Non comunicable disease

# Machine learning techniques to identify risk factors of breast cancer among women in Mashhad, Iran

ATIEH KHALEGHI[1], SEYYED MOHAMMAD TABATABAEI[2,3], ZEINAB SADAT HOSSEINI[4], MOSLEM TAHERI SOODEJANI[5], EHSAN MOSA FARKHANI[6], MARYAM YAGHOOBI[3,7]

[1] Data Scientist at AFRY AB, Sweden; [2] Department of Medical Informatics, Faculty of Health Mashhad University of Medical Sciences, Mashhad, Iran; [3] Clinical Research Development Unit, Imam Reza Hospital, University of Medical Sciences, Mashhad, Iran; [4] Faculty of Medicine, Islamic Azad University of Mashhad, Mashhad, Iran; [5] Research Center of Prevention and Epidemiology of Non-Communicable Disease, Department of Biostatistics and Epidemiology, Shahid Sadoughi University of Medical Sciences, Yazd, Iran; [6] Department of Epidemiology, Faculty of Health Mashhad University of Medical Sciences, Mashhad, Iran; [7] Department of Epidemiology, Faculty of Public Health, Iran University of Medical Sciences, Tehran, Iran

## Keywords

Random forest • Logistic regression • Decision tree • Principal component analysis • Breast cancer

## Summary

**Background**. *Low survival rates of breast cancer in developing countries are mainly due to the lack of early detection plans and adequate diagnosis and treatment facilities.*
**Objectives**. *This study aimed to apply machine learning techniques to recognize the most important breast cancer risk factors.*
**Methods**. *This case-control study included women aged 17-75 years who were referred to medical centers affiliated with Mashhad University of Medical Science between March 21, 2015, and March 19, 2016. The study had two datasets: one with 516 samples (258 cases and 258 controls) and another with 606 samples (303 cases and 303 controls). Written informed consent has been observed. Decision Tree (DT), Random Forest (RF), Logistic Regression (LR), and Principal Component Analysis (PCA) were applied using R studio software.*
**Results**. *Regarding the DT and RF, the most important features that impact breast cancer were family cancer, individual history of breast cancer, biopsy sampling, rarely consumption of a dairy, fruit, and vegetable meal, while in PCA and LR these features including family cancer, pregnancy number, pregnancy tendency, abortion, first menstruation, the age of first childbirth and childbirth number.*
**Conclusions**. *Machine learning algorithms can be used to extract the most important factors in the diagnosis of breast cancer in developing countries such as Iran.*

## Introduction

The most prevalent malignancy of women in the world is breast cancer [1]. Globally, there were 2.3 million new instances of breast cancer in 2020, and this disease claimed 685,000 lives [2]. By 2040, it is expected that there will be more than 3 million new cases and 1 million deaths due to breast cancer worldwide [3]. Breast cancer has been identified as the fifth-leading cause of mortality among Iranian women [4]. In Iran, 15,492 new cases of breast cancer were diagnosed in 2022, making up 11.3% of all new cancer cases [5]. The exact causes of breast cancer are still unknown, but there are both modifiable and non-modifiable risk factors associated with this malignancy [6]. Some of the main non-modifiable risk factors include aging [7], being a woman [8], having a family history of breast cancer [9], possessing certain genetic mutations [10], late menopause, and early menarche [11]. On the other hand, modifiable risk factors for breast cancer include obesity [8], alcohol, smoking [12], postmenopausal hormone therapy [13], and being single [14].
Early detection of breast cancer can lead to lower treatment costs and more effective treatment. Thus, knowledge about the main risk factors that affect this mysterious cancer is a crucial task [15].

Machine learning algorithms have a significant impact on the healthcare system by analyzing large data sets, and assisting to make decisions about a patient's health status, disease progression, and optimal treatment plans [16]. Several machine learning algorithms are used in the medical field for prediction and classification including Decision Trees (DTs), Logistic Regression (LR), random forests (RFs), Principal Component Analysis (PCA), and Support Vector Machines (SVMs) [16]. Decision trees use a tree-like structure to represent decisions and their potential outcomes [17], while logistic regression predicts the probability of an observation belonging to a specific class [18]. Random forest is a method that improves the accuracy and robustness of predictions by combining multiple decision trees [19]. Principal component analysis is an algorithm that reduces the number of dimensions in large datasets with many features, making the data simpler to analyze and interpret [20]. By accounting for time-varying information and managing multiple predictors, interactions, and missing values, these models can improve the accuracy of clinical risk predictions [21].
Unfortunately, the majority of women in developing countries know little about the principal factors that escalate the risk of breast cancer. Moreover, breast

cancer is diagnosed at a younger age in many developing countries such as Iran, Tunisia, and Pakistan [22, 23]. Determining the inhibitory policies and decisions from responsible people to diminish the occurrence of breast cancer requires having knowledge of the risk factors of this cancer. Reviews on risk factors of breast cancer from various published articles in developed countries may not provide a practical solution, since the findings of these studies cannot be applied to women in developing countries. To the best of our knowledge, despite the importance of risk factors associated with this disease, few studies have been conducted in Iran using machine learning methods. Thus, the aim of this study was to determine the most important risk factors of breast cancer using machine learning algorithms in Mashhad, northeastern Iran.

## Materials and methods

### Study population

In this case-control study, data was collected from two different data sets. All women aged 17-75 years who were referred to medical centers affiliated with Mashhad University of Medical Science (MUMS) between March 21, 2015, and March 19, 2016, were enrolled. From these women, the case group was composed of individuals who had positive mammography and were confirmed to have breast cancer through histology. Women without breast cancer were selected as the control group. The smaller data set contained 516 samples, divided into 258 cases and 258 controls. The larger data set contained 606 samples, divided into 303 cases and 303 controls. To account for the effect of age, frequency matching was performed for group matching. However, other factors were not matched in order to measure their effects [24].

### Data preprocessing

Data preprocessing involves several steps. In the parsing step, individual data elements were identified in the source files and isolated in the target files. These parsed data elements were then corrected using sophisticated data algorithms in R, and the data was standardized and transformed into its preferred format.

To handle missing data, imputation methods such as mice in R were used to fill in missing values. The missing data was also removed to compare the results of these two approaches, and the results were found to be the same. As a result, removing the missing data was chosen as the preferred method for proceeding with the analysis.

### Breast cancer prediction models

This paper proposes a method that combines feature selection techniques with machine learning methods to predict prospective cases. Four different methods were used to extract features. The first method is the information.gain decision tree algorithm, which selects the best combination of features based on their correlation with class attributes using a function based on entropy. The second method is the Elastic net algorithm, which balances between LASSO and ridge penalties to address over-regularization. This is achieved by setting a hyper-parameter called Alpha, where Alpha = 1 results in a LASSO model and Alpha = 0 results in a ridge model. This technique, known as Elastic net, is a regularization regression that uses the cv.glmnet function with the family set to 'binomial'.

The third method used in this paper is Random Forest, which uses the built-in random Forest function to extract important variables using the importance function. This feature selection method falls under the category of embedded methods. Each random forest contains 4 to 12 hundred decision trees, each built using a random sample of observations and features from the dataset. Each individual tree consists of a sequence of yes-no questions based on single or combined features.

The fourth method used is principal component analysis, which is a widely recommended method for reducing a dataset with many variables. This is achieved by projecting the data onto fewer variables using linear combinations of the original variables, using the prcomp function in the latest version of R Studio software.

## Results

When using the Information.gain algorithm on a small sample size of the data set, the 10 most important factors that had an effect on breast cancer were identified and ranked according to their coefficients. These factors, in order of importance, were: family cancer, scancerb (Individual history of breast cancer), sampling B (Biopsy sampling), rarely consumption of a dairy meal, fruit meal, and vegetable meal daily, table salt, fast food consumption, oil consumption and type of job. As shown in Figure 1, the results for this method using a large sample size of the data set are the same as those for a small sample size, with the exception of the type of job. For the larger sample size, age with a coefficient of 0.05 was identified as an important factor instead of type of job.

As shown in Figure 2, the Elastic net results identified the 10 most effective factors on breast cancer as family cancer, pregnancy number, pregnancy tendency, first menstruation, abortion, childbirth (the age when they gave birth to the first child), childbirth number, stillbirth, scancerb (Individual history of breast cancer), sampling B (Biopsy sampling).

The results for the Elastic net algorithm were the same for both large and small sample sizes of the dataset.

According to the Random Forest algorithm, for the small sample size, the 10 most important factors that affected breast cancer were identified as family cancer, sampling B, scancerb, rarely consumption of dairy, fruit, and vegetable meals daily, job type, table salt, fast food consumption, and first menstruation (Fig. 3).

**Fig. 1.** The 10 most important risk factors of breast cancer in the larger sample selected by Information.gain function from the Decision Tree algorithm.

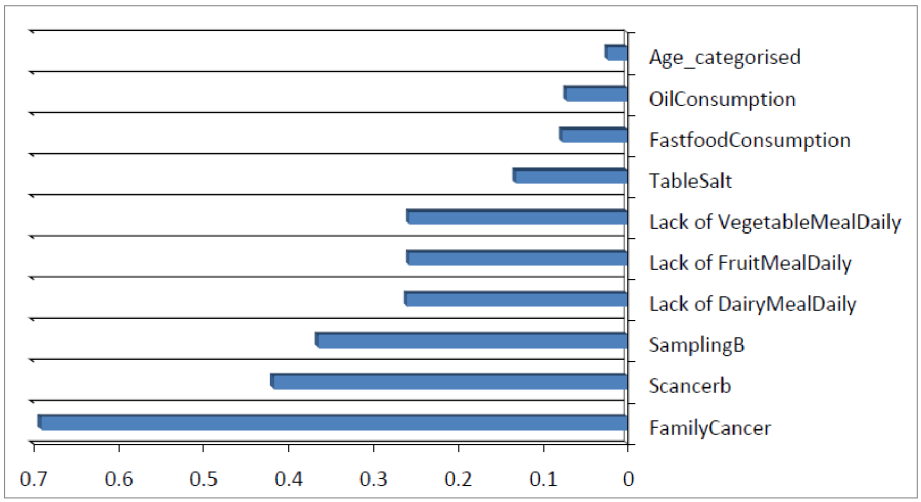
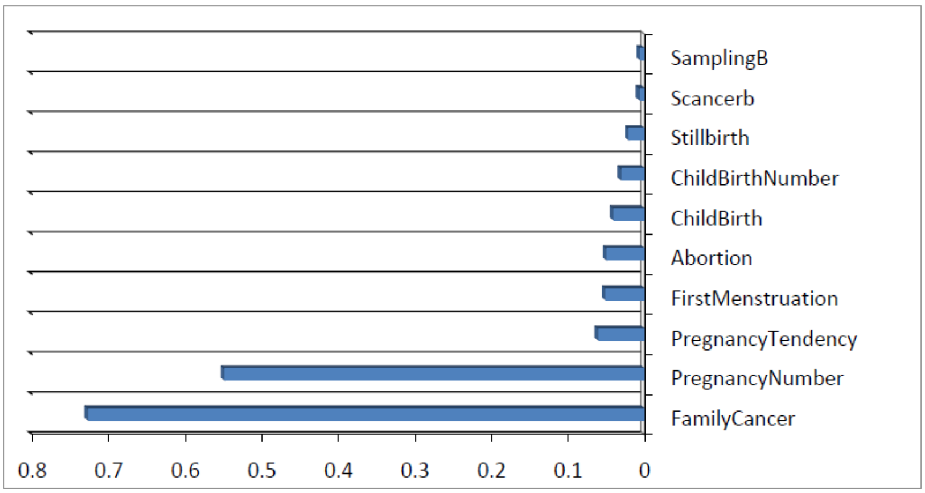



**Fig. 2.** The 10 most important risk factors of breast cancer selected by the Elastic Net algorithm.




For the large sample size of the dataset using the Random Forest algorithm, two different factors were identified as important: age and number of pregnancies, instead of type of job and age at first menstruation.

When using Principal component analysis, as shown in Figure 4, the 10 most important risk factors for breast cancer were identified as the same in both sample sizes including pregnancy tendency, first menstruation, pregnancy number, childbirth number, childbirth (the age of first giving birth), abortion, stillbirth, scancerb, sampling B and family cancer (Fig. 4).

## Discussion

This case-control study investigated the most important risk factors associated with breast cancer among Iranian women. Four machine learning methods were performed in this study including Decision Tree (DT), Random Forest (RF), Logistic Regression (LR), and Principal Component Analysis (PCA). Based on all 4 machine learning algorithms, family history of cancer, biopsy sampling, and individual history of breast cancer were the most important risk factors for breast cancer.

A case-control study indicated that family and personal history of breast cancer were the most important risk factors for breast cancer [25]. Women with two or more persons of breast cancer in their family among women younger than 50 years, or three or more persons at any age, but who do not have BRCA mutations, are about 4 times more likely to have breast cancer [26]. These studies are consistent with the findings of all four machine learning methods in our research. Thakur P et.al found that early menstruation increases the risk of breast cancer by more than two times [27]. Numerous other studies have also corroborated this finding [28, 29]. Based on the RF, LR, and PCA in our study, early menarche was among the top 10 associated risk factors of breast cancer. However, a cohort study found no association between early menstruation age and the risk of breast cancer [30].

**Fig. 3.** The 10 most important risk factors of breast cancer selected by the Random Forest algorithm.

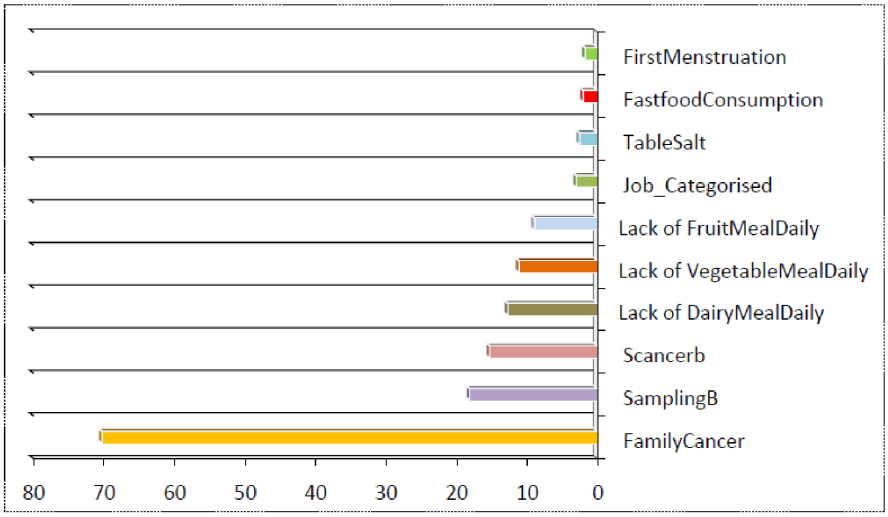
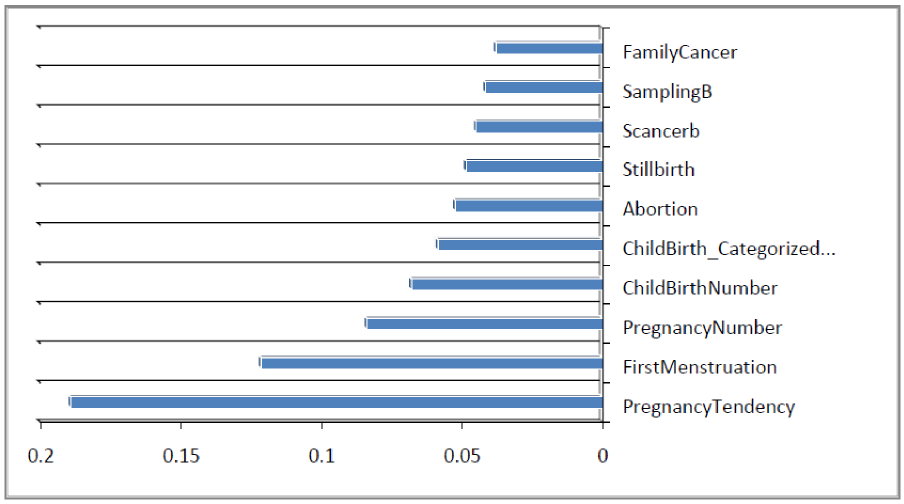



**Fig. 4.** The 10 most important risk factors of breast cancer selected by the Principal Component Analysis algorithm.




According to a case-control study, the risk of breast cancer is inversely related to the number of children born. Women who have had five or more children have a 50% lower risk of developing breast cancer compared to women who have never given birth [31]. Also, a systematic review and meta-analysis in the Eastern Mediterranean region found that having no history of live birth was among the factors that had the largest odd ratios for breast cancer [32]. Our observations showed that the number of pregnancies was only in LR and PCA models among the 10 risk factors associated with breast cancer.

Diet can play an important role in the prevention and management of many diseases, including breast cancer. A systematic review and meta-analysis showed that there was only an association between milk consumption and breast cancer among dairy products. Consumption of more than 450 grams of milk per day was associated with an increased risk of breast cancer. Also, a diet rich in vegetables, fruits, and soy products, and low in red meats, has been associated with a lower risk of breast cancer [33]. Our DT and RF models showed that low consumption of dairy products, fruits, and vegetables was associated with breast cancer risk. However, Patricia M et al. found that there was no significant relationship between dairy consumption and the risk of breast cancer [34].

It is important to note that this study has some limitations that should be considered. First, we only used four machine learning models in our study. Second, the data for this study was collected from hospitals affiliated with Mashhad University of Medical Sciences. As a result, there is uncertainty about whether the prediction model developed in this study can be applied to other datasets from different populations or regions.

## Conclusions

Our study showed that the use of machine learning methods can play an effective role in predicting breast

cancer risk factors in Iranian population with large sample size and different characteristics. Future research could involve using classification techniques to train machine learning models on the data. These models could then be used to make predictions or classifications based on the patterns and relationships identified in the data.

## Acknowledgements

We would like to thank the Clinical Research Development Unit, Imam Reza Hospital, Mashhad University of Medical Sciences, Mashhad, Iran, for their collaboration in conducting this study.

## Conflict of interest statement

The authors declare that they have no conflicting interests.

## Funding

This study had no specific funding and was done with the authors' interest.

## Ethical statement

Ethical issues (Including plagiarism, informed consent, misconduct, data fabrication and/or falsification, double publication and/or submission, redundancy, *etc.*) have been completely observed by the authors.
The study has ethical approval from Mashhad University of Medical Sciences (ID: IR.MUMS.REC.1397.125).

## Author's contributions

MT-S, E-MF, Z-SH, AK, S-MT, MY: study conception and design; MY, Z-SH: data collection; AK, S-MT: analysis and interpretation of results; AT, S-MT: draft manuscript preparation; MT-S, Z-SH, E-MF, AK, S-M, T, MY: reviewed the results and approved the final version of the manuscript.

### References

[1] Najaf Najafi M, Salehi M, Ghazanfarpour M, Hoseini ZS, Khadem-Rezaiyan M. The association between green tea consumption and breast cancer risk: a systematic review and meta-analysis. Phytother Res 2018;32:1855-64. https://doi.org/10.1002/ptr.6124.

[2] Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, Bray F. Global Cancer Statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 Countries. CA Cancer J Clin 2021;71:209-49.https://doi.org/10.3322/caac.21660.

[3] Arnold M, Morgan E, Rumgay H, Mafra A, Singh D, Laversanne M, Vignat J, Gralow JR, Cardoso F, Siesling S, Soerjomataram I. Current and future burden of breast cancer: global statistics for 2020 and 2040. Breast 2022;66:15-23. https://doi.org/10.1016/j.breast.2022.08.010.

[4] Mousavi SM, Gouya MM, Ramazani R, Davanlou M, Hajsadeghi N, Seddighi Z. Cancer incidence and mortality in Iran. Ann Oncol 2009;20:556-63. https://doi.org/10.1093/annonc/mdn642.

[5] InternationaAgency for Research on Cancer. Cancer Day. World Health Organization. Data visualisation tools for exploring the global cancer burden in 2022. Available at: https://gco.iarc.fr/today/data/factsheets/populations/364-iran-islamic-republic-of-fact-sheets.pdf (Accessed on: 13/7/2024).

[6] Łukasiewicz S, Czeczelewski M, Forma A, Baj J, Sitarz R, Stanisławek A. Breast cancer-epidemiology, risk factors, classification, prognostic markers, and current treatment strategies-an updated review. Cancers (Basel) 2021;13:4287. https://doi.org/10.3390/cancers13174287.

[7] Lei S, Zheng R, Zhang S, Wang S, Chen R, Sun K, Zeng H, Zhou J, Wei W. Global patterns of breast cancer incidence and mortality: a population-based cancer registry data analysis from 2000 to 2020. Cancer Commun (Lond) 2021;41:1183-94. https://doi.org/10.1002/cac2.12207.

[8] Youn HJ, Han W. A review of the epidemiology of breast cancer in asia: focus on risk factors. Asian Pac J Cancer Prev 2020;21:867-80. https://doi.org/10.31557/APJCP.2020.21.4.867.

[9] Terry MB, Liao Y, Whittemore AS, Leoce N, Buchsbaum R, Zeinomar N, Dite GS, Chung WK, Knight JA, Southey MC, Milne RL, Goldgar D, Giles GG, McLachlan SA, Friedlander ML, Weideman PC, Glendon G, Nesci S, Andrulis IL, John EM, Phillips KA, Daly MB, Buys SS, Hopper JL, MacInnis RJ. 10-year performance of four models of breast cancer risk: a validation study. Lancet Oncol 2019;20:504-17. https://doi.org/10.1016/S1470-2045(18)30902-1.

[10] Ibrahim SS, Hafez EE, Hashishe MM. Presymptomatic breast cancer in Egypt: role of BRCA1 and BRCA2 tumor suppressor genes mutations detection. J Exp Clin Canc Res 2010;29:1-10. https://doi.org/10.1186/1756-9966-29-82.

[11] Collaborative Group on Hormonal Factors in Breast Cancer. Menarche, menopause, and breast cancer risk: individual participant meta-analysis, including 118 964 women with breast cancer from 117 epidemiological studies. Lancet Oncol 2012;13:1141-51. https://doi.org/10.1016/S1470-2045(12)70425-4.

[12] Zeinomar N, Knight JA, Genkinger JM, Phillips KA, Daly MB, Milne RL, Dite GS, Kehm RD, Liao Y, Southey MC, Chung WK, Giles GG, McLachlan SA, Friedlander ML, Weideman PC, Glendon G, Nesci S; kConFab Investigators; Andrulis IL, Buys SS, John EM, MacInnis RJ, Hopper JL, Terry MB. Alcohol consumption, cigarette smoking, and familial breast cancer risk: findings from the Prospective Family Study Cohort (ProF-SC). Breast Cancer Res 2019;21:128. https://doi.org/10.1186/s13058-019-1213-1.

[13] Collaborative Group on Hormonal Factors in Breast Cancer. Type and timing of menopausal hormone therapy and breast cancer risk: individual participant meta-analysis of the worldwide epidemiological evidence. Lancet 2019;394:1159-68. https://doi.org/10.1016/S0140-6736(19)31709-X.

[14] Li M, Han M, Chen Z, Tang Y, Ma J, Zhang Z, Liu Z, Zhang N, Xi C, Liu J, Tian D, Wang X, Huang X, Chen J, Wang W, Zhai S. Does marital status correlate with the female breast cancer risk? A systematic review and meta-analysis of observational studies. PLoS One 2020;15:E0229899. https://doi.org/10.1371/journal.pone.0229899.

[15] Ekici S, Jawzal H. Breast cancer diagnosis using thermography and convolutional neural networks. Med Hypotheses 2020;137:109542. https://doi.org/10.1016/j.mehy.2019.109542.

[16] Javaid M, Haleem A, Singh RP, Suman R, Rab S. Significance of machine learning in healthcare: features, pillars and applica-

tions. Int J Intell Netw 2022;3:58-73. https://doi.org/10.1016/j.ijin.2022.05.002.

[17] Von Winterfeldt D, Edwards W. Decision analysis and behavioral research. (No Title) 1986.

[18] Worth AP, Cronin MT. The use of discriminant analysis, logistic regression and classification tree analysis in the development of classification models for human health effects. Journal of Molecular Structure: THEOCHEM 2003;622:97-111. https://doi.org/10.1016/S0166-1280(02)00622-X.

[19] Shaikhina T, Lowe D, Daga S, Briggs D, Higgins R, Khovanova N. Decision tree and random forest models for outcome prediction in antibody incompatible kidney transplantation. Biomed Signal Process Control 2019;52:456-62. https://doi.org/10.1016/j.bspc.2017.01.012.

[20] Jolliffe IT, Cadima J. Principal component analysis: a review and recent developments. Philos Trans A Math Phys Eng Sci 2016;374:20150202. https://doi.org/10.1098/rsta.2015.0202.

[21] Shipe ME, Deppen SA, Farjah F, Grogan EL. Developing prediction models for clinical use using logistic regression: an overview. J Thorac Dis 2019;11(Suppl 4):S574-84. https://doi.org/10.21037/jtd.2019.01.25.

[22] Nguyen H, Bui X-N. Predicting blast-induced air overpressure: a robust artificial intelligence system based on artificial neural networks and random forest. Natural Resources Research 2019;28:893-907. https://doi.org/10.1007/s11053-018-9424-1.

[23] Shaukat U, Ismail M, Mehmood N. Epidemiology, major risk factors and genetic predisposition for breast cancer in the Pakistani population. Asian Pac J Cancer Prev 2013;14:5625-9. https://doi.org/10.7314/apjcp.2013.14.10.5625.

[24] Wacholder S, McLaughlin JK, Silverman DT, Mandel JS. Selection of controls in case-control studies. I. Principles. Am J Epidemiol 1992;135:1019-28. https://doi.org/10.1093/oxfordjournals.aje.a116396.

[25] Rosner B, Tamimi RM, Kraft P, Gao C, Mu Y, Scott C, Winham SJ, Vachon CM, Colditz GA. Simplified breast risk tool integrating questionnaire risk factors, mammographic density, and polygenic risk score: development and validation. Cancer Epidemiol Biomarkers Prev 2021;30:600-7. https://doi.org/10.1158/1055-9965.EPI-20-0900.

[26] Metcalfe KA, Finch A, Poll A, Horsman D, Kim-Sing C, Scott J, Royer R, Sun P, Narod SA. Breast cancer risks in women with a family history of breast or ovarian cancer who have tested negative for a BRCA1 or BRCA2 mutation. Br J Cancer 2009;100:421-5. https://doi.org/10.1038/sj.bjc.6604830.

[27] Thakur P, Seam RK, Gupta MK, Gupta M, Sharma M, Fotedar V. Breast cancer risk factor evaluation in a Western Himalayan state: a case-control study and comparison with the Western World. South Asian J Cancer 2017;6:106-9. https://doi.org/10.4103/sajc.sajc_157_16.

[28] Kim Y, Yoo KY, Goodman MT. Differences in incidence, mortality and survival of breast cancer by regions and countries in Asia and contributing factors. Asian Pac J Cancer Prev 2015;16:2857-70. https://doi.org/10.7314/apjcp.2015.16.7.2857.

[29] Clavel-Chapelon F, Launoy G, Auquier A, Gairard B, Brémond A, Piana L, Lansac J, Renaud R. Reproductive factors and breast cancer risk. Effect of age at diagnosis. Ann Epidemiol 1995;5:315-20. https://doi.org/10.1016/1047-2797(95)00099-s.

[30] Khalis M, Charbotel B, Chajès V, Rinaldi S, Moskal A, Biessy C, Dossus L, Huybrechts I, Fort E, Mellas N, Elfakir S, Charaka H, Nejjari C, Romieu I, El Rhazi K. Menstrual and reproductive factors and risk of breast cancer: a case-control study in the Fez region, Morocco. PLoS One 2018;13:E0191333. https://doi.org/10.1371/journal.pone.0191333.

[31] Lambe M, Hsieh CC, Chan HW, Ekbom A, Trichopoulos D, Adami HO. Parity, age at first and last birth, and risk of breast cancer: a population-based study in Sweden. Breast Cancer Res Treat 1996;38:305-11. https://doi.org/10.1007/BF01806150.

[32] Namiranian N, Moradi-Lakeh M, Razavi-Ratki SK, Doayie M, Nojomi M. Risk factors of breast cancer in the Eastern Mediterranean Region: a systematic review and meta-analysis. Asian Pac J Cancer Prev 2014;15:9535-41. https://doi.org/10.7314/apjcp.2014.15.21.9535.

[33] Kazemi A, Barati-Boldaji R, Soltani S, Mohammadipoor N, Esmaeilinezhad Z, Clark CCT, Babajafari S, Akbarzadeh M. Intake of various food groups and risk of breast cancer: a systematic review and dose-response meta-analysis of prospective studies. Adv Nutr 2021;12:809-49. https://doi.org/10.1093/advances/nmaa147.

[34] Moorman PG, Terry PD. Consumption of dairy products and the risk of breast cancer: a review of the literature. Am J Clin Nutr 2004;80:5-14. https://doi.org/10.1093/ajcn/80.1.5.

Received on July 27, 2023. Accepted on March 18, 2024.

**Correspondence:** Maryam Yaghoobi, Clinical Research Development Unit, Imam Reza Hospital, University of Medical Sciences, Mashhad, Iran. E-mail: myaghooobi@yahoo.com

**How to cite this article:** Khaleghi A, Tabatabaei Sm, Sadat Hosseini Z, Taheri Soodejani M, Farkhani Em, Yaghoobi M. Machine learning techniques to identify risk factors of breast cancer among women in Mashhad, Iran. J Prev Med Hyg 2024;65:E221-E226. https://doi.org/10.15167/2421-4248/jpmh2024.65.2.3045