Toygar et al. BMC Medical Informatics and Decision Making (2025) 25:313
https://doi.org/10.1186/s12911-025-03155-9

BMC Medical Informatics and Decision Making

RESEARCH Open Access

A machine learning approach to predict self-efficacy in breast cancer survivors

İsmail Toygar[1*†], Su Özgür[2,3†], Gülcan Bağçivan[4,8], Ezgi Karaçam[5,9], Hilal Benzer[6], Ferda Akyüz Özdemir[1], Halise Taşkın Duman[1] and Özlem Ovayolu[7]

Abstract

Purpose To determine predictors of self-efficacy in breast cancer survivors and identify vulnerable groups.

Methods This descriptive study was conducted between November 2023 and April 2024 at three hospitals in Türkiye and involved 430 breast cancer survivors. Data were collected through face-to-face surveys using a patient identification form and the Breast Cancer Survivor Self-Efficacy Scale. This study identified patient characteristics that indicate a tendency towards higher self-efficacy using four machine learning models; Logistic Regression (LR), Random Forest (RF), Support Vector Machine (SVM), XGBoost (XGB).

Results The mean age of participants was 50.7 ± 11.5 years. Majority of the participants ($n = 425$) were female. AUC values were used as ranker for the machine learning models. The ranks of the models were as follows; logistic regression model (0.715), RF (0.710), SVM (0.704), and XGBoost (0.694). Education level ranked first in the LR (0.3874), RF (0.3290), and SVM (0.1250) models, and was the second most important variable in the XGB (0.2327) model. Conversely, the cancer stage stood out in the LR (0.2466) and RF (0.1935) models, ranking third and fourth, respectively, while it ranked third in SVM (0.0683) and fourth in XGB (0.1872). Additionally, comorbidity ranked third in importance in the LR (0.2213) and RF (0.1681) models, but second in SVM (0.0705) and seventh in XGB (0.1393).

Conclusion The study demonstrated that the self-efficacy of breast cancer survivors was associated with their sociodemographic and medical characteristics. These characteristics may assist healthcare professionals in enhancing the care provided to breast cancer survivors. It is of the utmost importance to consider the aforementioned patient group as being vulnerable with regard to breast cancer survivor self-efficacy. There is a clear need for a focus on this vulnerable cohort.

Keywords Breast cancer, Survivorship, Self-efficacy, Machine learning

†İsmail Toygar and Su Özgür are joint first authors.

*Correspondence:
İsmail Toygar
ismail.toygar1@gmail.com
[1]Fethiye Faculty of Health Sciences, Muğla Sıtkı Koçman University, Fethiye, Muğla 48330, Turkey
[2]Ege University Faculty of Medicine, EgeSAM-Translational Pulmonary Research Center, Bornova, İzmir 35100, Turkey
[3]Regional Hub for Cancer Registration in Northern Africa, Central and Western Asia, WHO/IARC GICR, İzmir, Turkey
[4]Koç University Faculty of Nursing, İstanbul 34010, Turkey
[5]Dr Sadi Konuk Training and Research Hospital, İstanbul 30110, Turkey
[6]Hasan Kalyoncu University Vocational School, Şahinbey, Gaziantep 27410, Turkey
[7]Gaziantep University Faculty of Health Sciences, Şahinbey, Gaziantep 27410, Turkey
[8]College of Nursing and Health Sciences, University of Massachusetts Dartmouth, North Dartmouth, MA, USA
[9]Istanbul University – Cerrahpaşa, Postgraduate Education Institute, İstanbul, Turkey

Introduction

Breast cancer is the most prevalent form of cancer in 157 of the 185 countries worldwide [1, 2]. In 2022, the World Health Organization (WHO) reported that breast cancer was responsible for 670,000 deaths globally. It is anticipated that the incidence of breast cancer will increase by over 40% by 2040, reaching 3 million cases annually, while mortality will rise by over 50%, reaching 1 million [3]. The incidence and prevalence of breast cancer have risen consistently in recent years, primarily as a result of notable breakthroughs in early detection, enhanced treatment modalities, and greater healthcare accessibility, which have facilitated earlier diagnoses and extended survival rates. Nevertheless, despite these developments, patients still face considerable challenges, including the emotional and financial burdens of prolonged treatments, the risk of recurrence, and the long-term physical and psychological side effects of cancer therapies, all of which can significantly impact their quality of life [4, 5]. Therefore, a more comprehensive understanding and management of the consequences of cancer and its treatment, including both emotional and physical needs among survivors, is crucial for the advancement of cancer care in breast cancer [6].

A growing body of evidence indicates that elevated levels of self-efficacy are linked to superior health outcomes, including diminished symptoms of anxiety and depression, enhanced quality of life, and an augmented capacity to cope with the adverse effects of treatment [7, 8]. Self-efficacy can be defined as an individual's belief in their capacity to engage in the requisite behaviors to attain specific objectives. It signifies the extent to which an individual is confident in their capacity to amass the cognitive, motivational, emotional, and behavioural resources indispensable for the attainment of a goal, the navigation of a particular situation, or the completion of a task [9, 10]. A sense of responsibility is reflected in self-efficacy, which persists throughout pathogenic processes. Additionally, it is associated with an improved quality of life, emotional well-being, and a stronger commitment to rehabilitation activities. Moreover, high self-efficacy levels are associated with enhanced personal well-being. The concept of self-efficacy in cancer patients is one that is subject to constant change, influenced by a number of factors including diagnosis, therapy and prognosis. A variety of cancer-related factors also exert an influence, including symptom, diagnosis, stage, treatment, spiritual beliefs, pain level, exhaustion, sadness and anxiety [9–11]. Mental and physical health are also significant factors that influence self-efficacy in individuals affected by cancer. A positive attitude, for instance, has been demonstrated to enhance self-efficacy while concurrently reducing stress and anxiety. The successful completion of tasks, such as observing the achievements of others with similar experiences, as well as receiving encouraging feedback and support from others, has been shown to enhance self-efficacy in cancer patients [9, 10].

The term "machine learning (ML)" is a broad one, encompassing a variety of models and strategies based on algorithmic modelling. In the field of epidemiology, the term 'regression' is typically used to refer to a range of frequentist regression models, including logistic, linear, and Cox proportional hazards models [12]. These models are commonly employed in epidemiological and biostatistical research. An increasing number of sociological, psychological and biological phenomena are being modelled using machine learning algorithms [13]. The application of machine learning to predict self-efficacy and associated factors in breast cancer entails the utilisation of sophisticated algorithms to analyse intricate datasets, which encompass clinical, demographic, and psychological variables. These models are capable of identifying patterns and relationships that may be overlooked by traditional statistical methods, thereby facilitating more accurate predictions of a patient's self-efficacy level [14]. The analysis of large datasets by ML allows for the identification of trends and variables that affect patient outcomes. This facilitates the development of more precise prognostic models, which can inform therapeutic choices and improve survival rates. By employing techniques such as logistic regression, decision trees, and neural networks, we can gain deeper insights into the factors influencing self-efficacy, leading to enhanced personalised interventions and support for breast cancer patients [15, 16].

Literature review

A growing body of research has explored the role of self-efficacy in breast cancer, demonstrating its significant impact on treatment adherence, emotional well-being, and overall health outcomes [17–20].

In a systematic review conducted by BorjAlilu et al. (2017) on the significant factors affecting the perception of self-efficacy in breast cancer and to delineate the role of self-efficacy in breast cancer, it was elucidated that the demographic variables that influence self-efficacy in breast cancer are diagnosis, treatment, body image, quality of life, bio-psycho-social-mental status, and physician-patient interaction. In the same review, it was demonstrated that self-efficacy affects a number of important health-related outcomes, including physical and mental health, pain management, quality of life, body image, clinician-patient communication, and health information-seeking behaviour [18]. The results of another systematic review conducted by Baik et al. (2020) demonstrated that elevated levels of cancer-relevant self-efficacy were associated with superior overall health-related quality of life, enhanced social, emotional, and functional well-being, and a reduced burden of breast cancer symptoms and cancer-specific distress [7].

The application of machine learning (ML) in breast cancer research has been extensive, with the objective of enhancing diagnostic accuracy, optimising treatment plans, predicting risk factors, prognosis, cancer recurrence and patient outcomes through the analysis of clinical, imaging and genomic data [21–24].

The utilisation of the ML for the purpose of predicting patient outcomes is comparatively limited in scope, as opposed to its deployment for the accurate prediction of diagnostic risk factors, prognoses and the optimisation of treatment options. The results of a recent scoping review conducted by Pezzolato et al. (2023) included twenty-one papers describing predictive models for psychological distress, quality of life, and adherence to medication in breast cancer [25]. The review demonstrated that treatment-related factors, such as side-effects, type of surgery or treatment received, Socio-demographic factors, such as younger age, lower income, and inactive occupational status, may predict poorer outcomes. Similarly, clinical factors, including advanced disease stage, comorbidities, and physical symptoms such as fatigue, insomnia, and pain, may also predict poorer outcomes. Additionally, psychological variables, such as anxiety, depression, and body image dissatisfaction, may also predict poorer outcomes.

Although machine learning (ML) has previously been employed to predict quality of life and psychological outcomes in patients, to the best of our knowledge, no study has focused on the use of ML to predict self-efficacy in breast cancer survivors. This study aimed to identify predictors of self-efficacy in breast cancer survivors using ML.

Methods

Study design, setting, time

This is a descriptive study conducted in two state and two university hospital in Türkiye between November 2023 - April 2024. Flow diagram of the study is presented in Fig. 1.

Sampling

The study population consisted of breast cancer survivors who admitted to hospital for treatment and follow-up. There is no reported method for calculating the sample in machine learning. Therefore, the aim was to reach the maximum number of patients in the sample of this study and to have a high reliability and consistency (area under the curve - AUC) of the prediction. A total of 430 breast cancer survivors who meet the inclusion criteria participated in this study.

Inclusion and exclusion criteria

Patients who are 18 years or older, diagnosed with breast cancer at least six months age, without cognitive or functional limitations, able to communicate in Turkish, and willing to participate in the study. Patients who did

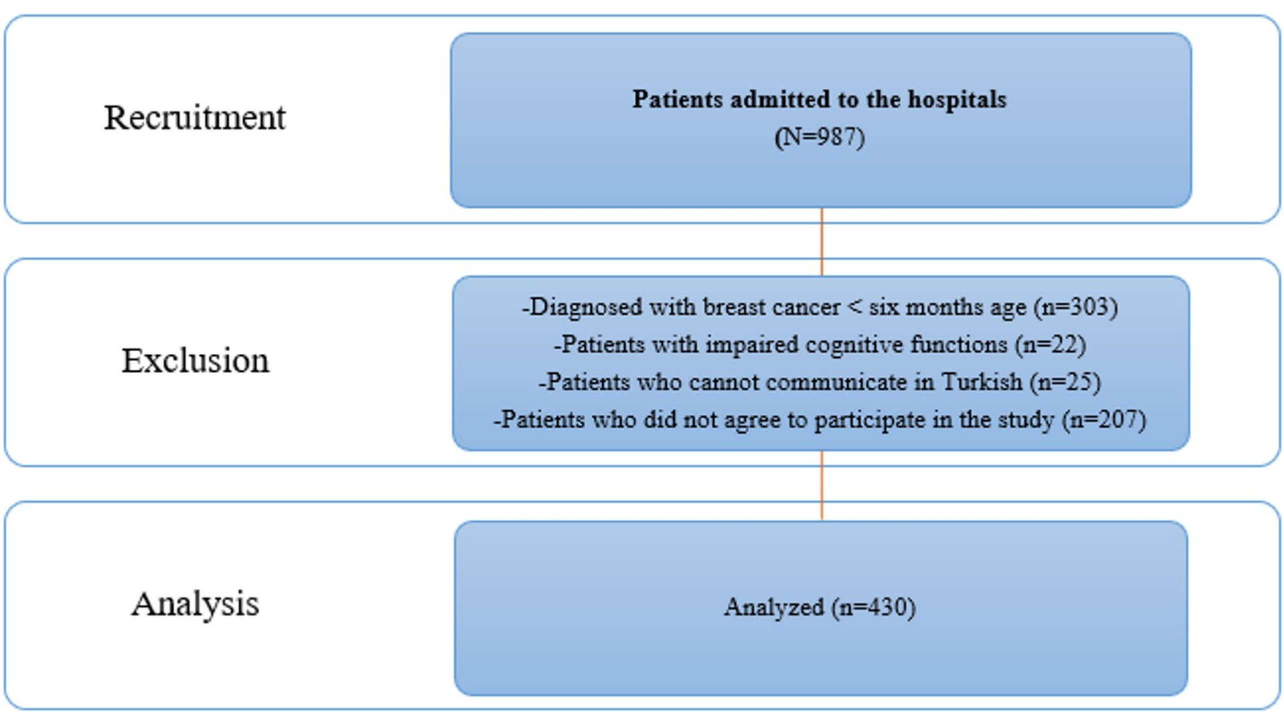

Fig. 1 Flow diagram of the study

not meet the inclusion criteria were excluded. The tool, breast cancer survivors self-efficacy scale, was translated into Turkish, and its validity and reliability were evaluated and confirmed for Turkish-speaking individuals. For this reason, those who were unable to communicate in Turkish were excluded from the study.

Dependent and independent variables of the study

The dependent variable of the study was breast cancer survivor self-efficacy and the independent variables were age, gender, time since diagnosis, disease stage, number of comorbidities, type of treatment, income level, education level, marital status, employment status, history of mastectomy and presence of lymphedema.

Measures

A socio-demographic and disease-related data collection form and a breast cancer survivor self-efficacy scale were used to collect the data.

Sociodemographic and disease-related data collection form: This form was developed by the researchers in accordance with the relevant literature [16, 26]. The form includes the patient's age, gender, duration since diagnosis, stage of disease, number of comorbidities, type of treatment, income level, education level, marital status, employment status, history of mastectomy and presence of lymphoedema.

Breast cancer survivor self-efficacy scale: The scale was developed by Champion et al. (2013) to assess the self-efficacy of breast cancer survivors. The Turkish validity and reliability study of the scale was conducted by Uslu et al. (2023) and the scale was reported as a valid and reliable measurement tool for the Turkish population. The scale consists of 11 items and one dimension and is scored as a five-point Likert scale. The score range of the scale is 11–55 and high scores indicate high levels of self-efficacy. The scale has no cut-off score. The original scale has one dimension. In the Turkish validity and reliability study of the scale, a two-factor structure was identified. These are 'self-care and coping' (items 2,3,5,6,6,7,7,8,9,10) and 'self-help seeking' (items 1,4,11). However, as with the original scale, it is recommended that the total score be assessed. The reliability coefficient of the scale for the total score was reported as Cronbach's $\alpha = 0.852$, Cronbach's $\alpha = 0.823$ for the first factor sub-dimension and Cronbach's $\alpha = 0.776$ for the second factor sub-dimension. The correlation coefficients of the scale items were reported to be between 0.50 and 0.61 [27, 28].

Data collection

The data were collected face-to-face. The data collection process for each patient lasted approximately ten minutes.

As the research was conducted across multiple centres, the following principles were adhered to in order to ensure consistency and standardisation in the explanation of the research purpose and the application process.

The research was explained to the participants in writing, with the informed consent form serving as the sole instrument of explanation. The same informed consent form was utilised in all participating centres.

In all centres, a suitable empty room was made available to the patient, who was then asked to complete the data collection forms. Patients were not constrained by time limitations and were permitted to complete the questionnaire at their own pace. Consequently, standardisation was ensured in the data collection process.

Statistical analysis

Basic statistics were performed using IBM SPSS V25 (IBM Corp. Released 2017. IBM SPSS Statistics for Windows, Version 25.0. Armonk, NY: IBM Corp.). In the comparison of means/medians between two independent groups for continuous data, the independent sample t-test/Mann-Whitney U test was employed. The chi-square test was utilized for evaluating categorical data. To predict self-efficacy in groups, participants were divided into two categories according to their self-efficacy score: below average and above average. Statistical significance was determined at the $p < 0.05$ level.

Also, machine learning analyses were performed using Ddsv4-series Azure Virtual Machines with a vCPU count of 32 and a memory capacity of 128 GiB.

Data Splitting Strategy: In this study, we divided the dataset into three subsets: training (70%), test (20%), and validation (10%). This approach is widely accepted in the machine learning literature, as it ensures a balanced methodology for training, tuning, and evaluating the model. The training set is used to develop the model, the validation set is employed to fine-tune the hyperparameters and prevent overfitting, and the test set is reserved for evaluating the final model's performance on unseen (has not encountered during the training process) data. This splitting strategy aligns with standard practices outlined in previous studies and has been shown to optimize model performance while ensuring robust generalization to new data. By allocating 70% of the data to training, the model has sufficient data to learn patterns effectively, while the 20% validation and 10% test subsets allow for accurate tuning and evaluation without compromising the integrity of the results [29].

Hyperparameter optimization was performed for all algorithms used in the study, including Logistic Regression, Random Forest, Support Vector Machine, and XGBoost. This process was carried out using Microsoft Azure Machine Learning's HyperDrive framework, which systematically explores defined parameter ranges

[30]. In the current study, authors tuned key hyperparameters that directly influence model performance -such as max_depth, learning_rate, and n_estimators for XGBoost, or C and gamma for SVM-within carefully selected search spaces. The optimal parameter combinations were identified based on performance on the validation dataset, ensuring that our models were both accurate and generalizable.

The results and parameters of the best model obtained from the analyses conducted in Azure Automated ML have been presented, showcasing the four (logistic regression (LR), random Forest (RF), support vector machine (SVM) and XGBoost (XGB) algorithms best-performing algorithms. While the results of the algorithms were presented with performance metrics, the models' variables were presented with their respective variable importance [31, 32]. In the determination of the best model performance, AUC (Area Under the ROC Curve) was used as the primary metric for selecting the best model because of its robustness and its ability to evaluate classification performance across various thresholds [33]. The current study also reported on accuracy, precision, recall, the F1 score and the Matthews correlation coefficient (MCC) in order to provide a more comprehensive assessment of performance. This combination of metrics ensured that we considered both overall accuracy and the balance between false positives and false negatives. The best model was selected as the one achieving the highest AUC on the validation dataset while also maintaining strong and balanced performance across the other metrics.

Machine learning models

Logistic regression

Logistic regression is a statistical method commonly used in machine learning for binary classification tasks. It models the relationship between one or more independent variables and a dependent variable by applying the logistic function, ensuring that predictions fall between 0 and 1. This allows the model to estimate the probability of an observation belonging to a specific class. Unlike linear regression, which predicts continuous values, logistic regression predicts categorical outcomes by transforming the linear combination of input features into probabilities. The algorithm is efficient, interpretable, and works well with datasets where the relationship between features and the target is approximately linear. Its simplicity makes it a widely used baseline model for evaluating more complex algorithms [29].

Random forest

Random Forest is an ensemble learning method widely used for both classification and regression tasks. It operates by constructing multiple decision trees during training and aggregating their outputs to make predictions, either by voting for classification or averaging for regression. This approach reduces the risk of overfitting that can occur with individual decision trees, resulting in better generalization to unseen data. Each tree is trained on a bootstrapped subset of the data, and feature selection is randomized at each split, increasing diversity among the trees and improving robustness. Random Forest is known for its high accuracy, ability to handle large datasets with many features, and robustness to noise and missing data [34].

Support vector machine

Support Vector Machine (SVM) is a supervised learning algorithm commonly used for classification and regression tasks. It works by identifying a hyperplane that best separates data points into different classes in a high-dimensional space. SVM aims to maximize the margin between the hyperplane (a decision boundary that separates classes in a multidimensional space) and the nearest data points from each class, known as support vectors, which play a crucial role in defining the decision boundary. This approach makes SVM effective in handling both linearly and non-linearly separable data, especially when combined with kernel functions that map data into higher dimensions. SVM is valued for its ability to handle high-dimensional datasets and its robustness against overfitting, especially in small or medium-sized datasets [34].

XGBoost

XGBoost (short for Extreme Gradient Boosting) is a machine learning algorithm widely used for classification and regression tasks. It is based on the gradient boosting framework, where models are built sequentially to minimize prediction errors. Unlike traditional boosting methods, XGBoost introduces improvements such as regularization techniques to reduce overfitting, efficient handling of missing data, and parallel computation for faster training. These features make it highly efficient and accurate for structured datasets. Additionally, it provides flexibility in defining objective functions and supports large-scale data, making it a popular choice in machine learning competitions and real-world applications [35].

Performance metrics

In this study, we utilized several performance metrics to evaluate the effectiveness of our model. These metrics are critical in assessing the model's ability to accurately classify instances and generalize to unseen data [31].

Precision: Precision is the ratio of correctly predicted positive instances to the total predicted positive instances. It measures the model's ability to avoid false positives and is defined as:

$$Precision = \frac{True\,Positives}{True\,Positives + False\,Positives}$$

Recall (Sensitivity): Recall quantifies the model's ability to correctly identify all relevant instances. It is calculated as:

$$Recall = \frac{True\,Positives}{True\,Positives + False\,Negatives}$$

Accuracy: Accuracy is the ratio of correctly predicted instances (both positive and negative) to the total number of instances. It is defined as:

$$Accuracy = \frac{True\,Positives + True\,Negatives}{Total\,Instances}$$

F1 Score: The F1 score is the harmonic mean of precision and recall, balancing the trade-off between false positives and false negatives. It is calculated as:

$$F_1\,Score = 2\,x\,\frac{Precision\,x\,Recall}{Precision + Recall}$$

Matthews Correlation Coefficient (MCC): MCC provides a comprehensive measure of model performance, especially for imbalanced datasets. It considers all elements of the confusion matrix and is given by

$$MCC = \frac{(TP\,x\,TN) - (FP\,x\,FN)}{\sqrt{(TP + FP)\ x\ (TP + FN)\,x\ (TN + FP)\,x\,(TN + FN)}}$$

Area Under the Curve (AUC): AUC represents the area under the Receiver Operating Characteristic (ROC) curve, providing a summary measure of the model's ability to distinguish between classes across various thresholds. A higher AUC indicates better performance.

Confusion matrix terms

True Positive (TP): Instances correctly classified as positive by the model.

False Positive (FP): Instances incorrectly classified as positive, which should have been negative.

False Negative (FN): Instances incorrectly classified as negative, which should have been positive.

True Negative (TN): Instances correctly classified as negative by the model.

The combination of these metrics allows for a thorough evaluation of model performance, addressing both overall accuracy and the balance between precision and recall [31].

Ethical consideration

Ethical permission (dated 15 November 2023 and approval number 143) was obtained from the Muğla Sıtkı Koçman University Health Sciences Ethical Committee, and written institutional permissions were obtained from centres for conducting the study. In addition, written and verbal consent were obtained from the participants. The study complied with the principles of the Declaration of Helsinki "Recommendations Guiding Physicians in Biomedical Research Involving Human Subjects", adopted by the 18th World Medical Assembly, Helsinki, Finland, June 1964 (and its successive amendments).

Results

The mean age of the patients participating in the study was 50.9 ± 12.0 years in the group whose self-efficacy score was below the mean and 50.6 ± 11.0 years in the group whose self-efficacy score was above the mean ($p = 0.782$). While 48.9% (208) of the women were in the group with a self-efficacy score below the mean and 51.1% (217) in the group with a self-efficacy score above the mean, 40.0% (2) of the men were in the group with a self-efficacy score below the mean and 60.0% (3) in the group with a self-efficacy score above the mean ($p = 0.999$). 47.7% (163) of patients in the group with a self-efficacy score below the mean and 52.3% (179) of individuals in the group with a self-efficacy score above the mean were married ($p = 0.336$). The distribution of individuals in both groups according to employment status was statistically significant. A significant difference was found ($p = 0.013$). The working frequency of individuals in the group whose self-efficacy score was above the mean was high. The distribution of education level was found to be different in the groups ($p = 0.003$). The frequency of illiteracy (60.7%) was high in the group whose self-efficacy score was below the mean, and the frequency of high school (66.7%) and university (57.4%) was high in the group whose self-efficacy score was above the mean. When income status was evaluated, the income of patients in the group whose self-efficacy score was below the mean was less than their expenses (55.8%), the income of 61.5% of patients in the group whose self-efficacy score was below the mean was equal to their expenses, and the income of 54.8% was more than their expenses ($p = 0.004$). The distribution of cancers by stage was similar between groups ($p = 0.323$). The median time to diagnosis for patients was 12.0 [1.0-264.0] in the group whose self-efficacy score was below the mean and 24.0 [1.0-300.0] in the group whose self-efficacy score was above the mean ($p = 0.102$). The frequency distribution of comorbidities was similar between groups ($p = 0.077$). The frequency of mastectomy ($p = 0.089$) and lymphedema ($p = 0.608$) was similar between groups (Table 1).

Table 1 Comparison of descriptive statistics of participants by self-efficacy score

Variables	Category	Group		Total	Test statistics; p-value
		Below the Mean	Above the Mean		
Gender	Female	208 (48.9)	217 (51.1)	425 (100)	$X^2=0.158$ $p=0.999$
	Male	2 (40.0)	3 (60.0)	5 (100)	
Age	Mean ± SD	50.9 ± 12.0	50.6 ± 11.0	50.7 ± 11.5	t = 0.277 $p=0.782$
	Med [Min-Max]	51.0 [20.0–80.0]	50.0 [24.0–80.0]	50.0 [20.0–80.0]	
Marital status	Married	163 (47.7)	179 (52.3)	342 (100)	$X^2=0.926$ $p=0.336$
	Single	47 (53.4)	41 (46.6)	88 (100)	
Work status	Housewife	146 (50.7)	142 (49.3)	288 (100)	$X^2=10.702$ $p=0.013$
	Working	9 (23.7)	29 (76.3)	38 (100)	
	Not Working	38 (52.8)	34 (47.2)	72 (100)	
	Retired	17 (53.1)	15 (46.9)	32 (100)	
Educational Level	Illiterate	71 (60.7)	46 (39.3)	117 (100)	$X^2=14.242$ $p=0.003$
	Primary Education	95 (49.0)	99 (51.0)	194 (100)	
	High School	24 (33.3)	48 (66.7)	72 (100)	
	University	20 (42.6)	27 (57.4)	47 (100)	
Income Level	Income Less Than Expenses	134 (55.8)	106 (44.2)	240 (100)	X2 = 11.232 $p=0.004$
	Equal to Income	57 (38.5)	91 (61.5)	148 (100)	
	Income Exceeds Expenses	19 (45.2)	23 (54.8)	42 (100)	
Disease duration (months)	Mean ± SD	32,0 ± 46.1	37,0 ± 45.5	30.6 ± 45.8	U = 2098.0 $p=0.102$
	Med [Min-Max]	12.0 [1.0-264.0]	24.0 [1.0-300.0]	18.0 [1.0-300.0]	
Cancer stage	Cured	1 (25.0)	3 (75.0)	4 (100)	$X^2=4.402$ $p=0.367$
	Stage 1	37 (44.0)	47 (56.0)	84 (100)	
	Stage 2	70 (48.3)	75 (51.7)	145 (100)	
	Stage 3	67 (55.8)	53 (44.2)	120 (100)	
	Stage 4	35 (45.5)	42 (54.5)	77 (100)	
Comorbidity	Absence	88 (54.3)	74 (45.7)	162 (100)	$X^2=3.128$ $p=0.077$
	Presence	122 (45.5)	146 (54.5)	268 (100)	
Treatment	Chemotherapy	87 (51.5)	82 (48.5)	169 (100)	$X^2=12.199$ $p=0.007$
	Immunotherapy + Surgery	9 (34.6)	17 (65.4)	26 (100)	
	Chemotherapy + Surgery	95 (45.5)	114 (54.5)	209 (100)	
	Other	17 (81.0)	4 (19.0)	21 (100)	
Mastectomy	Absence	104 (53.3)	91 (56.7)	195 (100)	$X^2=2.887$ $p=0.089$
	Presence	106 (45.1)	129 (54.9)	235 (100)	
Lymphedema	Absence	150 (48.1)	162 (51.9)	312 (100)	$X^2=0.263$ $p=0.608$
	Presence	60 (50.8)	58 (49.2)	118 (100)	

X^2: Chi square test statistics, t: Independent sample t test statistics, U: Mann Whitney U test statistics, $p<0.05$ Significance level

In the study, performance metrics such as accuracy, AUC, prediction, recall, F1 score, and Matthews correlation were presented for the four models demonstrating the best prediction performance.

The Accuracy values for Logistic Regression, Random Forest, and Support Vector Machine were approximately 0.647, indicating comparable overall performance. In contrast, XGBoost achieved a slightly lower Accuracy of 0.588, suggesting reduced overall classification performance. Using the logistic regression model obtained the highest AUC value (0.715), highlighting its superior ability to distinguish between classes. RF (0.710) and SVM (0.704) followed closely, while XGBoost had the lowest AUC (0.694), reflecting its comparatively reduced discriminatory power. XGBoost achieved the highest Precision (0.743), indicating its ability to minimize false positives effectively. RF and SVM demonstrated comparable Precision values (0.738 and 0.707, respectively), while Log Reg scored slightly lower (0.716). Logistic regression model achieved the highest F1-score (0.647), indicating a balanced trade-off between Precision and Recall. RF (0.618) and SVM (0.639) showed similar performance, while XGBoost had the lowest F1-score (0.588), reflecting challenges in balancing Precision and Recall. MCC values further emphasized the differences among the models. Log Reg achieved the highest MCC (0.298), indicating the most reliable predictions, whereas XGBoost had the lowest MCC (0.180), suggesting weaker performance in balancing true and false predictions. Recall values for Log Reg, RF, and SVM were identical (0.647), indicating their ability to identify true positives with the same effectiveness. XGBoost, however, had a

Table 2 Performans metrics of the prediction models

Metrics/Algorithm	TP	FP	FN	TN	Precision	Recall	Accuracy	F1 Score	MCC	AUC
Logistic Regression	13	5	8	17	0.72222	0.61905	0.64706	0.66667	0.39696	0.71528
Random Forest	12	4	9	18	0.75000	0.57143	0.61765	0.64865	0.40291	0.71007
Support Vector Machine	13	5	8	17	0.72222	0.61905	0.64706	0.66667	0.39696	0.70486
XGBoost	12	4	9	18	0.75310	0.57143	0.58824	0.64865	0.40291	0.69444

AUC: Area under the curve, TP: True Positive, FP: False Positive, FN: False Negative, TN: True Negative, MCC: Matthews Correlation Coefficient

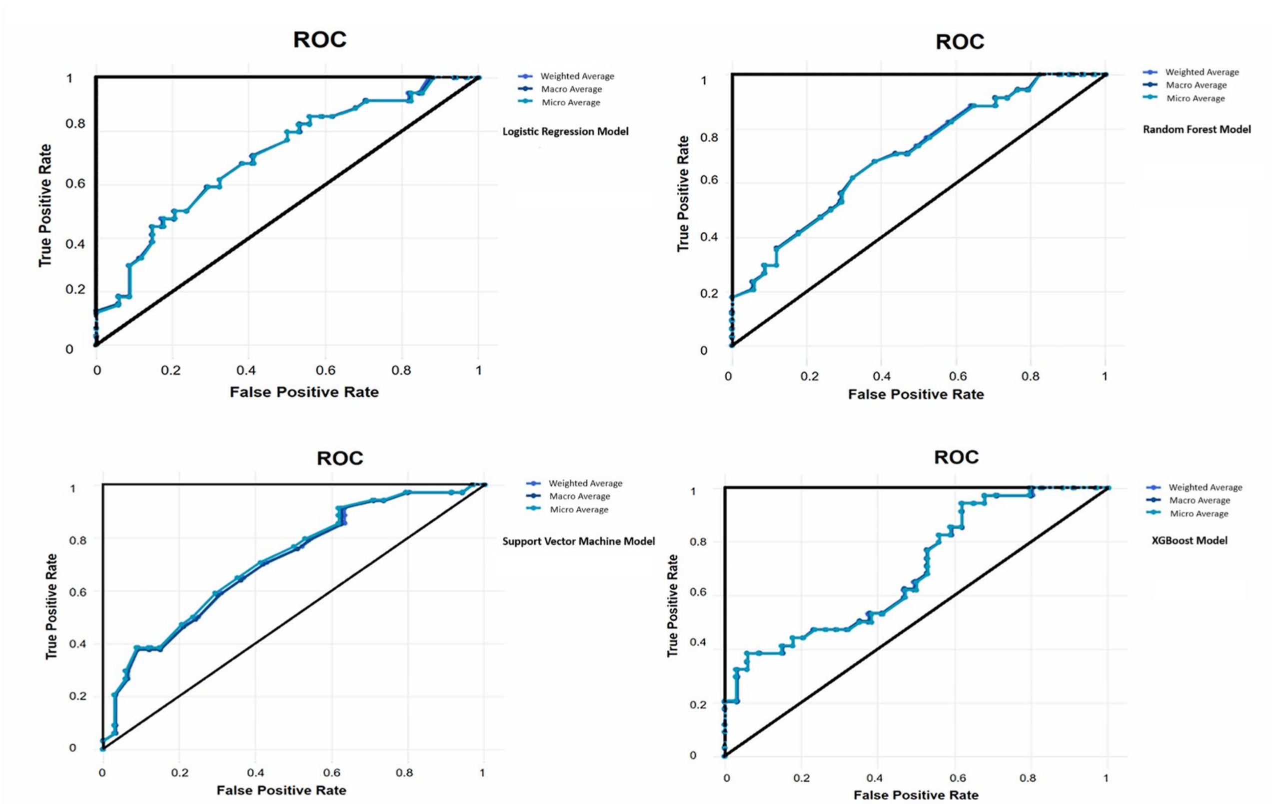

Fig. 2 ROC of the machine learning models

lower Recall (0.588), indicating it missed a higher proportion of actual positives compared to the other models. (Table 2; Fig. 2).

Education level ranked first in the LR (0.3874), RF (0.3290), and SVM (0.1250) models, and was the second most important variable in the XGB (0.2327) model. Conversely, the cancer stage stood out in the LR (0.2466) and RF (0.1935) models, ranking third and fourth, respectively, while it ranked third in SVM (0.0683) and fourth in XGB (0.1872). Additionally, comorbidity ranked third in importance in the LR (0.2213) and RF (0.1681) models, but second in SVM (0.0705) and seventh in XGB (0.1393) (Table 3).

Discussion

The current study used four machine learning models to determine the predictors of breast cancer survivor's self-efficacy. Machine learning-based logistic regression showed the best prediction performance. Education level, cancer stage, comorbidity, and age were the leading predictors in all models. Additionally, logistic regression model revealed that employment status, income level, mastectomy, treatment regimen, disease duration, and lymphedema were other predictors of the self-efficacy in breast cancer survivors.

The current study demonstrated a positive causal relationship between higher levels of education and higher levels of self-efficacy among breast cancer survivors. Those with higher levels of education often demonstrate greater health literacy, a more comprehensive understanding of their condition, treatment options and the healthcare system. This allows individuals to take a more active role in the decision-making process regarding their care [36]. This allows them to take a more active role in the decision-making process regarding their care [37]. This level of engagement can foster a sense of control and

Table 3 Variable importance of the prediction models

	Model 1-Logistic Regression		Model 2-Random Forest		Model 3-Support Vector Machine		Model 4-XGBoost	
	Variables	**Variable Importance**	**Variables**	**Variable Importance**	**Variables**	**Variable Importance**	**Variables**	**Variable Importance**
1	Education level	0.3874	Education level	0.3290	Education level	0.1250	Disease duration	0.2888
2	Cancer stage	0.2466	Cancer stage	0.1935	Comorbidity	0.0705	Education level	0.2327
3	Comorbidity	0.2213	Comorbidity	0.1681	Cancer stage	0.0683	Age	0.1972
4	Work status	0.1536	Income	0.1681	Age	0.0595	Cancer stage	0.1872
5	Age	0.1443	Disease duration	0.1321	Work status	0.0432	Treatment	0.1453
6	Income	0.0111	Age	0.1077	Income	0.0363	Income	0.1403
7	Mastectomy	0.1089	Work status	0.0775	Treatment	0.0299	Comorbidity	0.1393
8	Treatment	0.0649	Mastectomy	0.0619	Mastectomy	0.0256	Work status	0.0612
9	Disease duration	0.0236	Treatment	0.0595	Disease duration	0.0094	Mastectomy	0.0501
10	Lymphedema	0.0024	Marital status	0.0229	Lymphedema	0.0021	Lymphedema	0.0453

agency, which, in turn, can lead to higher levels of self-efficacy in patients with breast cancer. This is because they feel more able to cope with the challenges associated with the disease. In addition, the level of education influences a patient's access to information, resources and support networks, all of which are important in promoting self-efficacy [38, 39]. Patients with lower levels of education may face barriers in understanding medical information, navigating complex treatment regimens, and advocating for their needs within the healthcare system. Consequently, they may be prone to feelings of uncertainty, fear, and helplessness, which can diminish their self-efficacy and overall quality of life during the cancer journey or beyond. Income was another predictor of self-efficacy in the current study. Yuan et al. (2014) reported that self-efficacy of breast cancer patients was significantly influenced by socio-economic indicators (education, income, employment status and health insurance status) [40]. Our findings on education and income confirm the study reported in the literature. Individuals with lower socioeconomic status are at risk of having lower levels of breast cancer survivor self-efficacy. We recommend healthcare practitioners to be aware of this vulnerable group and implement interventions to increase self-efficacy in this group. Similar to our findings, Yuan et al. (2014) also reported that self-efficacy of the cancer patients significantly varies by socioeconomic status [40].

Cancer stage was also found to be a significant predictor of breast cancer survivors' self-efficacy. The lowest levels of self-efficacy were observed among those with stage 3 disease, while the highest levels were observed among those with stage 1 disease. While self-efficacy is stable over time in early-stage breast cancer patients, in advanced-stage breast cancer, self-efficacy has been reported to be associated with post-traumatic growth [41, 42]. On the other hand, self-efficacy can be influenced by individual, social and cognitive factors. That's why it's not possible to link levels of self-efficacy to cancer stage. With such a complex phenomenon, it's not possible to consider a single factor as determining self-efficacy levels. We recommend that all factors that influence self-efficacy are considered together.

Comorbidity was another predictor of the self-efficacy of breast cancer survivors in the current study. Perkins et al. (2009) reported that comorbidity were not associated with self-efficacy for physical activity in cancer survivors [43]. However, in the current study, self-efficacy was higher in those with comorbidities. We believed that the difference between the current study and other studies in the literature was related to the items of the scales used to measure self-efficacy. The scale developed by Champion et al. (2013) was used for the current study. The scale items includes adherence to disease, back to the work life, dealing with emotions, and asking for help [27]. We believe that coping mechanisms may have been strengthened because they may have previously experienced similar psychological problems due to chronic illness. In self-efficacy for physical activity, physiological capacity is more important than psychological problems, and chronic diseases can negatively affect physiological capacity. This may explain the difference between the literature and our study.

Working status is another important variable in all the prediction models. 76.3% of the survivors who were actively working had a score above the average, while most of the housewives, retired and unemployed survivors had a score below the average. There is a two-sided interaction between the employment status and self-efficacy. Self-efficacy plays a role in the return to work for breast cancer survivors. Breast cancer survivors with higher workability tend to report better self-efficacy. Factors such as lower levels of depression, financial difficulties and physical fatigue are associated with better workability and consequently higher self-efficacy [44]. In their 2023 study, Cheng and colleagues emphasised that returning to work has a positive impact on self-efficacy [45]. This can be facilitated by social engagement, financial stability and a reinvigorated perception of the normality of everyday life. Furthermore, survivors who believe themselves capable of meeting the demands of the workplace tend to display elevated levels of self-confidence, which serves to reinforce their capacity to effectively navigate health-related concerns. Conversely, prolonged periods of unemployment or withdrawal from the workforce may exacerbate feelings of dependency and a sense of diminished identity, which could have a detrimental impact on self-efficacy. Interventions designed to facilitate the reintegration of survivors into the workforce, such as targeted occupational therapy or psychosocial support, could play a pivotal role in enhancing self-efficacy.

Among the treatment options, those including the combination of surgery and immunotherapy or chemotherapy has higher level of self-efficacy. In the current study, breast conservative surgery and mastectomy considered as surgical approach. Enien et al. (2018) reported that patients treated with breast conservative surgery has higher level of quality of life [46]. We believed that surgical approaches would lead to higher self-efficacy in cancer survivors, thanks to their higher survival rate and fewer systemic complications compared.

The duration of the disease was found to be longer in those with higher levels of self-efficacy. Ghayth et al. (2023) reported that there is a positive relationship between the disease duration and self-efficacy in chronic conditions [47]. Self-management skills and coping strategies in chronic conditions develop over time [48, 49]. There is a positive way relationship between the self-management, coping mechanism and self-efficacy [49]. It was concluded that this was the reason for the observed differences between the groups. It is therefore recommended that healthcare practitioners be aware of this issue, namely that breast cancer survivors with a shorter disease duration may be vulnerable to a low level of self-efficacy.

The application of machine learning in healthcare research has the potential to transform the field, offering enhanced capabilities for analysing complex datasets and uncovering patterns that traditional statistical methods often fail to identify. In contrast to conventional regression-based methodologies, machine learning algorithms are capable of processing high-dimensional data, identifying non-linear relationships, and generating predictions with enhanced accuracy. In this study, the logistic regression model exhibited the most optimal predictive performance, with an AUC of 0.715, which surpassed the conventional benchmarks reported in previous studies utilising classical methods. Wiemken and Kelley (2020) emphasised that machine learning offers a substantial advantage in health outcomes research by employing algorithmic modelling to predict health behaviours and clinical outcomes with greater efficacy [12]. By applying machine learning to survivorship research, not only is the comprehension of predictors, such as education level and comorbidities, enhanced, but avenues for personalised care interventions are also opened, thereby improving the quality of life for breast cancer survivors.

Limitations

This study has several limitations that should be considered when interpreting the results. It is noteworthy that this study was conducted in only four hospitals within a single country, which may limit the generalisability of the findings. Furthermore, self-efficacy is influenced by numerous socio-cultural, psychological, and healthcare factors that were not included in this study. It would therefore be beneficial for future research to incorporate a broader range of variables and settings in order to validate these findings. The limited geographic and cultural diversity of the sample may have resulted in a selection bias, which could limit the generalisability of the findings to breast cancer survivors in other regions or healthcare settings. It is important to note that cultural norms, healthcare systems and access to resources can vary significantly between different populations, which may affect the predictors of self-efficacy in diverse populations.

Secondly, although the sample size is relatively large for a machine learning study ($n = 430$), it may still be insufficient for capturing rare predictors or complex interactions between variables. For example, the relatively low number of male participants limits the ability to draw meaningful conclusions about gender differences in self-efficacy. Furthermore, while the data were collected across multiple centres, the study relied on a convenience sample, which may have introduced selection bias. It is possible that patients who agreed to participate were more health-literate or motivated, leading to higher self-efficacy scores compared to the broader survivor population.

Thirdly, self-reported measures were employed to evaluate sociodemographic and disease-related variables, in addition to self-efficacy. These are vulnerable to recall and social desirability biases, which may result in inaccuracies in the reported data. This could have affected the models' capacity to accurately predict self-efficacy, as the input data may not fully reflect the participants' actual experiences or circumstances.

It is also important to note that the absence of external validation for the machine learning models represents a limitation of this study. The findings were based solely on the dataset collected for this study, and therefore their performance may differ when applied to other populations or settings. It would therefore be beneficial for future studies to validate these models using external datasets in order to confirm their predictive accuracy and reliability across diverse contexts.

Notwithstanding these constraints, the study offers valuable insights into the predictors of self-efficacy in breast cancer survivors and demonstrates the utility of machine learning in survivorship research. Extending the scope of future studies to include larger, more diverse populations and additional variables could further enhance the robustness and applicability of the findings.

Conclusion

This study employed four machine learning models to predict self-efficacy in breast cancer survivors, identifying significant predictors such as education level, cancer stage, comorbidities, and age. Although previous studies have employed machine learning in oncology research, including survivorship care, there is, to the best of our knowledge, a paucity of evidence concerning its specific application to the prediction of self-efficacy in breast cancer survivors. This serves to illustrate the novel contribution that this research makes to the existing body of knowledge.

The findings emphasise the necessity of incorporating sociodemographic and clinical variables into the development of interventions for breast cancer survivors. It is recommended that healthcare professionals prioritise the needs of vulnerable groups, including individuals with lower educational attainment, those diagnosed at more advanced stages of disease, and those with lower incomes or shorter disease durations. The findings of this study can inform the development of targeted interventions and policies designed to enhance self-efficacy and, subsequently, quality of life in this population.

Acknowledgements
Preliminary results of this study has been presented as poster presentation at MASCC/AFSOS/ISOO Annual Meeting 2024 and the abstract of this preliminary version has been published in the proceeding books.

Author contributions
All authors contributed to the study conception and design. Material preparation, data collection and analysis were performed by Su Özgür, İsmail Toygar, Gülcan Bağcivan, Hilal Benzer, Ezgi Bilmiç, Ferda Akyüz Özdemir, Halise Taşkın Duman and Özlem Ovayolu. The first draft of the manuscript was written by İsmail Toygar, Su Özgür, Gülcan Bagcivan, Ferda Akyüz Özdemir, Halise Taşkın Duman, and Ezgi Bilmiç. All authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Funding
The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.

Data availability
The datasets used and/or analysed during the current study are available from the corresponding author on reasonable request.

Declarations

Ethics approval
Ethical permission (dated 15 November 2023 and approval number 143) was obtained from the Muğla Sıtkı Koçman University Health Sciences Ethical Committee, and written institutional permissions were obtained from centres for conducting the study. In addition, written and verbal consent were obtained from the participants. The study complied with the principles of the Declaration of Helsinki "Recommendations Guiding Physicians in Biomedical Research Involving Human Subjects", adopted by the 18th World Medical Assembly, Helsinki, Finland, June 1964 (and its successive amendments).

Consent to participate
Informed consent was obtained from all individual participants included in the study.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 30 July 2024 / Accepted: 13 August 2025
Published online: 19 August 2025

References

1. Organization WH. Fact sheets: breast cancer. 2024 [cited 2024 4 June 2024]; Available from: https://www.who.int/news-room/fact-sheets/detail/breast-cancer
2. Coles CE, Earl H, Anderson BO, et al. The Lancet Breast Cancer Comm Lancet. 2024;403(10439):1895–950.
3. Arnold M, Morgan E, Rumgay H, et al. Current and future burden of breast cancer: global statistics for 2020 and 2040. Breast. 2022;66:15–23.
4. Society AC. Breast cancer facts & Figs. 2023–2024. 2024 05.12.2024]; Available from: https://www.cancer.org/research/cancer-facts-statistics/all-cancer-facts-figures/breast-cancer-facts-figures-2023-2024.html
5. Miller KD, Nogueira L, Devasia T, et al. Cancer treatment and survivorship statistics, 2022. Cancer J Clin. 2022;72(5):409–36.
6. Heidary Z, Ghaemi M, Rashidi H. Quality of life in breast cancer patients: A systematic review of the qualitative studies. Cancer Control. 2023;30:10732748231168318.
7. Baik SH, Oswald LB, Buitrago D, et al. Cancer-Relevant Self-Efficacy is related to better Health-Related quality of life and lower cancer-Specific distress and symptom burden among Latina breast cancer survivors. Int J Behav Med. 2020;27(4):357–65.
8. White LL, Cohen MZ, Berger AM, et al. Self-Efficacy for management of symptoms and symptom distress in adults with cancer: an integrative review. Oncol Nurs Forum. 2019;46(1):113–28.
9. Maddux JE, Gosselin JT. Self-efficacy. The Guilford Press; 2012.
10. Luszczynska A, Schwarzer R, Lippke S, et al. Self-efficacy as a moderator of the planning-behaviour relationship in interventions designed to promote physical activity. Psychol Health. 2011;26(2):151–66.
11. Ali R, Draman N, Yusoff M. Self-Efficacy for coping with breast cancer in North-Eastern state of Peninsular Malaysia. Asian Pac J Cancer Prev. 2020;21(10):2971–8.
12. Wiemken TL, Kelley RR. Machine learning in epidemiology and health outcomes research. Annu Rev Public Health. 2020;41(1):21–36.
13. Kino S, Hsu YT, Shiba K, et al. A scoping review on the use of machine learning in research on social determinants of health: trends and research prospects. SSM Popul Health. 2021;15:100836.
14. Obermeyer Z, Emanuel EJ. Predicting the Future - Big data, machine learning, and clinical medicine. N Engl J Med. 2016;375(13):1216–9.
15. Foster C, Fenlon D. Recovery and self-management support following primary cancer treatment. Br J Cancer. 2011;105(1):S21–8.
16. Phillips SM, McAuley E. Physical activity and fatigue in breast cancer survivors: a panel model examining the role of self-efficacy and depression. Cancer Epidemiol Biomarkers Prev. 2013;22(5):773–81.
17. Hack TF, Ruether JD, Pickles T, et al. Behind closed doors II: systematic analysis of prostate cancer patients' primary treatment consultations with radiation

oncologists and predictors of satisfaction with communication. Psychooncology. 2012;21(8):809–17.
18. BorjAlilu S, Kaviani A, Helmi S et al. Exploring the role of self-efficacy for coping with breast cancer: A systematic review. Archives Breast Cancer. 2017:42–57.
19. Rottmann N, Dalton SO, Christensen J, et al. Self-efficacy, adjustment style and well-being in breast cancer patients: a longitudinal study. Qual Life Res. 2010;19(6):827–36.
20. Choi YY, Rha SY, Park JS, et al. Cancer coping self-efficacy, symptoms and their relationship with quality of life among cancer survivors. Eur J Oncol Nurs. 2023;66:102373.
21. Sood T, Bhatia R, Khandnor P. Cancer detection based on medical image analysis with the help of machine learning and deep learning techniques: A systematic literature review. Curr Med Imaging. 2023;19(13):1487–522.
22. Jalloul R, Chethan HK, Alkhatib R. A review of machine learning techniques for the classification and detection of breast cancer from medical images. Diagnostics (Basel). 2023;13(14):2460.
23. Lu D, Long X, Fu W, et al. Predictive value of machine learning for breast cancer recurrence: a systematic review and meta-analysis. J Cancer Res Clin Oncol. 2023;149(12):10659–74.
24. Stark GF, Hart GR, Nartowt BJ, et al. Predicting Breast Cancer Risk Using Personal Health Data Mach Learn Models PLoS One. 2019;14(12):e0226765.
25. Pezzolato M, Spada GE, Fragale E, et al. Predictive models of psychological distress, quality of life, and adherence to medication in breast cancer patients: A scoping review. Patient Prefer Adherence. 2023;17:3461–73.
26. Awick EA, Phillips SM, Lloyd GR, et al. Physical activity, self-efficacy and self-esteem in breast cancer survivors: a panel model. Psychooncology. 2017;26(10):1625–31.
27. Champion VL, Ziner KW, Monahan PO, et al. Development and psychometric testing of a breast cancer survivor self-efficacy scale. Oncology nursing forum. NIH Public Access; 2013.
28. Uslu Y, Er S, Subaşi SD, et al. Turkish validity and reliability study of breast cancer self-efficacy survivor scale. Gümüshane Üniversitesi Saglik Bilimleri Dergisi 2023;12:3.
29. Hastie T. The elements of statistical learning: data mining, inference, and prediction. Springer; 2009.
30. Microsoft. Hyperparameter tuning in azure machine learning. 2025 [cited 2025 11.08.2025]; Available from: https://learn.microsoft.com/en-us/azure/machine-learning/how-to-tune-hyperparameters?view=azureml-api-2
31. Özgür S, Taner S, Bozcuk GG, et al. Exploring the predictive role of inflammatory markers in neuropathic bladder-related kidney damage with machine learning. J Pediatr Res. 2024;11(1).
32. Ozgur S, Altinok YA, Bozkurt D, et al. Performance evaluation of machine learning algorithms for sarcopenia diagnosis in older adults. In Healthcare. MDPI; 2023.
33. Microsoft. Primary metric selection in azure machine learning. 2025 [cited 2025 11.08.2025]; Available from: https://learn.microsoft.com/en-us/azure/machine-learning/how-to-tune-hyperparameters?view=azureml-api-2#primary-metric
34. Ozgur S, Kocaslan Toran M, Toygar I, et al. A machine learning approach to determine the risk factors for fall in multiple sclerosis. BMC Med Inf Decis Mak. 2024;24(1):215.
35. Özgür S, Orman M. Application of deep learning technique in next generation sequence experiments. J Big Data. 2023;10(1):160.
36. Liu YB, Liu L, Li YF, et al. Relationship between health literacy, health-related behaviors and health status: A survey of elderly Chinese. Int J Environ Res Public Health. 2015;12(8):9714–25.
37. Seo J, Goodman MS, Politi M, et al. Effect of health literacy on Decision-Making preferences among medically underserved patients. Med Decis Making. 2016;36(4):550–6.
38. Suka M, Odajima T, Okamoto M, et al. Relationship between health literacy, health information access, health behavior, and health status in Japanese people. Patient Educ Couns. 2015;98(5):660–8.
39. Fischer CS, Beresford L. Changes in support networks in late middle age: the extension of gender and educational differences. J Gerontol B Psychol Sci Soc Sci. 2015;70(1):123–31.
40. Yuan C, Wei C, Wang J, et al. Self-efficacy difference among patients with cancer with different socioeconomic status: application of latent class analysis and standardization and decomposition analysis. Cancer Epidemiol. 2014;38(3):298–306.
41. Manne SL, Ostroff JS, Norton TR, et al. Cancer-specific self-efficacy and psychosocial and functional adaptation to early stage breast cancer. Ann Behav Med. 2006;31(2):145–54.
42. Mystakidou K, Parpa E, Tsilika E, et al. Self-Efficacy and its relationship to posttraumatic stress symptoms and posttraumatic growth in cancer patients. J Loss Trauma. 2015;20(2):160–70.
43. Perkins HY, Baum GP, Taylor CL, et al. Effects of treatment factors, comorbidities and health-related quality of life on self-efficacy for physical activity in cancer survivors. Psychooncology. 2009;18(4):405–11.
44. Rogers LQ, McAuley E, Courneya KS, et al. Correlates of physical activity Self-efficacy among breast cancer survivors. Am J Health Behav. 2008;32(6):594–603.
45. Cheng AS, Lee S, Li N, et al. Chinese translation and cross-cultural adaptation of the return-to-work self-efficacy scale among Chinese female breast cancer survivors. Int J Environ Res Public Health. 2023;20(5):4225.
46. Enien MA, Ibrahim N, Makar W, et al. Health-related quality of life: impact of surgery and treatment modality in breast cancer. J Cancer Res Ther. 2018;14(5):957–63.
47. Ghayth EI, Mohammed Shoukr EM, Fathy DM, et al. Long-Term conditions, Multimorbidity burden, and chronic disease Self-Efficacy among geriatric patients: A correlational study. Zagazig Nurs J. 2023;19(2):73–92.
48. Schulman-Green D, Jaser S, Martin F, et al. Processes of self-management in chronic illness. J Nurs Scholarsh. 2012;44(2):136–44.
49. Bakan G, Inci FH. Predictor of self-efficacy in individuals with chronic disease: Stress-coping strategies. J Clin Nurs. 2021;30(5–6):874–81.

Publisher's note