Original Article

J Adv Med Biomed Res. 2024; 32(154): 350-360

Identification of Effective Factors in Breast Cancer Survival in Isfahan Using Machine Learning Techniques

Hossein Bagherian [1], Shaghayegh Haghjooy Javanmard [2], Azam Mosayebi [2], Pegah Noorshargh [2], Saeedeh Arabzedeh [2], Mehran Sharifi [3], Mohammad Sattari *[1]

1. Health Information Technology Research Center, Isfahan University of Medical Sciences, Isfahan, Iran
2. Applied Physiology Research Center, Cardiovascular Research Institute, Isfahan University of Medical Sciences, Isfahan, Iran
3. Cancer Prevention Research Center, Isfahan University of Medical Sciences, Isfahan, Iran

Article Info

10.30699/jambs.32.154.350

Received: 2024/01/02;
Accepted: 2024/12/09;
Published Online: 31 Dec 2024;

Corresponding Information:
Mohammad Sattari,
Health Information Technology Research Center,Isfahan University of Medical Sciences,Isfahan,Iran
E-Mail
msatatrimng.mui@gmail.com

ABSTRACT

Background & Objective: Breast cancer is a leading cause of female mortalities worldwide. This study has used machine learning techniques to determine the most critical factors influencing the survival rate of breast cancer patients in Isfahan.

Materials & Methods: A list of variables influencing the survival of breast cancer patients was initially extracted from the data sets of two Isfahan hospitals for this analytical investigation, leading to the extraction of 16 critical factors based on the opinions of oncologists. In the next step, the missing values were identified and deleted or corrected, followed by converting some features into numerical ranges. Ultimately, the key variables influencing the survival rate of breast cancer patients were determined by applying 11 machine learning algorithms.

Results: Forward selection is more accurate than other techniques. Of the 15 input features, 13 were extracted as influential survival rates at least once using different techniques, with BC-ER-PR-HER2 ranking first among the features. The six first features, including Bc-ER-PR-HER2, lymph node dissection, behavior, primary surgery procedure, the exact number of nodes examined, and the exact number of positive nodes, were determined as the best combination for identifying breast cancer patients. Even though cancer behavior patterns differ in various societies, there are still similarities in risk factors.

Conclusion: Forward selection combined with principal component analysis using support vector machines, neural networks, and random forests can be the best model for breast cancer prediction. Neural networks, random forests, and support vector machines are very good at predicting breast cancer survival.

Keywords: Breast cancer, Survival analysis, Data mining

Introduction

Breast cancer is a leading cause of female mortalities worldwide **(1).** Recent research has reported the increasing annual prevalence of this condition **(2)**, expecting 2.1 million newly diagnosed cases by 2030. However, tremendous advancements have been made in early identification and prompt treatment in the last 20 years, subsequently decreasing mortality from the disease. Meanwhile, all of the features of other malignancies can also be present in breast cancer, which often affects the mammary glands or occasionally originates from the breast's supporting tissues. Since breast cancer is a complex illness, its precise cause is yet unknown. However, environmental and genetic factors can be broadly categorized as factors involved in breast cancer development **(3)**. The most important factors linked to a lower disease survival rate are the type of tumor pathology, the tumor's grade or severity, negative estrogen and

progesterone receptors, socioeconomic status (low education and economic status), fertility status, and high body mass index. The most significant factors influencing death, according to Hapochka et al. **(4)**, are the kind of tumor pathology, higher tumor intensity, negative receptors, different treatments, and the severity of the illness stage. The incidence and mortality of breast cancer can be reduced in large part by identifying the risk factors associated with the disease and evaluating the effect of these factors on the survival rate of patients **(5)**. Survival analysis is a fundamental approach to identifying the features related to disease survival rates.

The waiting period until an event happens is measured using analytical techniques to calculate survival rates **(6)**. According to studies, the survival rate of breast cancer patients is taken into account while assessing and endorsing various treatment modalities **(7)**. Patients with breast cancer have varying odds of surviving depending on their unique clinical features, generally experiencing a higher chance of survival than those with other cancers, particularly in the case of an early diagnosis **(8)**. In the meantime, higher survival rates highlight new concerns, such as establishing suitable screening programs, fighting against the illness, early recurrence identification, and long-term care enhancement. Machine learning methods have recently evolved into useful research instruments for medical researchers to employ several methods for tumor survival or recurrence prediction **(9)**. Ganggayah et al. employed six machine learning algorithms to identify the variables influencing breast cancer patient's chances of survival and determine which of the four selected variables (malignancy type, tumor size, removal of all axillary lymph nodes, and positive lymph nodes) had the greatest impact on patient fatalities. Since breast cancer is a complex illness, prompt actions can be made to greatly lower patient mortality if the influential factors are accurately identified. There is no method for cancer survival analysis that performs appropriately for a given dataset. Thus, this study sought to use machine learning techniques to determine the most critical factors influencing the survival of breast cancer patients.

Materials and Methods

The research methodology included three main steps: data set preparation, preprocessing, and applying machine learning techniques.

2.1 Data Set

Data from two hospitals, Alzahra and Seyedolshohada, in Isfahan were used in this study. The Isfahan-based Applied Physiology Research Center is the owner of the data set. The patient data in the dataset were organized as a matrix, with patient features in the columns and patient records in the rows. Due to the size of the matrix and the lack of relevance for all of its aspects, 16 features were selected based on comparable studies and professional judgment.

The Isfahan Breast Cancer Database included various features from which 16 were extracted by consulting experts and reviewing texts. These 16 features include vital status, surgical margin, BC-PR-ER-HER2, lymph node dissection, lymph vascular invasion (LVI), Nstage, behavior, age at the time of diagnosis, diagnosis date, primary surgery procedure, surgery, grade, radiotherapy, hormone therapy, the exact number of nodes examined, and the exact number of positive nodes. Table 1 shows the values of these features. Some features such as Nstages are numerical, and others such as vital status are binominal. Vital status has two values, with 1 corresponding to survival and 0 representing death. The dataset included 7505, of whom 3727 remained after data cleaning. As Table 1 demonstrates, most patients were in the 40–50 age range. Moreover, 98% of patients had surgery, and 49% were diagnosed as grade 2. In addition, 87% of patients removed their lymph nodes, and 54 had lymph node invasion.

2. 2 Pre-Processing

Data cleaning was performed in three steps: missing values, attribute conversion, and target selection. The values extracted from the data set contained missing values**.** The records with many missing values were deleted due to the likelihood of producing undesirable results in different stages of data mining. Besides, some values were converted into numerical ranges, as the variation in numerical values can affect the performance of data mining. For example, the values are between 1 and 40in features such as the exact number of nodes examined, necessitating conversion into five consecutive intervals of 1 to 5, 6 to 10, and so on. Also, an attribute such as "date of diagnosis," which could have varying values, was translated into one-year time increments. The "vital status" attribute was considered the target attribute, which could contain either die or survive values.

2.3 Applying machine learning techniques

The RapidMiner program 9.1 was employed to classify the remaining data. The training data set and the test data set were separated before the data classification. A template of survival based on candidate traits was extracted from a training data set, and the test data set was then used to apply this template. The data set was divided into training and testing sets using the 10-fold cross-validation technique, which resulted in ten groups from the 3727 samples. Nine groups, comprising 90% of the primary data set, served as the train data set over the experiment's ten iterations. The remaining 10% of the total data set formed the test dataset. Using various machine learning approaches, all conceivable combinations of five features out of a total of sixteen and fifteen were examined, yielding a total of 120 combinations. Techniques like forward selection, polynomial regression, AdaBoost-De, Bayesian boosting, PCA, random forest, linear regression, decision tree, naïve Bayes, neural network, and support vector machine were employed in this study. The two basic feature selection methods used in

this study were forward selection and principal component analysis for comparison with classification methods. Each technique was implemented separately for each combination to choose the technique with the highest accuracy for each combination.

The Random Forest algorithm makes decisions based on a random selection of many trees. The decision tree has a distinct nature as opposed to the random forest. Through deep scrolling, the movement begins at the root node and ends at the leaf nodes. The other nodes are independent variables, and the leaf node is the target (dependent) variable. A dependent variable is predicted by linear regression using several independent factors. The target variable and overall diagnosis are the dependent factors, while factors such as height, weight, and age form the independent variables.

The conditional probability of independent variables concerning the dependent variable is measured by Naïve Bayes . The input, processing, and output layers make up a neural network. The output and input layers are represented by the dependent (target) and independent variables, respectively. The goal of the support vector machine is to find a linear relationship between the independent and dependent variables with a high degree of confidence.

AdaBoost contains a set of weak classifiers interconnected in a series, where each weak classifier tries to improve the classification of samples wrongly classified by the previous poor classifier. The classifier used with Adaboost is a decision tree, which is why this method is called AdaBoost-De. Bayesian boosting is a boosting algorithm based on Bayesian theory applicable to this study because the target attribute is binary.

The above techniques are among the most important and widely used classification methods, each implemented in the training data set. Then the most significant features affecting the survival rate were extracted using the mentioned classification techniques. Finally, features with a value greater than a threshold were chosen as the most influential on breast cancer survival in female patients in Isfahan. The threshold in this study was 33% of the total number of the three techniques. Forward Selection **(10)** and principal component analysis **(11)** were feature selection methods.

Evaluation Metric

The test data set served as the foundation to evaluate the employed methodologies **(12-13)**. First, a confusion matrix was computed assuming a true class of survival and a false class of death. The amount of records accurately identified as survival records was indicated by the term "true-positive." Records accurately categorized as death records were identified as true negatives. False-negative records were those mistakenly classified as death records, whereas false-positive records were those mistakenly classified as survival records. Subsequently, metrics for the effectiveness of the feature selection and classification processes, including accuracy **(14)**, were computed using the confusion matrix. The superior outcome could be determined by how accurate this criterion was. The formula below was used to calculate this metric:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Sensitivity and specificity were the other metrics **(15)**, signifying better results when closer to one. These criteria were calculated as follows:

$$Sensitivity = \frac{TP}{TP + FN}$$
$$Specificty = \frac{TN}{TN + FP}$$

Table 1. Features, values, and value proportion of Isfahan Breast Cancer Dataset (Alzahra/Seyedolshohada)

Features	Values	Proportion
Surgical Margin	Positive	0.07
	Negative	0.93
BC-ER-PRHER2	1 (ER negative, PR negative, HER2 negative (Triple negative)	0.03
	2 (ER negative, PR negative, HER2 positive)	0.02
	3 (ER negative, PR positive, HER2 negative)	0.51
	4 (ER negative, PR positive, HER2 positive)	0.16
	5 (ER positive, PR negative, HER2negative)	0.16
	6 (ER positive, PR negative, HER2 positive)	0.11
	7 (ER positive, PR positive, HER2 negative)	0.01
Lymph Node Dissection	Yes	0.87
	No	0.13
Lymph Vascular Invasion (LVI)	Yes	0.54
	No	0.46
N Stage	1-4	0.51
	5-8	0.24
	9-12	0.17

Behavior	13-16	0.9
	1 (Uncertain)	0.06
	2 (in situ)	0.2
	3 (Malignant, primary site)	0.97
	6 (Malignant, metastatic site)	0.04
Date at the Diagnose	2015	0.04
	2016	0.22
	2017	0.25
	2018	0.32
	2019	0.16
	2020	0.01
Primary Surgery Procedure **Total MRM** **Radical MRM** **BCS**	2 (Partial mastectomy. NOS)	0.47
	-4 (Lumpectomy or excisional biopsy)	0.26
	7 (Total (simple) mastectomy. NOS)	0.14
	8 (Modified radical mastectomy. NOS)	0.06
	12(Mastectomy. NOS)	0.06
	13(Surgery. NOS)	0.01
Surgery	Done	0.98
	Not – done	0.02
Grade	1 (well-differentiated)	0.10
	2 (moderately differentiated)	0.49
	3 (poorly differentiated)	0.41
Radiotherapy	Done	0.9
	Not-done	0.1
Hormonotherapy	Done	0.64
	Not-done	0.36
Exact Number of Nodes Examined	1-10	0.6
	11-20	0.35
	21-30	0.04
	31-40	0.01
Exact Number of Nodes Positive	1-10	0.88
	11-20	0.1
	21-30	0.01
	31-40	0.01
Age at the Diagnose	<30	0.03
	30-40	0.20
	40-50	0.34
	50-60	0.22
	60-70	0.14
	70-80	0.05
	80-90	0.01
	>90	0.01
Vital Status	Alive	0.98
	Dead	0.02

Results

The comparison results of nine classification techniques are listed in Table 2. The results show that Forward Selection and PCA could detect more accurately than other techniques. The accuracy, sensitivity, and specificity of several approaches are depicted in Figure 1. In terms of accuracy, the support vector machine method performed the best among the classification techniques, followed by the random forest and neural network methods as the second and third-best classification techniques, respectively.

Five characteristics influencing the survival rate were derived independently for each categorization method, as presented in Table 2. In Table 3, if the number of occurrences of attribute Bc-ER-PR-HER2 equals 7, seven techniques have chosen this feature as one of the selected features. Five techniques, including forward selection, PCA (principal component analysis), support vector machine, random forest, and neural net, had the highest performance in terms of accuracy.

Of the 15 input features, 13 were extracted at least once by techniques as an influencing survival attribute. According to Table 5, behavior was the first important feature chosen by 10 out of the 11 and all of the highest accuracy techniques.

The Bc-ER-PR-HER2 and Lymph Node Dissection were considered by 9 and 7 techniques, respectively. Four of the five highest accurate techniques chose Bc-ER-PR-HER2, and the same number chose lymph node dissection. The primary surgery procedure was the fourth feature influencing breast cancer, followed by the exact number of nodes examined and positive nodes as the fifth and sixth features chosen by 2 of the 3 highest accurate techniques and 5 of the 11 techniques, respectively. The seventh feature was age at the diagnosis selected by 4 of the 11 techniques.

The seven selected features included behavior, Bc-ER-PR-HER2, lymph node dissection, primary surgery The patients who had malignant cancers, whose primary procedure was partial mastectomy, and survived three years form 12% of the total population. However, patients with malignant cancers, whose primary procedure was total mastectomy, and survived three years form 8% of the total population.

procedure, the exact number of nodes examined, the exact number of positive nodes, and age of diagnosis, which can be considered the right combination for breast cancer patient identification. Two features, including radiotherapy and lymph vascular invasion (LVI), were not selected as features affecting survival by any of the techniques. Only one technique selected the features of hormone therapy and the date of diagnosis as features influencing survival rate.

The relation between significant factors and survival is shown in Table 4, revealing that lymph node removal and lymph node dissection increase the chance of survival in breast cancer patients. Also, the chances of survival are lower in malignant behavior and metastatic sites where breast cancer is more invasive compared to tumors with other behaviors. The mean relationship of the malignant, primary site, and ER negative_PR positive_HER2 negative with survival rate is 53% and 95%, respectively.

Table 5 shows the relationship between survival (years) and malignancy, lymph nodes, and primary procedure. Patients whose cancers showed malignant behavior, removed their lymph nodes, and survived two years form 11% of the total population, while patients whose cancers showed malignant behavior, did not remove their lymph nodes, and survived two years form 9% of the total population.

The z-value for 1- to 4-year survival was 0 and the p-value was also 0, indicating insignificant results at $p < .05$. The z-value for 5-year survival is 2.85774. The value of p is 0.00424. The result is significant at $p < .05$.

Table 2. The combinations achieved by the classification and feature selection techniques

Technique	Selected Features
Forward Selection	Bc-ER-PR-HER2 Lymph Node Dissection Behavior Exact number of nodes examined Exact number of nodes positive
PCA	Nstage Primary Surgery Procedure Bc-ER-PR-HER2 Surgical Margin Behavior
SVM	Date at the diagnosis Exact Number of Nodes examined Exact Number of Nodes Positive Behavior Lymph Node Dissection
Random Forest	Behavior Lymph Node Dissection Hormonotherapy Bc-ER-PR-HER2 Surgery

Neural Net	Behavior Primary Surgery Procedure Bc-ER-PR-HER2 Grade Lymph Node Dissection
CHAID	Behavior Primary Surgery Procedure Nstage Bc-ER-PR-HER2 Age at the Diagnose
ID3	Exact Number of Nodes Examined Exact Number of Nodes Positive Age at the Diagnose Bc-ER-PR-HER2 Primary Surgery Procedure
Polynominal Regression	Exact Number of Nodes Examined Exact number of Nodes Positive Behavior Lymph Node Dissection Age at the Diagnosis
Bayesian Boosting	Age at the diagnosis Exact Number of Nodes Positive Bc-ER-PR-HER2 Primary Surgery Procedure Exact Number of Nodes Examined
Adaboost-de	Grade Surgery Behavior Bc-ER-PR-HER2 Primary Surgery Procedure
Naive Bayes	Bc-ER-PR-HER2 Lymph Node Dissection Behavior Primary Surgery Procedure Grade

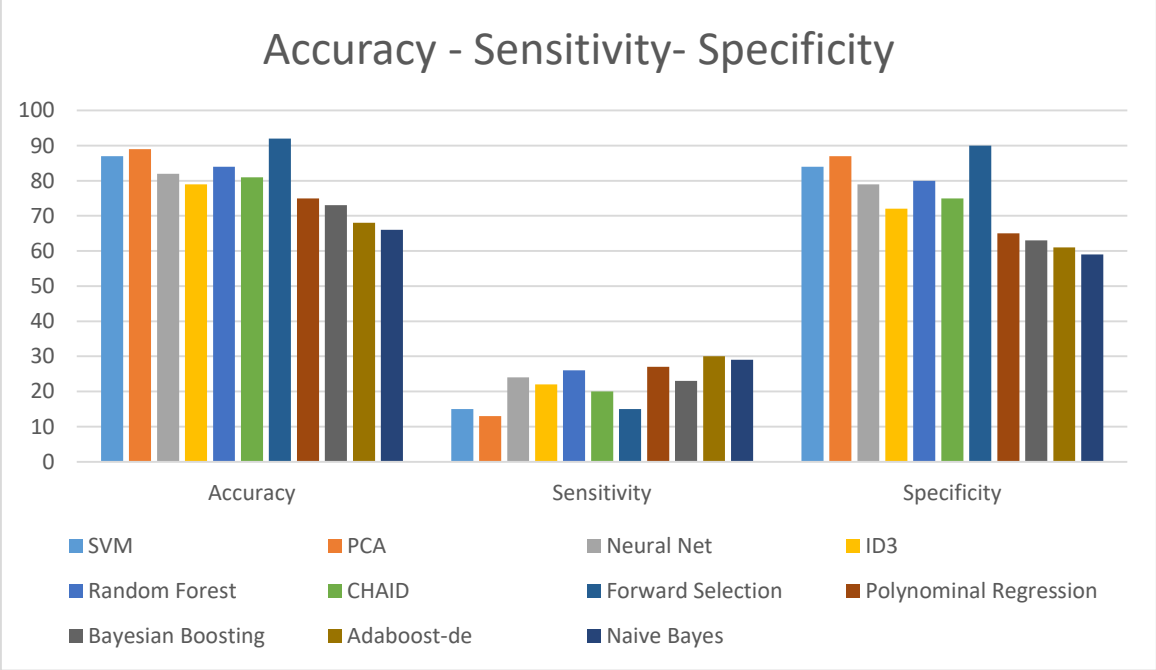

Figure 1. Accuracy, sensitivity, and specificity of different methods

Table 3. **Comparison between features resulting in the performance**

Features	The occurrence in the top Five highest accuracies	Total occurrence
Behavior	5	10
Bc-ER-PR-HER2	4	9
Lymph Node Dissection	4	7
Primary Surgery Procedure	2	6
Exact Number of Nodes Examined	2	5
Exact Number of Nodes Positive	2	5
Age at the Diagnosis	0	4
Grade	1	3
Nstage	2	2
Surgery	1	2
Date at the Diagnosis	1	1
Hormonotherapy	0	1
Surgical Margin	1	1

Table 4. **Patient survival depending on the significant features**

Significant Feature		One year-Survival	Two year-Survival	Three year-Survival	Four year-Survival	Five – year survival	Survival-Mean
Behavior	Uncertain	100%	90%	40%	10%	5%	49%
	In situ	90%	78%	25%	15%	10%	43.6%
	Malignant, primary site	95%	90%	50%	25%	5%	53%
	Malignant, metastatic site	65%	65%	38%	15%	1%	36.8%
Bc-ER-PR-HER2	(ER negative_PR negative_HER2 negative)	97%	95%	93%	90%	87%	92%
	(ER negative_PR negative_HER2 positive)	93%	90%	88%	85%	83%	88%
	(ER negative_PR positive_HER2 negative)	99%	97%	95%	94%	92%	95%
	(ER negative_PR positive_HER2 positive)	99%	98%	95%	93%	91%	95%
	(ER positive_PR negative_HER2negative)	96%	94%	91%	88%	83%	90%
	(ER positive_PR negative_HER2 positive)	95%	93%	90%	87%	85%	90%
	(ER positive_PR positive_HER2 negative)	89%	85%	81%	78%	75%	82%
Lymph Node Dissection	Yes	99%	96%	94%	91%	90%	94%
	No	97%	93%	87%	85%	83%	89%
Primary Surgery Procedure	Partial mastectomy	97%	93%	90%	87%	84%	90%
	Lumpectomy or excisional biopsy	99%	94%	92%	89%	85%	92%
	Total (simple) mastectomy.	97%	93%	89%	86%	82%	89%
	Modified radical mastectomy.	98%	95%	93%	89%	86%	92%
	Mastectomy	97%	95%	91%	87%	83%	91%
	Surgery	99%	96%	93%	91%	89%	94%
Exact Number of Nodes Examined	1 to 10	99%	97%	94%	91%	89%	94%
	11 to 20	98%	96%	95%	93%	91%	95%
	21 to 30	99%	97%	96%	95%	92%	96%
	31 to 40	75%	73%	71%	70%	68%	71%
Exact Number of Nodes Positive	1 to 10	99%	97%	94%	91%	89%	94%
	11 to 20	97%	92%	89%	88%	83%	90%
	21 to 30	80%	78%	75%	74%	70%	75%
	31 to 40	70%	69%	66%	62%	59%	65%

Age at the Diagnosis	<30	99%	99%	98%	97%	97%	98%
	30-40	97%	97%	97%	96%	96%	97%
	40-50	95%	94%	93%	93%	91%	93%
	50-60	95%	92%	89%	88%	85%	90%
	60-70	96%	94%	93%	92%	90%	93%
	70-80	73%	61%	59%	52%	47%	58%
	80-90	50%	49%	45%	40%	38%	44%
	>90	45%	42%	40%	30%	20%	35%

Table 5. Extracted rules (malignant, and lymph node dissection, primary procedure, and survival year)

Rules	Survival (year) – fraction of patients				
	1-year	2-year	3-year	4-year	5-year
Rule 1(If Behavior=Malignant, Primary Site and Lymph Node Dissection = Yes)	18%	13%	11%	7%	5%
Rule 2 (If Behavior=Malignant, Primary Site and Lymph Node Dissection = No)	15%	11%	9%	6%	4%
Rule 3 (If Behavior=Malignant, Primary Site and Primary Procedure = Partial Mastectomy)	17%	14%	12%	9%	7%
Rule 4 (If Behavior=Malignant, Primary Site and Primary Procedure = Total Mastectomy)	14%	12%	8%	5%	4%

Discussion

The survival rate is a significant indicator for policymakers and physicians in providing a proper method for breast cancer diagnosis and treatment by estimating the disease prognosis. Several studies have used the five-year survival rate for breast cancer. According to the GLOBOCAN report, 86% of American women with breast cancer survive the disease for five years **(16)**. Finland and Sweden had the highest survival rates (82 and 83%, respectively), according to Sant et al. **(17)**. Domestic research has also shown that survival rates ranged from 77% to 55% in university facilities **(18-19)** and 89% to 81% in private centers **(20-21)**. Seven key characteristics affecting breast cancer patient's chances of survival, including behavior, Bc-ER-PR-HER2, lymph node dissection, primary surgery approach, number of nodes exactly investigated, number of positive nodes precisely identified, and the age at diagnosis, were extracted from this study. The treatment method should be related to tumor behavior; otherwise, it will adversely affect the survival of patients **(22).**

The degree of a patient's suffering is closely correlated with the number of removed lymph nodes **(17)**, with the greater number of excised nodes leading to more pain the patient feels. These conditions, which can potentially be fatal, alter patients' quality of life. Results indicated that when lymph nodes were removed, the survival rate rose from the first to the fifth year compared to when lymph nodes were not removed, as validated by numerous research **(16,18,22)**. These results highlight the significance of applying targeted topical therapies, including radiation, to patients whose lymph nodes are involved. Conversely, the results highlight the importance of early detection and screening for breast cancer. Suitable health policies may help with correct diagnosis and treatment. In the meantime, one of the most successful methods is to raise awareness to accomplish timely illness control and enhance patient survival. The findings revealed that individuals who underwent surgery had a higher survival rate. However, the five-year survival rate suggests a modest decline in patients undergoing many procedures. Other variables, including disease metastasis and patient age, may affect the five-year survival. Studies show that surgery can lower the patient's death rate.

According to Rapiti, women who received initial breast cancer surgery had a 50% lower death rate than those who did not have surgery **(23)**. These findings are consistent with those found by Khan et al. in their 2002 retrospective research of 1,623 patients, revealing that primary breast cancer surgery lowered the chance of mortality by 39% **(24)**. However, some studies have shown the opposite. For example, research conducted on 129 patients undergoing breast mastectomy showed that different surgical techniques and anesthesia were effective in the recurrence of the disease **(25)**, probably attributed to the disease progression rate and the proper use of surgical techniques.

It seems that partial mastectomy survival rates in tumors with malignant behavior are higher than those in total mastectomy. One study showed that women with partial mastectomy were more stressed than those who had undergone total mastectomy **(26)**. Women with early-stage breast cancer typically have partial mastectomy **(27)**. Additionally, the survival rate of patients with breast cancer is influenced by the use of procedures that shorten the time between diagnosis and surgery **(28).**

Patients under 40 have a low chance of breast cancer, but their survival rate is lower than that of older patients because young individuals typically have more invasive cancer **(29)**. Women who are 45 to 49 years old have higher survival rates than other age groups **(30)**. Breast cancer is rare in people under thirty. Three important factors are involved in survival prediction, including the age and stage of cancer occurrence and the rate of cancer progress **(31)**. Breast cancer and the number of positive nodes are directly correlated **(32)**. Forward selection, PCA, support vector machines, neural networks, and random forests are the five strategies that outperform other approaches. The three primary data mining techniques have been identified in several studies as support vector machines, neural networks, and random forests **(33-34)**. The SVM performs well in various pattern recognition methods **(35)** and contributes significantly to generalizing previously observed test data Consequently, it can affect survival rate, where test data that have not yet been observed become important **(36)**. Clinicians can utilize support vector machines, neural networks, and random forests as a base to assess how well various treatments perform and how they affect patient survival. Physicians can determine the patient's survival rate by entering the values of the most significant risk factors, such as Bc-ER-PR-HER2, into a decision support system. They can select the best course of action by carefully considering the decision support system's output, hence improving patient survival.

Conclusion

The survival rate is an important metric that helps doctors and legislators develop appropriate protocols for the detection and management of breast cancer. Seven key characteristics that affect breast cancer patient's chances of survival (behavior, Bc-ER-PR-HER2, lymph node dissection, primary surgery approach, number of nodes exactly investigated, number of positive nodes precisely identified, and age of diagnosis) are extracted from this study. The survival rate of patients across various nations can be estimated using these seven features. Five methods, including random forest, support vector machine, neural network, principal component analysis, and forward selection, outperformed the others. The optimum model for breast cancer prediction includes forward selection, principal component analysis with support vector machine, neural network, and random forest.

Acknowledgments

None.

Authors' Contribution

MS did data analysis.MS and HB wrote the manuscript.All authors reviewed the manuscript.

Conflict of Interest

The authors declare that they have no conflict of interest.

Funding

It is supported by Isfahan University of Medical Sciences.

Ethics Approval and consent to participate

This article contains no experiments involving human participants or animals conducted by any of the writers. Article extracted from a research project "Predicting the probability of survival in breast cancer patients using data mining techniques" with code IR.MUI.RESEARCH.REC.1398.714

References

1. Prabha S, Sujatha C. Proposal of index to estimate breast similarities in the mammograms using fuzzy C means and anisotropic diffusion filter based fuzzy C means clustering. Infrared Physics Technol. 2018;93:316-25. https://doi.org/10.1016/j.infrared.2018.08.018

2. Collaboration GBoDC. 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: a systematic analysis for the Global Burden of Disease Study Global Burden of Cancer, 1990 to 2016Global Burden of Cancer, 1990 to 2016. JAMA Oncol. 2018;4:1553-68. https://doi.org/10.1001/jamaoncol.2018.2706 PMid:29860482 PMCid:PMC6248091

3. Fazeli Z, Najafian Zade M, Eshtati B, Almasi Hashiani A. Five-year evaluation of epidemiological, geographical distribution and survival analysis of breast cancer in Markazi province, 2007-2011. J Arak Unive Med Sci.

4. Hapochka DO , Zaletok SP, Gnidyuk MI. Relationship between NF-κB, ER, PR, Her2/neu,

Ki67, p53 expression in human breast cancer. Experiment Oncol, 2012.

5. Kasaeian A, Abadi A, Mehrabi Y, Mousavi Jarrahi A. Estimating relative survival of breast cancer patients referring to Imam Khomeini cancer institute during. 2015; 16(14): 5853-8 https://doi.org/10.7314/APJCP.2015.16.14.5853 PMid:26320462

6. Sabouri, S. Determining related factors to survival of colorectal cancer patients using cox regression. Mashhad Unive Med Sci J.2018; 1082-1092.

7. Haghighat S. Survival rate and its correlated factors in breast cancer patients referred to Breast Cancer Research Center. Iran Quarter J Breast Dis. 2013; 6 (3) :28-36

8. Momenyan S, Baghestani AR, Momenyan N, Naseri P, Akbari ME. Survival prediction of patients with breast cancer: comparisons of decision tree and logistic regression analysis. Int J Cancer Manag. 2018. https://doi.org/10.5812/ijcm.9176

9. Moeinzadeh F, Rouhani MH, Mortazavi M, Sattari M. Prediction of chronic kidney disease in Isfahan with extracting association rules using data mining techniques. Tehran Univ Med J. 2021; 79(6):459-67.

10. Blanchet FG, Legendre P, Borcard D. Forward selection of explanatory variables. Ecology. 2008; 89(9):2623-32. https://doi.org/10.1890/07-0986.1 PMid:18831183

11. Wold S, Esbensen K, Geladi P. Principal component analysis. Chemometrics and intelligent laboratory systems. 1987; 2(1-3):37-52. https://doi.org/10.1016/0169-7439(87)80084-9

12. Hosseini A, Eshraghi MA, Taami T, Sadeghsalehi H, Hoseinzadeh Z, Ghaderzadeh M, Rafiee M. A mobile application based on efficient lightweight CNN model for classification of B-ALL cancer from non-cancerous cells: a design and implementation study. Informat Med Unlock. 2023; 39:101244. https://doi.org/10.1016/j.imu.2023.101244

13. Ghaderzadeh M, Aria M, Hosseini A, Asadi F, Bashash D, Abolghasemi H. A fast and efficient CNN model for B-ALL diagnosis and its subtypes classification using peripheral blood smear images. Int J Intell Sys. 2022; 37(8):5113-33. https://doi.org/10.1002/int.22753

14. BS ISO 5725-1: "Accuracy (trueness and precision) of measurement methods and results - Part 1: General principles and definitions.", p.1 (1994)

15. Altman DG, Bland JM. Diagnostic tests. 1: Sensitivity and specificity. BMJ. 1994.308 (6943): 1552. https://doi.org/10.1136/bmj.308.6943.1552 PMid:8019315 PMCid:PMC2540489

16. Li C. Breast Cancer Epidemiology. New York: Springer 2010. https://doi.org/10.1007/978-1-4419-0685-4

17. Sant M, Francisci S, Capocaccia R, Verdecchia A, Allemani C, Berrino F. Time trends of breast cancer survival in Europe in relation to incidence and mortality. Int J Cancer 2006; 119(10): 2417-22. https://doi.org/10.1002/ijc.22160 PMid:16964611

18. Rezaianzadeh A, Peacock J, Reidpath D, Talei A, Hosseini SV, Mehrabani D. Survival analysis of 1148 women diagnosed with breast cancer in Southern Iran. BMC Cancer 2009; 9168. https://doi.org/10.1186/1471-2407-9-168 PMid:19497131 PMCid:PMC2699348

19. Vahdaninia M, Montazeri A. Breast cancer in Iran: a survival analysis. Asian Pac J Cancer Prev 2004; 5(2): 223-5.

20. Mousavi SM, Gouya MM, Ramazani R, Davanlou M, Hajsadeghi N, Seddighi Z. Cancer incidence and mortality in Iran. Ann Oncol 2009; 20(3): 556-63. https://doi.org/10.1093/annonc/mdn642 PMid:19073863

21. Khan U, Shin H, Choi JP, Kim M. wFDT: weighted fuzzy decision trees for prognosis of breast cancer survivability. InProceedings of the 7th Australasian Data Mining Conference. 2008;87:141-152.

22. Seedhom AE, Kamal NN. Factors affecting survival of women diagnosed with breast cancer in El-Minia Governorate, Egypt. Int J Prev Med 2011; 2(3): 131-8. 21.

23. Rapiti E, Verkooijen HM, Vlastos G, Fioretta G, Neyroud- Caspar I, Sappino AP, Chappuis PO, Bouchardy C, Complete excision of primary breast tumor improves survival of patients with metastatic

breast cancer at diagnosis. J Clin Oncol 2006; 24: 2743-9.
https://doi.org/10.1200/JCO.2005.04.2226
PMid:16702580

24. Khan SA, Stewart AK, Morrow M. Does aggressive local therapy improve survival in metastatic breast cancer? Surgery 2002; 132: 27.
https://doi.org/10.1067/msy.2002.127544
PMid:12407345

25. Exadaktylos AK, Buggy DJ, Moriarty DC, Mascha E, Sessler DI. Can anesthetic technique for primary breast cancer surgery affect recurrence or metastasis? Anesthesiol. 2006; 105: 660-4.
https://doi.org/10.1097/00000542-200610000-00008
PMid:17006061 PMCid:PMC1615712

26. Brisson J, Deschenes L. Psychological distress after initial treatment for breast cancer: a comparison of partial and total mastectomy. J Clin Epidemiol. 1989 ;42(8):765-7
https://doi.org/10.1016/0895-4356(89)90074-7
PMid:2760668

27. Covelli AM. Choosing Mastectomy: A qualitative exploration of the increasing mastectomy rates in women with early-stage breast cancer [Doctoral dissertation].

28. Canadian Partnership Against Cancer. Pan-Canadian standards for breast cancer surgery. https://s22457.pcdn.co/wp-content/uploads/2019/05/BreastCancer-Surgery-Standards-EN-April-2019.pdf; 2019. Accessed October 1, 2019.

29. Anders CK, Johnson R, Litton J, Phillips M, Bleyer A. Breast cancer before age 40 years. In Seminars in oncology. 2009 ; 36, (3): 237-249. WB Saunders.
https://doi.org/10.1053/j.seminoncol.2009.03.001
PMid:19460581 PMCid:PMC2894028

30. Chen HL, Zhou MQ, Tian W, Meng KX, He HF. Effect of age on breast cancer patient prognoses: a population-based study using the SEER 18 database. PloS one. 2016 ; 11(10):e0165409.
https://doi.org/10.1371/journal.pone.0165409
PMid:27798652 PMCid:PMC5087840

31. Singletary SE. Rating the risk factors for breast cancer. Annals of surgery. 2003 Apr; 237(4):474.
https://doi.org/10.1097/01.SLA.0000059969.64262.87
PMid:12677142 PMCid:PMC1514477

32. Katz A, Smith BL, Golshan M, Niemierko A, Kobayashi W, Raad RA, et al. Nomogram for the prediction of having four or more involved nodes for sentinel lymph node-positive breast cancer. J Clin Oncol,2008; 26(13): 2093-2098.
https://doi.org/10.1200/JCO.2007.11.9479
PMid:18445838

33. Nijhawan R, Raman B, Das J. Meta-classifier approach with ANN, SVM, rotation forest, and random forest for snow cover mapping. InProceedings of 2nd International Conference on Computer Vision & Image Processing 2018; (279-287). Springer, Singapore.
https://doi.org/10.1007/978-981-10-7898-9_23

34. Lazri M, Ameur S. Combination of support vector machine, artificial neural network and random forest for improving the classification of convective and stratiform rain using spectral features of SEVIRI data. Atmospher Res. 2018; 203: 118-29.
https://doi.org/10.1016/j.atmosres.2017.12.006

35. Byun H, Lee SW. A survey on pattern recognition applications of support vector machines. Int J Pattern Recog Artif Intell. 2003; 17(3): 459-486.
https://doi.org/10.1142/S0218001403002460

36. Huang MW, Chen CW, Lin WC, Ke SW, Tsai CF. SVM and SVM ensembles in breast cancer prediction. PloS one. 2017 ; 12(1):e0161501.
https://doi.org/10.1371/journal.pone.0161501
PMid:28060807 PMCid:PMC5217832

How to Cite This Article:

Hossein Bagherian, Shaghayegh Haghjooy Javanmard, Azam Mosayebi, Pegah Noorshargh, Saeedeh Arabzedeh, Mehran Sharifi, Mohammad Sattari .Identification of Effective Factors in Breast Cancer Survival in Isfahan Using Machine Learning Techniques. J Adv Med Biomed Res. 2024; 32(154): 350-360.

Download citation:
BibTeX | RIS | EndNote | Medlars | ProCite | Reference Manager | RefWorks

Send citation to:
Mendeley Zotero RefWorks