Iranian Journal of Breast Diseases. 2026; 19(1):64-76

Original Article

Prediction of Breast Cancer Metastasis Using Tree-Based Machine Learning Models: A Retrospective Analysis of Iranian Women

Maryam Khademi[1]✉, Pooneh Khodabakhsh[2], Ziba Heidarpoor[3], Sahba Paktinat[2], Alireza Atashi[4,5]

[1]Department of Applied Mathematics, Islamic Azad University South Tehran Branch, Tehran, Iran
[2]Department of IT and Computer Engineering, Islamic Azad University South Tehran Branch, Tehran, Iran
[3]Doctor of Medicine (MD), Islamic Azad University, Tehran Medical Branch, Tehran, Iran
[4]Medical Informatics Department, Breast Cancer Research Center, Iranian National Cancer Institute, ACECR, Tehran, Iran
[5]Department of Artificial Intelligence, School of Advanced Technologies in Medicine, Tehran University of Medical Sciences, Tehran, Iran

Received: 2025/06/05
Accepted: 2025/10/08

***Corresponding Author:**
maryam_khademi@iau.ac.ir

Ethics Approval:
IR.ACECR.AVICENNA.REC.1404.002

Abstract

Introduction: Breast cancer metastasis is a major cause of cancer-related death. Accurate prediction helps doctors make better decisions. This study developed and evaluated tree-based machine learning models to predict metastasis in Iranian women, using real clinical data with substantial missing values.

Methods: They looked at clinical records of 8,148 breast cancer patients in Tehran from 1997 to 2020. Variables with over 50% missing data were removed, leaving 4,310 complete cases. They compared Decision Tree, Random Forest, and XGBoost (which handles missing data well) against K-NN and Naïve Bayes (which need data imputation). They used stratified 10-fold cross-validation to check for overfitting and class imbalance, and then tested the best models on a separate hold-out set. Performance was measured using AUC, sensitivity, specificity, accuracy, and F1 score.

Results: Tree-based models worked better than the others. XGBoost had the best discrimination (AUC = 0.96, accuracy = 99.4%, F1 = 0.96), and Decision Trees were highly interpretable (sensitivity = 94%, specificity = 96.9%). Even though key predictors such as tumor size were excluded, other variables, such as hormone receptor status and age at menarche, allowed for strong predictions. K-NN had very low sensitivity (6%), and Naïve Bayes was inconsistent.

Conclusion: Decision trees and similar models can reliably predict breast cancer metastasis using incomplete, imbalanced real-world data, provided they are properly validated. These models are good for places with fewer resources. Future work should focus on better data collection and imputation methods. This study demonstrates the utility of interpretable machine learning for cancer applications in underrepresented populations.

Keywords: Breast Cancer, Metastasis, Machine Learning, Decision Trees, XGBoost, Iran

[Downloaded from ijbd.ir on 2026-11-25]

[DOI: 10.66224/ijbd.19.1.64]

Copyright © 2026 Khademi et al. Published by Breast Cancer Research Center, ACECR
This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License (https://creativecommons.org/licenses/by-nc/4.0/) Non-commercial uses of the work are permitted, provided the original work is properly cited.

Introduction

Breast cancer metastasis is a major global health challenge, responsible for most cancer deaths. Accurately identifying high-risk patients is crucial for timely intervention [1]. Machine learning (ML) shows promise in oncology for improving prognostic accuracy. However, many high-performing ML models, like deep learning, are "black boxes," hindering clinical trust. In contrast, tree-based models (Decision Trees, ensemble variants) offer interpretability through explicit rules, which is critical in oncology [2.3]. Existing ML models often lack generalizability, as they are frequently trained on Western or East Asian data and may not represent other populations like Iranian women, who have distinct epidemiological and clinical characteristics that influence metastatic risk [4]. Real-world data from resource-limited settings is often incomplete, posing challenges for ML algorithms that typically require complete datasets or extensive imputation, which can introduce bias. Tree-based methods have the advantage of natively handling missing values and maintaining interpretability [5]. This study aims to develop and validate interpretable tree-based ML models to predict breast cancer metastasis in Iranian women using real-world clinical data with substantial missingness, emphasizing transparency, robustness, and population specificity.

Materials and Methods

The dataset included 8,148 breast cancer patients from Tehran (1997-2020) with 18 variables. Variables with over 50% missing values were excluded, removing predictors like tumor size and HER2 status, resulting in 4,310 complete cases for model development. The retained variables included biologically relevant features, such as hormone receptor status, age at menarche, and menopausal status, which are known to correlate with metastasis. Feature selection involved expert review and L1-regularized logistic regression. The data was split into training (70%) and testing (30%) sets. Stratified sampling was used to split the data, and 10-fold cross-validation was used to address class imbalance, preserving the original data distribution. Five ML algorithms were evaluated: Decision Tree [6], Random Forest, and XGBoost (with native missing-data handling) [7], and K-Nearest Neighbors (K-NN) [8] and Naïve Bayes (after imputation) [9]. Performance was assessed using stratified 10-fold cross-validation and the independent test set, measuring AUC, accuracy, sensitivity, specificity, and F1-score, with emphasis on sensitivity. Figure 1 shows the general schema for a predictive model.

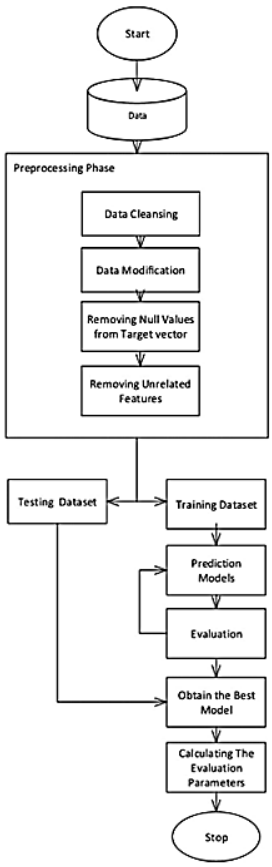

Fig 1: General schema for predictive model

[Downloaded from ijbd.ir on 2026-11-25]

[DOI: 10.66224/ijbd.19.1.64]

Results

As shown in Table 1, tree-based models demonstrated robust, reproducible performance with low fold-to-fold variability. XGBoost achieved the highest discriminative performance (mean AUC = 0.96), while Decision Trees offered a good balance of predictive performance (AUC = 0.95) and interpretability. Traditional classifiers performed poorly. K-Nearest Neighbors had extremely low sensitivity (6%), making it clinically unsuitable, likely due to imputation and class imbalance. Despite excluding key predictors such as tumor size and HER2 status, the tree-based models' high performance suggests that the retained variables (hormone receptor status, reproductive factors) captured sufficient prognostic information. This demonstrates their robustness to missing key predictors.

Table 1: Performance of models for predicting breast cancer metastasis

Classifier	AUC	Accuracy	Sensitivity	Specificity	F1 Score
K-NN (K=4)	0.53	95%	0.06	0.99	0.11
Naïve Bayes	0.94	99%	0.89	0.99	0.88
Decision Tree (Information Gain)	0.95	97%	0.94	0.96	0.81
Random Forest	0.95	99%	0.92	1.0	0.96
XGBoost	0.96	99%	0.92	1.0	0.96

Discussion

This study demonstrates the feasibility of applying ML to real-world clinical oncology data with substantial missingness. The exclusion of tumor size (94.4% missing) highlights data collection limitations, possibly indicating a Missing Not At Random (MNAR) mechanism, which supports the use of models with native missing-data handling over imputation [10]. Retained variables, such as hormone receptor status and age at menarche, provided sufficient prognostic information, indicating that correlated features can compensate for missing primary variables, and that tree-based models leverage them effectively. Tree-based algorithms outperformed K-NN and Naïve Bayes because they can handle missing values [6,7]. While ensemble models like XGBoost had the highest performance, simpler Decision Trees offered greater interpretability [2]. These findings suggest that meaningful predictive models can be developed in resource-limited settings with imperfect data [11]. However, models trained on Iranian data may not generalize to other populations, necessitating external validation. Improved data completeness for variables like tumor size and HER2 status could further enhance performance, building on previous decision tree studies [12].

Conclusion

Interpretable tree-based ML models can reliably predict breast cancer metastasis using real-world clinical data with substantial missingness. These models can aid clinical decision-making by identifying patients who need intensified follow-up or treatment, thereby requiring appropriate clinician training. The exclusion of key prognostic variables highlights data-collection deficiencies that need to be prioritized. Future work should focus on improving documentation, developing hybrid imputation models, and integrating validated models into clinical decision support systems. Clinical deployment must ensure transparency, patient privacy, and external validation to maintain safety and prevent bias, making interpretable ML valuable, especially in resource-limited settings.

References

1. Zhang M, Deng H, Hu R, Chen F, Dong S, Zhan S, et al. Patterns and prognostic implications of distant metastasis in breast Cancer based on SEER population data. *Scientific Reports.* 2025; 15: 26717. doi:10.1038/s41598-025-12883-x.
2. Yang B, Gül M, Chen Y. Comparative analysis of deep learning and tree-based models in power demand prediction: Accuracy, interpretability, and computational efficiency. J Build Phys.

[Downloaded from ijbd.ir on 2026-11-25]

[DOI: 10.66224/ijbd.19.1.64]

2025;49(1):127-69. doi: 10.1177/17442591251333144.
3. Pramanik S, Kumar Bandyopadhyay S. Identifying disease and diagnosis in females using machine learning. In Encyclopedia of Data Science and Machine Learning, pp. 3120-43, 2023. doi: 10.4018/978-1-7998-9220-5.ch187
4. Kasraian L, Ashkani-Esfahani S, Foruozandeh H. Reasons for the under-representation of Iranian women in blood donation. Hematol Transfus Cell Ther. 2021;43(3):256-62. doi: 10.1016/j.htct.2020.03.009.
5. Bhardwaj A, Antil Y, Srivastava MVP, Vinny PW, Vishnu VY, Garg R. A novel machine learning framework for stroke type identification in resource constrained settings with robustness to missing data. Sci Rep. 2025;15(1):31207. doi: 10.1038/s41598-025-16660-8.
6. Costa V. G., Pedreira C.E. Recent advances in decision trees: an updated survey. Artificial Intelligence Review. 2023:56(5):4765-800. doi:10.1007/s10462-022-10275-5
7. Idris N.F., Ismail M.A. A review of homogenous ensemble methods on the classification of breast cancer data. Przegląd Elektrotechniczny, 2024:101-4. doi:10.15199/48.2024.01.21.
8. Suyal M, Goyal P. A review on analysis of k-nearest neighbor classification machine learning algorithms based on supervised learning. International Journal of Engineering Trends and Technology. 2022; 70(7):43-8. doi:10.14445/22315381/IJETT-V70I7P205.
9. Bafjaish S. S., Comparative analysis of Naive Bayesian techniques in health-related for classification task. Journal of Soft Computing and Data Mining. 2020;1(2):1-10. doi:10.30880/jscdm.2020.01.02.001.
10. Bell ML, Floden L, Rabe BA, Hudgens S, Dhillon HM, Bray VJ, et al. Analytical approaches and estimands to take account of missing patient-reported data in longitudinal studies. Patient Relat Outcome Meas. 2019;10:129-40. doi: 10.2147/PROM.S178963.
11. Agrawal M. Cancer prediction using machine learning algorithms. International Journal of Science and Research (IJSR). 2020;9(8):281-6.
12. Razavi M, Wang L, Karssemeijer N, Linsen L, Frese U, Hahn H. et al, Novel Morphological Features for Non-mass-like Breast Lesion Classification on DCE-MRI. Lecture Notes in Computer Science. 2016:305-12. doi:10.1007/978-3-319-47157-0_37.

[Downloaded from ijbd.ir on 2026-11-25]

[DOI: 10.66224/ijbd.19.1.64]

مقاله‌پژوهشی

مجله علمی
بیماری‌های پستان ایران
۱۴۰۵؛۱۹(۱):٦٤-٧٦

پیش‌بینی متاستاز سرطان پستان با استفاده از مدل‌های یادگیری ماشین مبتنی بر درخت: یک تحلیل گذشته‌نگر در میان زنان ایرانی

مریم خادمی[1]✉، پونه خدابخش[2]، زیبا حیدرپور[3]، صهبا پاک طینت[2]، علیرضا آتشی[4,5]

[1] گروه ریاضی کاربردی، دانشگاه آزاد اسلامی واحد تهران جنوب، تهران، ایران
[2] گروه مهندسی فناوری اطلاعات و کامپیوتر، دانشگاه آزاد اسلامی واحد تهران جنوب، تهران، ایران
[3] دکترای پزشکی، دانشگاه آزاد اسلامی واحد پزشکی تهران، تهران، ایران
[4] گروه انفورماتیک پزشکی، مرکز تحقیقات سرطان پستان، پژوهشگاه ملی سرطان، جهاد دانشگاهی (ACECR)، تهران، ایران
[5] گروه هوش‌مصنوعی، دانشکده فناوری‌های نوین در پزشکی، دانشگاه علوم پزشکی تهران، تهران، ایران

چکیده

مقدمه: متاستاز سرطان پستان یکی از علل اصلی مرگ‌ومیر ناشی از سرطان است. پیش‌بینی دقیق پیشرفت متاستاتیک برای تصمیم‌گیری بالینی ضروری است. هدف این مطالعه، توسعه و اعتبارسنجی مدل‌های یادگیری ماشین مبتنی بر درخت برای پیش‌بینی متاستاز سرطان پستان در زنان ایرانی با استفاده از داده‌های بالینی واقعی دارای میزان بالای داده‌های مفقود بوده است.

روش بررسی: این مطالعه‌ی گذشته‌نگر شامل سوابق بالینی ۸۰۱۴۸ بیمار مبتلا به سرطان پستان بود که بین سال‌های ۱۹۹۷ تا ۲۰۲۰ در تهران تحت درمان قرار گرفتند. پس از حذف متغیرهایی که بیش از ۵۰٪ داده‌ی مفقود داشتند و رکوردهای مرتبط با آنها، ۴۰۳۱۰ نمونه‌ی کامل باقی‌ماند (برای مثال، اندازه‌ی تومور دارای ۹۴/۴٪ داده‌ی مفقود بود.) سه مدل درخت تصمیم، جنگل تصادفی و XGBoost (با قابلیت ذاتی در برخورد با داده‌های ناقص) با دو الگوریتم مرجع K-NN و Naïve Bayes، با استفاده از اعتبارسنجی متقابل ده‌تایی مقایسه شدند. عملکرد مدل‌ها با شاخص‌های AUC، حساسیت، ویژگی و امتیاز F1 ارزیابی شد.

یافته‌ها: مدل‌های مبتنی بر درخت عملکرد بهتری نسبت به روش‌های سنتی داشتند؛ XGBoost بالاترین تمایز را نشان داد (AUC=0.96، دقت99.4٪=، F1=0.96) و درخت تصمیم بیشترین قابلیت تفسیر بالینی را ارائه داد (حساسیت 94=٪، ویژگی96.9=٪). علی‌رغم حذف متغیرهای کلیدی مانند اندازه‌ی تومور و وضعیت HER2، متغیرهای باقی‌مانده مانند گیرنده‌های هورمونی و سن شروع قاعدگی توانستند پیش‌بینی‌های دقیقی ارائه دهند. الگوریتم K-NN از نظر بالینی عملکرد ضعیفی داشت (حساسیت=٪۶) در حالی‌که Naïve Bayes ناپایداری نسبی نشان داد (حساسیت=89.01).

نتیجه‌گیری: مدل‌های درخت تصمیم و تجمیع‌شده‌ی آن‌ها می‌توانند به‌طور قابل‌اعتمادی متاستاز را در داده‌های واقعی ناقص پیش‌بینی کنند و از این رو برای محیط‌های با منابع محدود گزینه‌های مناسبی به‌شمار می‌روند. پژوهش‌های آینده باید بر استانداردسازی جمع‌آوری داده‌ها و توسعه‌ی رویکردهای ترکیبی جایگزینی داده‌های مفقود تمرکز داشته باشند. این مطالعه بر اهمیت استفاده از مدل‌های یادگیری ماشین قابل‌تفسیر در کاربردهای انکولوژی، به‌ویژه در جمعیت‌های کمتر مورد مطالعه، تأکید دارد.

واژه‌های کلیدی: سرطان پستان، متاستاز، یادگیری ماشین، درخت تصمیم، XGBoost، ایران

تاریخ ارسال: ۱۴۰۴/۰۳/۱۵
تاریخ پذیرش: ۱۴۰۴/۰۷/۱۶

* **نویسنده مسئول:**
maryam_khademi@iau.ac.ir

[Downloaded from ijbd.ir on 2026-11-25]

[DOI: 10.66224/ijbd.19.1.64]

مقدمه

سرطان پستان همچنان یکی از علل اصلی بروز بیماری و مرگ‌ومیر در سراسر جهان به‌شمار می‌رود، به‌طوری که متاستاز عامل اصلی اکثریت مرگ‌های ناشی از سرطان است (۱). پیش‌بینی دقیق پیشرفت متاستاتیک برای توسعه‌ی راهبردهای درمانی هدفمند حیاتی است.

پیشرفت‌های اخیر در یادگیری ماشین، پیش‌آگهی سرطان را متحول ساخته‌اند و امکان پیش‌بینی‌های دقیق‌تر و شخصی‌سازی‌شده‌تر را فراهم کرده‌اند (۲). با این حال، زنان ایرانی همچنان در متون علمی موجود کمتر مورد بررسی قرار گرفته‌اند، در حالی که ویژگی‌های بالینی و اپیدمیولوژیک متمایزی دارند. روش‌های مختلف یادگیری ماشین، از الگوریتم‌های سنتی نظیر نزدیک‌ترین همسایه (K-Nearest Neighbors) و بیز ساده (Naïve Bayes) (۳) تا مدل‌های پیچیده‌تری مانند شبکه‌های عصبی مصنوعی و ماشین‌های بردار پشتیبان (۴)، در کاربردهای انکولوژی از تفسیر داده‌های ژنتیکی تا تحلیل ریزمحیط تومور عملکرد قابل‌توجهی نشان داده‌اند (۵).

الگوی استاندارد در این‌گونه مدل‌های پیش‌بینی معمولاً شامل داده‌های جامع از متغیرهای اجتماعی-جمعیتی، سبک‌زندگی و بالینی است (۶). علی‌رغم توان بالای این مدل‌ها، یکی از چالش‌های اساسی مطرح‌شده در مرورهای اخیر، مسئله‌ی قابلیت تفسیر است. مدل‌هایی مانند معماری‌های یادگیری عمیق به‌صورت «جعبه‌سیاه» عمل می‌کنند؛ یعنی پیش‌بینی را ارائه می‌دهند بدون آنکه منطق تصمیم‌گیری آنها شفاف باشد. این ابهام مانع پذیرش بالینی مدل‌ها می‌شود، زیرا پزشکان نیاز دارند تا منطق پیش‌بینی را درک کنند تا بتوانند به نتایج مدل اعتماد نمایند. این کمبود اعتماد، به‌ویژه هنگام به‌کارگیری مدل‌ها در جمعیت‌های کمتر مورد مطالعه با ویژگی‌های خاص، نگرانی‌هایی درباره‌ی سوگیری و دقت مدل ایجاد می‌کند (۷، ۸).

با وجود حجم گسترده‌ی پژوهش‌ها در این زمینه، همچنان شکاف قابل‌توجهی در تعمیم‌پذیری مدل‌ها برای جمعیت‌های خاص با ویژگی‌های بالینی و اپیدمیولوژیک متفاوت وجود دارد. به‌عنوان نمونه، زنان ایرانی در متون علمی موجود به‌طور جدی کمتر مورد توجه قرار گرفته‌اند (۹). علاوه بر این، داده‌های بالینی واقعی، به‌ویژه در محیط‌های دارای محدودیت منابع، معمولاً با نرخ بالای داده‌های مفقود مشخص می‌شوند (۱۰) که این امر محدودیتی جدی برای مدل‌هایی است که به داده‌های کامل نیاز دارند.

برای پر کردن این شکاف‌ها، این مطالعه از الگوریتم‌های مبتنی بر درخت تصمیم برای پیش‌بینی متاستاز در میان زنان ایرانی استفاده کرده است. این الگوریتم‌ها مزایای ویژه‌ای در قابلیت تفسیر بالینی دارند، چرا که مسیرهای تصمیم‌گیری را از طریق قوانین طبیعی و قابل فهم برای پزشکان ارائه می‌کنند. ساختار گرافیکی و خروجی‌های قاعده‌محور آنها، درک فوری منطق پیش‌بینی را ممکن می‌سازد؛ ویژگی‌ای که در تصمیم‌گیری‌های پزشکی بسیار ارزشمند است. در مقابل، معماری‌های یادگیری عمیق با وجود توان بالا، از شبکه‌های عصبی پیچیده تشکیل‌شده‌اند که فرآیند تصمیم‌گیری در آنها غیرقابل تفسیر باقی می‌ماند، و این امر چالشی اساسی برای کاربردهای بالینی محسوب می‌شود (۷).

در این پژوهش، چندین الگوریتم مبتنی بر درخت و روش‌های مرجع یادگیری ماشین با استفاده از یک مجموعه‌داده‌ی ملی شامل ۸،۱۴۸ بیمار مبتلا به سرطان پستان که بین سال‌های ۱۹۹۷ تا ۲۰۲۰ در کلینیک‌های تهران درمان شده‌اند، مقایسه گردیدند. با تأکید بر قابلیت تفسیر و مقاومت در برابر داده‌های ناقص، هدف این مطالعه شناسایی مدلی است که ضمن حفظ دقت پیش‌بینی، برای استفاده‌ی بالینی در میان زنان ایرانی مبتلا به سرطان پستان نیز عملی و قابل‌اعتماد باشد.

مواد و روش‌ها

در این پژوهش از روش‌های یادگیری ماشین بر روی داده‌های بالینی تاریخی ۸۱۴۸ زن ایرانی مبتلا به سرطان پستان استفاده شده است. داده‌ها از چندین مرکز درمانی سطح سوم در تهران گردآوری و توسط گروه پژوهشی انفورماتیک در مؤسسه‌ی سرطان معتمد بین سال‌های ۱۹۹۷ تا ۲۰۲۰ تجمیع شده‌اند. هدف از تحلیل، پیش‌بینی وقوع متاستاز در بیماران بود. مجموعه‌داده‌ی اولیه شامل ۸۱ متغیر بالینی، جمعیت‌شناختی و پاتولوژیک بود.

مطابق با جدول ۱، تحلیل جامع داده‌های مفقود نشان داد که چندین متغیر کلیدی دارای درصد بالایی از داده‌های گمشده بودند؛ به‌طور خاص، متغیر اندازه‌ی تومور دارای ۹۴/۳۶٪ داده‌ی مفقود بود. بر اساس راهنمایی متخصصان

[Downloaded from ijbd.ir on 2026-11-25]

[DOI: 10.66224/ijbd.19.1.64]

بالینی و معیارهای آماری، آستانه‌ای معادل ۵۰٪ برای داده‌های مفقود تعیین شد؛ به این معنا که هر متغیری که بیش از این میزان داده‌ی مفقود داشت، به‌همراه رکوردهای مرتبط حذف گردید تا یکپارچگی آماری تحلیل حفظ شود.

جدول ۱: متغیرهای مجموعه داده

Table 1: Distribution of Dataset variables

Characteristic	Category	Value
Total Patients (N)		4310
	Mean (SD)	47.3 (11.1)
Age at Diagnosis	Median [Min, Max]	46.0 [21.0, 89.0]
Age at Menarche	Mean (SD)	13.4 (1.5)
	Median [Min, Max]	13.0 [8.0, 20.0]
Education Level	Elementary	486 (11.3%)
	High School	1462 (33.9%)
	Some College	1303 (30.2%)
	University Degree	807 (18.7%)
Marital Status	Single	242 (5.6%)
	Married	3491 (81.0%)
	Other	504 (11.7%)
Breastfeeding Duration	Median [IQR] (months)	42.0 [21.0, 72.0]
	Never Breastfed (0 months)	248 (5.8%)
Oral Contraceptive Use	Never Used	2404 (55.8%)
	Used (Any Duration)	1906 (44.2%)
	Median Duration (users only)	30 months
Smoking History	Non-Smoker	3692 (85.7%)
	Current Smoker	145 (3.4%)
	Former Smoker	334 (7.7%)
Personal History of Cancer	Yes	1724 (40.0%)
Other Personal History	Yes	31 (0.7%)
Body Mass Index (BMI)	Mean (SD)	28.2 (4.8) kg/mÂ²
	<18.5 (Underweight)	33 (0.8%)
	18.5-24.9 (Normal)	1031 (23.9%)
	25-29.9 (Overweight)	1667 (38.7%)
	>= 30 (Obese)	1267 (29.4%)
Overall Stage	1	519 (12.0%)
	2	1727 (40.1%)
	3	1130 (26.2%)
	4	310 (7.2%)
Tumor Size	<2 cm	814 (18.9%)
	2-5 cm	2180 (50.6%)
	>5 cm	608 (14.1%)
Surgical Margin Status	Negative	2564 (59.5%)
	Close	134 (3.1%)
	Positive	116 (2.7%)
	Not Reported/Unknown	1489 (34.5%)
Estrogen Receptor (ER)	Negative	1125 (26.1%)
	Positive	2497 (57.9%)
	Unknown	688 (16.0%)
Progesterone Receptor (PR)	Negative	1281 (29.7%)
	Positive	2302 (53.4%)
	Unknown	727 (16.9%)
HER2 Status	0 (Negative)	1426 (33.1%)
	1+ (Negative)	321 (7.4%)
	2+ (Borderline)	340 (7.9%)
	3+ (Positive)	915 (21.2%)
	Unknown	1308 (30.3%)
Metastasis	Absent	3961 (91.9%)
	Present	349 (8.1%)

[Downloaded from ijbd.ir on 2026-11-25]

[DOI: 10.66224/ijbd.19.1.64]

پس از مرحله‌ی پیش‌پردازش، متغیرهایی که از آستانه‌ی ۵۰٪ داده‌ی مفقود فراتر رفته بودند، به‌همراه رکوردهای مربوطه حذف شدند؛ در نتیجه، مجموعه‌داده‌ی نهایی شامل ۴،۳۱۰ مورد کامل بود که مبنای تمام مدل‌سازی‌های بعدی قرار گرفت) جدول ۲).

جدول ۲: خطر مؤثر در بروز سرطان پستان

Table 2: Influential risk factors of breast cancer

Category	Risk factors	Missing values	Proportion of Missing Values (n= 4310)
None Modifiable Risk Factors	Age of Menarche	185	0.042923
	Age of Diagnosis	17	0.003944
	Estrogen Receptor (ER)	688	0.159629
	Progesterone Receptor (PR)	727	0.168677
	History of Other Cancers	110	0.025522
	History of Breast Cancer	94	0.02181
Modifiable Risk Factors	Body Mass Index (BMI)	312	0.07239
	Smoking	127	0.029466
	Education level	179	0.041531
	Marital Status	66	0.015313
	Oral contraceptive pill (OCP) use	2168	0.503016
	Breastfeeding	501	0.116241
Clinical Pathological Risk Factors	Location of tumor	1810	0.419954
	Cancer Stage	548	0.127146
	Margin of Tumor	1489	0.345476
	Tumor size	4067	0.943619
	Cancer Type	546	0.126682
	Her2 marker	1308	0.30348

در جدول ۲، متغیرهای باقی‌مانده بر اساس سه حوزه‌ی اصلی دسته‌بندی شدند، ۱) عوامل غیرقابل‌تغییر (مانند سن تشخیص و وضعیت گیرنده‌های هورمونی)، ۲) عوامل قابل‌تغییر (مانند شاخص توده بدنی، مصرف سیگار و قرص‌های ضدبارداری)، ۳) عوامل کلینیکی-پاتولوژیک (مانند نوع سرطان و مرحله‌ی بیماری). قابل‌ذکر است که وضعیت گیرنده‌های ER و PR با وجود نزدیک بودن به آستانه‌ی حذف، به‌دلیل ارزش پیش‌آگهی تثبیت‌شده‌شان در سرطان پستان حفظ شدند.

همان‌طور که در شکل ۱ نشان داده شده است، فرآیند انتخاب ویژگی در دو مرحله انجام شد، در گام نخست، متخصصان بالینی متغیرهای زیستی محتمل را شناسایی کردند. در گام دوم، رگرسیون لجستیک با منظم‌سازی L1 برای انتخاب مؤثرترین ویژگی‌های پیش‌بینی‌کننده به کار رفت. سپس داده‌ها به‌صورت تصادفی به نسبت ۷۰٪ برای آموزش و ۳۰٪ برای آزمون مستقل تقسیم شدند. پنج الگوریتم یادگیری ماشین با رویکردهای متفاوت در برخورد با داده‌های ناقص پیاده‌سازی شدند:

- درخت تصمیم (۱۱) و روش‌های تجمیعی مانند جنگل تصادفی و (۱۲) XGBoost از سازوکارهای ذاتی برای مدیریت داده‌های ناقص (نظیر انشعابات جایگزین و تقسیم‌بندی‌های حساس به خلأ) استفاده کردند.
- در مقابل، الگوریتم‌های (۱۳) K-NN و Naïve Bayes (۱۴) نیاز به داده‌های کامل داشتند و داده‌های مفقود در آن‌ها از طریق میانه یا مد جایگزین شدند.

تمام مدل‌ها با استفاده از اعتبارسنجی متقابل طبقه‌بندی‌شده ده‌تایی-(10-fold stratified cross validation) جهت کنترل عدم توازن کلاسی توسعه یافتند. عملکرد نهایی براساس شاخص‌های AUC، Sensitivity، Specificity و F1 Score با تأکید ویژه بر Sensitivity به‌دلیل اهمیت بالینی تشخیص موارد متاستاز، ارزیابی گردید.

[Downloaded from ijbd.ir on 2026-11-25]

[DOI: 10.66224/ijbd.19.1.64]

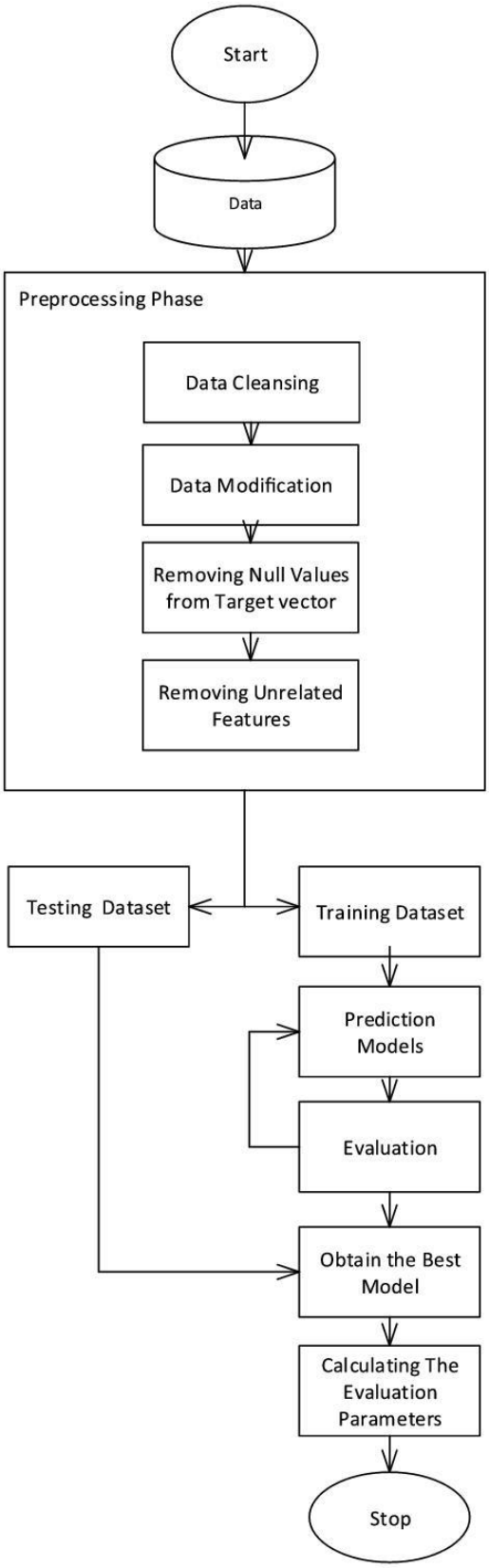

Fig 1: General schema for predictive model

شکل ۱: طرح کلی مدل پیشنهادی

[Downloaded from ijbd.ir on 2026-11-25]

[DOI: 10.66224/ijbd.19.1.64]

نتایج

در این مطالعه داده‌های مربوط به ۴،۳۱۰ بیمار مبتلا به سرطان پستان مورد بررسی قرار گرفت که میانگین سنی آنان ۴۷ سال بود. از میان این بیماران، ۱،۵۷۰ زن یائسه بودند که سن شروع یائسگی در آن‌ها بین ۲۹ تا ۶۵ سال متغیر بود. از مجموع بیماران، ۵۳٪ دارای گیرنده‌ی پروژسترون مثبت (+PR) و ۵۸٪ دارای گیرنده‌ی استروژن مثبت (+ER) بودند.

مطابق با جدول ۳، عملکرد مدل‌ها بسته به نوع الگوریتم تفاوت قابل‌توجهی نشان داد. روش‌های مبتنی بر درخت، به‌ویژه در مواجهه با چالش داده‌های ناقص، پایداری و دقت بالاتری نسبت به سایر مدل‌ها از خود نشان دادند. در این میان، الگوریتم XGBoost بالاترین میزان تمایز (AUC=0.96) و دقت کلی (Accuracy=99.4%) را به‌دست آورد، در حالی که درخت تصمیم با وجود دقت اندکی کمتر، بهترین قابلیت تفسیر بالینی را ارائه کرد (AUC=0.95، Sensitivity=94%، Specificity=96.9%).

تحلیل اهمیت متغیرها نشان داد که وضعیت گیرنده‌های هورمونی (ER و PR) با وجود میزان نسبتاً بالای داده‌های مفقود (حدود ۳۹٪) از مهم‌ترین پیش‌بینی‌کننده‌ها در مدل‌ها بودند. نکته‌ی مهم آن است که متغیرهای حذف‌شده مانند اندازه‌ی تومور و وضعیت HER2، که از شناخته شده‌ترین عوامل مؤثر در متاستاز هستند، در این مجموعه‌داده به‌دلیل درصد بالای داده‌های مفقود حذف شدند. بنابراین، عملکرد فعلی مدل‌ها می‌تواند به‌عنوان حد پایین (Lower Bound) از آنچه در شرایط داده‌ی کامل قابل دستیابی است، تلقی شود.

جدول ۳: عملکرد مدل‌ها در پیش‌بینی متاستاز سرطان پستان

Table 3: Performance of models for predicting breast cancer metastasis

Classifier	AUC	Accuracy	Sensitivity	Specificity	F1 Score
K-NN (K=4)	0.53	95%	0.06	0.99	0.11
Naïve Bayes	0.94	99%	0.89	0.99	0.88
Decision Tree (Information Gain)	0.95	97%	0.94	0.96	0.81
Random Forest	0.95	99%	0.92	1.0	0.96
XGBoost	0.96	99%	0.92	1.0	0.96

همان‌طور که در جدول ۳ می‌توان دید، روش‌های سنتی در مواجهه با داده‌های ناقص عملکرد مطلوبی نداشتند. الگوریتم K-NN به‌طور کامل شکست خورد) با Sensitivity تنها ۶٪) ، در حالی که مدل Naïve Bayes عملکردی نسبتاً مناسب اما غیر پایدار از خود نشان داد (٪۸۹ Sensitivity =) که در تکرارهای مختلف اعتبارسنجی متقابل نوسانات چشمگیری داشت. این مسئله احتمالاً ناشی از دشواری جایگزینی داده‌ها در شرایطی است که حدود ۴۰٪ از مقادیر کلیدی پیش‌بینی‌کننده ناقص هستند.

بحث

یافته‌های این پژوهش چند بینش کلیدی در خصوص مدل‌سازی پیش‌بینانه در داده‌های واقعی انکولوژی ارائه می‌دهد. نخست، ضرورت حذف متغیر اندازه‌ی تومور، که احتمالاً مهم‌ترین عامل پیش‌آگهی در سرطان پستان است، به‌دلیل ۹۳٪ داده‌ی مفقود، چالش‌های بنیادین کیفیت داده در نظام‌های بالینی را برجسته می‌سازد. این الگوی داده‌های ناقص احتمالاً بازتاب‌دهنده‌ی ناهماهنگی در ثبت یا اندازه‌گیری اطلاعات است نه فقدان واقعی داده، که منجر به ایجاد وضعیتی موسوم به «داده‌های مفقود تصادفی‌نبود (MNAR)» می‌شود، وضعیتی که روش‌های استاندارد جایگزینی داده نمی‌توانند آن را به‌درستی مدیریت کنند (۱۵).

با این وجود، حتی در این شرایط، متغیرهای باقی‌مانده مانند وضعیت گیرنده‌های هورمونی و سن شروع قاعدگی همچنان ارزش پیش‌بینی‌کنندگی خود را حفظ کردند. این یافته نشان‌می‌دهد که متغیرهای ثانویه می‌توانند تا حدی کمبود اطلاعات اولیه را جبران کنند، زمانی که داده‌های اصلی در دسترس نیستند.

عملکرد برتر مدل‌های مبتنی بر درخت با پیش‌بینی‌های نظری در این حوزه همخوانی دارد. توانایی ذاتی این مدل‌ها در مدیریت داده‌های ناقص از طریق سازوکارهایی مانند

[Downloaded from ijbd.ir on 2026-11-25]

[DOI: 10.66224/ijbd.19.1.64]

انشعابات جایگزین (surrogate splits) در درخت تصمیم و الگوریتم‌های حساس به خلأ (sparsity-aware) در XGBoost، مزیت مشخص نسبت به مدل‌هایی دارد که به داده‌های کامل وابسته هستند. با این حال، قابلیت تفسیر بالینی درخت تصمیم با هزینه‌ی اندکی در دقت همراه بود، تقریباً ۲ تا ۳ درصد کاهش در مقدارAUC نسبت به مدل‌های تجمیعی (ensemble)، که نشان‌دهنده‌ی یک تبادل عملی میان دقت و قابلیت تفسیر در کاربردهای پزشکی است.

در میان ویژگی‌های پاتولوژیک بالینی، متغیر اندازه‌ی تومور نقش بسیار مهمی داشت و مشاهده شد که حدود ۲۴٪ از تومورها در ربع فوقانی خارجی پستان (UOQ) قرار داشتند. عوامل دیگر مؤثر شامل وضعیت -HER2 مثبت و نوع بافت تومور بودند، که اکثریت موارد متعلق به کارسینوم مجرایی مهاجم (IDC) بودند (۱۶). سایر ویژگی‌های مهم نیز شامل وضعیت گیرنده‌های هورمونی و نوع سرطان بودند. مطالعات اپیدمیولوژیک پیشین هر دو گروه عوامل غیرقابل‌تغییر (مانند وضعیت گیرنده‌های هورمونی) و قابل‌تغییر (مانند عوامل سبک‌زندگی) را به‌عنوان عوامل خطر شناسایی کرده‌اند (۱۷-۲۲). مجموعه‌داده‌ی حاضر در مقایسه با آمار جهانی تفاوت‌هایی از نظر میزان شیوع گیرنده‌های مثبت هورمونی نشان داد.

مطالعات پیشین پیرامون عوامل خطر متاستاز اغلب بر متاستاز استخوانی تمرکز داشته‌اند، زیرا این نوع متاستاز از شیوع بالاتری برخوردار است. برای مثال، مطالعه‌ی یزدانی و همکاران بر نقش سن، تهاجم غدد لنفاوی و اندازه‌ی تومور در متاستاز استخوانی تأکید داشت (۲۳). یافته‌های ما این نتایج را گسترش می‌دهد و پیش‌بینی متاستاز را به‌صورت کلی‌تر در انواع مختلف متاستازها بررسی می‌کند.

مطالعات مرتبط دیگر نیز از الگوریتم‌های یادگیری ماشین برای پیش‌بینی بقای بیماران یا احتمال متاستاز کلی استفاده کرده‌اند، اما با مقادیر حساسیت پایین‌تر. به‌عنوان نمونه، مطالعه‌ی تپاک (Tapak) تنها یکی از نمونه‌های این دسته است (۲۴). همچنین رضوی و همکاران استفاده از مدل درخت تصمیم را برای پیش‌بینی متاستاز سرطان پستان پیشنهاد داده بودند، هرچند داده‌ی آن مطالعه محدود بود (۲۵). یافته‌های کنونی ما اثربخشی مدل درخت تصمیم را در یک مجموعه‌داده‌ی واقعی با وجود چالش داده‌های مفقود تأیید می‌کند و برخی از محدودیت‌های مطالعات قبلی را برطرف می‌سازد.

از محدودیت‌های این پژوهش می‌توان به از دست رفتن بخشی از اطلاعات به‌دلیل حذف رکوردهای ناقص اشاره کرد که ناگزیر منجر به کاهش حجم داده‌ها شد. با این حال، استفاده از مدل درخت تصمیم این امکان را فراهم ساخت که داده‌های باقی‌مانده بدون نیاز به جایگزینی مصنوعی حفظ شوند و یکپارچگی مجموعه‌داده برقرار بماند. لازم به ذکر است که نتایج به‌دست‌آمده محدود به پایگاه داده‌ی فعلی هستند و ممکن است قابلیت تعمیم به سایر جمعیت‌ها را نداشته‌باشند. برای پژوهش‌های آینده، پیشنهاد می‌شود داده‌ها از چشم‌انداز بین‌المللی و با حجم نمونه‌ی بیشتر گردآوری شوند تا پیش‌بینی متاستاز و عود سرطان با دقت و تعمیم‌پذیری بالاتری انجام گیرد.

نتایج این مطالعه پیامدهای مهمی برای عمل بالینی و روش‌شناسی پژوهش دارد، نخست، بر ضرورت استانداردسازی ثبت متغیرهای پیش‌آگهی کلیدی مانند اندازه‌ی تومور و وضعیتHER2 تأکید می‌کند. دوم، نشان‌می‌دهد که حتی در شرایط وجود داده‌های ناقص، می‌توان با طراحی دقیق مدل‌های مناسب، الگوهای بالینی مفیدی استخراج کرد، موضوعی که برای محیط‌های با منابع محدود اهمیت ویژه‌ای دارد. سوم، نتایج نشان‌می‌دهد که رویکردهای رگرسیون سنتی ممکن است برای داده‌های واقعی انکولوژی که با میزان بالای داده‌ی مفقود مواجه هستند، مناسب نباشند، مگر آن‌که از روش‌های پیشرفته‌ی مدیریت داده‌ی ناقص بهره گیرند.

نتیجه‌گیری

این تحلیل جامع از کاربرد یادگیری ماشین در پیش‌بینی متاستاز سرطان پستان با داده‌های واقعی و دارای مقادیر مفقود، به سه نتیجه‌ی اصلی منتهی شد، نخست، روش‌های مبتنی بر درخت، به‌ویژه XGBoost و درخت تصمیم، نشان دادند که در برابر چالش داده‌های ناقص از پایداری و مقاومت چشمگیری برخوردارند، در حالی‌که قابلیت تفسیر بالینی خود را نیز حفظ می‌کنند. دوم، حذف متغیرهای پیش‌بینی‌کننده‌ی کلیدی مانند اندازه‌ی تومور و وضعیت HER2 به‌دلیل میزان بالای داده‌ی مفقود، هم‌زمان نشان‌دهنده‌ی محدودیت سیستم‌های فعلی ثبت داده‌های بالینی و نیز فرصتی برای بهبود کیفیت داده‌ها در آینده

[Downloaded from ijbd.ir on 2026-11-25]

[DOI: 10.66224/ijbd.19.1.64]

است. سوم، عملکرد قابل‌قبول مدل‌ها با وجود داده‌های ناقص بیانگر آن است که چنین رویکردهایی می‌توانند در محیط‌های بالینی با منابع محدود نیز کاربرد فوری داشته‌باشند و در عین حال، پایه‌ای برای توسعه‌ی مدل‌های پیشرفته‌تر با داده‌های کامل‌تر فراهم سازند.

پیشنهاد می‌شود جهت تحقیقات آینده، تمرکز بر محورهای زیر باشد،

۱. اجرای مداخلات بیمارستانی برای بهبود مستندسازی متغیرهای پیش‌آگهی کلیدی،
۲. توسعه‌ی رویکردهای ترکیبی مدل‌سازی که روش‌های مبتنی بر درخت را با جایگزینی انتخابی چندگانه داده‌های مفقود تلفیق نماید،
۳. طراحی سیستم‌های پشتیبان تصمیم بالینی که بتوانند به‌طور شفاف عدم قطعیت ناشی از داده‌های ناقص را در فرآیند تصمیم‌گیری در نظر بگیرند.
۴. به‌کارگیری این مدل‌های پیش‌بینانه در عمل بالینی، نیازمند همکاری نزدیک میان دانشمندان داده و پزشکان است تا تفسیر و استفاده‌ی مناسب از نتایج، با در نظر گرفتن محدودیت‌های ذاتی داده‌ها، تضمین شود.

سپاس‌گزاری

این مطالعه مطابق با اصول اخلاقی انجام شده و توسط کمیته‌ی اخلاق مؤسسه‌ی سرطان معتمد تأیید شده است. کد تأیید اخلاقی: (IR.ACECR.AVICENNA.REC.1404.002). تمام داده‌های بیماران به‌صورت ناشناس مورد استفاده قرار گرفتند و رضایت آگاهانه‌ی کتبی به‌دلیل ماهیت گذشته‌نگر پژوهش لغو گردید.

تعارض منافع

هیچ تعارض منافعی وجود ندارد.

References

[1] Zhang M, Deng H, Hu R, Chen F, Dong S, Zhan S, et al. Patterns and prognostic implications of distant metastasis in breast Cancer based on SEER population data. *Scientific Reports*.2025; 15: 26717. doi:10.1038/s41598-025-12883-x

[2] Shirzad M, Shaban M, Mohammadzadeh V, Rahdar A, Fathi-karkan S, Hoseini ZS, et al. Artificial Intelligence-Assisted Design of Nanomedicines for Breast Cancer Diagnosis and Therapy: Advances, Challenges, and Future Directions. BioNanoScience. 2025;15(3):354. doi:10.1007/s12668-025-01980-w

[3]Agrawal M. Cancer prediction using machine learning algorithms. International Journal of Science and Research (IJSR). 2020;9(8):281-6.

[4] Abdul kareem, S. A., & Rasheed, Z. F. (2023). A A Machine Learning Model for Cancer Disease Diagnosis using Gene Expression Data. Journal of Kufa for Mathematics and Computer. 2023;10(2): 179-85. doi:10.31642/JoKMC/2018/100227

[5] Sharma A, Rani R. Machine learning perspective in cancer research. In Handbook of research on disease prediction through data analytics and machine learning. IGI Global Scientific Publishing. 2021:142-63. doi:10.4018/978-1-7998-2742-9.ch008.

[6] Sharma A, Rani R. A systematic review of applications of machine learning in cancer prediction and diagnosis. Archives of Computational Methods in Engineering. 2021;28(7):4875-96. doi:10.1007/s11831-021-09556-z

[7] Yang B, Gül M, Chen Y. Comparative analysis of deep learning and tree-based models in power demand prediction: Accuracy, interpretability, and computational efficiency. J Build Phys. 2025;49(1):127-69. doi: 10.1177/17442591251333144.

[8] Pramanik S, Kumar Bandyopadhyay S. Identifying disease and diagnosis in females using machine learning. *In Encyclopedia of Data Science and Machine Learning,* pp. 3120-43, 2023. doi: 10.4018/978-1-7998-9220-5.ch187

[9]Kasraian L, Ashkani-Esfahani S, Foruozandeh H. Reasons of under-representation of Iranian women in blood donation. Hematol Transfus Cell Ther. 2021;43(3):256-62. doi: 10.1016/j.htct.2020.03.009.

[10]Bhardwaj A, Antil Y, Srivastava MVP, Vinny PW, Vishnu VY, Garg R. A novel machine learning framework for stroke type identification in resource constrained settings with robustness to missing data. Sci

[Downloaded from ijbd.ir on 2026-11-25]

[DOI: 10.66224/ijbd.19.1.64]

Rep. 2025;15(1):31207. doi: 10.1038/s41598-025-16660-8.
[11] Costa V. G., Pedreira C.E. Recent advances in decision trees: an updated survey. Artificial Intelligence Review. 2023:56(5):4765-800. doi:10.1007/s10462-022-10275-5
[12] Idris N.F., Ismail M.A. A review of homogenous ensemble methods on the classification of breast cancer data. Przegląd Elektrotechniczny, 2024:101-4. doi:10.15199/48.2024.01.21
[13] Suyal M, Goyal P. A review on analysis of k-nearest neighbor classification machine learning algorithms based on supervised learning. International Journal of Engineering Trends and Technology. 2022; 70(7):43-8. doi:10.14445/22315381/IJETT-V70I7P205
[14] Bafjaish S. S., Comparative analysis of Naive Bayesian techniques in health-related for classification task. Journal of Soft Computing and Data Mining. 2020;1(2):1-10. doi:10.30880/jscdm.2020.01.02.001
[15] Bell ML, Floden L, Rabe BA, Hudgens S, Dhillon HM, Bray VJ, et al. Analytical approaches and estimands to take account of missing patient-reported data in longitudinal studies. Patient Relat Outcome Meas. 2019;10:129-40. doi: 10.2147/PROM.S178963.
[16] Arafat HM, Omar J, Muhamad R, Al-Astani TAD, Shafii N, Al Laham NA, et al. Breast Cancer Risk From Modifiable and Non-Modifiable Risk Factors among Palestinian Women: A Systematic Review and Meta-Analysis. Asian Pac J Cancer Prev. 2021;22(7):1987-95. doi: 10.31557/APJCP.2021.22.7.1987.
[17] Youn HJ, Han W. A Review of the Epidemiology of Breast Cancer in Asia: Focus on Risk Factors. Asian Pac J Cancer Prev. 2020;21(4):867-80. doi: 10.31557/APJCP.2020.21.4.867.
[18] Ho PJ, Lau HSH, Ho WK, Wong FY, Yang Q, Tan KW, et al. Incidence of breast cancer attributable to breast density, modifiable and non-modifiable breast cancer risk factors in Singapore. Sci Rep. 2020;10(1):503. doi: 10.1038/s41598-019-57341-7.
[19] Daly AA, Rolph R, Cutress RI, Copson ER. A Review of Modifiable Risk Factors in Young Women for the Prevention of Breast Cancer. Breast Cancer (Dove Med Press). 2021;13:241-57. doi: 10.2147/BCTT.S268401.
[20] Dadziak M, Olko P, Zapala M A, Hunek A, Chmielarz K, Wiśiewska-Skomra J, et al. The non-modifiable risk factors for breast cancer development in women. *Journal of Education, Health and Sport*. Online. 2023.;25(1): 134-46. doi: 10.12775/JEHS.2023.25.01.012.
[21] Vishwakarma G, Mehta A, Saifi M, Garg D, Paliwal D. Modifiable (Sleeping Pattern and Stress) and Non-Modifiable Risk Factors Associated with Breast Cancer: A Matched Case-Control Study in Delhi, India. Asian Pac J Cancer Prev. 2022;23(7):2469-76. doi: 10.31557/APJCP.2022.23.7.2469.
[22] Bastos D. R. d. Risk factors related to breast cancer development. Mastology. 2019;29(4):218–23. doi:10.29289/2594539420190000461
[23] Yazdani A, Dorri S, Atashi A, Shirafkan H, Zabolinezhad H. Bone Metastasis Prognostic Factors in Breast Cancer. Breast Cancer (Auckl). 2019;13:1178223419830978. doi: 10.1177/1178223419830978.
[24]Tapak L, Shirmohammadi-khoram N, Amini P, Poorolajal J. P. Prediction of survival and metastasis in breast cancer patients using machine learning classifiers. Clinical Epidemiology and Global Health. 2019;7(3):293-9.
[25]Razavi M, Wang L, Karssemeijer N, Linsen L, Frese U, Hahn H. et al, Novel Morphological Features for Non-mass-like Breast Lesion Classification on DCE-MRI. Lecture Notes in Computer Science. 2016:305-12. doi:10.1007/978-3-319-47157-0_37.
[26]Jakkanwar B. S. Review on Multiple Cancer Disease Prediction And Identification using Machine Learning Techniques. International Journal for Research in Applied Science and Engineering Technology. 2023;11(6):1333-7. doi:10.22214/ijraset.2023.53112
[27] Mao L, Wang H, Hu LS, Tran NL, Canoll PD, Swanson KR, Li J. Knowledge-Informed Machine Learning for Cancer Diagnosis and Prognosis: A Review. IEEE Trans Autom Sci Eng. 2025;22:10008-28. doi:10.1109/tase.2024.3515839.

[Downloaded from ijbd.ir on 2026-11-25]
[DOI: 10.66224/ijbd.19.1.64]