www.jcimcr.org

JCIMCR OPEN ACCESS Journal of Clinical Images and Medical Case Reports

ISSN 2766-7820

Research Article

Open Access, Volume 6

Derivation and external validation of machine learning-based model for detection of breast cancer recurrence

Shaghayegh Khodabakhshian[1]; Elmira Keramatfar[2]*; Mostafa Taheri[3]; Mahdieh Karkehabadi[3]

[1]Student Research Center, Jask Department of Education, Jask, Iran.

[2]Faculty of Medicine, Friedrich Schiller University Jena, 07743 Jena, Germany.

[3]Gastroenterology and Liver Disease Research Center, Research Institute for Gastroenterology and Liver Diseases, Shahid Beheshti University of Medical Sciences, Tehran, Iran.

***Corresponding Author: Elmira Keramatfar**
Faculty of Medicine, Friedrich Schiller University Jena, 07743 Jena, Germany.
Email: kiarashbabaie1990@gmail.com

Received: Jul 12, 2025
Accepted: Aug 08, 2025
Published: Aug 15, 2025
Archived: www.jcimcr.org

DOI: www.doi.org/10.52768/2766-7820/3739

Abstract

Background: Breast cancer is the most prevalent cancer among women worldwide and the leading cause of cancer-related fatalities in females. Various factors contribute to breast cancer, including lifestyle choices and genetic predispositions. The diagnosis and treatment of breast cancer can significantly affect the physical and emotional well-being of women due to treatment side effects, fear of mortality, and feelings of social stigma.

Objective: The objective of this study is to propose a rule-based classification method with machine learning techniques for predicting different types of Breast cancer survival.

Methods: This study of 833 breast cancer patients analyzed factors including ER, HER-2, and PR status, days to surgery, tumor size, mitotic grade, and tumor location. It found that 750 patients experienced no recurrence. Machine learning models were employed to identify significant prognostic factors for breast cancer survival. The derived classifier was extensively evaluated using a five-fold cross-validation scheme.

Results: This study's P-values indicate that the results are statistically significant ($P<0.05$), meaning the observed effect is unlikely due to random variation alone. This research shows that "Days to Surgery" has a more substantial impact than the other variables. The MSE value indicated that Days to Surgery is the most important feature of this disease. This study uses RF to create a classification model predicting female patients' breast tumor types (Benign/Malignant).

Conclusions: AI models have emerged as a crucial resource for forecasting and identifying cancer. Recent progress in machine learning (ML) has greatly enhanced the ability to detect cancer at an early stage. In this study, the methods of DT, SVM, RF, NB, LR, and KNN models are employed as the classification to predict the nature of breast cancer with other attributes.

Keywords: Breast cancer; Machine learning; Classification; Predictive medicine.

Introduction

Breast cancer is the most prevalent cancer among women worldwide and the leading cause of cancer-related fatalities in females [1,2]. Various factors contribute to breast cancer, including lifestyle choices and genetic predispositions [3]. The diagnosis and treatment of breast cancer can significantly affect the physical and emotional well-being of women due to treatment side effects, fear of mortality, and feelings of social stigma [4]. The mortality rate of breast cancer (BC) has declined in recent decades due to advanced therapies and improved management of each patient's personalized risk profile. With these techniques and treatments, the focus is shifting toward minimizing the negative effects of oncological treatments to enhance the quality of life for breast cancer patients. The molecular subtype of breast cancer is an independent prognostic factor [5,6].

Assessing the expression of estrogen receptor (ER) and progesterone receptor (PR), as well as the overexpression of human epidermal growth factor receptor 2 (HER-2), is utilized to guide therapy decisions [7]. Breast cancer can be classified into distinct molecular subtypes based on receptor expression, which include luminal A, luminal B, HER-2, and triple-negative [8]. Metabolic changes are observed in various molecular subtypes and histological types of breast cancer [9]. Breast MRI has shown high sensitivity, but its specificity varies from 37% to 97% [10]. Consequently, multiple biopsy tests need to be conducted as supplements. Recently, specialized methods such as dynamic contrast-enhanced magnetic resonance imaging (DCE-MRI) and diffusion-weighted magnetic resonance imaging (DW-MRI) have advanced significantly [11]. These techniques now offer quantitative measurements of tissue properties that are closely related to assessing tumor progression and responses to treatment [12]. Angiogenesis is essential for the growth, progression, and spread of tumors. In the case of breast cancer, contrast-enhanced breast magnetic resonance imaging (MRI) is an effective imaging technique for visualizing tumor angiogenesis. The pattern of contrast enhancement in the tumor is directly related to the micro vessel density, which is a key characteristic of tumor angiogenesis [13].

Massafra et al. conducted a study using an ensemble machine-learning approach. They combined the predictions of three baseline models through a voting mechanism and employed a grid search procedure. Their method successfully predicts the occurrence of invasive disease events in breast cancer patients at 5- and 10-year intervals [14]. Mikhailova and colleagues utilized six different machine-learning models and validated the potential of machine learning for forecasting breast cancer recurrence [15]. Random forest (RF) is a type of machine-learning algorithm that utilizes several decision trees to recognize, categorize, and forecast target data [16]. An RF model might be utilized to forecast a decline in quality of life for patients with thyroid cancer following thyroidectomy [17]. It has been established that the random forest (RF) algorithm can achieve satisfactory levels of sensitivity and specificity. Consequently, this study sought to develop and evaluate a machine learning model utilizing the RF algorithm to enhance the classification and prediction of events, which are defined as either the first tumor recurrence occurring locally, regionally, or at a distant site; the identification of a secondary malignant tumor; or mortality from any cause [18-20].

Predictive medicine leverages advanced bioinformatics, genomics, and artificial intelligence to assess an individual's risk of developing diseases and tailor preventive or therapeutic strategies accordingly [21-25]. By analyzing genetic variations, biomarkers, and patient history, predictive models can identify predispositions to conditions such as cancer, cardiovascular diseases, and neurodegenerative disorders [26-29]. High-throughput sequencing and machine learning algorithms enable early diagnosis, prognosis estimation, and personalized treatment plans based on a patient's molecular profile [30-33]. Additionally, pharmacogenomics—a key component of predictive medicine—optimizes drug selection and dosage by predicting individual responses to medications, reducing adverse effects and enhancing treatment efficacy [34-36]. This approach is revolutionizing healthcare by shifting from a reactive to a proactive model, ultimately improving patient outcomes and reducing medical costs [37-39].

Material and methods

Data collection and preparation

This study included 833 breast cancer patients who underwent surgery at Gangnam Severance Hospital between January 2007 and July 2011. Of the selected patients, 750 did not experience recurrence. The inclusion criteria for participation were as follows: (a) a confirmed diagnosis of pathological breast cancer, (b) no prior history of cancer in the breast or any other part of the body, (c) availability of breast MRI imaging before surgery, and (d) absence of distant metastasis at the time of diagnosis. Considerable effort was dedicated to data collection and model development. Each patient's characteristics were incorporated as indicators, which were appropriately combined to create a variable aimed at improving predictive accuracy. After cleaning the data, preparing the methodologies, and determining the cognitive status at various assessment time points, we applied the proposed models.

Feature selection

After examining the patient's medical records, factors such as age, tumor stage, histological grade, estrogen receptor (ER) status, progesterone receptor (PR) status, and human epidermal growth factor receptor type 2 (HER-2) status were evaluated. The features selected for this study include:

- Days to surgery
- Estrogen receptor (ER) status
- Progesterone receptor (PR) status
- Tumor mitotic grade
- Tumor location
- Tumor grade (tubule formation)
- Tumor grade (nuclear characteristics)
- HER-2 status
- Type of definitive surgery
- Tumor size staging

- Node staging
- Adjuvant radiation therapy

HER-2 status: HER-2 is a protein that is overactivated in certain types of breast cancer, signifying that cancer cells grow more rapidly. HER-2 is a type 2 human epidermal growth factor receptor protein found on the surface of breast cells. It plays a crucial role in regulating cell growth and division. HER-2 is categorized using scores of 0, 1+, 2+, and 3+. Tumors with scores of 0 and 1+ are classified as HER-2 negative, while tumors with a score of 3+ are classified as HER-2 positive.

Tumor grade: Tumor grade indicates how abnormal the cancer cells are compared to normal cells. This is assessed by examining the cancer cells under a microscope and evaluating their differentiation, growth rate, and structure. In breast cancer, tumor grades range from Grade 1 to Grade 3, which provides insight into the potential aggressiveness of the disease. Higher grades (Grade 3) represent more aggressive tumors that tend to grow and spread more quickly, whereas lower grades (Grade 1) indicate tumors that grow more slowly and are less aggressive.

Tumor size: The size of a tumor plays a vital role in the staging of breast cancer, which helps identify the extent of cancer spread. Staging is categorized as T1, T2, T3, or T4, based on the tumor's physical dimensions and whether it has infiltrated nearby tissues. Tumors that are smaller and classified as T1 are generally linked to earlier cancer stages and a more favorable prognosis. In contrast, larger tumors classified as T3 or T4 typically signal a more advanced stage of the disease and may require more intensive treatment. Along with lymph node involvement and metastasis, tumor size is essential in assessing the overall stage of breast cancer, which can range from Stage 0 to Stage IV.

PR (Progesterone Receptor) and ER (Estrogen Receptor) Status: PR and ER are types of hormone receptors. The positivity of these receptors indicates that hormones contribute to tumor growth, suggesting that hormonal treatments may be effective. The status of estrogen receptors (ER) and progesterone receptors (PR) is essential for understanding breast cancer biology and guiding clinical management. Tumors that express either ER or PR, or both (referred to as ER+/PR+), are categorized as hormone receptor-positive (HR+). This category accounts for about 70% of breast cancer cases and typically exhibits more favorable outcomes, as these tumors usually respond positively to hormonal treatments like tamoxifen or aromatase inhibitors [40].

Tumor grade (Mitotic grade): Tumor grade, particularly mitotic grade, serves as a vital prognostic factor in breast cancer, reflecting the extent of cellular proliferation and tumor aggressiveness. The Nottingham Histologic Score (NHS) evaluates mitotic count, tubule formation, and nuclear pleomorphism to categorize tumors into three grades: 1, 2, and 3. Grade 3, marked by elevated mitotic activity, correlates with less favorable outcomes [41]. The mitotic grade is crucial for treatment protocols, emphasizing its role in tailoring therapeutic approaches.

Molecular subtyping: Molecular subtyping of breast cancer categorizes the disease into four primary types: Luminal A, Luminal B, HER2-enriched, and triple-negative breast cancer (TNBC). This classification plays a crucial role in determining prognosis and treatment strategies. Luminal A tumors are characterized by positivity for hormone receptors (estrogen receptor-positive and progesterone receptor-positive), negativity for HER2, and low Ki-67 levels (a proliferation marker), which typically lead to the most favorable outcomes. These subtypes form the basis for modern clinical guidelines, underscoring the importance of personalized management approaches for each subtype to improve patient survival.

Statistical analysis

We employed a data-centric methodology utilizing machine learning models along with baseline information to categorize individuals. In this research, the classifiers implemented included Support Vector Machine (SVM), Logistic Regression (LR), Decision Tree (DT), K Nearest Neighbor (KNN), Naive Bayes (NB), and Random Forest (RF) [42-44]. In machine learning, the selection of classes significantly impacts the performance of the model [11,45]. We assessed classical classification algorithms—SVM, Naive Bayes, KNN, Decision Trees, and Logistic Regression—on feature groups with more than two features and averaged their performance [46]. We utilized the 5-fold cross-validation method to apply and evaluate our classifiers [47]. This technique involves splitting the original dataset into two parts: a training sample used to develop the model, and a test set used for evaluation. This study uses RF to create a classification model predicting female patients' breast tumor types (Benign/Malignant). The study confirms the effectiveness of the classification model by evaluating its prediction accuracy. It compares the Random Forest (RF) model with other machine learning models, including Decision Trees (DT), Support Vector Machines (SVM), Logistic Regression (LR), and Neural Networks (NN). The results demonstrate that RF outperforms the other four models in classification accuracy. Assessing the MSE showed that the impact of each feature on this disease is greater than that of the other features.

Data availability

The data for this study is publicly available in the Cancer Imaging Archive (TCIA), a platform that hosts a large archive of medical cancer images accessible for public use. The direct link to our publicly available data is: https: //www.cancerimagingarchive.net/collection/duke-breast-cancer-mri/.

Results

This study focuses on the comparative assessment and prediction of breast cancer outcomes using six different classification models. The models implemented in this research include Decision Tree (DT), Support Vector Machine (SVM), Random Forest (RF), Naive Bayes (NB), Logistic Regression (LR), and K-Nearest Neighbors (KNN). The predictive results aim to reduce the rate of misdiagnoses and help develop appropriate treatment plans.

The study involved the examination of 833 patients diagnosed with breast cancer. Key attributes such as estrogen receptor (ER) status, days to surgery, HER-2 status, tumor size staging, tumor mitotic grade, tumor location, and progesterone receptor (PR) status were considered for the patients. The primary features of this disease were selected, and additional significant characteristics were identified through machine learning analysis.

Table 1 presents the most significant features, as determined by their P-values. The P-values in this study indicate that the results are statistically significant ($P<0.05$), suggesting that the observed effects are unlikely to be due to random variation. Table 1 shows that the P-values for several features, including tumor grade mitotic, days to surgery, and PR, are significant.

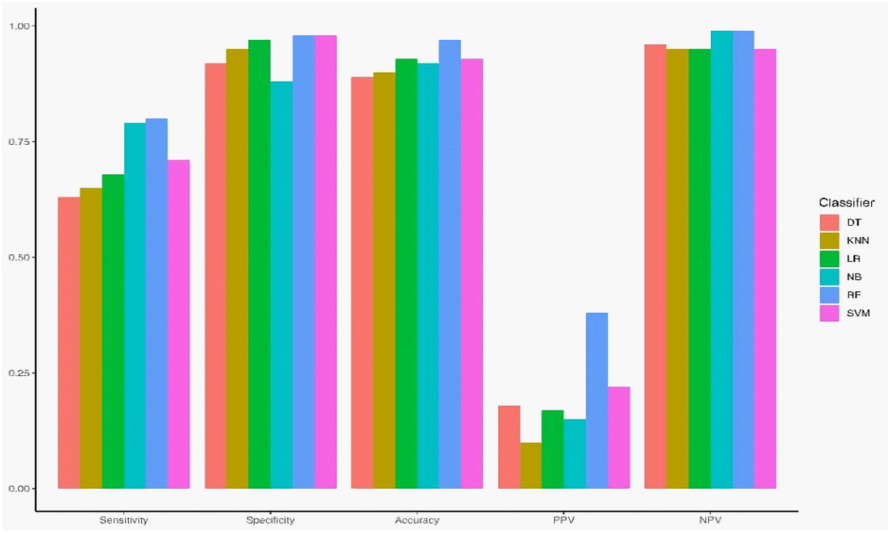

Figure 1: Displaying and comparing various classified algorithms for breast cancer analysis.

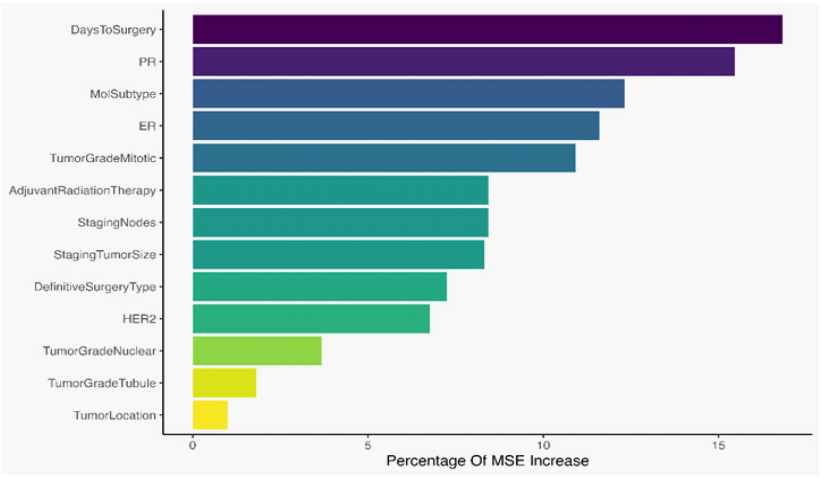

Figure 2: The effects of each feature on breast cancer.

Table 1: Overview of the participants according to their baseline diagnosis.

	Level	Nonrecurrence	Recurrence	p
n		**750**	**83**	
Days to surgery		85.52 (71.76)	119.64 (87.33)	<0.001
ER (positive)		575 (76.7)	51 (61.4)	0.004
PR (positive)		507 (67.6)	37 (44.6)	<0.001
HER2 (negative)		623 (83.1)	68 (81.9)	0.914
Tumor Location (right)		372 (49.6)	39 (47.0)	0.618
Breast conservation surgery		405 (54.0)	31 (37.3)	0.006
Adjuvant Radiation Therapy		537 (71.6)	58 (69.9)	0.841
Mol subtype	0	498 (66.4)	44 (53.0)	0.031
	1	87 (11.6)	9 (10.8)	
	2	40 (5.3)	6 (7.2)	
	3	125 (16.7)	24 (28.9)	
Staging tumor size	1	346 (46.1)	24 (28.9)	<0.001
	2	324 (43.2)	40 (48.2)	
	3	68 (9.1)	13 (15.7)	
	4	12 (1.6)	6 (7.2)	
Staging nodes	0	455 (60.7)	41 (49.4)	<0.001
	1	222 (29.6)	18 (21.7)	
	2	44 (5.9)	15 (18.1)	
	3	29 (3.9)	9 (10.8)	

Tumor grade tubule	1	58 (7.7)	2 (2.4)	0.012
	2	132 (17.6)	7 (8.4)	
	3	560 (74.7)	74 (89.2)	
Tumor grade nuclear	1	48 (6.4)	3 (3.6)	0.006
	2	343 (45.7)	25 (30.1)	
	3	359 (47.9)	55 (66.3)	
Tumor grade mitotic	1	465 (62.0)	38 (45.8)	<0.001
	2	169 (22.5)	18 (21.7)	
	3	116 (15.5)	27 (32.5)	

Figure 1 displays the highest RF value obtained from the calculations of accuracy, sensitivity, and specificity. The RF is highlighted as the most important indicator in Figure 1. These findings suggest that RF surpasses other models concerning classification accuracy.

This research shows that "Days to Surgery" has a more substantial impact than the other variables. MSE value was calculated, and Days to Surgery was identified as the most important feature of this disease, with a value of 16.82 compared to other features. In Figure 2, the most important features identified are Days to Surgery, PR, and Mol Subtype, listed in that order. The analysis of the mean squared error (MSE) indicated that the impact of each feature on the disease is more pronounced compared to other features.

Discussion

This study developed machine learning models using breast cancer data from Gangnam Severance Hospital to identify key prognostic factors for breast cancer survival. All algorithms (decision tree, random forest, Naive Bayes, k-Nearest Neighbors, logistic regression, and support vector machine) produced similar accuracies, with random forest achieving the highest. The decision trees and survival curves used for validation demonstrated that the identified important variables are valuable as a decision support tool for clinicians.

This study analyzed breast cancer features and identified key characteristics using a classification model. The derived classifier was extensively evaluated using a five-fold cross-validation scheme. The key features identified in this study include Days to Surgery, molecular subtype (ER, PR), and Tumor Grade Mitotic. The risk of recurrence is a significant concern, influenced by factors such as tumor biology, surgical margins, and adjuvant therapies. Receptor discordance between primary tumors and synchronous axillary lymph node metastases (ALNM) exists before treatment, potentially reflecting intratumoral heterogeneity and breast cancer phenotypic plasticity. Currently, ER, PR and HER2 expression instability have limited clinical consequences for neoadjuvant therapy [48].

The main finding of this breast cancer study, determined through P-value and MSE calculations, is the "Days to Surgery." Figure 2 indicates that calculating the MSE value is a key feature of PR. This study highlights the significance of PR and Days to Surgery in breast cancer. The time from diagnosis to surgical intervention, referred to as Days to Surgery (DTS), significantly impacts breast cancer outcomes. Research has shown mixed results regarding the relationship between prolonged DTS and survival rates. The length of time between diagnosis and surgery influences survival outcomes in early-stage breast cancer, making it pertinent to strive for a reduction in this interval. While the impact on overall and disease-specific survival is minimal, it is important to set achievable and sensible targets for the timing of surgical procedures to provide this group with a limited, yet meaningful, improvement in survival [49].

In addition to the MSE calculated from Figure 2, key features include tumor-grade mitotic and molecular subtypes. The mitotic index is an important factor in determining the tumor grade in breast cancer. It reflects how actively the cells are dividing and serves as a strong prognostic indicator. Histological grading systems, such as the Nottingham Grading System, combine the mitotic count with the degree of tubule formation and nuclear pleomorphism to classify tumors into three grades: 1 (low), 2 (intermediate), and 3 (high). A higher mitotic score, defined as 10 or more mitoses per high-power field, is linked to more aggressive tumor behavior, a higher risk of recurrence, and poorer survival outcomes. A high ratio of atypical to typical mitoses is linked to adverse outcomes, especially in luminal breast cancer and triple-negative breast cancer subtypes. Furthermore, atypical mitoses correlate with aggressive tumor characteristics, such as increased tumor size, elevated tumor grade, and a negative response to chemotherapy. The ratio of atypical to typical mitoses is a noteworthy prognostic factor in breast cancer, offering critical insights into tumor characteristics and treatment efficacy [50].

Molecular subtyping of breast cancer, determined through gene expression profiling, has significantly transformed prognosis and treatment strategies. The main subtypes include Luminal A, Luminal B, HER2-enriched, and Triple-Negative/Basal-like. Triple-negative breast cancer (TNBC), characterized by the absence of estrogen receptors (ER), progesterone receptors (PR), and HER2, tends to be more aggressive, with higher recurrence rates. TNBC patients often rely on chemotherapy or emerging immunotherapies. Highlighting substantial genomic heterogeneity within TNBC molecular subtypes allows for a better understanding of disease biology and identifies potential therapeutic targets, paving the way for novel anticancer therapies [51].

Random Forest (RF) is highlighted as the most important indicator in Figure 1. The assessment of accuracy and specificity, as well as the sensitivity metrics for each model, are detailed in Figure 1, which holds significant relevance for breast cancer predictions. RF, an ensemble machine-learning algorithm, has become a powerful tool in breast cancer research for risk prediction, classification, and biomarker discovery. This method could also be applied to detect other cancers, offering doctors guidance for early diagnosis and valuable clinical applications in breast tumor diagnosis [52].

There is increasing evidence suggesting that the performance of stand-alone machine learning (ML) models is on par with that of human readers and that ML is capable of handling triage tasks at a scale and speed that human readers cannot achieve. While only retrospective studies have been carried out, it is plausible that algorithms could match or even surpass the accuracy of a reader in the real-time breast screening pro-

cess. Nevertheless, stronger prospective data is essential for understanding how algorithm thresholds are established and is needed to explore the relationship between human readers and algorithms, as well as the impact on reader performance and patient outcomes over time.

Conclusion

AI models have emerged as a crucial resource for forecasting and identifying cancer. Recent progress in machine learning (ML) has greatly enhanced the ability to detect cancer at an early stage. In this study, the methods of DT, SVM, RF, NB, LR, and KNN models are employed as the classification to predict the nature of breast cancer with other attributes. Breast cancer prediction via machine learning, deep learning, and data mining. We aim to identify the most effective algorithm for predicting breast cancer occurrences. Researchers should consider the disparity between positive and negative data, as this can introduce bias in favor of either positive or negative predictions. Additionally, it is crucial to address the unequal representation of breast cancer images compared to the affected patches to ensure accurate diagnosis and prediction of breast cancer. The effectiveness of any method depends on the number of features and the methodology employed. Features are used as input for prediction. The success of any approach relies on the quantity of features and the strategies used. Features serve as inputs for making predictions. The results of predictions aid in minimizing misdiagnosis rates and encourage the development of efficient treatment strategies. Scientists can investigate various constraints in the management of this illness.

Declarations

Ethical approval: This retrospective study was approved by the institutional review board of Gangnam Severance Hospital:

- Dr. Ji Hyun Youk, Department of Radiology, Gangnam Severance Hospital, Yonsei University College of Medicine, Eonju-ro, Gangnam-Gu, Seoul, South Korea.
- Dr. Jung Choi, Department of Radiology, Chonbuk National University Medical School and Hospital, Institute of Medical Science, Research Institute of Clinical Medicine.
- Dr. HyeMi Choi, Department of Statistics, Institute of Applied Statistics, Chonbuk National University, Dukjin-Dong, Jeonju, Jeonbuk.

and the requirement for informed consent was waived.

References

1. Lima SM, Kehm RD, Terry MB. Global breast cancer incidence and mortality trends by region, age-groups, and fertility patterns. EClinicalMedicine. 2021; 38.
2. Yang H, Pawitan Y, Fang F, Czene K, Ye W. Biomarkers and disease trajectories influencing women's health: results from the UK biobank cohort. Phenomics. 2022; 2(3): 184-93.
3. Rojas K, Stuckey A. Breast cancer epidemiology and risk factors. Clinical obstetrics and gynecology. 2016; 59(4): 651-72.
4. You J, Lu Q. Social constraints and quality of life among Chinese-speaking breast cancer survivors: a mediation model. Quality of Life Research. 2014; 23: 2577-84.
5. Gaudet MM, Gierach GL, Carter BD, Luo J, Milne RL, Weiderpass E, et al. Pooled analysis of nine cohorts reveals breast cancer risk factors by tumor molecular subtype. Cancer research. 2018; 78(20): 6011-21.
6. Plevritis SK, Munoz D, Kurian AW, Stout NK, Alagoz O, Near AM, et al. Association of screening and treatment with breast cancer mortality by molecular subtype in US women, 2000-2012. Jama. 2018; 319(2): 154-64.
7. Hammond MEH, Hayes DF, Dowsett M, Allred DC, Hagerty KL, Badve S, et al. American Society of Clinical Oncology/College of American Pathologists guideline recommendations for immunohistochemical testing of estrogen and progesterone receptors in breast cancer. Journal of clinical oncology. 2010; 28(16): 2784-95.
8. Deyarmin B, Kane JL, Valente AL, van Laar R, Gallagher C, Shriver CD, et al. Effect of ASCO/CAP guidelines for determining ER status on molecular subtype. Annals of surgical oncology. 2013; 20: 87-93.
9. Cappelletti V, Iorio E, Miodini P, Silvestri M, Dugo M, Daidone MG. Metabolic footprints and molecular subtypes in breast cancer. Disease markers. 2017; 2017(1): 7687851.
10. Schelfout K, Van Goethem M, Kersschot E, Colpaert C, Schelfhout A, Leyman P, et al. Contrast-enhanced MR imaging of breast lesions and effect on treatment. European Journal of Surgical Oncology (EJSO). 2004; 30(5): 501-7.
11. Jalali S, Dadkhah K, Ghazi MM. Peritoneal Metastasis Prediction in Gastric Cancer: A Machine Learning Approach. medRxiv. 2025: 2025.04. 11.25325702.
12. Delille J-P, Slanetz PJ, Yeh ED, Halpern EF, Kopans DB, Garrido L. Invasive ductal breast carcinoma response to neoadjuvant chemotherapy: noninvasive monitoring with functional MR imaging—pilot study. Radiology. 2003; 228(1): 63-9.
13. Weidner N, Semple JP, Welch WR, Folkman J. Tumor angiogenesis and metastasis—correlation in invasive breast carcinoma. New England Journal of Medicine. 1991; 324(1): 1-8.
14. Massafra R, Comes MC, Bove S, Didonna V, Diotaiuti S, Giotta F, et al. A machine learning ensemble approach for 5-and 10-year breast cancer invasive disease event classification. Plos one. 2022; 17(9): e0274691.
15. Mikhailova V, Anbarjafari G. Comparative analysis of classification algorithms on the breast cancer recurrence using machine learning. Medical & Biological Engineering & Computing. 2022; 60(9): 2589-600.
16. Song Y, Yin Z, Zhang C, Hao S, Li H, Wang S, et al. Random forest classifier improving phenylketonuria screening performance in two Chinese populations. Frontiers in Molecular Biosciences. 2022; 9: 986556.
17. Liu YH, Jin J, Liu YJ. Machine learning–based random forest for predicting decreased quality of life in thyroid cancer patients after thyroidectomy. Supportive Care in Cancer. 2022: 1-7.
18. Abolghasemi S, Torbati MB, Pakzad P, Ghafouri-Fard S. Gene expression analysis of SOCS, STAT and PIAS genes in lung cancer patients. Pathology-Research and Practice. 2023; 249: 154760.
19. Koushki EH, Abolghasemi S, Mollica A, Aghaeepoor M, Moosavi SS, Farshadfar C, et al. Structure-based virtual screening, molecular docking and dynamics studies of natural product and classical inhibitors against human dihydrofolate reductase. Network Modeling Analysis in Health Informatics and Bioinformatics. 2020; 9: 1-21.
20. Samarin ZE, Abolghasemi S, Dehnavi E, Akbarzadeh A, Hadian A, Khodabandeh M, et al. Response Surface Optimization of the Expression Conditions for Synthetic Human Interferon α-2b Gene in Escherichia coli. Indian Journal of Pharmaceutical Sciences. 2018; 80(3).

21. Houri H, Aghdaei HA, Firuzabadi S, Khorsand B, Soltanpoor F, Rafieepoor M, et al. High prevalence rate of microbial contamination in patient-ready gastrointestinal endoscopes in Tehran, Iran: an alarming sign for the occurrence of severe outbreaks. Microbiology Spectrum. 2022; 10(5): e01897-22.

22. Khorsand B, Khammari A, Shirvanizadeh N, Zahiri J, Arab SS. OligoCOOL: a mobile application for nucleotide sequence analysis. Biochemistry and Molecular Biology Education. 2019; 47(2): 201-6.

23. Khorsand B, Savadi A, Naghibzadeh M. Parallelizing assignment problem with DNA strands. Iranian Journal of Biotechnology. 2020; 18(1): e2547.

24. Babak Khorsand and Nazanin Naderi and Seyedeh Sara Karimian and Maedeh Mohaghegh and Alireza Aghaahmadi and Seyedeh Negin Hadisadegh and Mina Owrang and Hamidreza H. Comprehensive transcriptomic analysis of hepatocellular Carcinoma: Uncovering shared and unique molecular signatures across diverse etiologies. Biochemistry and Biophysics Reports. 2025; 43: 102123.

25. Hesami Z, Atashrooz M, Sardarzehi R, Looha MA, Khorsand B, Houri H. The oro- and nasopharyngeal microbiota as a revolutionary perspective on mental disorders and related psychopathology: a systematic review and meta-analysis. Journal of Translational Medicine. 2025; 23(1): 726.

26. Irankhah L, Khorsand B, Naghibzadeh M, Savadi A. Analyzing the performance of short-read classification tools on metagenomic samples toward proper diagnosis of diseases. Journal of bioinformatics and computational biology. 2024; 22(5): 2450012.

27. Kharaghani AA, Harzandi N, Khorsand B, Rajabnia M, Kharaghani AA, Houri H. High prevalence of Mucosa-Associated extended-spectrum β-Lactamase-producing Escherichia coli and Klebsiella pneumoniae among Iranain patients with inflammatory bowel disease (IBD). Annals of Clinical Microbiology and Antimicrobials. 2023; 22(1): 86.

28. Razavi SA, Khorsand B, Salehipour P, Hedayati M. Metabolite signature of human malignant thyroid tissue: A systematic review and meta-analysis. Cancer Medicine. 2024; 13(8): e7184.

29. Sadeghnezhad E, Sharifi M, Zare-maivan H, Khorsand B, Zahiri J. Cross talk between energy cost and expression of Methyl Jasmonate-regulated genes: from DNA to protein. Journal of Plant Biochemistry and Biotechnology. 2019; 28: 230-43.

30. Haghzad T, Khorsand B, Razavi SA, Hedayati M. A computational approach to assessing the prognostic implications of BRAF and RAS mutations in patients with papillary thyroid carcinoma. Endocrine. 2024; 86(2): 707-22.

31. Khorsand B, Asadzadeh Aghdaei H, Nazemalhosseini-Mojarad E, Nadalian B, Nadalian B, Houri H. Overrepresentation of Enterobacteriaceae and Escherichia coli is the major gut microbiome signature in Crohn's disease and ulcerative colitis; a comprehensive metagenomic analysis of IBDMDB datasets. Frontiers in cellular and infection microbiology. 2022; 12: 1015890.

32. Soltanyzadeh M, Khorsand B, Baneh AA, Houri H. Clarifying differences in gene expression profile of umbilical cord vein and bone marrow-derived mesenchymal stem cells; a comparative in silico study. Informatics in Medicine Unlocked. 2022; 33: 101072.

33. Zareei S, Khorsand B, Dantism A, Zareei N, Asgharzadeh F, Zahraee SS, et al. PeptiHub: a curated repository of precisely annotated cancer-related peptides with advanced utilities for peptide exploration and discovery. Database. 2024; 2024: baae092.

34. Khorsand B, Savadi A, Zahiri J, Naghibzadeh M. Alpha influenza virus infiltration prediction using virus-human protein-protein interaction network. Mathematical Biosciences and Engineering. 2020; 17(4): 3109-29.

35. Khorsand B, Savadi A, Naghibzadeh M. Comprehensive host-pathogen protein-protein interaction network analysis. BMC bioinformatics. 2020; 21: 1-22.

36. Khorsand B, Savadi A, Naghibzadeh M. SARS-CoV-2-human protein-protein interaction network. Informatics in medicine unlocked. 2020; 20: 100413.

37. Samandari Bahraseman MR, Khorsand B, Esmaeilzadeh-Salestani K, Sarhadi S, Hatami N, Khaleghdoust B, et al. The use of integrated text mining and protein-protein interaction approach to evaluate the effects of combined chemotherapeutic and chemopreventive agents in cancer therapy. Plos one. 2022; 17(11): e0276458.

38. Shiralipour A, Khorsand B, Jafari L, Salehi M, Kazemi M, Zahiri J, et al. Identifying key lysosome-related genes associated with drug-resistant breast cancer using computational and systems biology approach. Iranian Journal of Pharmaceutical Research: IJPR. 2022; 21(1): e130342.

39. Samandari-Bahraseman MR, Hajibarati M, Khorsand B, Soltani N, Esmaeilzadeh-Salestani K, Loit E. Deciphering the biosynthesis pathway of gamma terpinene cuminaldehyde and para cymene in the fruit of Bunium persicum. Scientific Reports. 2025; 15(1): 22438.

40. Olesen F, Hansen RP, Vedsted P. Delay in diagnosis: the experience in Denmark. British journal of cancer. 2009; 101(2): S5-S8.

41. Elston CW, Ellis IO. Pathological prognostic factors in breast cancer. I. The value of histological grade in breast cancer: experience from a large study with long-term follow-up. Histopathology. 1991; 19(5): 403-10.

42. Hourfar H, Taklifi P, Razavi M, Khorsand B. Machine Learning-Driven Identification of Molecular Subgroups in Medulloblastoma via Gene Expression Profiling. Clinical Oncology. 2025: 103789.

43. Khorsand B, Vaghf A, Salimi V, Zand M, Ghoreishi SA. Enhancing ischemic stroke management: leveraging machine learning models for predicting patient recovery after Alteplase treatment. Brain Injury. 2025: 1-7.

44. Khorsand B, Rajabnia M, Jahanian A, Fathy M, Taghvaei S, Houri H. Enhancing the accuracy and effectiveness of diagnosis of spontaneous bacterial peritonitis in cirrhotic patients: a machine learning approach utilizing clinical and laboratory data. Advances in Medical Sciences. 2025; 70(1): 1-7.

45. Hematpour A, Habibi P, Alavimanesh S, Dadkhah K, Babaie K. Machine learning approach to predict protein-protein interactions between human and hepatitis E virus: revealing links to hepatocellular carcinoma. bioRxiv. 2025: 2025.02. 23.639757.

46. Hesami Z, Sabzehali F, Khorsand B, Alipour S, Sadeghi A, Asri N, et al. Microbiota as a State-of-the-art Approach in Precision Medicine for Pancreatic Cancer Management: A Comprehensive Systematic Review. iScience. 2025: 112314.

47. Khorsand B, Hesami Z, Alipour S, Farmani M, Houri H. Harnessing artificial intelligence for detection of pancreatic cancer: a machine learning approach. Clinical and Experimental Medicine. 2025; 25(1): 228.

48. Ding M, Li M, Liu Q, Xu L. Biomarker conversion from primary breast cancer to synchronous axillary lymph node metastasis and neoadjuvant therapy response: a single-center analysis. Journal of Cancer Research and Clinical Oncology. 2024; 150(6): 297.

49. Bleicher RJ, Ruth K, Sigurdson ER, Beck JR, Ross E, Wong Y-N, et al. Time to surgery and breast cancer survival in the United States. JAMA oncology. 2016; 2(3): 330-9.

50. Lashen A, Toss MS, Alsaleem M, Green AR, Mongan NP, Rakha E. The characteristics and clinical significance of atypical mitosis in breast cancer. Modern Pathology. 2022; 35(10): 1341-8.

51. Bareche Y, Venet D, Ignatiadis M, Aftimos P, Piccart M, Rothe F, et al. Unravelling triple-negative breast cancer molecular heterogeneity using an integrative multiomic analysis. Annals of oncology. 2018; 29(4): 895-902.

52. Huang Z, Chen D. A breast cancer diagnosis method based on VIM feature selection and hierarchical clustering random forest algorithm. IEEE Access. 2021; 10: 3284-93.