and has a negative correlation with patients prognosis.³⁵ The protein encoded by *CLINT1* is involved in clathrin-coated vesicle formation and vesicular transportation between Golgi and endosomes.^{36,37} *CLINT1*, as a tumor suppressor gene, associates with distant metastasis-free survival for BC patients.³⁸ *ACBD3* is a Golgi protein that has a role in the maintenance of Golgi apparatus structure. The protein encoded by this gene also regulates several biological processes such as apoptosis and steroidogenesis.³⁹ The expression level of this gene is associated with tumorigenesis, metastasis, and poor prognosis of BC.⁴⁰ *NEK2* has a role in the regulation of mitosis and centrosome splitting. The protein encoded by this gene as a serine/threonine kinase controls the correct separation of chromosomes through the cell cycle.⁴¹ Upregulation of *NEK2* is associated with tumorigenesis and poor prognosis for BC.⁴² Type II Collagen encoded by *COL2A1* is one of the major components of the extracellular matrix (ECM) and therefore has an essential role in tumor progression.⁴³ The expression level of this gene was associated with chemoresistance in HER2+ BC patients.⁴⁴ Moreover, miRNA-301 promotes proliferation, invasion, and chemoresistance by targeting *COL2A1*, *FOXF2*, *PTEN*, and *BBC3*.⁴⁵ The protein encoded by *WFDC1* is a small protein with protease inhibitor function. Copy number variation and loss of heterozygosity for *WFDC1* are common in several cancers such as BC.⁴⁶ *RACGAP1* has an essential role in controlling the growth and differentiation of tumor cells.⁴⁷ The expression level of this gene is associated with poor prognosis in BC patients.⁴⁸ *ACTL6A* is involved in regulating several cellular functions, such as cell cycle histone acetylation and chromatin remodeling. Expression level of this gene correlates with poor prognosis in triple-negative BC.⁴⁹ *SHC1*, as an adaptor protein, has a role in the activation of several signaling pathways

such as PI3K and RAS/MAPK. In BC patients, upregulation of this gene is associated with a worse predicted outcome.⁵⁰

These results indicated that SNU13, CLINT1, ACBD3, NEK2, COL2A1, WFDC1, and RACGAP1 have an essential role in several biological processes such as proliferation, apoptosis, invasion, stemness, and differentiation, therefore potentially playing a role in regulating the tumor genesis, distant metastasis, response to treatment, and prognosis of BC.

Our findings also revealed that there were interactions between RACGAP1, NEK2, and CCNA2 proteins in breast cancer: a) Interaction between RACGAP1 and NEK2: The collaboration observed between RACGAP1, a GTPase-activating protein, and NEK2, a serine/threonine kinase, suggests their potential involvement in breast cancer progression. Dysregulation of both genes has been associated with aggressive tumor behavior and unfavorable clinical outcomes. RACGAP1's role in cell division and cytokinesis, along with NEK2's function in mitotic progression and centrosome duplication, indicates their joint impact on promoting invasive behavior and metastasis. Further investigation is necessary to unveil the underlying mechanisms governing this interaction. B) Interaction between RACGAP1 and CCNA2: The interaction between RACGAP1 and CCNA2 implies their interplay in breast cancer development. CCNA2, a critical regulator of the cell cycle, has been linked to aggressive tumor behavior and poor prognosis. Considering RACGAP1's involvement in cell division, its interaction with CCNA2 potentially influences cell cycle progression, contributing to tumor growth and metastasis. Unraveling the precise mechanisms and functional consequences of this interaction holds promise for enhancing our understanding of breast cancer pathogenesis. C) Interaction between NEK2 and CCNA2: The interaction between NEK2 and CCNA2 suggests their collaboration in breast cancer pathogenesis. Both genes are involved in cell cycle regulation and play essential roles in facilitating DNA replication and cell division. Dysregulation of NEK2 and CCNA2 has been associated with aggressive tumor behavior and adverse clinical outcomes. This interaction likely affects cell cycle progression and may contribute to the development of high-risk groups and an elevated risk of distant metastasis in breast cancer patients.

In this study, we utilized the Cox-Boost model due to its distinct advantages in handling survival data, particularly in the context of heart failure research. The Cox-Boost model extends the traditional Cox proportional hazards model by incorporating boosting techniques to improve predictive accuracy and variable selection. One of the key advantages of the Cox-Boost model is its ability to handle high-dimensional datasets, which is critical in our analysis given the complexity of factors influencing heart failure outcomes. Unlike conventional methods that often face challenges with multicollinearity and overfitting, the Cox-Boost model effectively identifies important predictors while maintaining robustness against data noise.⁵¹ Additionally, this model offers a flexible framework that allows

for the inclusion of unpenalized mandatory covariates, which receive a rapid coefficient build-up during the boosting steps, while other optional covariates are subjected to penalization. This approach ensures that essential covariates are incorporated efficiently, while the penalization helps manage the selection of less critical variables, reducing the risk of overfitting.²⁴ By leveraging these strengths, the Cox-Boost model provides a more nuanced understanding of survival patterns among heart failure patients, ultimately supporting more informed clinical decision-making and enabling personalized treatment strategies.²⁴ This rationale underscores our choice of the Cox-Boost model as a fitting and powerful tool for our study's objectives. The references provided in the search results offer valuable insights into the Cox-Boost model and its applications in survival analysis, discussing the advantages of the Cox-Boost model in handling high-dimensional data and its ability to incorporate mandatory covariates. We also utilized censored quantile regression to calculate the probability of survival in different deciles. Time to event data are commonly skewed and semi-parametric approaches, such as quantile regression models are better choices for model fitting.⁵² Future experimental validation studies are recommended.

The present study provides valuable insights for biomarker research on breast cancer metastasis-free survival. By utilizing component-wise likelihood-based boosting and random forest techniques, key metastasis-associated genes were identified and validated through Kaplan-Meier plots. From this analysis, eleven genes emerged as potential prognostic markers: SNU13, CLINT1, ACBD3, NEK2, COL2A1, WFDC1, RACGAP1, CCNA2, DTL, ARL6IP5, and PDCCD4. These genes warrant further investigation to clarify their molecular mechanisms before being implemented for evaluating relapse/metastasis-free survival in breast cancer patients. Among these, RACGAP1, NEK2, and CCNA2 are particularly significant. STRING database analysis revealed interactions among these genes, emphasizing their potential role in breast cancer metastasis and patient outcomes. Further investigation into these interactions could provide critical insights into the mechanisms driving aggressive tumor behavior. However, additional experimental and clinical studies are essential to validate these findings. Keeping up with the latest scientific developments and fostering collaborations with experts in the field will be key to advancing our understanding of the interactions between RACGAP1, NEK2, and CCNA2 in breast cancer prognosis and metastasis.

Conclusion

Based on the findings of this study, the Cox-Boost model effectively identified high-risk and low-risk survival subgroups of breast cancer patients. From this analysis, a set of key genes emerged as potential prognostic markers for distant metastasis in lymph-node-negative breast cancer, including SNU13, CLINT1, ACBD3, NEK2, COL2A1, WFDC1, RACGAP1,

CCNA2, DTL, ARL6IP5, and PDCD4. It is recommended that further molecular studies be conducted to validate the role of these genes in tumorigenesis, invasion, metastasis, and epithelial-mesenchymal transition in breast cancer.

Abbreviations

BC: Breast Cancer

ER: Estrogen

GEO: Gene expression Omnibus

Acknowledgements

We would like to appreciate the Research and Technology Deputy of the Hamadan University of Medical Sciences and the Research and Technology Deputy of the Hamedan University of Technology for technical support for their approval and support of this work.

Authors' Contributions

P. A., O. H. and L. T. made a substantial contribution to the concept or design of the work; or acquisition, analysis or interpretation of data, P. A., O. H., and L. T., Y. Z., S. A., and I.D. drafted the article or revised it critically for important intellectual content. All authors approved the version to be published.

Data Availability

The data underlying this article are available in <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=gse2034>.

Ethics Approval

This study was approved by the Ethical Committee of Hamadan University of Medical Sciences (IR.UMSHA.REC.1402.309).

Consent for Publication

Not applicable.

Consent for Participation in the Study

Not applicable.

ORCID iDs

Payam Amini  <https://orcid.org/0000-0001-8675-0045>

Leili Tapak  <https://orcid.org/0000-0002-4378-3143>

Yasaman Zohrab Beigi  <https://orcid.org/0000-0003-3640-9222>

Supplemental Material

Supplemental material for this article is available online.

REFERENCES

- Iacoviello L, Bonaccio M, de Gaetano G, Donati MB. Epidemiology of breast cancer, a paradigm of the "common soil" hypothesis. *Semin Cancer Biol.* 2021;72:4-10.
- Siegel RL, Miller KD, Jemal A. Cancer statistics, 2019. *CA Cancer J Clin.* 2019;69:7-34.
- Yadav S, Hu C, Hart SN, et al. Evaluation of germline genetic testing criteria in a hospital-based series of women with breast cancer. *World J Clin Oncol.* 2020;38:1409-1418.
- Feruza X. Current concepts of breast cancer risk factors. *Int J Philos Stud Soc Sci.* 2021;1:57-66.
- McAndrew NP, Bottalico L, Mesaros C, et al. Effects of systemic inflammation on relapse in early breast cancer. *NPJ Breast Cancer.* 2021;7:7-10.
- Das DK. *Insights Into a Phased Approach to Breast Cancer Early Detection Programs.* LWW; 2021.
- Arnold M, Morgan E, Rumgay H, et al. Current and future burden of breast cancer: Global statistics for 2020 and 2040. *Breast.* 2022;66:15-23.
- McCormack V, McKenzie F, Foerster M, et al. Breast cancer survival and survival gap apportionment in sub-Saharan Africa (ABC-DO): a prospective cohort study. *Lancet Glob Health.* 2020;8:e1203-e1212.
- Elobaid YE, Aw TC, Grivna M, Nagelkerke N. Breast cancer screening awareness, knowledge, and practice among Arab women in the United Arab Emirates: a cross-sectional survey. *PLoS One.* 2014;9:e105783.
- Bleicher RJ, Ruth K, Sigurdson ER, et al. Time to surgery and breast cancer survival in the United States. *JAMA Oncol.* 2016;2:330-339.
- Pei J, Wang Y, Li Y. Identification of key genes controlling breast cancer stem cell characteristics via stemness indices analysis. *J Transl Med.* 2020;18:74-15.
- Tungskrutthai S, Petpiroon N, Chanvorachote P. Molecular mechanisms of breast cancer metastasis and potential anti-metastatic compounds. *Anticancer Res.* 2018;38:2607-2618.
- Wang L, Zhang S, Wang X. The metabolic mechanisms of breast cancer metastasis. *Front Oncol.* 2020;10:602416.
- De Talhouet S, Peron J, Vuilleumier A, et al. Clinical outcome of breast cancer in carriers of BRCA1 and BRCA2 mutations according to molecular subtypes. *Sci Rep.* 2020;10:7073-7079.
- Ren M, Orozco A, Shao K, et al. Germline variants in hereditary breast cancer genes are associated with early age at diagnosis and family history in Guatemalan breast cancer. *Breast Cancer Res Treat.* 2021;189(2):533-539.
- Wood ME, McKinnon W, Garber J. Risk for breast cancer and management of unaffected individuals with non-BRCA hereditary breast cancer. *Breast J.* 2020;26:1528-1534.
- Ursu R, Truica RA, Cojocaru A, et al. Genetic factors involved in breast cancer. *Rom -med J.* 2022;69:10-12.
- Asghar N, Khalil U, Ahmad B, et al. Improved nonparametric survival prediction using CoxPH, Random Survival Forest & DeepHit Neural Network. *BMC Med Inform Decis Mak.* 2024;24:120.
- Zemmour C, Bertucci F, Finetti P, et al. Prediction of early breast cancer metastasis from DNA microarray data using high-dimensional cox regression models. *Cancer Inform.* 2015;14s2:CIN.S17284.
- De Bin R. Boosting in Cox regression: a comparison between the likelihood-based and the model-based approaches with focus on the R-packages CoxBoost and mboost. *Comput Stat.* 2016;31:513-531.
- Spooner A, Chen E, Sowmya A, et al. A comparison of machine learning methods for survival analysis of high-dimensional clinical data for dementia prediction. *Sci Rep.* 2020;10:20410-10.
- Yan D, Shen M, Du Z, et al. Developing ZNF gene signatures predicting radio-sensitivity of patients with breast cancer. *J Oncol.* 2021;2021:9255494.
- Wang Y, Klijn J, Zhang Y, et al. Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *Lancet.* 2005;365:671-679.
- Binder H, Schumacher M. Allowing for mandatory covariates in boosting estimation of sparse high-dimensional survival models. *BMC Bioinformatics.* 2008;9:10.
- Schönlau M. Random forests. In: Härdle WK (ed.) *Applied Statistical Learning: With Case Studies in Stata.* Springer; 2023;183-204.
- Fei Z, Zheng Q, Hong HG, Li Y. Inference for high-dimensional censored quantile regression. *J Am Stat Assoc.* 2023;118:898-912.
- McGuire A, Brown JA, Kerin MJ. Metastatic breast cancer: the potential of miRNA for diagnosis and treatment monitoring. *Cancer Metastasis Rev.* 2015;34:145-155.
- Osborne CK, Schiff R. Mechanisms of endocrine resistance in breast cancer. *Annu Rev Med.* 2011;62:233-247.
- Nagao Y, Hisaoka M, Matsuyama A, et al. Association of microRNA-21 expression with its targets, PDCD4 and TIMP3, in pancreatic ductal adenocarcinoma. *Mod Pathol.* 2012;25:112-121.
- Sánchez-Navarro I, Gámez-Pozo A, Pinto A, et al. An 8-gene qRT-PCR-based gene expression score that has prognostic value in early breast cancer. *BMC Cancer.* 2010;10:1-10.
- Chen H, Bai J, Ye J, et al. JWA as a functional molecule to regulate cancer cells migration via MAPK cascades and F-actin cytoskeleton. *Cell Signal.* 2007;19:1315-1327.
- Chen X, Feng J, Ge Z, et al. Effects of the JWA gene in the regulation of human breast cancer cells. *Mol Med Rep.* 2015;11:3848-3853.
- Hamma T, Ferré-D'Amaré AR. Structure of protein L7Ae bound to a K-turn derived from an archaeal box H/ACA sRNA at 1.8 p resolution. *Structure.* 2004;12:893-903.
- Rothé B, Back R, Quinternet M, et al. Characterization of the interaction between protein Snu13p/15.5K and the Rsa1p/NUFIP factor and

- demonstration of its functional importance for snoRNP assembly. *Nucleic Acids Res.* 2014;42:2015-2036.
35. Zhang H, Han B, Han X, et al. Comprehensive analysis of splicing factor and alternative splicing event to construct subtype-specific prognosis-predicting models for breast cancer. *Front Genet.* 2021;12:736423-736423.
 36. Dodd ME, Hatzold J, Mathias JR, et al. The ENTH domain protein clint1 is required for epidermal homeostasis in zebrafish. *Development.* 2009;136:2591-2600.
 37. Leventis PA, Da Sylva TR, Rajwans N, et al. Liquid facets-related (lqfR) is required for egg chamber morphogenesis during *Drosophila* oogenesis. *PLoS One.* 2011;6:e25466.
 38. Lee H, Lee S, Jeong D, Kim SJ. Ginsenoside Rh2 epigenetically regulates cell-mediated immune pathway to inhibit proliferation of MCF-7 breast cancer cells. *J Ginseng Res.* 2018;42:455-462.
 39. Fan J, Liu J, Culty M, Papadopoulos V. Acyl-coenzyme A binding domain containing 3 (ACBD3; pap7; GCP60): an emerging signaling molecule. *Prog Lipid Res.* 2010;49:218-234.
 40. Huang Y, Yang L, Pei YY, et al. Overexpressed ACBD3 has prognostic value in human breast cancer and promotes the self-renewal potential of breast cancer cells by activating the Wnt/beta-catenin signaling pathway. *Exp Cell Res.* 2018;363:39-47.
 41. Xia J, He Y, Meng B, et al. NEK2 induces autophagy-mediated bortezomib resistance by stabilizing Beclin-1 in multiple myeloma. *Mol Oncol.* 2020;14:763-778.
 42. Marina M, Saavedra HI. Nek2 and Plk4: prognostic markers, drivers of breast tumorigenesis and drug resistance. *Front Biosci.* 2014;19:352-365.
 43. Nissen NI, Karsdal M, Willumsen N. Collagens and cancer associated fibroblasts in the reactive stroma and its relation to cancer biology. *J Exp Clin Cancer Res.* 2019;38:115-115.
 44. Hanker AB, Estrada MV, Bianchini G, et al. Extracellular matrix/Integrin signaling promotes resistance to combined inhibition of HER2 and PI3K in HER2(+) breast cancer. *Cancer Res.* 2017;77:3280-3292.
 45. Shi W, Gerster K, Alajez NM, et al. MicroRNA-301 mediates proliferation and invasion in human breast cancer. *Cancer Res.* 2011;71:2926-2937.
 46. Watson JE, Kamkar S, James K, et al. Molecular analysis of WFDC1/ps20 gene in prostate cancer. *Prostate.* 2004;61:192-199.
 47. Yeh CM, Sung WW, Lai HW, et al. Opposing prognostic roles of nuclear and cytoplasmic RACGAP1 expression in colorectal cancer patients. *Hum Pathol.* 2016;47:45-51.
 48. Pliarchopoulou K, Kalogeras KT, Kronenwett R, et al. Prognostic significance of RACGAP1 mRNA expression in high-risk early breast cancer: a study in primary tumors of breast cancer patients participating in a randomized Hellenic Cooperative Oncology Group trial. *Cancer Chemother Pharmacol.* 2013;71:245-255.
 49. Jian Y, Huang X, Fang L, et al. Actin-like protein 6A/MYC/CDK2 axis confers high proliferative activity in triple-negative breast cancer. *J Exp Clin Cancer Res.* 2021;40:56-56.
 50. Wright KD, Miller BS, El-Meanawy S, et al. The p52 isoform of SHC1 is a key driver of breast cancer initiation. *Breast Cancer Res.* 2019;21:74-74.
 51. Binder H, Binder MH. Package 'CoxBoost'. *Citeseer*; 2015.
 52. Portnoy S. Censored regression quantiles. *J Am Stat Assoc.* 2003;98:1001-1012.