Epidemiology and Health System Journal

doi:10.34172/ehsj.26085 2024 Winter;11(1):x-x http://ehsj.skums.ac.ir

Original Article

# Survivability Prediction of Breast Cancer Patients Using Three Data Mining Methods: A Comparative Study

**Maryam Jalali**[1], **Navid Reza Ghasemi**[2], **Samane Nematolahi**[3*], **Najaf Zare**[4]

[1]Colorectal Research Center, Shiraz University of Medical Sciences, Shiraz, Iran
[2]Project Managers at Gas Company, Bam, Kerman, Iran
[3]Noncommunicable Diseases Research Center, Bam University of Medical Sciences, Bam, Kerman, Iran
[4]Infertility Research Center, Shiraz University of Medical Sciences, Shiraz, Iran

***Corresponding Author:**
Samane Nematolahi,
Email: samane.nematolahi@yahoo.com

**Received:** November 14, 2023
**Accepted:** January 24, 2024
**ePublished:** x xx, 2024

**Abstract**

**Background and aims:** Breast cancer (BC) is the leading cause of mortality among women. Early diagnosis is crucial for effective treatment. This study applied suitable data mining methods that provide rules and present influential prognostic factors on the survival time of BC patients.

**Methods:** The dataset consisted of 1574 women diagnosed between January 2002 and December 2012 at the Cancer Registry Center of Nemazi hospital in Fars Province, Iran. Patients were classified based on prognostic factors using three popular data mining methods, including decision tree (J48), Naïve Bayes (NB), and nominal logistic regression (NLR). The Weka software was considered to compare these methods using sensitivity, specificity, and accuracy metrics. The outcome of the study was the median survival time, which was categorized into three classes.

**Results:** In total, 212 women (13.5%) died of BC, whose mean age was 49.74 years old. Overall survival rates at 2, 3, 5, and 10 years were 0.98, 0.94, 0.87, and 0.76, respectively. The mean and median survival times were 4.81 and 4.27 years. Sensitivity, specificity, and accuracy for J48 and NB were 0.480, 0.570, and 0.572, as well as 0.483, 0.610, and 0.584, respectively. In addition, the corresponding values were 0.488, 0.584, and 0.579 for NLR, respectively. Further, J48 showed that the Nottingham Prognostic Index (NPI) was the most influential prognostic factor.

**Conclusion:** This paper sought to improve the accuracy of BC classification using data mining methods. Comparing multiple prediction models gave us an insight into the relative prediction abilities of different data mining methods. The results suggested NB as the best classifier due to its higher accuracy and specificity. Finally, J48 identified the NPI as the most effective prognostic factor.

**Keywords:** Breast neoplasm, Data mining techniques, Prognostic factors

## Introduction

Breast cancer (BC) ranks as the most prevalent cancer among women worldwide, standing as the fifth leading cause of death. In 2020, an estimated 684 996 women lost their lives to BC, with 2 261 419 newly diagnosed cases.[1,2] Moreover, there has been a decreasing trend in the number of women succumbing to BC since 2007, resulting in an annual 1% drop in the death rate from 2013 to 2018, particularly among women under the age of 50.[1]

Several factors influence the risk of developing BC, including determinants of endogenous hormones, such as the early or late onset of menarche and menopause, the late age of first pregnancy, and genetic predisposition. The other influential factors are high intake of the exogenous hormone, physical activity, a healthy diet, smoking, anthropometric characteristics, and family history.[3,4]

Each female's recovery chance is associated with various variables, including tumor size, the involved lymph node numbers, and other tumor characteristics, implying that estimating the survival chance for the patients might be difficult.[5]

Although this cancer ranks as the second main cause of cancer death in women, its survival rate is notably high.[6] Therefore, determining accurate prognostic factors affecting a patient's survival is crucial.

Survival analysis, a statistical technique used when the response variable is the time to an event, plays a pivotal role in estimating recovery chances for patients. The Cox proportional hazard model is widely used[7]; this regression model assumes linear associations between predictors and survival outcomes. Data mining methods offer an alternative by considering all potential interactions and effect modifications between variables.[8,9]

Recently, data mining methods have been extensively applied in BC diagnosis and treatment.[10] They play a crucial role in reducing the frequency of false positives and false-negative results, aiding physicians in decision-making, and assisting researchers in identifying disease patterns and predicting outcomes when dealing with numerous variables.[11]

Recent studies analyzing medical survival data have shown that the decision tree (DT; c5) exhibits the highest accuracy, surpassing 90% when compared to artificial neural networks (ANNs) and logistic regression methods.[12] An overview of data mining techniques for BC predictions revealed that the C4.5 algorithm is the most accurate, achieving over 80% accuracy.[12] Another study focusing on machine learning algorithms for predicting BC survival rates concluded that the J48 DT model was more sensitive, logistic regression was more accurate, and ANNs had the highest specificity.[12] Recent research has demonstrated that the Naïve Bayes (NB) technique surpasses other classifiers in terms of accuracy.[13]

Mohammed et al found that the J48 DT was more accurate, efficient, and effective in predicting BC risks based on evaluation criteria.[14] In a comparative study on data mining techniques, it was discovered that DTs and ANN methods could classify data with high accuracy.[15]

Our study has aimed to apply suitable data mining methods to survival data, providing rules and presenting influential prognostic factors on the survival time of BC patients. Notably, our study differs from previous ones as the target variable has three classes, whereas the majority of recent studies have focused on classifying patients into two categories.

## Materials and Methods

A BC cohort study was conducted at the Nemazi hospital Cancer Registry Centre from January 2002 to December 2012. The inclusion criteria involved a BC diagnosis during the study period with no other type of cancer involvement. Ultimately, 1574 patients were included in the study. Patient medical information encompassed nipple involvement (NI), skin involvement (SI), lymphatic and vascular invasion (LVI; LV involvement), progesterone receptor, estrogenic receptor, age, node total, nuclear grade, disease stage, tumor size, marital status, and education. The survival time for each case was computed as the difference between the time that each case entered the study and the time of death for cases that were followed until death, or from the baseline to the closing date of follow-up for living patients.[7] In this paper, to predict tumor nature for improved treatment in BC patients and determine influential prognostic factors on BC survivability rates, we focused on three classification algorithms in Weka data mining tools, including J48 decision tree, NB, and nominal logistic regression (NLR), ultimately seeking the most accurate classifier techniques. The target variable was categorized into three classes based on the median survival time for data mining purposes. Patients who survived more than the median survival time and remained alive until the end of the study were classified as the "Above-median" class. The "Below-median" class covered patients who survived less than 4 years with BC, while the "Undetermined category" consisted of BC patients who were alive and had less than 4 years in the study[16] (Algorithm 1).

### *Data Mining Methods*

The NB technique is based on the famous Bayes theorem, assuming strong (naïve) independence between features. In other words, knowing the value of one attribute reveals nothing about the value of another attribute.

The maximum likelihood function in the NB method has a closed-form expression, making it less expensive than other types of classifiers that use iterative approximation. NB, while not a Bayesian method, creates statistical predictive models based on Bayes' theorem.[17]

One notable advantage of NB is its requirement for a small training sample for parameter estimation.

The DT (J48) is one of the most popular learning models for the powerful classification of observations. The tree models are called classification trees and regression trees when the outcome variable is categorical and continuous, respectively. The most important features of J48 include DT pruning, handling missing values, continuous ranges of attributes, rule derivation, and the like.

In this method, mathematical algorithms such as the Gini index, the chi-square test, and the like are used to allocate the input observations into subgroups. This process is repeated until the tree is completed.[12,18]

NLR, developed by Joseph Berkson, is a generalization of linear regression. In this model, the log odds for the value of the outcome are a linear combination of predictor variables.

NLR is utilized when the dependent variable has more than two categories. The benefits of the model include a strong statistical foundation, a probabilistic model for completely explaining the observations, high efficiency, interpretability, and the lack of requiring too many computational resources.[19]

### *Evaluation Criteria (Performance Parameters)*

The criteria for comparing the results of various data mining tools include sensitivity, specificity, and accuracy. The related definitions and formulas are provided as follows:

**Algorithm 1**

Setting the survivability dependent variable for 4 years' threshold (median of survival time).

```
if $Time \le 4$ years and alive then
    the record is pre-classified as "undetermined"
else if $Time \le 4$ years and dead then
    the record is pre-classified as "below median"
else if $Time > 4$ years and dead then
    the record is pre-classified as "above median"
else
    ignore the record
end if
```

Sensitivity (Recall/true positive rate) presents the ability of a classifier to identify the actual positive results.

$Sensitivity = TP / (TP + FN)$

Specificity (True negative rate) is clear by name; it is the proportion of actual negatives accurately recognized by the classifier.

$Specificity = TN / (TN + FP)$

Accuracy refers to the ability of the classifiers to correctly predict the class label.

$Accuracy = (TP + TN) / (TP + TN + FP + FN)$

where TP, FP, TN, and FN denote true positives, false positives, true negatives, and false negatives, respectively.[19]

In this study, a 10-fold cross-validation procedure was employed to calculate the unbiased prediction accuracy of the applied prediction models. In this technique, the primary sample is partitioned into 10 subsamples of equal size randomly. One of these subsamples is used for testing, and the remaining are utilized for training. This process is repeated 10 times so that every subsample is considered the validation dataset. Finally, a single measure of accuracy for the model is computed by averaging the results from the folds. All classification techniques were run in the WEKA software.

### *Prognostic Factors*

This study investigated the effects of nipple and SI, LVI, progesterone and estrogen receptor (ER) status, age, total nodes, nuclear grade, disease stage, tumor size, education, marital status, and Nottingham prognostic index (NPI) on the status of a patient's survival. The patient's age was initially recorded as a continuous variable, but it was discretized into two classes ($<35$ and $>35$ years old). Marital status was classified as single or married. Education level groups were illiterate, primary school, high school, and university. The Tumour, Node, Metastasis (TNM) staging system categorized patients into different stages at diagnosis. Nodal status classes were 0, 1–3, and $>3$ nodes involved. Tumor size was grouped into 3 categories ($<3$, 3–5, and $>5$ cm). The nuclear grade was classified into 3 levels (i.e., well-differentiated, poorly differentiated, and undifferentiated levels). Estrogen and progesterone receptor (PR) groups were positive or negative. Nipple, skin, and LV involvement were categorized as involved or free.

The NPI for each BC patient is the sum of the values of tumor size multiplied by 0.2, nuclear grade (1–3), and nodal status (1–3). The original NPI was grouped into 3 classes (good, moderate, and poor, with cut-off values of $\leq 3.4$, 3.5–5.4, and $>5.4$).[7]

### *Variable Selection*

To determine significant prognostic factors for improving classification accuracy, univariate Cox regression models were applied, and factors with $P<0.20$ were selected for entry into the final model. Prognostic factors included age, tumor size, SI, ER status, PR status, NI, nodal status, nuclear grade, and LVI. According to the NPI formula, it is evident that nuclear grade, tumor size, and nodal status are correlated with NPI. Additionally, there is a strong correlation between disease stages and nodal status. Consequently, NI, SI, ER status, PR status, LVI, age, and NPI were selected for entry into the classification models.[6] The selected variables and their classes are listed in Table 1.

## Results

By December 2012, 212 women (13.5%) had died due to BC. The mean age at diagnosis was 49.74 years old. Overall survival rates at 2, 3, 5, and 10 years were 0.98, 0.94, 0.87, and 0.76, respectively. The mean and median survival times were 4.81 and 4.27 years. Sensitivity, specificity, and accuracy for J48 were 0.480, 0.570, and 0.572. In addition, the corresponding values for NB and NLR were 0.483, 0.610, and 0.584, as well as 0.488, 0.584, and 0.579, respectively.

In this study, the values of TP, FP, TN, and FN were computed for each class in the confusion matrix, separately due to having three classes. Finally, by using a weighted average, the accuracy, sensitivity, and specificity were computed for every data mining technique, considering that the assigned weight for each class is its size (Table 2).[20]

Specificity and sensitivity can be computed by using confusion matrix information (TP, FP, TN, and FN).[21] The results of our approaches are reported in Table 3.

Table 3 presents the evaluation criteria of the three classification algorithms used for the BC data set. The values show that the best classifier was NB due to the highest values of accuracy and specificity (58.4% and 61%). The second place was filled by NLR, and the third was J48. The J48 data mining method gives some additional information by building classification models in the form of tree structures. The most influential predictor is at the top of the DT. We concluded that the NPI was the most influential predictor in the survival status of BC patients. Other influential predictors were LVI, PR, ER, age, NI, and stage. The reported numbers in the DT in Figure 1 demonstrate the prediction process. For example, 3 (24/10) in the box implies that at this path, the prediction was level 3, and the (24/10) means that 24 observations in the dataset ended up at this path and 10 were incorrectly

**Table 1.** Dataset Attributes

| Nominal Variable Name | Number of Classes | Description |
|---|---|---|
| Nipple involvement | 2 | Free or involved |
| Skin involvement | 2 | Free or involved |
| Lymphatic and vascular invasion | 2 | Free or involved |
| Progesterone receptor | 2 | Positive or negative |
| Estrogenic receptor | 2 | Positive or negative |
| Stage | 3 | I, II, or III |
| Age | 2 | $>35$ or $\leq 35$ |
| Nottingham Prognostic Index | 3 | Good, moderate, or poor |
| Class (dependent variable) | 3 | Undetermined, below median, or above median |

**Table 2.** Survivability Class Instances

| Class | Number of Instances | Percentage |
|---|---|---|
| Undetermined | 672 | 42.7 |
| Below median | 166 | 10.5 |
| Above median | 736 | 46.8 |
| Total | 1,574 | 100 |

**Table 3.** Evaluation Criteria in Three Data Mining Methods

| Classification Technique | Accuracy | Specificity | Sensitivity |
|---|---|---|---|
| J48 | 0.572 | 0.570 | 0.480 |
| Naive Bayes | 0.584 | 0.610 | 0.483 |
| Nominal logistic regression | 0.579 | 0.584 | 0.488 |

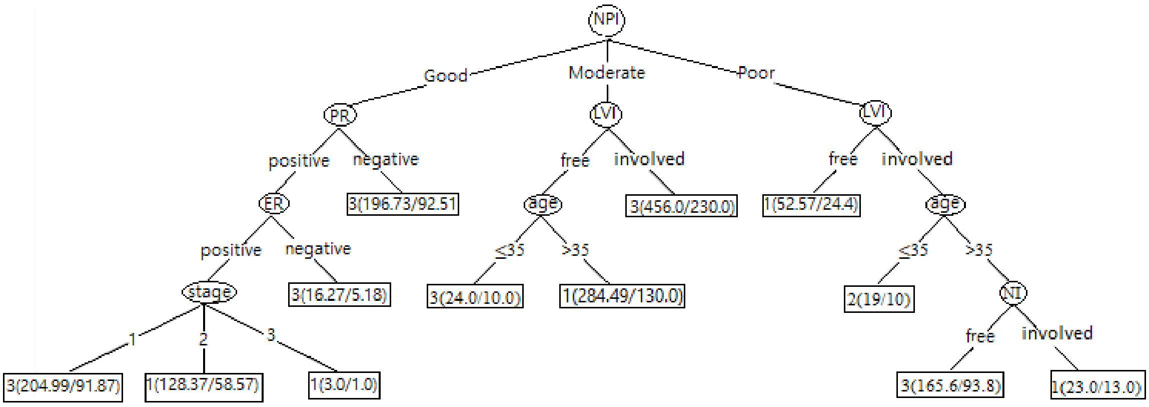

**Figure 1.** Decision Tree of Breast Cancer Patients for Medical Application

classified; in other words, 14 and 10 had the label 3 and label 1 or 2, respectively.

Some leaves in the DT had float numbers; when the instance has missing attributes, then the classifier (J48) does not know the way of the tree for that attribute. Thus, the classifier will divide the instance according to the probability and percentage of the instance.

## Discussion

In this paper, the Nemazi hospital BC dataset was employed to evaluate the accuracy of three popular prediction models, including one from statistics (NLR) and two from machine learning (J48, NB). NB was the best classifier, while J48 was the least effective classification method.

Based on our results, it is confirmed that applying advanced data mining methods leads to high predictive accuracy. However, some issues exist regarding data collection, data mining methods, and predictive ability that should be taken into consideration. Two crucial aspects of predictive accuracy are data size and quality.[22] Medical data typically exhibit heterogeneity, making the application of data mining techniques challenging. Issues such as missing data, redundancy, imprecision, and inconsistency can influence the results of data mining methods. Additionally, data-gathering methods may introduce noise.[23,24]

Although data mining methods have some drawbacks, such as not fulfilling classical statistical conditions,[25] they can potentially be essential tools in medicine, exploring aspects of diseases and providing valuable information for future research.[26,27] The acceptable and good predictive accuracy of our applied models is just one factor emphasizing the importance of data mining methods in the medical field.

Recent studies on the survival of BC patients have employed different analytical methods, including data mining methods and the Cox regression model, the traditional statistical model for analyzing survival data. The most common data mining techniques used in previous studies were J48, NB, and ANN. A review of past studies revealed that in most of them, the J48 algorithm was more accurate than the other data mining methods.[7,8,19,21,28-31] Recent research has shown that NB was the most accurate method,[13] which is consistent with the results of our study.

The secondary objective of our study was to determine prognostic factors. Based on our results, using suitable predictive attributes led to the development of a model for accurately predicting outcomes. In medicine, these models can be applied in prediction, diagnosis, and treatment.[27,32] Important influential attributes were identified based on J48. The first and most important was NPI. It is a traditional and widely used method for predicting the survival of BC[33] and is also used to provide a basis for assessing newly designed methods for prognosis in BC, including microarray techniques[34]. Other important predictors were LVI, PR, ER, age, NI, and stage. Our findings are in line with those of previous studies. In a recent study on survival prediction using DTs and logistic regression analysis, LVI and stage were identified as influential factors.[35] Tanha et al investigated the relationship among prognostic indices of BC using classification techniques in 2020. In this study, PR, ER, and age were identified as significant prognostic factors.[36] NI was one of the significant independent prognostic discriminants in pathologic findings from the National Surgical Adjuvant Breast Project.[37]

Enhancing accuracy and precision in prediction is feasible through various measures. These measures encompass modifying the size of variables, diminishing the number of features, or choosing the most dependable features through applying robust algorithms such as principal component analysis for feature selection. Moreover, modifying the techniques utilized for data preprocessing, adjusting runtime parameters, and employing ensemble methods with varying parameters can potentially enhance precision and accuracy rates.[38]

Considering the large amount of data available in medical databases and the potential significant association between symptoms and diagnosis, applying data mining algorithms to explore these relationships is advantageous. However, data mining is not intended to replace medical professionals but rather to enhance their efforts in saving human lives. Medical researchers and statisticians must examine the availability of their biological data concerning variables associated with cancer survivability prediction. Variables in this study were selected using the literature on computational biology and the available BC dataset, along with the researcher's domain knowledge. Data quality has the potential to determine the outcome of a machine learning method, either leading to success or failure. This crucial step accounts for a significant portion, ranging from 60 to 80%, of the overall data mining or machine learning procedure.[39]

## Limitations of Data Mining Methods

Despite the potential of data mining to offer valuable insights and assistance to medical professionals through pattern identification, there are limitations to what it can do. Not all patterns discovered through data mining can be deemed "noteworthy"; a noteworthy pattern must possess logical reasoning and be actionable. For instance, while data mining can be useful in diagnosing or suggesting treatment, it is not a proper replacement for a physician's intuition and interpretive abilities.

## Conclusion

In this paper, it was attempted to improve the accuracy of BC classification by utilizing data mining methods. Comparing multiple prediction models for BC survivability gave us insight into the relative prediction abilities of different data mining methods. The results suggested NB as the best classifier due to its higher accuracy and specificity. Our experimental results revealed significant relationships between different prognostic indices in the BC dataset. J48 identified the NPI as the most effective prognostic factor. As a future insight, there is still a need for a comprehensive investigation employing data mining methods to determine designs that yield a higher level of precision and accuracy. To make considerable strides forward in the prognosis and medication of BC, continued investigation and assistance between data scientists, medical experts, and researchers is crucial.

### Acknowledgments

The study was approved by the Ethics Committee, and confirmation was taken from BAM University of Medical Sciences (ethical code: IR.MUBAM.REC.1401.046). The authors are grateful to Sedighe Tahmasebi at Nemazi Hospital for their cooperation in receiving data.

### Authors' Contribution

**Conceptualization:** Samane Nematolahi.
**Data curation:** Samane Nematolahi and Navid Reza Ghasemi.
**Formal analysis:** Samane Nematolahi and Navid Reza Ghasemi.
**Investigation:** Samane Nematolahi and Navid Reza Ghasemi.
**Methodology:** Samane Nematolahi
**Project administration:** Samane Nematolahi
**Resources:** Samane Nematolahi.
**Software:** Samane Nematolahi and Navid Reza Ghasemi.
**Supervision:** Najaf Zare.
**Validation:** Samane Nematolahi, Navid Reza Ghasemi, and Maryam Jalali.
**Visualization:** Samane Nematolahi, Navid Reza Ghasemi, and Maryam Jalali.
**Writing–original draft:** Samane Nematolahi and Maryam Jalali.
**Writing–review & editing:** Samane Nematolahi, Maryam Jalali, and Najaf Zare.

### Competing Interests

The authors declare that there is no conflict of interests.

### Ethical Approval

The study was approved by the Ethics Committee, and confirmation was taken from BAM University of Medical Sciences (ethical code: IR.MUBAM.REC.1401.046). Confidentiality of their data was emphasized, and informed written consent was obtained from each of the patients before enrollment. All methods were performed in accordance with the Declaration of Helsinki.

### Funding

Not applicable.

### References

1. Zafar T, Naik AQ, Kumar M, Shrivastava VK. Epidemiology and risk factors of breast cancer. In: Shakil Malik S, Masood N, eds. Breast Cancer: From Bench to Personalized Medicine. Singapore: Springer; 2022. p. 3-29. doi: 10.1007/978-981-19-0197-3_1.
2. Nourelahi M, Zamani A, Talei A, Tahmasebi S. A model to predict breast cancer survivability using logistic regression. Middle East J Cancer. 2019;10(2):132-8. doi: 10.30476/mejc.2019.78569.
3. lkhomovna KD. Current understanding of breast cancer risk factors. International Journal of Culture and Modernity. 2021;6:31-7. doi: 10.51699/ijcm.v6i.48.
4. Pashayan N, Antoniou AC, Ivanus U, Esserman LJ, Easton DF, French D, et al. Personalized early detection and prevention of breast cancer: ENVISION consensus statement. Nat Rev Clin Oncol. 2020;17(11):687-705. doi: 10.1038/s41571-020-0388-9.
5. Nematolahi S, Ayatollahi SM. A comparison of breast cancer survival among young, middle-aged, and elderly patients in southern Iran using Cox and empirical Bayesian additive hazard models. Epidemiol Health. 2017;39:e2017043. doi: 10.4178/epih.e2017043.
6. Giordano SH. Breast cancer in men. N Engl J Med. 2018;378(24):2311-20. doi: 10.1056/NEJMra1707939.
7. Nematolahi S, Rezaianzadeh A, Zare N, Akrami M, Tahmasebi S. Prognostic factors of breast cancer survival in the Islamic Republic of Iran: an additive empirical Bayesian approach. East Mediterr Health J. 2018;23(11):721-8. doi: 10.26719/2017.23.11.721.
8. Du M, Haag DG, Lynch JW, Mittinty MN. Comparison of the tree-based machine learning algorithms to Cox regression in predicting the survival of oral and pharyngeal cancers: analyses based on SEER database. Cancers (Basel). 2020;12(10):2802.

doi: 10.3390/cancers12102802.

9. Shah C, Shaikh M, Shah D, Samdani K. A review on big data practices in healthcare. In: 2019 IEEE International Conference on System, Computation, Automation and Networking (ICSCAN). Pondicherry, India: IEEE; 2019. doi: 10.1109/icscan.2019.8878687.
10. Eltalhi S, Kutrani H. Breast cancer diagnosis and prediction using machine learning and data mining techniques: a review. IOSR J Dent Med Sci. 2019;18(4):85-94. doi: 10.9790/0853-1804208594.
11. Mosayebi A, Mojaradi B, Bonyadi Naeini A, Khodadad Hosseini SH. Modeling and comparing data mining algorithms for prediction of recurrence of breast cancer. PLoS One. 2020;15(10):e0237658. doi: 10.1371/journal.pone.0237658.
12. Maleki Birjandi S, Khasteh SH. A survey on data mining techniques used in medicine. J Diabetes Metab Disord. 2021;20(2):2055-71. doi: 10.1007/s40200-021-00884-2.
13. Chaurasia V, Pal S, Tiwari BB. Prediction of benign and malignant breast cancer using data mining techniques. J Algorithm Comput Technol. 2018;12(2):119-26. doi: 10.1177/1748301818756225.
14. Mohammed SA, Darrab S, Noaman SA, Saake G. Analysis of Breast Cancer Detection Using Different Machine Learning Techniques. In: Tan Y, Shi Y, Tuba M, eds. Data Mining and Big Data. Singapore: Springer; 2020. p. 108-17. doi: 10.1007/978-981-15-7205-0_10.
15. Alhasani AT, Alkattan H, Subhi AA, El-Kenawy ES, Eid MM. A comparative analysis of methods for detecting and diagnosing breast cancer based on data mining. Journal of Artificial Intelligence and Metaheuristics. 2023;4(2):8-17. doi: 10.54216/jaim.040201.
16. Kusiak A, Dixon B, Shah S. Predicting survival time for kidney dialysis patients: a data mining approach. Comput Biol Med. 2005;35(4):311-27. doi: 10.1016/j.compbiomed.2004.02.004.
17. Jaitha A. An Introduction to the Theory and Applications of Bayesian Networks [dissertation]. Claremont McKenna College; 2017.
18. Gupta B, Rawat A, Jain A, Arora A, Dhami N. Analysis of various decision tree algorithms for classification in data mining. Int J Comput Appl. 2017;163(8):15-9.
19. Delen D, Walker G, Kadam A. Predicting breast cancer survivability: a comparison of three data mining methods. Artif Intell Med. 2005;34(2):113-27. doi: 10.1016/j.artmed.2004.07.002.
20. Kaya Keleş M. Breast cancer prediction and detection using data mining classification algorithms: a comparative study. Teh Vjesn. 2019;26(1):149-55. doi: 10.17559/TV-20180417102943.
21. Grandini M, Bagli E, Visani G. Metrics for multi-class classification: an overview. ArXiv [Preprint]. August 13, 2020. Available from: https://arxiv.org/abs/2008.05756.
22. Chiu CY, Verma B. Relationship between data size, accuracy, diversity and clusters in neural network ensembles. Int J Comput Intell Appl. 2013;12(4):1340005. doi: 10.1142/s1469026813400051.
23. Nayak P. A survey on medical data by using data mining techniques. Int J Adv Res Ideas Innov Technol. 2017;3(6):1330-5.
24. Saraswat P, Raj S. Data pre-processing techniques in data mining: a review. Int J Innov Res Comput Sci Technol. 2022;10(1):122-5. doi: 10.55524/ijircst.2022.10.1.22.
25. Wu WT, Li YJ, Feng AZ, Li L, Huang T, Xu AD, et al. Data mining in clinical big data: the frequently used databases, steps, and methodological models. Mil Med Res. 2021;8(1):44. doi: 10.1186/s40779-021-00338-z.
26. Richards G, Rayward-Smith VJ, Sönksen PH, Carey S, Weng C. Data mining for indicators of early mortality in a database of clinical records. Artif Intell Med. 2001;22(3):215-31. doi: 10.1016/s0933-3657(00)00110-x.
27. Boeri C, Chiappa C, Galli F, De Berardinis V, Bardelli L, Carcano G, et al. Machine learning techniques in breast cancer prognosis prediction: a primary evaluation. Cancer Med. 2020;9(9):3234-43. doi: 10.1002/cam4.2811.
28. bin Othman MF, Yau TM. Comparison of different classification techniques using WEKA for breast cancer. In: Ibrahim F, Osman NA, Usman J, Kadri NA, eds. 3rd Kuala Lumpur International Conference on Biomedical Engineering 2006. Berlin, Heidelberg: Springer; 2007.p. 520-3. doi: 10.1007/978-3-540-68017-8_131.
29. Mosquim Júnior S, de Oliveira J. Comparative study on data mining techniques applied to breast cancer gene expression profiles. In: Proceedings of the 10th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2017) - BIOINFORMATICS. Porto, Portugal: SciTePress; 2017. doi: 10.5220/0006170201680175.
30. Verma D, Mishra N. Analysis and prediction of breast cancer and diabetes disease datasets using data mining classification techniques. In: 2017 International Conference on Intelligent Sustainable Systems (ICISS). Palladam, India: IEEE; 2017. p. 533-8. doi: 10.1109/iss1.2017.8389229.
31. Williams K, Idowu PA, Balogun JA, Oluwaranti AI. Breast cancer risk prediction using data mining classification techniques. Trans Netw Commun. 2015;3(2):1-11. doi: 10.14738/tnc.32.662.
32. Burke HB, Goodman PH, Rosen DB, Henson DE, Weinstein JN, Harrell FE Jr, et al. Artificial neural networks improve the accuracy of cancer survival prediction. Cancer. 1997;79(4):857-62. doi: 10.1002/(sici)1097-0142(19970215)79:4<857::aid-cncr24>3.0.co;2-y.
33. Blamey RW, Ellis IO, Pinder SE, Lee AH, Macmillan RD, Morgan DA, et al. Survival of invasive breast cancer according to the Nottingham Prognostic Index in cases diagnosed in 1990-1999. Eur J Cancer. 2007;43(10):1548-55. doi: 10.1016/j.ejca.2007.01.016.
34. Edén P, Ritz C, Rose C, Fernö M, Peterson C. "Good Old" clinical markers have similar power in breast cancer prognosis as microarray gene expression profilers. Eur J Cancer. 2004;40(12):1837-41. doi: 10.1016/j.ejca.2004.02.025.
35. 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;11(7):e9176. doi: 10.5812/ijcm.9176.
36. Tanha J, Salarabadi H, Aznab M, Farahi A, Zoberi M. Relationship among prognostic indices of breast cancer using classification techniques. Inform Med Unlocked. 2020;18:100265. doi: 10.1016/j.imu.2019.100265.
37. Fisher ER, Costantino J, Fisher B, Redmond C. Pathologic findings from the National Surgical Adjuvant Breast Project (Protocol 4). Discriminants for 15-year survival. National Surgical Adjuvant Breast and Bowel Project Investigators. Cancer. 1993;71(6 Suppl):2141-50. doi: 10.1002/1097-0142(19930315)71:6+<2141::aid-cncr2820711603>3.0.co;2-f.
38. Edeki C, Pandya S. Comparison of data mining techniques used to predict cancer survivability. Int J Comput Sci Inf Secur 2012. 2012;10(6):1-119.
39. Witten IH, Frank E, Hall MA, Pal CJ. Practical machine learning tools and techniques. In: Data Mining. Vol 2. Amsterdam, The Netherlands: Elsevier; 2005.p. 403-13.