Breast Cancer Classification and Identification of Important Risk Factors with Bagged Cart

İpek Balikçi Çiçek [1*], Zeynep Küçükakçali [2]
[1-2] Inonu University, Faculty of Medicine, Department of Biostatistics and Medical Informatics, Malatya, Turkey

* Corresponding Author: **İpek Balikçi Çiçek**

Article Info

ISSN (online): 2582-8940
Volume: 05
Issue: 04
October-December 2024
Received: 25-09-2024
Accepted: 29-10-2024
Page No: 139-143

Abstract
Objective: The objective of this study is to classify breast cancer using the Bagged Classification and Regression Trees (CART) method and identify possible risk factors that play a role in the development of the disease. By evaluating the classification performance of the model, this study aims to contribute to the development of early detection and preventive strategies.
Materials and Methods: The study used an open-access dataset that includes variables such as menopausal status, tumor grade, lymph nodes, progesterone and estrogen receptors, age, and tumor size. The dataset was divided into training (80%) and testing (20%) subsets, and a 10-fold cross-validation approach was used to improve generalization. Bagged CART was employed for classification, and model performance was evaluated using metrics like accuracy, balanced accuracy, sensitivity, specificity, positive predictive value, negative predictive value, and F1-score.
Results: The Bagged CART model achieved high classification performance, with an accuracy of 99.6%, sensitivity of 100%, and specificity of 99.2%. Variable importance analysis revealed that recurrence-free survival time was the most critical factor in the classification process, followed by progesterone and estrogen receptors, lymph nodes, tumor size, and age. Variables such as tumor grade and hormonal treatment status had lower importance.
Conclusion: The findings indicate that Bagged CART is highly effective in breast cancer classification, with a strong emphasis on factors such as recurrence-free survival time, hormone receptor status, and tumor characteristics. The model's high accuracy and reliability suggest that it could contribute to enhancing early detection and preventive strategies, ultimately improving patient outcome.

Keywords: Breast cancer, Bagged CART, machine learning, classification, recurrence-free survival time, hormone receptors, early detection, risk factors, predictive modeling

Introduction

Breast cancer continues to be a significant public health concern, being the most commonly diagnosed cancer among women globally, with an incidence rate that has surpassed lung cancer. In 2020, approximately 2.3 million new cases were reported, accounting for 1 in every 8 cancers diagnosed, and leading to 685,000 deaths, making it one of the leading causes of cancer mortality worldwide. The disease's prevalence is particularly concerning, as it constitutes 1 in 4 cancer cases and 1 in 6 cancer deaths among women, ranking first in incidence and mortality in numerous countries [1]. The complexity of breast cancer, characterized by its heterogeneous nature and varying biological behaviors, underscores the urgent need for innovative approaches to enhance early detection and diagnosis. Traditional screening methods, such as mammography, have proven effective but are not without limitations. Challenges such as false positives and negatives can lead to unnecessary patient anxiety and increased healthcare costs [2]. Recent advancements in artificial intelligence (AI) and machine learning (ML) have emerged as promising solutions to improve the accuracy and efficiency of breast cancer diagnostics.

AI technologies, particularly those utilizing deep learning algorithms, have shown potential in analyzing extensive datasets, including imaging studies and laboratory results, to identify patterns that may elude human observers [3]. For instance, studies have demonstrated that AI can significantly enhance the interpretation of mammograms by reducing false positives and improving the detection of malignant lesions [3, 4]. Several studies have explored the efficacy of AI in breast cancer detection. For example, research by Shen *et al*. demonstrated that deep learning methods could improve breast cancer detection on screening mammography, achieving higher accuracy rates compared to traditional methods [1]. Similarly, Salim *et al*. conducted an external evaluation of commercial AI algorithms, finding that these systems could independently assess screening mammograms with sufficient diagnostic performance, thereby reducing the workload on radiologists while maintaining high sensitivity and specificity [5]. Moreover, Riveira-Martin *et al*. highlighted the robustness of AI systems across different mammography vendors, indicating that AI can maintain accuracy regardless of variations in imaging technology [6]. The integration of AI into mammography screening has the potential to address inter-reader variability, a significant issue in traditional screening practices where sensitivity can vary widely among radiologists. AI systems can standardize readings, thereby reducing the likelihood of missed cancers and improving overall diagnostic accuracy [3]. Furthermore, studies have shown that AI can assist in triaging mammograms, effectively prioritizing cases that require immediate attention and potentially reducing the interval cancer rate [7, 8]. In conclusion, the application of AI and ML in breast cancer detection represents a transformative approach that could significantly enhance early diagnosis and improve patient outcomes. As research continues to validate these technologies, their integration into routine clinical practice could lead to more accurate, efficient, and accessible breast cancer screening programs, ultimately contributing to a reduction in mortality rates associated with this prevalent malignancy [4, 9]. The aim of this study is to classify breast cancer using the machine learning-based Bagged CART method and to identify possible risk factors that play a role in the occurrence of the disease. In this context, while evaluating the classification performance of the model with the Bagged CART method, it is aimed to contribute to the development of early detection and preventive strategies by analyzing risk factors.

Materials and Method

Data description

The variables included in the open access dataset used in the study are given below.

1) Meno (menopausal status: 0 = premenopausal, 1 = postmenopausal) It pertains to a woman's hormonal condition and her attainment of menopause. Menopause is a natural phenomenon that generally transpires in women between the ages of 45 and 55, signifying the conclusion of their reproductive phase. The menopausal status is crucial in breast cancer as hormonal elements significantly influence the disease's development, progression, and treatment.

2) Grade -- In breast cancer, "grade" denotes an assessment of the morphological deviation of cancer cells from normal cells as observed microscopically. The grade indicates the aggressiveness or the probability of cancer proliferation and metastasis. It aids in establishing the treatment strategy and forecasting the patient's prognosis.

3) Nodes -- They assist in assessing the extent of cancer cell dissemination into organs and inform the selection of therapeutic interventions, including surgery, radiation, chemotherapy, or targeted treatments.A lower value is preferable.

4) pgr (progesterone receptors (fmol/l)) -- Progesterone receptors (PR) are proteins located on the surface of breast cancer cells. Their existence signifies that the cancer cells exhibit responsiveness to the hormone progesterone, which is endogenously synthesized in the body.

5) ER (estrogen receptors (fmol/l)) Estrogen receptors (ER) assist in ascertaining the sensitivity of cancer cells to hormone therapy.

6) Hormone (hormonal treatment, 0 = no, 1 = yes) -- Hormonal characteristics indicate how hormones respond to cancer. Therapy. If discovered positive indicates 1, and if found negative signifies that hormones are unresponsive to therapy.

7) rfstime (recurrence-free survival time) -- The rfstime feature in breast cancer denotes the duration a patient survives without exhibiting any signs or symptoms of cancer following the conclusion of initial treatment.

8) Age: Age of the patient in years.

9) size: Tumor size in millimeters.

10) Status: Patient status indicator, where 0 represents being alive without recurrence, and 1 indicates recurrence or death.

Modeling

In the modeling phase of the study, Bagged CART was employed to analyze the specified dataset. The dataset was divided into training and testing segments, with 80% of the data allocated for training purposes and 20% reserved for testing. To enhance the reliability of the modeling outcomes, a 10-fold cross-validation approach was utilized, which allowed for better generalization of the models by iteratively validating the performance on different.

The evaluation of the model performances was conducted using multiple assessment metrics, including accuracy, balanced accuracy, sensitivity, specificity, positive predictive value, negative predictive value, and F1 score.

Bagged CART (Classification and Regression Trees)

The concept of bagged CART (Classification and Regression Trees) has gained significant traction in various fields, particularly in machine learning and data analysis. Bagging, or bootstrap aggregating, is a powerful ensemble technique that enhances the predictive performance of decision trees by reducing variance and mitigating the risk of overfitting. This method involves generating multiple versions of a predictor and using these to obtain an aggregated prediction, which typically leads to improved accuracy compared to individual models. The foundational work on CART by Breiman *et al*. established a robust framework for constructing decision trees, which has since been enhanced through bagging techniques to create more reliable predictive models [10-12]. Bagging operates by creating multiple subsets of the training data through random sampling with replacement, allowing for the construction of several decision trees. Each tree is trained independently, and their predictions are combined, usually by averaging for regression tasks or majority voting for classification tasks. This process effectively stabilizes the predictions and reduces the model's sensitivity to noise in the training data. The effectiveness of bagged CART has been

demonstrated across various applications, including credit scoring, medical diagnosis, and environmental modeling, showcasing its versatility and robustness in handling complex datasets [13]. In the context of medical applications, bagged CART has been utilized for predicting patient outcomes and treatment responses. For instance, studies have shown that bagging can enhance the accuracy of models predicting the efficacy of treatments in patients with chronic conditions, such as liver cirrhosis and ascites. The standard CART procedure in these contexts often involves multiple steps, including data collection, model training, and validation, with bagging providing a significant boost in predictive performance [14]. The integration of bagging with CART algorithms allows for a more nuanced understanding and prediction of patient outcomes, ultimately leading to better clinical decision-making. Moreover, the application of bagged CART extends beyond healthcare into fields such as finance and environmental science. In financial modeling, for example, bagging has been employed to improve the accuracy of credit risk assessments, where the stability of predictions is crucial for effective risk management. The ensemble approach helps in addressing issues of data imbalance and enhances the model's ability to generalize across different datasets [15]. Similarly, in environmental studies, bagged CART has been used to predict ecological outcomes based on complex environmental data, demonstrating its efficacy in handling high-dimensional datasets with numerous predictor variables [16, 17]. The advantages of bagged CART are not limited to its predictive accuracy; it also provides a framework for interpreting complex models. The decision trees generated through CART are inherently interpretable, allowing practitioners to understand the decision-making process behind predictions. This interpretability is crucial in fields such as healthcare, where understanding the rationale behind a model's predictions can inform treatment decisions and patient management strategies [18]. Furthermore, the use of bagging helps to enhance this interpretability by providing a consensus view from multiple trees, thereby reducing the likelihood of erratic predictions that may arise from a single model. Despite its strengths, the bagged CART approach is not without limitations. One notable challenge is the computational cost associated with training multiple trees, especially in large datasets. However, advancements in computational power and optimization techniques have mitigated these concerns, making bagged CART a feasible option for many real-world applications [19]. Additionally, while bagging improves stability and accuracy, it may not always outperform other ensemble methods, such as boosting, in certain contexts. Therefore, practitioners must carefully consider the specific characteristics of their data and the goals of their analysis when selecting an appropriate modeling strategy [20]. In conclusion, bagged CART represents a significant advancement in the field of machine learning, providing a robust framework for enhancing the predictive performance of decision trees. Its application across various domains, including healthcare, finance, and environmental science, underscores its versatility and effectiveness in addressing complex predictive challenges. As the field continues to evolve, ongoing research into optimizing bagged CART and exploring its integration with other machine learning techniques will likely yield further improvements in predictive modeling capabilities.

Results

Metric	Value
Accuracy	0.996
Balanced accuracy	0.996
Sensitivity	1
Specificity	0.992
Positive predictive value	0.99
Negative predictive value	1
F1-score	0.995

The statistics of the model's classification performance on the training data show very high accuracy and consistency. The accuracy rate of 99.6% indicates that the model is almost error-free in the classification task, while the balanced accuracy value shows that the model can distinguish both classes equally well despite the imbalance of the classes. A sensitivity of 1 indicates that the model can detect positive examples completely, while a selectivity of 99.2% indicates that the model correctly distinguishes negative examples. In addition, a positive prediction value of 99% indicates that most of the examples that the model classifies as positive are indeed correct, while a negative prediction value of 100% indicates that it makes no mistakes in the examples it classifies as negative. Finally, an F1-score of 99.5% confirms that the overall performance of the model is balanced and strong in terms of both sensitivity and precision. These results clearly show that the model's classification ability on the training data offers high accuracy and reliability.

Variables	Variable Importance Values
Rfstime	100
Progesterone receptors	64.571
Estrogen receptors	55.569
Nodes	55.515
Size	54.371
Age	51.475
Grade=2	2.456
Hormon=yes	0.733
Grade=3	0.271

The variable importance table clearly highlights which variables play a crucial role in the model's classification process. The recurrence-free survival time variable, with the highest importance score of 100, emerges as the most critical factor influencing the model's decision-making. This is followed by progesterone receptors (64.571), estrogen receptors (55.569), nodes (55.515), size (54.371), and age (51.475), indicating that these variables significantly contribute to the model's predictive performance. On the other hand, variables such as grade=2 (2.456), hormon=yes (0.733),and grade=3 (0.271) have much lower importance, playing a minimal role in the decision-making process.

Discussion

The model's performance metrics reveal a highly successful classification capability. The accuracy and balanced accuracy values are high at 0.996, indicating that the model is generally accurate. In particular, the sensitivity value was 1.0, meaning that the model perfectly detected the positive class and gave no false negatives. The specificity was high at 0.992, indicating that the model was able to correctly distinguish the negative class and minimize false positives. The positive predictive value of 0.99 and negative predictive value of 1.0 indicate that the model predicts both classes with confidence

and has a very low error rate. Finally, the F1-score of 0.995 indicates that the model handles both classes in a balanced way and its overall performance is very high. These findings show that the model performs robustly and reliably, and the high performance indicates that the model offers overall high accuracy and reliability.

The variable importance ranking table, which shows which variables play a more critical role in the classification process of the model, provides important information. The variable with the highest importance is relapse-free survival time (rfstime), which stands out as the most decisive factor in the model's decision-making process with a score of 100. This is followed by progesterone receptors (64.571), estrogen receptors (55.569), nodes (55.515), tumor size (54.371) and age (51.475). These variables significantly affect the prediction performance of the model and play a major role in making accurate classifications. On the other hand, variables such as Grade=2 (2.456), hormone=yes (0.733) and Grade=3 (0.271) are of much lower importance and have a limited impact on the model's decision-making process. These findings suggest that the model focuses primarily on health status and biological factors in classification, but some clinical characteristics play a less important role.

The findings from the variable importance table underscore the critical role of various biological factors in the classification process of the model, particularly in the context of breast cancer prognosis. The recurrence-free survival time variable, which received the highest importance score, signifies its paramount influence on the model's decision-making. This aligns with existing literature that emphasizes the significance of survival metrics in cancer prognosis, as recurrence-free survival is a key indicator of treatment efficacy and patient outcomes [21, 22]. The strong emphasis on this variable suggests that models predicting cancer outcomes must prioritize survival metrics to enhance their predictive accuracy. Following the recurrence-free survival time, the next most important variables include progesterone receptors, estrogen receptors, nodes, size, and age. The prominence of hormone receptor status, particularly progesterone and estrogen receptors, is well-documented in the literature. For instance, the presence of estrogen and progesterone receptors is known to correlate with better responses to hormonal therapies, which can significantly influence survival outcomes in breast cancer patients [23, 24]. Moreover, the number of affected lymph nodes and tumor size are established prognostic factors, with higher node involvement and larger tumor size typically indicating a worse prognosis [25, 26]. The variable age also plays a crucial role, as younger patients often present with more aggressive disease characteristics, which can adversely affect survival rates [27, 28]. Conversely, the variables with much lower importance scores, such as tumor grade and hormonal treatment status, indicate a minimal contribution to the model's decision-making process. This finding suggests that while tumor grade is traditionally considered a significant prognostic factor, its relative importance may vary depending on the context of the model and the specific patient population being studied. For instance, while higher tumor grades are generally associated with poorer outcomes, their predictive power may be overshadowed by other more influential variables such as hormone receptor status and survival metrics [29, 30]. The low importance of the variable "hormon=yes" further emphasizes that the mere presence of hormonal treatment does not necessarily equate to improved outcomes, as the effectiveness of such treatments can be highly variable based on individual tumor biology [28]. In summary, the analysis of variable importance reveals that recurrence-free survival time, hormone receptor status, and tumor characteristics are pivotal in predicting breast cancer outcomes. The findings suggest that models should focus on these critical factors to enhance their predictive capabilities and ultimately improve patient management strategies. The nuanced understanding of how these variables interact and their relative importance can inform clinical decision-making and guide future research directions in breast cancer prognosis. Moreover, the identification of key variables that significantly contribute to predictive performance can aid in refining models for better clinical applicability. As machine learning and statistical modeling continue to evolve in oncology, understanding the relative importance of various predictors will be essential for developing more accurate and personalized treatment approaches.

References

1. Shen L, Margolies LR, Rothstein JH, Fluder E, McBride R, Sieh W. Deep learning to improve breast cancer detection on screening mammography. Scientific Reports. 2019;9(1):12495.
2. Houssami N, Kirkpatrick-Jones G, Noguchi N, Lee CI. Artificial Intelligence (AI) for the early detection of breast cancer: A scoping review to assess AI's potential in breast screening practice. Expert Review of Medical Devices. 2019;16(5):351–62.
3. Kim H-E, Kim HH, Han B-K, Kim KH, Han K, Nam H, *et al*. Changes in cancer detection and false-positive recall in mammography using artificial intelligence: A retrospective, multireader study. The Lancet Digital Health. 2020;2(3):e138–48.
4. Freeman K, Geppert J, Stinton C, Todkill D, Johnson S, Clarke A, *et al*. Use of artificial intelligence for image analysis in breast cancer screening programmes: Systematic review of test accuracy. BMJ. 2021;374:n1872.
5. Salim M, Wåhlin E, Dembrower K, Azavedo E, Foukakis T, Liu Y, *et al*. External evaluation of 3 commercial artificial intelligence algorithms for independent assessment of screening mammograms. JAMA Oncology. 2020;6(10):1581–8.
6. Riveira-Martin M, Rodríguez-Ruiz A, Martí R, Chevalier M. Multi-vendor robustness analysis of a commercial artificial intelligence system for breast cancer detection. Journal of Medical Imaging. 2023;10(5):051807.
7. Yala A, Schuster T, Miles R, Barzilay R, Lehman C. A deep learning model to triage screening mammograms: A simulation study. Radiology. 2019;293(1):38–46.
8. Aboutalib SS, Mohamed AA, Berg WA, Zuley ML, Sumkin JH, Wu S. Deep learning to distinguish recalled but benign mammography images in breast cancer screening. Clinical Cancer Research. 2018;24(23):5902–9.
9. Elhakim MT, Stougaard SW, Graumann O, Nielsen M, Lång K, Gerke O, *et al*. Breast cancer detection accuracy of AI in an entire screening population: A retrospective, multicentre study. Cancer Imaging. 2023;23(1):127.
10. Hong W, Xu Q, Zhou L. Large unbalanced credit scoring using Lasso-logistic regression ensemble. PLOS One. 2015;10(2):e0117844.

11. Ramosaj B, Pauly M. Who wins the Miss contest for imputation methods? Our vote for Miss BooPF. [Conference paper]. 2017.
12. Жебко ОО, Калінінa І, Hannichenko T. Intelligent classification system based on ensemble methods. System Technologies. 2023;3(146):61–75.
13. Amroussia M, Viedma O, Achour H, Abbes C. Predicting spatially explicit composite burn index (CBI) from different spectral indices derived from Sentinel 2A: A case of study in Tunisia. Remote Sensing. 2023;15(2):335.
14. Sone K, Taguchi A, Kawata A, Eguchi S, Miyamoto Y, Tanikawa M, *et al*. Transiently elevated D-dimer levels post-concentrated ascites reinfusion therapy cannot be used to predict deep vein thrombosis-pulmonary embolism. Journal of Obstetrics and Gynaecology Research. 2022;48(3):817–23.
15. Yang Z, Zhang S. Financial risk analysis of enterprises based on random forest algorithm. Applied Machine Learning. 2023.
16. Farrell AV, Wang G, Rush SA, Martin JA, Belant JL, Butler A, *et al*. Machine learning of large-scale spatial distributions of wild turkeys with high-dimensional environmental data. Ecology and Evolution. 2019;9(10):5938–49.
17. Zhu Y, Tao T, Yu K, Qu X, Li S, Wickert J, *et al*. Machine learning-aided sea ice monitoring using feature sequences extracted from spaceborne GNSS-reflectometry data. Remote Sensing. 2020;12(22):3751.
18. Miyazaki Y, Kawakami M, Kondo K, Tsujikawa M, Honaga K, Suzuki K, *et al*. Improvement of predictive accuracies of functional outcomes after subacute stroke inpatient rehabilitation by machine learning models. PLOS One. 2023;18(5):e0286269.
19. Jing X, Jing Z, editors. Feature extraction and classification method of electrooculogram based on variational modal decomposition. 2nd International Conference on Signal Image Processing and Communication (ICSIPC 2022); 2022: SPIE.
20. Yamamoto K, Nagao S, Tsu T, Matsushima T, Ishido Y, Narita M, *et al*. Quality of life assessment of cell-free and concentrated ascites reinfusion therapy during initial treatment for advanced ovarian cancer: A prospective cohort study. Journal of Obstetrics and Gynaecology Research. 2021;47(4):1536–43.
21. Zhao Y, Dong X, Li R, Ma X, Song J, Li Y, *et al*. Evaluation of the pathological response and prognosis following neoadjuvant chemotherapy in molecular subtypes of breast cancer. OncoTargets and Therapy. 2015;8:1511–21.
22. Kheradmand AA, Ranjbarnovin N, Khazaeipour Z. Postmastectomy locoregional recurrence and recurrence-free survival in breast cancer patients. World Journal of Surgical Oncology. 2010;8:1–7.
23. Zdravković KZ, Zdravković J, Stošić M, Stojanović I, Popović RM, Zlatković M. Prognostic significance of estrogen receptor, progesterone receptor and human epidermal growth factor receptor 2 in patients with breast cancer. Medicinski Pregled. 2017;70(1–2):5–11.
24. Liu X-Y, Ma D, Xu X-E, Jin X, Yu K-D, Jiang Y-Z, *et al*. Genomic landscape and endocrine-resistant subgroup in estrogen receptor-positive, progesterone receptor-negative, and HER2-negative breast cancer. Theranostics. 2018;8(22):6386–98.
25. Erić I, Petek Erić A, Kristek J, Koprivčić I, Babić M. Rak dojke u mladih žena: Patološka i imunohistokemijska obilježja. Acta Clinica Croatica. 2018;57(3):497–501.
26. Yang H, Wang R, Zeng F, Zhao J, Peng S, Ma Y, *et al*. Impact of molecular subtypes on metastatic behavior and overall survival in patients with metastatic breast cancer: A single-center study combined with a large cohort study based on the Surveillance, Epidemiology, and End Results database. Oncology Letters. 2020;20(4):1–11.
27. Wan J, Gao Y, Zeng K, Yin Y, Zhao M, Wei J, *et al*. The levels of the sex hormones are not different between type 1 and type 2 endometrial cancer. Scientific Reports. 2016;6(1):39744.
28. Vetter M, Stadlmann S, Bischof E, Georgescu Margarint EL, Schötzau A, Singer G, *et al*. Hormone receptor expression in primary and recurrent high-grade serous ovarian cancer and its implications in early maintenance treatment. International Journal of Molecular Sciences. 2022;23(22):14242.
29. Need EF, Selth LA, Trotta AP, Leach DA, Giorgio L, O'Loughlin MA, *et al*. The unique transcriptional response produced by concurrent estrogen and progesterone treatment in breast cancer cells results in upregulation of growth factor pathways and switching from a luminal A to a basal-like subtype. BMC Cancer. 2015;15:1–17.
30. Luo Y, Li Q, Fang J, Pan C, Zhang L, Xu X, *et al*. The loss of progesterone receptor predicts more aggressive biological features and worse outcomes in estrogen receptor-positive and HER2-negative inflammatory breast cancer. [Unpublished study]. 2023.