



## UvA-DARE (Digital Academic Repository)

### Applications of different machine learning approaches in prediction of breast cancer diagnosis delay

Dehdar, S.; Salimifard, K.; Mohammadi, R.; Marzban, M.; Saadatmand, S.; Fararouei, M.; Dianati-Nasab, M.

**DOI**

[10.3389/fonc.2023.1103369](https://doi.org/10.3389/fonc.2023.1103369)

**Publication date**

2023

**Document Version**

Final published version

**Published in**

Frontiers in Oncology

**License**

CC BY

[Link to publication](#)

**Citation for published version (APA):**

Dehdar, S., Salimifard, K., Mohammadi, R., Marzban, M., Saadatmand, S., Fararouei, M., & Dianati-Nasab, M. (2023). Applications of different machine learning approaches in prediction of breast cancer diagnosis delay. *Frontiers in Oncology*, 13, Article 1103369. <https://doi.org/10.3389/fonc.2023.1103369>

**General rights**

It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

**Disclaimer/Complaints regulations**

If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: <https://uba.uva.nl/en/contact>, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.

*UvA-DARE is a service provided by the library of the University of Amsterdam (<https://dare.uva.nl>)*



## OPEN ACCESS

## EDITED BY

Alireza Sadjadi,  
Tehran University of Medical Sciences, Iran

## REVIEWED BY

Cong Jiang,  
Harbin Medical University, China  
Md. Rakibul Islam,  
Daffodil International  
University, Bangladesh

## \*CORRESPONDENCE

Khodakaram Salimifard

✉ salimifard@pgu.ac.ir

Mostafa Dianati-Nasab

✉ m.dianatinasab@maastrichtuniversity.nl

## SPECIALTY SECTION

This article was submitted to  
Cancer Epidemiology and Prevention,  
a section of the journal  
Frontiers in Oncology

RECEIVED 03 December 2022

ACCEPTED 30 January 2023

PUBLISHED 16 February 2023



## CITATION

Dehdar S, Salimifard K, Mohammadi R,  
Marzban M, Saadatmand S, Fararouei M  
and Dianati-Nasab M (2023) Applications of  
different machine learning approaches in  
prediction of breast cancer diagnosis delay.  
*Front. Oncol.* 13:1103369.  
doi: 10.3389/fonc.2023.1103369

## COPYRIGHT

© 2023 Dehdar, Salimifard, Mohammadi,  
Marzban, Saadatmand, Fararouei and  
Dianati-Nasab. This is an open-access article  
distributed under the terms of the [Creative  
Commons Attribution License \(CC BY\)](https://creativecommons.org/licenses/by/4.0/). The  
use, distribution or reproduction in other  
forums is permitted, provided the original  
author(s) and the copyright owner(s) are  
credited and that the original publication in  
this journal is cited, in accordance with  
accepted academic practice. No use,  
distribution or reproduction is permitted  
which does not comply with these terms.

# Applications of different machine learning approaches in prediction of breast cancer diagnosis delay

Samira Dehdar<sup>1</sup>, Khodakaram Salimifard<sup>1\*</sup>, Reza Mohammadi<sup>2</sup>,  
Maryam Marzban<sup>3</sup>, Sara Saadatmand <sup>1</sup>, Mohammad Fararouei<sup>4</sup>  
and Mostafa Dianati-Nasab <sup>5\*</sup>

<sup>1</sup>Computational Intelligence & Intelligent Optimization Research Group, Business and Economic School, Persian Gulf University, Bushehr, Iran, <sup>2</sup>Business Analytics Section, Amsterdam Business School, University of Amsterdam, Amsterdam, Netherlands, <sup>3</sup>Department of Public Health, School of Public Health, Bushehr University of Medical Science, Bushehr, Iran, <sup>4</sup>Department of Epidemiology, School of Public Health, Shiraz University of Medical Sciences, Shiraz, Iran, <sup>5</sup>Department of Complex Genetics and Epidemiology, School of Nutrition and Translational Research in Metabolism, Maastricht University, Maastricht, Netherlands

**Background:** The increasing rate of breast cancer (BC) incidence and mortality in Iran has turned this disease into a challenge. A delay in diagnosis leads to more advanced stages of BC and a lower chance of survival, which makes this cancer even more fatal.

**Objectives:** The present study was aimed at identifying the predicting factors for delayed BC diagnosis in women in Iran.

**Methods:** In this study, four machine learning methods, including extreme gradient boosting (XGBoost), random forest (RF), neural networks (NNs), and logistic regression (LR), were applied to analyze the data of 630 women with confirmed BC. Also, different statistical methods, including chi-square, p-value, sensitivity, specificity, accuracy, and area under the receiver operating characteristic curve (AUC), were utilized in different steps of the survey.

**Results:** Thirty percent of patients had a delayed BC diagnosis. Of all the patients with delayed diagnoses, 88.5% were married, 72.1% had an urban residency, and 84.8% had health insurance. The top three important factors in the RF model were urban residency (12.04), breast disease history (11.58), and other comorbidities (10.72). In the XGBoost, urban residency (17.54), having other comorbidities (17.14), and age at first childbirth (>30) (13.13) were the top factors; in the LR model, having other comorbidities (49.41), older age at first childbirth (82.57), and being nulliparous (44.19) were the top factors. Finally, in the NN, it was found that being married (50.05), having a marriage age above 30 (18.03), and having other breast disease history (15.83) were the main predicting factors for a delayed BC diagnosis.

**Conclusion:** Machine learning techniques suggest that women with an urban residency who got married or had their first child at an age older than 30 and those without children are at a higher risk of diagnosis delay. It is necessary to educate them about BC risk factors, symptoms, and self-breast examination to shorten the delay in diagnosis.

#### KEYWORDS

breast cancer (BC), random forest (RF), neural networks (NN), delay, machine learning, extreme gradient boosting, logistic regression

## 1 Introduction

Breast cancer (BC), the most frequently diagnosed cancer (1) and the second leading cause of death among women (2), accounts for nearly 35% of new cancer cases (3). In 2021, BC was recognized as the leading cause of mortality among women all over the world, with more than 685,000 deaths and 2.3 million new cases, equivalent to 11.7% of all identified cancer cases (1), causing 15% of all cancer deaths, mainly in less-developed countries (4).

Specifically, developing countries are suffering from an increasing number of BC cases with an increasing range of young women at risk of cancer (5). In recent years in Asian countries, including Iran, both the incidence and mortality of BC have had notable growth (6–9). Also, studies have declared that the average age of BC in Iranian women is almost a decade earlier than the world average (10, 11). Also, in Iran, delays in diagnosis and treatment (12, 13) and cancer detection at more advanced stages compared to Western countries have been reported (14).

The prolonged interval from the detection of initial symptoms until the histological diagnosis is defined as a diagnosis delay (15), which might happen for two main reasons: 1) patients' delay, which refers to the duration between noticing the first symptom and announcing it to the medical consultant, and 2) providers' delay, which is identified as the time interval between the first announcement of symptoms to the start of treatment (16). Longer delays lead to more advanced stages of cancer (17) and consequently a lower chance of survival (18, 19). Clinically, a 90-day or more delay in diagnosis is considered a delayed BC diagnosis (20).

Several studies have found that various factors are associated with BC diagnosis delays. Effective sociodemographic factors include age (21), education (22, 23), socioeconomic status (24, 25), marital status (22, 26), place of residence (27, 28), and family history (26, 29). Other important factors for a delayed presentation that lead to diagnosis delay are lack of knowledge regarding the disease (25, 30), lack of breast self-examination (23, 28), ignorance (25, 26, 31), stress of cancer treatment and consequences (32), and absence of qualified healthcare service (30, 31).

Machine learning, a subfield of artificial intelligence, uses a wide range of optimization, probabilistic, and statistical methods that allow computers to “learn” from past examples and to distinguish hard-to-

detect patterns from complicated datasets. In the medical field, clinics and hospitals record and keep massive databases of patients' symptoms and diagnoses. Therefore, researchers use this knowledge to develop classification models that can make inferences based on historical cases (33).

This study aimed to analyze the importance of a variety of factors to predict BC diagnosis delay by employing four different machine learning methods, including random forest (RF), neural network (NN), logistic regression (LR), and extreme gradient boosting (XGBoost).

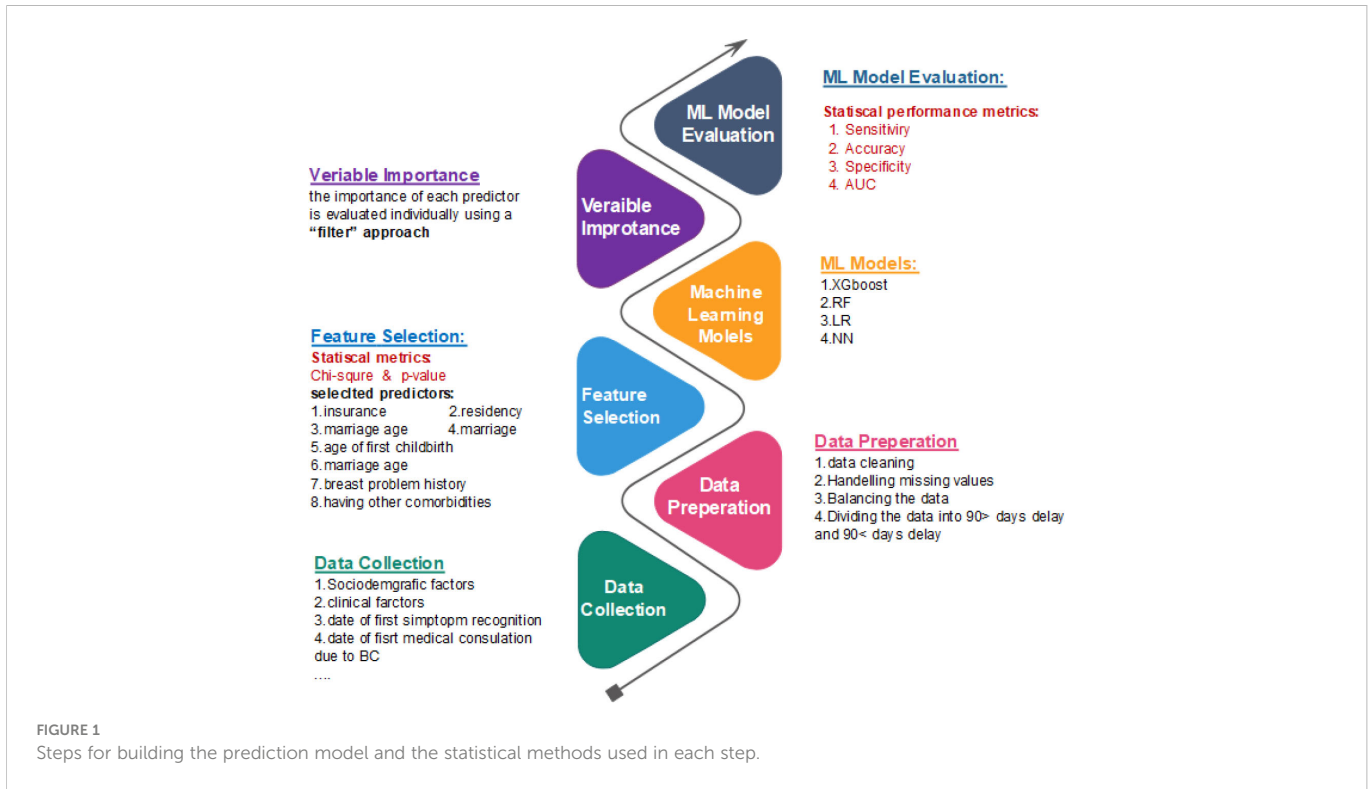
## 2 Materials and methods

In this study, a six-step methodology was applied to build a prediction model. Figure 1 illustrates an overview of the steps taken and the statistical methods that were used in each step. Different statistical methods, including chi-square, p-value, sensitivity, specificity, accuracy, and area under the receiver operating characteristic (ROC) curve (AUC), were utilized in this paper.

### 2.1 Data

In this study, 630 women with confirmed BC (incident or new cases) were assessed to identify the factors related to delayed diagnosis of BC. The data were obtained partly from the patients' hospital records and partly from an interview-administered questionnaire that was completed during the study period while the patients were visiting the center. Literate patients read and gave signed informed consent. Verbal consent was obtained from illiterate patients. Ethical approval was obtained from the Shiraz University of Medical Sciences ethics committee (23). A trained nurse was hired to interview the patients by using a validated questionnaire (23). The questionnaire and interview procedures were evaluated and revised during a pilot study on 50 patients. Accordingly, using the test–retest method, the questionnaire's reliability was estimated to be good (Cronbach alpha = 0.76) (23). Furthermore, other data, including self-reported date or type of initial signs and symptoms of BC noticed by the patients, date of first symptom recognition, and the month and year of their first medical consultation due to BC, were also collected. These dates were used as a reference to questions about whether or not the patients had perceived symptoms, the period before the first consultation, and socioeconomic factors at the moment of the first medical

**Abbreviations:** BC, breast cancer; XGBoost, extreme gradient boosting; RF, random forest; NN, neural network; LR, logistic regression.



consultation. Even though a standard questionnaire was used to collect both clinical and sociodemographic factors, some factors were put aside due to the missing data (such as body mass index (BMI) and menopause status).

Patients were divided into two categories: those 1) with less than 90 days’ delay in diagnosis and 2) with more than 90 days’ delay in diagnosis. Different features were analyzed in both groups, including age, marriage, residency, insurance, age at first childbirth, marriage age, having other comorbidities, and other breast disease histories. The main reason for the delay in diagnosis was also obtained from patients. In the second phase, clinical data including the stage of disease, tumor size, and lymph node status, was gathered by reviewing patients’ medical records (23). In this study, patients’ age was considered a continuous variable. The age at first marriage was divided into five categories (20, 20–25, 25–30, >30, and not married), and the age at first childbirth was divided into four (20, 20–25, > 30, single, or not having a child). Both sociodemographic and clinical data are shown in Table 1.

## 2.2 Machine learning methods

To optimize the hyperparameters for all the algorithms (RF, NN, XGBoost, and LR) in the train set, the grid search method in the Caret package (Kuhn, 2008) in the R programming language was used. Table 2 shows the parameter values for each applied machine learning (ML) algorithm.

### 2.2.1 Random forest

The RF algorithm is known as a highly stated machine learning method for classification problems (33). The algorithm has been

reported to originate one of the greatest accuracies (34). Computing the missing data and investigating multi-dimensional data are possible by RF algorithm (35). The significance of variables used for classification in RF can also be tuned in (35). The RF is a combined classification method based on the decision tree model. K decision trees are generated based on K diverse training data extracted from the main dataset. Decision trees build the final RF model (36). In such combined methods as RF, a “strong learner” is constructed by consuming numerous “weak learners” (37).

In this paper, to make the parameters appropriate for using the RF method, the number of trees was set to 200, and the minimum size of terminal nodes was set to one.

### 2.2.2 Logistic regression

Utilizing binary variables for classification problems can be performed by LR. This model generally demonstrates the probability of an event occurrence by measuring the correlation between a dependent binary variable and a minimum of one independent variable (38). The distribution of the odds is outlined in an S-shaped function (Figure 2) to achieve an output between 0 and 1 (39). As LR is mathematically bound to generate probabilities in the range of [0, 1], in case values are below 0.5, they will be assumed as 0; otherwise, they will be considered 1 (40).

The logistic function is shown in Equation 1:

$$S(z) = \frac{1}{1 + e^{-z}}, \tag{1}$$

where  $S(z)$  represents the probabilities in the range of [0, 1],  $z$  is the input, and  $e$  is a natural constant (41). In this paper, a multivariable LR with 20 predictors was used to define factors affecting BC diagnosis delay. The iteratively reweighted least-

TABLE 1 Sociodemographic and clinical factors.

Sociodemographic data	
Age	Year
Education	Primary and lower, middle school, high school, college
Age at first marriage	Year
Marital status	Single, married
Occupation	Employed, housewife
Menopausal status	Pre-menopausal, post-menopausal
Residency	Rural, urban
Health insurance	Yes, no
Daily exercise	<10, 10–20, >20 min
BMI (kg/m <sup>2</sup> )	Underweight, normal, overweight, obese
Smoking	Yes, no
X-ray history	Yes, no
Chronic disease	Yes, no
Delay time	Day
Family history of BC	Yes, no
Age at first pregnancy	Year
History of BD	Yes, no
Status of knowledge and regular practice of BSE	Yes, no
Clinical data	
Type of first symptom	Lump, discharge, pain, and others
Location of tumor	Right, left
Tumor type	Ductal, lobular/medullary, and others

BMI, body mass index; BD, breast disease; BSE, breast self-examination.

squares method was applied to make this method fit the available data (42).

### 2.2.3 Neural networks

The NN method is used in a vast variety of issues as a result of its superior implementation in classification problems. NN is one of the most reputable machine learning algorithms (43). This method is inspired by biological neural networks (44). The NN method is made up of a three-layered feedforward network. The notion of weights among hidden layers, the output–input layer in the network, leads to learning (45). The output of a neuron in NN achieves in two steps, using the following formulas (46):

Step 1:  $x_{ij}$  stands for the  $i$ th input to node  $j$ , and  $W_{ij}$  indicates the weight related to the  $i$ th input to node  $j$ .

$$\sum_i W_{ij}x_{ij} \tag{2}$$

Step 2:  $e$  is a natural constant, and  $x$  is the input of the function.

$$y = \frac{1}{1 + e^{-x}} \tag{3}$$

TABLE 2 Parameter values of the four applied ML algorithms.

Algorithm	Parameter	Value/setting
LR	Fitting method	Iteratively reweighted least squares
NN	Hidden layer	1
	Input layer	1
	Output layer	1
	Fitting method	Entropy
	Maximum number of iterations	100
	Maximum number of weights	1,000
RF	Number of trees to grow	200
	Minimum size of terminal nodes	1
XGBoost	Max depth	1
	Number of rounds	150
	Minimum child weight	1
	Eta	0.3
	Subsample ratio of columns	0.8
	Subsample	0.5

ML, machine learning; LR, logistic regression; NN, neural network; RF, random forest; XGBoost, extreme gradient boosting.

In this study, this method was utilized by setting one input layer including 20 variables, a hidden layer, and one output layer. The entropy fitting method was used to fit the NN to the dataset. The maximum number of iterations and the maximum number of weights were set to 100 and 1,000, respectively.

### 2.2.4 Extreme gradient boosting

XGBoost is a powerful boosting algorithm in the machine learning system (33). XGBoost is a kind of regression tree capable of supporting both regression and classification. XGBoost and decision trees have similar decision-making rules (47). With the use of an appropriate data structure, the XGBoost algorithm is able to optimize, predict, and classify a system with the highest accuracy (19). This algorithm organizes the data to reduce the lookup time to a minimum. It also leads to cutting down the model’s training time and,

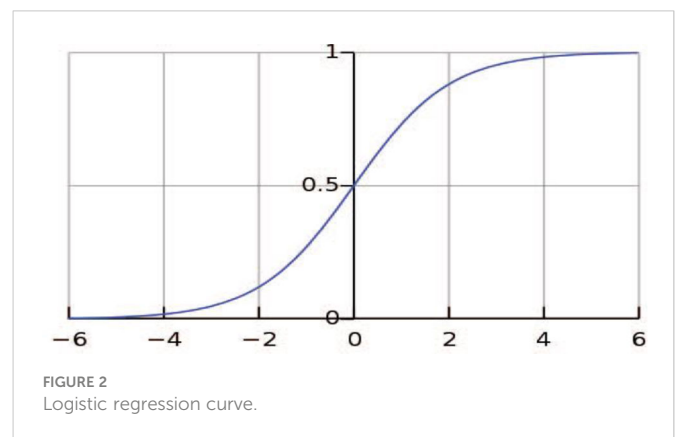


FIGURE 2 Logistic regression curve.

at the same time, improves the accuracy of the classification (48). The XGBoost algorithm is thriving as a result of its high scalability in any type of scenario (49).

In this paper, the number of rounds was set to 150 with a max depth of 1, an eta of 0.3, and a minimum child weight of 1. The subsample ratio of columns was considered to be 0.8, and the subsample was 0.5.

## 2.3 Feature selection

Feature selection is a practical, data-filtering evaluation procedure (50). In feature selection strategies, a subset of features from the primary dataset is picked by evaluating the relevance of the data to show inter-group impacts (51). Feature selection is not dependent on any machine learning algorithms. Instead, features are selected on the basis of their scores in various statistical tests for their correlation with the outcome variable. Chi-square is a statistical test applied to groups of categorical features to evaluate the likelihood of correlation or association between them using their frequency distribution.

To decide which features must be taken into consideration in building the prediction model, chi-square was calculated for 20 variables. Seven variables, including insurance, residency, marriage age, age of first childbirth, marriage, breast problem history, having other comorbidities, and marriage age, were chosen as the machine learning targets. For age, as the only continuous variable in the dataset, a p-value was calculated, so age is considered the eighth selected feature to construct the prediction model. The outcome of evaluating the chi-square for variables is shown in Table 3.

Insurance (p-value: 0.046), residency (p-value: 0.023), marriage age (p-value: 0.038), marital status (p-value: 0.006), breast problem history (p-value: 0.009), and having other comorbidities (p-value: 0.009) were found to be BC delay predictive factors when utilizing the chi-square method, and other features including patient or doctor delay (0.42), tumor type (0.41), location of the tumor (left/right breast) (0.11), first symptoms (0.93), education (0.07), income (0.10), job (0.52), family history (0.38), awareness of breast self-examination (0.20), daily exercise (0.19), chest X-ray history (0.07), and smoking (0.07) were omitted after measuring the amount of chi-square (higher than 0.05).

## 2.4 Variable importance

The importance of each predictor is evaluated individually using a “filter” approach. The filter method ranks each feature based on some univariate metrics and then selects the highest-ranking features. In this study, age was found to be of the highest importance in all methods conducted. Putting age aside, urban residency was the most effective variable in the RF and XGBoost methods, while in the NN method, it was found to be the least important one. Despite the fact

that insurance is expected to increase patients’ willingness to attend doctor appointments and undergo mammography, preventing delayed diagnosis, it has gained a low level of importance in all methods. Variables of importance in the four ML models are shown in Table 4.

## 3 Results

Among 630 BC patients, 204 (32%) had a diagnosis delay of more than 90 days. Among patients with a diagnosis delay of more than 90 days, 29.90% were between 40 and 50 years old, 88.72% were ever married, and 72.05% had urban residency. Only 15.19% of patients in this category did not have insurance, 52.45% were married when they were younger than 20 years, and 35.78% had given birth to their first child before they were 20 years old.

Among 426 patients who had a diagnosis delay of fewer than 90 days, 35.21% were between 40 and 50 years old, 54.47% were married at an age younger than 20 years, and 43.90% had their first experience of childbirth when they were younger than 20 years; 84.27% had a history of other breast comorbidities, and 80.75% had urban residency. The study population is shown in Table 5.

### 3.1 Evaluation metrics

Different performance measures were utilized to analyze each indicator’s importance in delayed BC diagnosis, as described in this part. Specificity, sensitivity, and ROC curves are commonly used in binomial classification tests to measure the performance of the statistics. The proportions of negatives are scaled by “specificity”, while the extent of actual positives is scaled by “sensitivity”. The specificity and sensitivity are calculated by Equations 4 and 5, respectively.

$$\text{Specificity} = \frac{TN}{(TN + FP)}, \quad (4)$$

$$\text{Sensitivity} = \frac{TP}{(TP + FN)}, \quad (5)$$

where *TP* means true-positive rate; *TN*, true-negative rate; *FP*, false-positive rate; and *FN*, false-negative rate.

The performance measures for the four machine learning methods are reported in Table 6. As shown, LR has the best performance in terms of accuracy, while NN, LR, and XGBoost have been able to have more considerable sensitivity.

AUC shows how qualified a parameter is at discerning among a couple of diagnostic categories. Figure 3 illustrates a comparative analysis of four different classification methods on the ROC curve. According to the ROC curve, RF has the highest AUC, while NN and LR have the second and third highest AUC, respectively.

TABLE 3 Outcome of chi-square method of selected variables.

Variables	Insurance	Residency	Marriage age	Age of first childbirth	Marriage age	Marriage	Breast problem history	Other comorbidities
<b>p-Value</b>	0.046	0.023	0.038	P < 0.005	0.038	0.006	0.009	<b>0.007</b>



TABLE 4 Variable importance.

Variable importance				
	XGBoost	RF	NN	LR
Age	100	100	100	100
Urban residency	17.54	12.04	0.63	43.19
Ever married	5.276	0.42	50.05	<0.001
Marriage age (20–25)	6.81	6.01	5.28	33.20
Marriage age (25–30)	3.65	3.32	15.09	38.90
Marriage age (>30)	2.32	0.71	18.03	34.73
Nulliparous	4.71	5.11	14.24	44.19
Age at first childbirth (20–30)	13.13	8.88	0.14	31.30
Age at first childbirth (>30)	11.42	6.80	15.36	82.57
Other breast disease history	12.37	11.58	15.83	42.99
Having other comorbidities	17.14	10.72	6.34	49.41
Health insurance	3.01	4.37	<0.001	14.61

XGBoost, extreme gradient boosting; RF, random forest; NN, neural network; LR, logistic regression.

TABLE 5 Statistics of BC patients based on model variables.

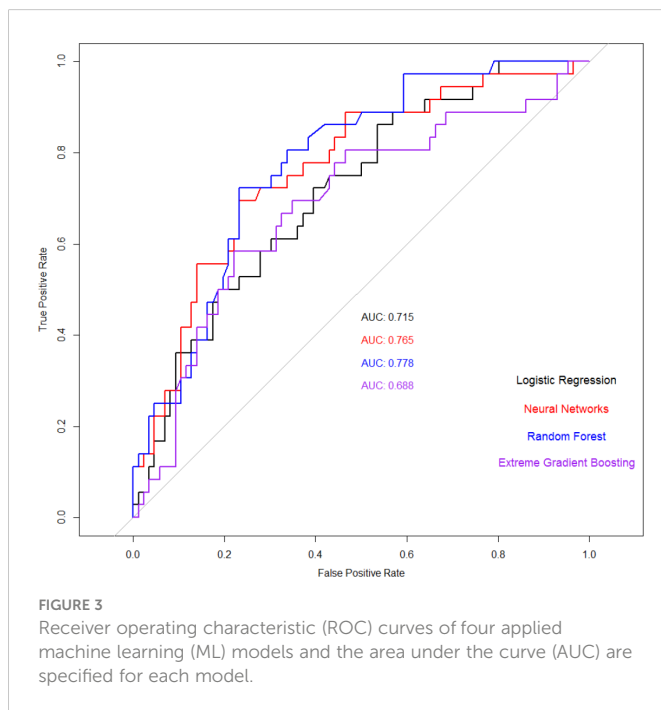
Variables		Delay in diagnosis (<90 days) (n = 426) (68%)	Delay in diagnosis (>90 days) (n = 204) (32%)	Total (n = 630)
<b>Age (years)</b>	<40	110 (25.82%)	51 (25.00%)	161 (25.56%)
	40–50	150 (35.21%)	61 (29.90%)	211 (33.49%)
	50–60	113 (26.52%)	56 (27.45%)	169 (26.83%)
	>60	53 (12.44%)	36 (17.64%)	89 (14.12%)
<b>Marriage</b>	Single (never married)	22 (5.16%)	23 (11.27%)	45 (7.14%)
	Ever married	405 (94.84%)	181 (88.72%)	585 (92.86%)
<b>Place of residence</b>	Rural	81 (19.25%)	57 (27.94%)	139 (22.06%)
	Urban	344 (80.75%)	147 (72.05%)	491 (77.94%)
<b>Insurance</b>	No	39 (9.15%)	31 (15.19%)	70 (11.11%)
	Yes	387 (90.85%)	173 (84.80%)	560 (88.89%)
<b>Age at first childbirth</b>	<20	187 (43.90%)	73 (35.78%)	260 (41.27%)
	20–30	156 (36.62%)	56 (27.45%)	212 (33.65%)
	>30	36 (8.45%)	37 (18.13%)	73 (11.59%)
	Single or no child	47 (11.03%)	38 (18.62%)	85 (13.49%)
<b>Marriage age</b>	<20	232 (54.47%)	107 (52.45%)	339 (53.81%)
	20–25	94 (22.06%)	37 (18.13%)	131 (20.79%)
	25–30	56 (13.14%)	24 (11.76%)	80 (12.70%)
	>30	23 (5.40%)	14 (6.86%)	37 (5.87%)
	Not married	21 (4.93%)	22 (10.78%)	43 (6.83%)
<b>Other comorbidities</b>	No	280 (65.73%)	111 (54.41%)	391 (62.06%)
	Yes	146 (34.27%)	93 (45.59%)	239 (37.94%)
<b>Other breast disease history</b>	No	359 (84.27%)	154 (75.49%)	513 (81.42%)
	Yes	67 (15.73%)	50 (24.51%)	117 (18.57%)

BC, breast cancer.

TABLE 6 Performance measures of four ML models.

Measure	RF	NN	LR	XGBoost
Accuracy	0.6967	0.7213	0.729	0.7131
Sensitivity	0.8372	0.8721	0.8721	0.8721
Specificity	0.3611	0.3611	0.3889	0.3333
AUC	0.788	0.765	0.715	0.688

ML, machine learning.



## 4 Discussion

The results show 32% of patient delay among women in Iran, which is a moderate amount in comparison with that in other developing countries, such as Pakistan (88.8%) (52), Uganda (89%) (53), Nigeria (81.6%) (26), and China (34%) (53). However, in developed countries, the situation is quite different. In the USA, the patient delay was reported to be 17.5% in white patients and 26.4% in African American patients (52). In the UK, 8.4% of BC patients postponed looking for treatment for more than 3 months (54), and in Malaysia, the patient delay was reported to be 33.1% (50). Therefore, compared to the reported amount in surveys from developed countries, the current study showed a more intense patient delay.

In this study, four machine learning methods, including XGBoost, RF, NN, and LR, were applied to analyze the variables' importance. In all methods, "age" was found to be of the highest importance. Putting age aside, urban residency (17.54), having other comorbidities (17.14), and age at first childbirth (>30) (13.13) were found to be the top three important variables in the XGBoost method. In the RF method, the outcome was almost identical to the XGBoost method, where the top three essential predictors (leaving "age" out) were urban residency (12.04), other breast disease history (11.58), and having other comorbidities (10.72). Conducting the NN method, being married (50.05), marriage age (>30) (18.03), and other breast

disease history (15.83) were found to be the top three effective risk factors. Considering the top three important predictors in the LR method, the only factor in common with the RF and XGBoost methods was having other comorbidities (49.41). With the use of this method, the outcome highlighted the first childbirth age, the age at the first childbirth at >30 (82.57), and being nulliparous (44.19) as the top three among the study variables.

In a study by Mirfarhadi et al. (55), 232 patients with confirmed BC in Iran were studied, and LR was applied to identify the main risk factors for BC diagnosis delay. Among the 16 factors that were studied in this paper, including age, place of residence, education level, marital status, number of children, monthly income, having insurance coverage, having complementary insurance, family history of BC, history of mammography, and stage of disease, the most important factors were found to be the stage of disease, primary insurance, and lack of complimentary insurance. Passing over the stage of disease and history of mammography, other factors were similar to the current study, whereas the same method "LR" showed a completely different outcome. Implementing the LR method in the current study, age, age at first childbirth, and having other comorbidities were found to be the most important factors in BC delayed diagnosis. In the analysis of 283 women with BC, taking similar factors such as age, place of residence, education level, medical payment method (insurance), monthly income, method of symptom discovery, knowledge of BC symptoms, family support, health values, internal and external health locus of control, and perceived health competence into consideration, the main BC delay predictors announced were knowledge of BC symptoms, external health locus of control, breast self-examination/physical examination, perceived health competence, family support, pain stimulation, and age.

In Senegal, data collected from patients within 7 years was studied (56) to analyze the association between sociodemographic factors and BC delay. In this study, no associations were detected between sociodemographic factors and BC delay, and the only relevant factor was found to be a negative history of family BC. In the UK (57) and Malaysia (58), which are also known as developed countries, the most important sociodemographic factor correlated to BC delay risk was found to be "marital status", as reported in (56, 59), and married women had a shorter delay than single and separated/divorced women. The results show that in developed countries, socioeconomic factors have little effect on the risk of delayed BC diagnosis. This can be a result of governmental planning and support, something that is not actually seen in less-developed countries. In a study in China (60), 1,431 women with diagnosed BC were studied to assess the correlation between variables including demographic data, clinical and tumor characteristics, and BC delay by employing multivariate LR and Kaplan–Meier regression models, and it was



directly reported that there was no association between age and BC delay. In contrast, 7 years later, another study (61) in the same country declared age as the main factor affecting BC diagnosis delay. In this study, multiple linear regression was utilized to measure the impact of sociodemographic characteristics, medical history, and knowledge of BC; residency and disclosure of symptom were the most important factors, excluding age as the vital factor. In another developing country, Ethiopia, age was declared as the main factor correlating with BC diagnosis delay (25). In this study, bivariable and multivariable LRs were conducted to assess the prevalence and factors associated with BC diagnosis delay. In this study, educational status, occupation, and residency also were announced as important factors regarding BC diagnosis delay.

In (56–58, 62, 63), and (60), different types of LR have been employed to assess the association between various sociodemographic and clinical factors and the risk of BC diagnosis delay.

The main strength of this paper is utilizing four different machine learning methods and comparing the outcomes, whereas in other papers, only one or two methods were used. We used a wide range of variables that might influence the rate of progression of BC. Recruiting participants who visited the biggest referral center in the southern part of Iran makes the results generalizable to the city's population.

The generalizability of the data might be pointed out as a limitation of this study, as the data were collected from one referral center in the south of Iran (no other parts of the country); however, this center is considered the source point for diagnosis and treatment of patients; also, some factors that could have affected the outcome, such as BMI and menopause status, had to be omitted due to the missing data. Future studies can consider a larger dataset that is collected from different centers in different cities to achieve more generalized outcomes and build more reliable models.

## 5 Conclusion

Early diagnosis plays a significant role in increasing the survival rate of BC patients. The diagnosis of cancer by pathologists is costly, and the outcome might vary greatly depending on the pathological process. Also, due to the human brain's limited ability to integrate large amounts of data, the accuracy of the diagnosis cannot be guaranteed, and it is impossible to avoid misdiagnosis. Artificial intelligence models are superb at handling large amounts of data. With the use of machine learning, which is a subset of artificial intelligence, an accurate and quick diagnosis of BC is possible. Machine learning techniques suggest that women with an urban residency who got married or had their first child at an age older than 30 and those who are nulliparous are at a higher risk of diagnosis

delay, and it is necessary to be educated about BC symptoms and self-breast examination.

## Data availability statement

The original contributions presented in the study are included in the article/supplementary material. Further inquiries can be directed to the corresponding authors.

## Ethics statement

The ethics code was obtained from the ethics committee of Shiraz University of Medical Sciences.

## Author contributions

Conceptualization: SD, KS, and RM. Data: MD-N, MF, and SD. Methodology: SD, MD-N, SS, RM, and KS. Formal analysis and investigation: MD-N, SS, KS, and RM. Writing—original draft preparation: SD. Writing—review and editing: MF, MD-N, RM, and MM. All authors contributed to the article and approved the submitted version.

## Acknowledgments

We would like to thank all the participants and wish them all happiness.

## Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

## References

1. Siegel RL, Miller KD, Wagle NS, Jemal A. Cancer statistics, 2023. *CA Cancer J Clin* (2023) 73(1):17–48.
2. US Cancer Statistics Working Group. *United states cancer statistics: 1999–2012 incidence and mortality web-based report*. Atlanta (GA: Department of health and human services, centers for disease control and prevention, and national cancer institute) (2015).
3. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2017. *CA. Cancer J Clin* (2017) 67(1):7–30. doi: 10.3322/caac.21387
4. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. "Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries." *CA. Cancer J Clin* (2018) 68(6):394–424. doi: 10.3322/caac.21492

5. An Y, Wang J, Zhang L, Zhao H, Gao Z, Huang H, et al. PASCAL: A pseudo cascade learning framework for breast cancer treatment entity normalization in Chinese clinical text. *BMC Med Inform. Decis. Mak.* (2020) 20(1):204. doi: 10.1186/s12911-020-01216-9
6. Torre LA, Bray F, Siegel RL, Ferlay J, Lortet-Tieulent J, Jemal A. Global cancer statistics, 2012. *CA. Cancer J Clin* (2015) 65(2):87–108. doi: 10.3322/caac.21262
7. Fitzmaurice C, Alsharif UC, El Bcheraoui I, Khalil R, Charara M, Moradi-Lakeh A, et al. Burden of cancer in the Eastern Mediterranean region, 2005–2015: Findings from the global burden of disease 2015 study. *Int J Public Health* (2018) 63(S1):151–64. doi: 10.1007/s00038-017-0999-9
8. Fitzmaurice C, Akinyemiju TF, Al Lami FH, Alam T, Alizadeh-Navaei R, Allen C, Alsharif U, et al. Global, regional, and national cancer incidence, mortality, years of life lost, years lived with disability, and disability-adjusted life-years for 29 cancer groups, 1990 to 2016. *JAMA Oncol* (2018) 4(11):1553. doi: 10.1001/jamaoncol.2018.2706
9. Luzzati T, Parenti A, Rughi T. Economic growth and cancer incidence. *Ecol Econ* (2018) 146:381–96. doi: 10.1016/j.ecolecon.2017.11.031
10. Akbari ME, Sayad S, Sayad S, Khayamzadeh M, Shojaei L, Shorameji Z, et al. Breast cancer status in Iran: Statistical analysis of 3010 cases between 1998 and 2014. *Int J Breast Cancer* (2017) 2017:2481021. doi: 10.1155/2017/2481021
11. Mousavi SM, Montazeri A, Mohaghegh MA, Jarrah AM, Harirchi I, Najafi M, et al. Breast cancer in Iran: an epidemiological review. *Breast J* 13(4):383–91. doi: 10.1111/j.1524-4741.2007.00446.x
12. Taghavi A, Fazeli Z, Vahedi M, Baghestani Reza A, Pourhoseingholi A, Barzegar F, et al. Increased trend of breast cancer mortality in Iran. *Asian Pac. J Cancer Prev* (2012) 13(1):367–70. doi: 10.7314/apjcp.2012.13.1.367
13. Bustami RT, Shulkin DB, O'Donnell N, Whitman ED. Variations in time to receiving first surgical treatment for breast cancer as a function of racial/ethnic background: a cohort study. *JRSM Open* (2014) 5(7):2042533313515863. doi: 10.1177/2042533313515863
14. Montazeri A, Vahdaninia M, Harirchi I, Harirchi Mahmood A, Sajadian A, Khaleghi F, et al. Breast cancer in Iran: need for greater women awareness of warning signs and effective screening methods. *Asia Pac. Fam. Med* (2008) 7(1):1–7. doi: 10.1186/1447-056X-7-6
15. Foroozani E, Ghasvand R, Mohammadianpanah M, Afrashteh S, Bastam D, Kashefi F, et al. Determinants of delay in diagnosis and end stage at presentation among breast cancer patients in Iran: A multi-center study. *Sci Rep* (2020) 10(1):21477. doi: 10.1038/s41598-020-78517-6
16. Sinn H-P, Helmchen B, Wittekind CH. TNM-klassifikation beim mammakarzinom. *Pathologe* (2010) 31(5):361–6. doi: 10.1007/s00292-010-1307-0
17. Caplan L. Delay in breast cancer: Implications for stage at diagnosis and survival. *Front Public Heal* (2014) 2:87. doi: 10.3389/fpubh.2014.00087
18. Burgess C, Ramirez A, Richards M, Love S. Who and what influences delayed presentation in breast cancer? *Br J Cancer* (1998) 77(8):1343–8. doi: 10.1038/bjc.1998.224
19. Harirchi I, Ghaemmaghami F, Karbaksh M, Moghimi R, Mazaherie H. Patient delay in women presenting with advanced breast cancer: An Iranian study. *Public Health* (2005) 119(10):885–91. doi: 10.1016/j.puhe.2004.11.005
20. Nosarti C, Crayford T, Roberts JV, Elias E, McKenzie K, David AS. Delay in presentation of symptomatic referrals to a breast clinic: Patient and system factors. *Br J Cancer* (2000) 82(3):742–8. doi: 10.1054/bjoc.1999.0990
21. Maghous A, Rais F, Ahid S, Benhmidou N, Bellahamou K, Loughlimi H, et al. Factors influencing diagnosis delay of advanced breast cancer in Moroccan women. *BMC Cancer* (2016) 16(1):356. doi: 10.1186/s12885-016-2394-y
22. Lopes TCR, Gravena AAF, de Oliveira Demitto M, Borghesan DHP, DellAgnolo CM, Brischiliar SCR, et al. Delay in diagnosis and treatment of breast cancer among women attending a reference service in Brazil. *Asian Pac. J Cancer Prev* (2017) 18(11):3017–23. doi: 10.22034/APJCP.2017.18.11.3017
23. Dianatinasab M, Fararouei M, Mohammadianpanah M, Zare-Bandamiri M. Impact of social and clinical factors on diagnostic delay of breast cancer. *Med (Baltimore)*. (2016) 95(38):e4704. doi: 10.1097/MD.0000000000004704
24. Sathwara JA, Balasubramaniam G, Bobdey SC, Jain A, Saoba S. Sociodemographic factors and late-stage diagnosis of breast cancer in India: A hospital-based study. *Indian J Med Paediatr Oncol* (2017) 38(03):277–81. doi: 10.4103/ijmpo.ijmpo\_15\_16
25. Gebremariam A, Addissie A, Worku A, Assefa M, Kantelhardt EJ, Jemal A. Perspectives of patients, family members, and health care providers on late diagnosis of breast cancer in Ethiopia: A qualitative study. *PLoS One* (2019) 14(8):e0220769. doi: 10.1371/journal.pone.0220769
26. Ibrahim NA, Oludara MA. Socio-demographic factors and reasons associated with delay in breast cancer presentation: A study in Nigerian women. *Breast* (2012) 21(3):416–8. doi: 10.1016/j.breast.2012.02.006
27. Pace LE, Mpunga T, Hategekimana V, Dusengimana J-MV, Habineza H, Bigirimana JB, et al. Delays in breast cancer presentation and diagnosis at two rural cancer referral centers in Rwanda. *Oncologist* (2015) 20(7):780–8. doi: 10.1634/theoncologist.2014-0493
28. Grosse Frie K, Kamaté B, Traoré CB, Ly M, Mallé B, Coulibaly B, et al. Factors associated with time to first healthcare visit, diagnosis and treatment, and their impact on survival among breast cancer patients in Mali. *PLoS One* (2018) 13(11):e0207928. doi: 10.1371/journal.pone.0207928
29. Khan MA, Hanif S, Iqbal S, Shahzad MF, Shafique S, Khan MT. Presentation delay in breast cancer patients and its association with sociodemographic factors in north Pakistan. *Chin J Cancer Res* (2015) 27(3):288–93. doi: 10.3978/j.issn.1000-9604.2015.04.11
30. Asoogo C, Duma SE. Factors contributing to late breast cancer presentation for health care among women in Kumasi, Ghana. *Curationis* (2015) 38(1):1–7. doi: 10.4102/curationis.v38i1.1287
31. Akuoko CP, Armah E, Sarpong T, Quansah DY, Amankwa I, Boateng D. Barriers to early presentation and diagnosis of breast cancer among African women living in sub-Saharan Africa. *PLoS One* (2017) 12(2):e0171024. doi: 10.1371/journal.pone.0171024
32. Chintamani, Tuteja A, Khandelwal R, Megha T, Bama R, Jain S, et al. Patient and provider delays in breast cancer patients attending a tertiary care centre: A prospective study. *JRSM Short Rep* (2011) 2(10):1–4. doi: 10.1258/shorts.2011.011006
33. Shahbazi Z, Hazra D, Park S, Byun YC. Toward improving the prediction accuracy of product recommendation system using extreme gradient boosting and encoding approaches. *Symmetry (Basel)* (2020) 12(9):1566. doi: 10.3390/sym12091566
34. Ganggayah MD, Taib NA, Har YC, Lio P, Dhillon SK. Predicting factors for survival of breast cancer patients using machine learning techniques. *BMC Med Inform. Decis. Mak.* (2019) 19(1):48. doi: 10.1186/s12911-019-0801-4
35. Calix RA, Gupta R, Gupta M, Jiang K. (2017). Deep gramulator: Improving precision in the classification of personal health-experience tweets with deep learning. in: *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, Kansas City, MO, USA. pp. 1154–9. doi: 10.1109/BIBM.2017.8217820
36. Fan H, Ma Z, Li H, Wang D, Liu J. Enhanced answer selection in CQA using multi-dimensional features combination. *Tsinghua Sci Technol* (2019) 24(3):346–59. doi: 10.26599/TST.2018.9010050
37. Breiman L. Random forests. *Mach Learn* (2001) 45:5–32. doi: 10.1023/A:1010933404324
38. Uddin S, Khan A, Hossain ME, Moni MA. Comparing different supervised machine learning algorithms for disease prediction. *BMC Med Inform. Decis. Mak.* (2019) 19(1):281. doi: 10.1186/s12911-019-1004-8
39. Gupta A, Katarya R. Social media based surveillance systems for healthcare using machine learning: A systematic review. *J Biomed Inform.* (2020) 108:103500. doi: 10.1016/j.jbi.2020.103500
40. Liu L. (2018). Research on logistic regression algorithm of breast cancer diagnose data by machine learning. in: *2018 International Conference on Robots & Intelligent System (ICRIS)*. pp. 157–60. doi: 10.1109/ICRIS.2018.00049
41. Cramer JS. *The Origins of Logistic Regression: Tinbergen Institute Discussion Papers*. No 02-119/4, Tinbergen Institute (2002).
42. Wolke R, Schwetlick H. Iteratively reweighted least squares: Algorithms, convergence analysis, and numerical comparisons. *SIAM J Sci Stat Comput* (1988) 9(5):907–21. doi: 10.1137/0909062
43. Schmidhuber J. Deep learning in neural networks: An overview. *Neural Networks* (2015) 61:85–117. doi: 10.1016/j.neunet.2014.09.003
44. Hastie T, Tibshirani R, Friedman J. *The elements of statistical learning*. New York, NY: Springer New York (2009).
45. Dalwinder S, Birmohan S, Manpreet K. Simultaneous feature weighting and parameter determination of neural networks using ant lion optimization for the classification of breast cancer. *Biocybern. Biomed Eng.* (2020) 40(1):337–51. doi: 10.1016/j.bbe.2019.12.004
46. Larose DT, Larose CD. *Discovering knowledge in data*. Hoboken, NJ, USA: John Wiley & Sons, Inc. (2014).
47. Chen M, Liu Q, Chen S, Liu Y, Zhang C-H, Liu R. XGBoost-based algorithm interpretation and application on post-fault transient stability status prediction of power system. *IEEE Access* (2019) 7:13149–58. doi: 10.1109/ACCESS.2019.2893448
48. Weldegebriel HT, Liu H, Haq AU, Bugingo E, Zhang D. A new hybrid convolutional neural network and eXtreme gradient boosting classifier for recognizing handwritten Ethiopian characters. *IEEE Access* (2020) 8:17804–18. doi: 10.1109/ACCESS.2019.2960161
49. Chen T, Guestrin C. (2016). XGBoost. in: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. (Ramraj S, Nishant Uzir, Sunil R, and Shatadeep Banerjee). pp. 785–94. doi: 10.1145/2939672.2939785
50. Bruch S, Ernst L, Schulz M, Ziegowski L, Tolba RH. Best variable identification by means of data-mining and cooperative game theory. *J Biomed Inform* (2021) 113:103625. doi: 10.1016/j.jbi.2020.103625
51. Wang L, Wang Y, Chang Q. Feature selection methods for big data bioinformatics: A survey from the search perspective. *Methods* (2016) 111:21–31. doi: 10.1016/j.jymeth.2016.08.014
52. George P, Chandwani S, Gabel M, Ambrosone CB, Rhoads G, Bandera EV, et al. Diagnosis and surgical delays in African American and white women with early-stage breast cancer. *J Women's Heal* (2015) 24(3):209–17. doi: 10.1089/jwh.2014.4773
53. Odongo J, Makumbi T, Kalungi S, Galukande M. Patient delay factors in women presenting with breast cancer in a low income country. *BMC Res Notes* (2015) 8(1):467. doi: 10.1186/s13104-015-1438-8
54. Forbes LJJ, Warburton F, Richards MA, Ramirez AJ. Risk factors for delay in symptomatic presentation: A survey of cancer patients. *Br J Cancer* (2014) 111(3):581–8. doi: 10.1038/bjc.2014.304
55. Mirfarhadi N, Ghanbari A, Khalili M, Rahimi A. Predictive factors for diagnosis and treatment delay in Iranian women with breast cancer. *Nurs. Midwifery Stud* (2016) 6(2):1735–8639. doi: 10.5812/nmsjournal.27452

56. Gueye M, Gueye SMK, Diallo M, Thiam O, Mbodji A, Diouf A, et al. Sociodemographic factors associated with delays in breast cancer. *Open J Obstet. Gynecol.* (2017) 07(04):455–63. doi: 10.4236/ojog.2017.74047
57. Neal RD, Allgar VL. Sociodemographic factors and delays in the diagnosis of six cancers: analysis of data from the 'National survey of NHS patients: Cancer. *Br J Cancer* (2005) 92(11):1971–5. doi: 10.1038/sj.bjc.6602623
58. Ghazali SM, Othman Z, Cheong KC, Lim KH, Mahiyuddin WR, Kamaluddin MA, et al. Non-practice of breast self examination and marital status are associated with delayed presentation with breast cancer. *Asian Pacific J Cancer Prev* (2013) 14(2):1141–5. doi: 10.7314/APJCP.2013.14.2.1141
59. Zhang H, Wang G, Zhang J, Lu Y, Jiang X. Patient delay and associated factors among Chinese women with breast cancer. *Med (Baltimore)*. (2019) 98(40):e17454. doi: 10.1097/MD.00000000000017454
60. Huo Q, Cai C, Zhang Y, Kong X, Jiang L, Ma T, et al. Delay in diagnosis and treatment of symptomatic breast cancer in China. *Ann Surg Oncol* (2015) 22(3):883–8. doi: 10.1245/s10434-014-4076-9
61. Li B, Xia L, Yang J, Wen M, Yu M, Mou E, et al. Enhancing social support and knowledge perception decreases patient delay in breast cancer. *Gland Surg* (2021) 10(7):2220–31. doi: 10.21037/ggs-21-227
62. Gulzar F, Akhtar MS, Sadiq R, Bashir S, Jamil S, Baig SM. Identifying the reasons for delayed presentation of Pakistani breast cancer patients at a tertiary care hospital. *Cancer Manage Res* (2019) 11:1087–96. doi: 10.2147/CMAR.S180388
63. Jedy-Agba E, McCormack V, Adebamowo C, Dos-Santos-Silva I. Stage at diagnosis of breast cancer in sub-Saharan Africa: A systematic review and meta-analysis. *Lancet Glob Health* (2016) 4(12):e923–35. doi: 10.1016/S2214-109X(16)30259-5