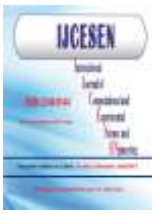
International Journal of Computational and Experimental Science and ENgineering (IJCESEN)

Vol. 10-No.4 (2024) pp. 911-917
http://www.ijcesen.com



ISSN: 2149-9144

Research Article

A Machine Learning Approach to Early Detection and Malignancy Prediction in Breast Cancer

Tuğçe ÖZNACAR[1]*, Neyhan ERGENE[2]

[1]Ankara Medipol University, Faculty of Medicine, Department of Biostatistics, Ankara-Turkiye
* **Corresponding Author Email:** tugce.sencelikel@ankaramedipol.edu.tr - **ORCID:** 0000-0003-0364-0401

[2]Ankara Medipol University, Department of Physiology, Ankara-Turkiye
Email: neyhan.ergene@ankaramedipol.edu.tr - **ORCID:** 0000-0003-2021-2564

Article Info:

DOI: 10.22399/ijcesen.516
Received : 15 October 2024
Accepted : 08 November 2024

Keywords :

Machine Learning
Artificial intelligence
Breast Cancer
AdaBoost

Abstract:

Breast cancer is the most common cancer among women, making early detection crucial for effective treatment. Traditional diagnostic methods often face limitations, leading to potential errors in diagnosis. This study explores the transformative potential of artificial intelligence (AI) and machine learning (ML) in breast cancer diagnosis, particularly through models like AdaBoost, SVM, Random Forest, and logistic regression. By analyzing key variables—such as age, tumor size, and menopausal status—this research aims to accurately differentiate between malignant and benign lesions. The findings reveal that the AdaBoost model significantly outperforms others, achieving an impressive AUC of 93.60% and a precision rate of 95.65%. This indicates its exceptional ability to accurately classify cases, minimizing false positives and ensuring reliable detection of true positives. With an F1 score of 86.27%, AdaBoost effectively balances precision and recall, positioning it as a valuable tool in clinical settings. Overall, this study underscores the importance of integrating AI-driven approaches in breast cancer diagnosis, enhancing accuracy and improving patient outcomes while reducing unnecessary invasive procedures. The promising results advocate for the adoption of these advanced techniques in healthcare, paving the way for more personalized and effective treatment strategies.

1. Introduction

Breast cancer stands out as the most prevalent cancer type among women worldwide, affecting millions of lives each year [1]. Early detection is critical in the treatment of this disease, as the effectiveness of therapeutic options varies significantly depending on the stage of cancer [2]. In this context, the rapid and accurate differentiation between malignant and benign lesions has become an essential requirement.

While traditional diagnostic methods, particularly imaging techniques and biopsies, are vital, human factors and limitations in these processes can lead to diagnostic errors [3]. In recent years, the application of artificial intelligence (AI) and machine learning (ML) in the medical field holds the promise of a revolutionary shift in breast cancer diagnosis. Specifically, deep learning algorithms, trained on large datasets, have demonstrated high accuracy in distinguishing between malignant and benign lesions [4].

AI systems have emerged as supportive tools in various domains such as image analysis, biomarker identification, and disease prognosis. This not only accelerates diagnostic processes but also reduces the rate of misdiagnosis, enabling better patient outcomes [5]. The advancements offered by AI in breast cancer diagnosis represent not just technical progress but also a fundamental shift in patient care. AI applications, with their ability to rapidly analyze large volumes of data, are enhancing the decision-making processes of physicians. For example, AI algorithms have the potential to detect subtle differences in mammography and ultrasound images, identifying malignant lesions at an early stage that may have previously been overlooked [6]. Such developments can guide more accurate treatment pathways from the outset, protecting patients from unnecessary biopsies and invasive procedures. Additionally, AI systems can analyze

individual data, such as patient history and genetic information, contributing to the creation of personalized treatment strategies. As a result, diagnostic accuracy increases, and patient satisfaction and quality of life improve significantly [7]. For these reasons, the integration of AI applications is becoming a critical element shaping the future of breast cancer diagnosis. This study will explore the role of AI in breast cancer diagnosis, focusing on novel approaches for differentiating between malignant and benign lesions.

The aim of this study is to predict malignant and benign breast lesions using machine learning algorithms, including AdaBoost, SVM, Random Forest, and logistic regression. Accurate and early differentiation in breast cancer diagnosis positively impacts patient treatment outcomes. This study seeks to distinguish between malignant and benign lesions through analyses performed on variables such as age, menopausal status, tumor size (cm), number of invasive lymph nodes, breast type, metastasis status, breast region, and medical history.

The evaluation of the employed machine learning models using performance metrics such as accuracy, precision, F1, and AUC will provide valuable insights for clinical applications and highlight the potential of AI in medical diagnostic processes. This research aims to emphasize the effectiveness of these algorithms in breast cancer diagnosis and their role in improving patient care in healthcare settings.

2. Material and Methods

2.1 Data Source

The breast cancer dataset used in this study was obtained from the publicly available Kaggle database, containing the medical records of 213 patients. Out of these patients, 120 were diagnosed with benign lesions, while 93 were diagnosed with malignant lesions.

2.2 Predictor Variable

The predictor variables in the breast cancer dataset include age, menopausal status, tumor size (cm), number of invasive lymph nodes, breast type, metastasis status, breast region, and medical history.

2.3 Data Splitting

The dataset was split into training and test sets with a 7:3 ratio.

2.4 Model Development

In this study, four different machine learning models—Adaptive Boosting (AdaBoost), Support Vector Machine (SVM), Random Forest, and Binary Logistic Regression—were applied to analyze the breast cancer data. These algorithms were developed using Python version 3.10.12 to ensure compatibility with the latest libraries and features.

Below, the key features and critical hyperparameters of each model are presented in detail. For hyperparameter optimization, the Grid Search method was employed. This method helped create a systematic workflow, ensuring the best model performance by addressing class imbalance and determining the optimal hyperparameters. The goal of this process was to enhance the effectiveness of each model by identifying the most suitable hyperparameter combinations. The grid search framework, k-fold cross-validation with k=5 was employed to identify the necessary hyperparameters that would yield the best model performance. The all process is shown in Figure 1.

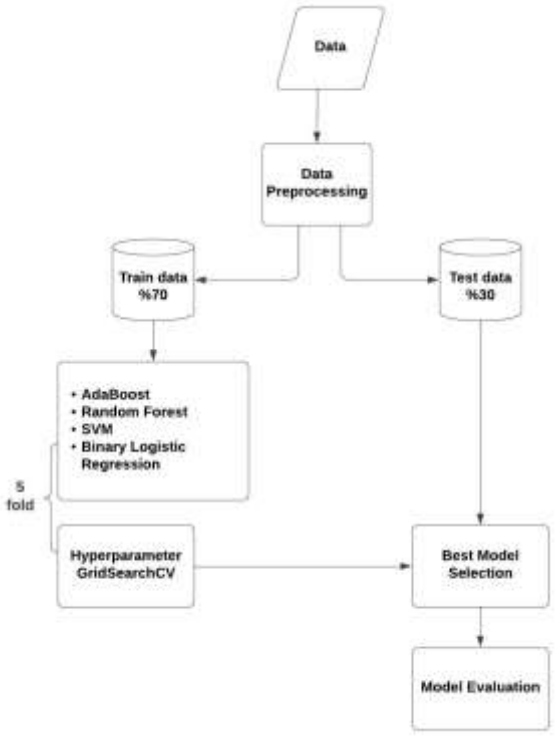


***Figure 1:**. Flow Chart*

2.4.1 AdaBoost

AdaBoost is an ensemble learning method that combines weak learners to create a strong classifier. Developed by Freund and Schapire in 1995, this algorithm improves performance by sequentially adding new learners that focus on the errors of

previous learners. AdaBoost dynamically updates the weights of each sample, aiming to minimize the errors made by previous classifiers. This approach enhances the overall accuracy of the model while allowing classifiers to become more complex over time.

The algorithm starts by assigning equal weights to all samples. After training each weak learner, the algorithm increases the weights of the misclassified samples. This adjustment encourages subsequent learners to focus more on correcting those errors, leading to improved model performance.

AdaBoost Hyperparameters:

- **n_estimators:** Determines the total number of weak learners.
- **learning_rate:** Controls the contribution of each weak learner.
- **base_estimator:** Specifies the type of weak learner used in the ensemble.
- **algorithm:** Defines the algorithm used during the classification process. There are two main options: "SAMME": Considers the contribution of each weak learner directly during boosting. "SAMME.R": Uses the weak learners' probability estimates to improve performance, often leading to better results when using probabilistic classifiers [8].

2.4.2 SVM

SVM, is a widely used method in supervised learning, particularly effective for classification problems. Developed in the mid-1990s by Vapnik and colleagues, SVM aims to find the hyperplane that best separates data points. Its ability to perform effectively even in high-dimensional spaces allows SVM to excel in complex classification tasks.

SVM utilizes the concept of maximum margin to classify data points. This approach seeks to identify the optimal separation line (or hyperplane) that maximizes the distance between the class boundaries, thereby enhancing the model's generalization capability and minimizing the risk of overfitting. Additionally, SVM offers the ability to transform non-linearly separable data into a higher-dimensional space through a method known as the "kernel trick." This transformation facilitates the classification of more complex data.

SVM's robust performance, even in high-dimensional spaces, makes it a powerful tool for a variety of classification problems.

SVM hyperparameters:

- **C (Regularization Parameter):** This parameter controls the complexity of the model.
- **kernel:** Determines the type of kernel function used to transform the data points.
- **gamma:** A hyperparameter used for RBF and polynomial kernels.
- **degree:** This parameter specifies the degree of the polynomial when using a polynomial kernel.
- **class_weight:** Used to determine the importance of different classes [9].

2.4.3 Random Forest

- Random Forest is an algorithm that belongs to supervised learning methods and stands out as one of the ensemble learning techniques. Developed by Leo Breiman, this method aims to create a powerful classifier by combining multiple decision trees. Random Forest allows for the independent construction of each decision tree, which are then aggregated to produce a final prediction. This process enhances the model's generalization ability and reduces the risk of overfitting.
- The Random Forest algorithm selects a random subset of training data for each tree and builds decision trees on these samples. Additionally, a random subset of features is chosen at each node to evaluate the best split. These two stages increase the diversity of the model, resulting in more stable and accurate predictions. By aggregating the predictions of many trees, Random Forest mitigates the variance associated with individual decision trees and improves overall model performance.
- Random Forest hyperparameters:
- **n_estimators:** This parameter specifies the total number of decision trees in the model.
- **max_depth:** Determines the maximum depth of each decision tree.
- **min_samples_split:** Defines the minimum number of samples required to split a node.
- **min_samples_leaf:** Specifies the minimum number of samples that must be present in a leaf node.
- **max_features:** Determines the maximum number of features to consider when evaluating each tree.
- **bootstrap:** Specifies the sampling method used for creating the trees. When set to True, the data samples for the trees are drawn using bootstrapping (sampling with replacement); when set to False, the entire dataset is used [10].

2.4.4 Binary Logistic Regression

Binary logistic regression is a statistical model used in binary classification problems, typically applied when the dependent variable is divided into two categories. This model is based on the assumption that independent variables (features) influence an outcome, utilizing a logistic function to determine

the effect of each independent variable on the dependent variable.
L2 regularization is a technique applied to control the model's complexity, minimizing the sum of the squares of the regression coefficients to prevent overfitting. This helps enhance the model's generalizability, allowing for more reliable predictions [11].

2.5 Performance metrics comparison of machine learning algorithms

Four machine learning algorithms were used to compare the accuracy, precision, recall and F1 scores, area under the curve (AUC) of the model metric values.

2.6 SHAP and Interpretability of Machine Learning Models

The Shapley Additive Explanations (SHAP) method is employed to enhance the interpretability of machine learning models, making their decision-making mechanisms more comprehensible. By utilizing SHAP values and feature importance plots, this approach visualizes the effects of the model's inputs, providing insights into how individual features contribute to the overall predictions [12].

2.7 Statistical Analysis

In the study, median (minimum-maximum) values were reported for numerical variables as descriptive statistics, while categorical data were presented as number and percentage n (%). The differences between benign and malignant groups in terms of numerical variables were analyzed using the Mann-Whitney U test, and for categorical variables, the Pearson Chi-square test was applied. A p-value of less than 0.05 ($p<0.05$) was considered statistically significant. All analyses were conducted using Python version 3.10.12.

3. Results and Discussions

The performance evaluations presented in the table compare various classification metrics for the AdaBoost, Support Vector Machine (SVM), Random Forest, and Logistic Regression models using the examined breast cancer dataset. Among these, the AdaBoost model demonstrates superior performance, particularly in the AUC (Area Under the Curve) metric (Table 2).
With an AUC value of 93.60%, AdaBoost is the strongest model in terms of classification capability. AUC measures the model's ability to distinguish between positive and negative classes, and in this context, AdaBoost outperforms all other models. Additionally, it achieves a high precision rate of 95.65%, indicating that the majority of samples classified as positive by the model are indeed positive, resulting in a very low false positive rate. The recall value of 78.57% shows that AdaBoost has a strong ability to detect true positives. Furthermore, the model achieves an F1 score of 86.27%, reflecting its capacity to balance precision and recall effectively.
In comparison, the SVM and Random Forest models exhibited moderate performance but fell short of AdaBoost, particularly in AUC values. SVM provides a balanced performance with an accuracy of 75% and recall, but its precision rate of 70% indicates a deficiency in reducing false positives. The Random Forest model garnered attention with an AUC and precision rate of 91.67%, yet it did not reach the overall performance level of AdaBoost.
The Logistic Regression model demonstrated effective performance with an accuracy of 89.06%, but its recall rate of 75% suggests that it may miss some true positives. The exceptionally high precision rate of 99.99% indicates very few false positives; however, this imbalance slightly limits the model's overall success.
There were significant differences between benign and malignant tumors in terms of tumor size, age, history, presence of inv-nodes, presence of menopause, and breast quadrant (Table 1).

Table 1: *Intergroup comparison results*

	Bening	Malign	p
Tumor Size (cm)	3.0 (1.0 - 7.0)	6.0 (1.0 - 14.0)	**<0.001**[a]
Age	33.0 (13 - 69)	47.0 (25 - 77)	**<0.001**[a]
Breast Left/right	66 (55.0%)/ 51 (42.5%)	41 (44.1%)/ 49 (52.7%)	0.287[b]
History No / Yes	80 (66.7%)/ 39 (32.5%)	44 (47.3%)/ 48 (51.6%)	**0.018**[b]
Inv-Nodes No / Yes	119 (99.2%)/ 1 (0.8%)	20 (21.5%)/ 72 (77.4%)	**<0.001**[b]
Menopause No / Yes	21 (17.5%) / 99 (82.5%)	50 (53.8%) / 43 (46.2%)	**<0.001**[b]
Metastasis No / Yes	116 (96.7%) / 4 (3.3%)	23 (24.7%) / 69 (74.2%)	**<0.001**[b]
Year 2019/2020	56 (46.7%) / 64 (53.3%)	45 (48.4%) / 47 (50.5%)	0.496[b]
Breast Quadrant			**<0.001**[b]
Lower inner	35 (29.2%)	9 (9.7%)	
Lower outer	36 (30.0%)	18 (19.4%)	
Upper inner	24 (20.0%)	21 (22.6%)	
Upper outer	24 (20.0%)	43 (46.2%)	

a: Mann-Whitney U test; median(min-max)

b: Pearson chi-Square test; n(%)

Table 2: *Evaluation metrics of algorithms*

	AdaBoost	SVM	Random Forest	Logistic Regression
Accuracy	0.8906	0.75	0.8750	0.8906
Recall	0.7857	0.75	0.7857	0.75
Precision	0.9565	0.70	0.9167	0.9999
F1	0.8627	0.7241	0.8462	0.8571
AUC	0.9360	0.7669	0.9167	0.9226

As age increases, the model's predicted partial dependence value initially fluctuates but reaches its highest level in the 60s. This indicates that age has a positive effect on the model up to a certain point, after which a declining trend is observed (Figure 2). As tumor size increases, the model's predicted dependence value generally rises.

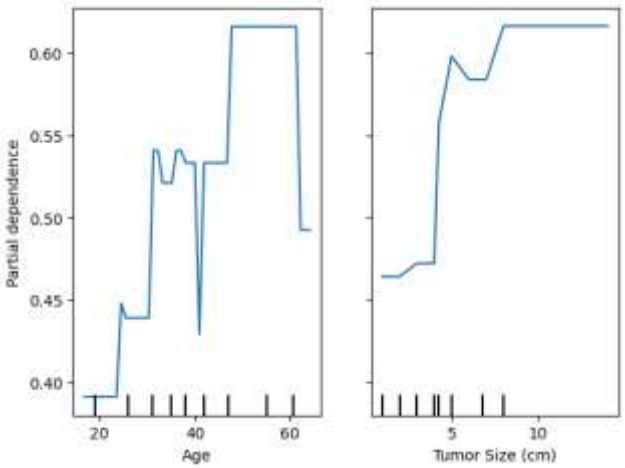


Figure 2*: Age and Tumor size partial dependence graph*

Notably, tumor sizes exceeding 5 cm significantly enhance the model's predictions, peaking around 10 cm. When examining the SHAP summary plot and SHAP importance graph, tumor size, age, and metastasis status emerge as the most important features (Figure 3-4).

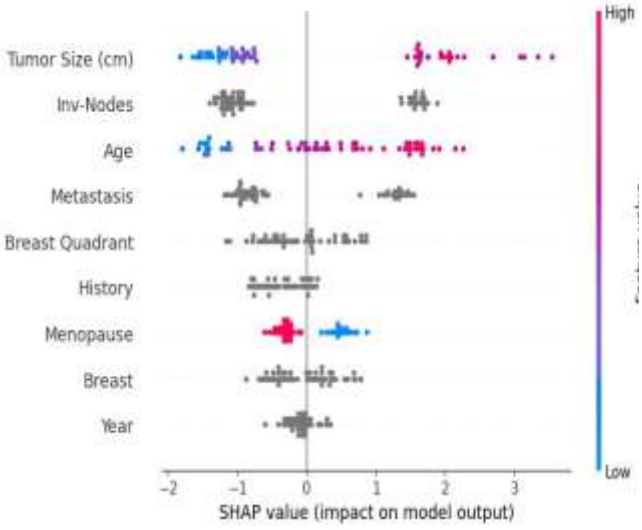


Figure 3**:** *SHAP Summary Plot*

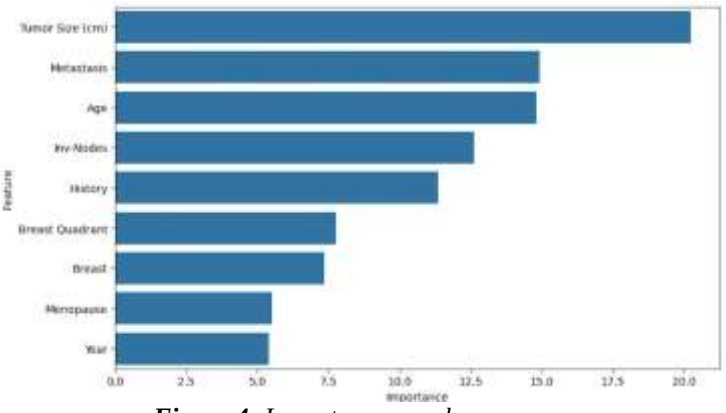


Figure 4**:** *Importance graph*

3.1 Discussion

This study compared the performance of AdaBoost, SVM, Random Forest, and Logistic Regression algorithms on breast cancer data. The findings indicate that the AdaBoost algorithm outperformed other models in terms of the AUC metric (%93.60). AUC is a metric used to measure the model's capacity to accurately distinguish between positive and negative classes, and AdaBoost's high success in this context supports similar studies in the literature [13][14]. In a study by Zuo et al. on a different breast cancer dataset, an AUC value of %98.7 was reported for the AdaBoost algorithm, highlighting its high discriminative ability in classification tasks[13]. Chai et al. examined the relationship between depression and breast cancer, reporting that the AdaBoost algorithm achieved the best modeling results with an AUC value of %84 [14].

The %95.65 precision rate achieved by AdaBoost indicates that the model maintains a very low risk of producing false positives. This finding reveals the model's significant success in distinguishing true positive classes and limiting positive results that could lead to unnecessary treatments or misdiagnosis.

Montazeri et al. (2016) evaluated machine learning algorithms to predict breast cancer survival using large datasets, obtaining the best result with the Random Forest algorithm at %91.67 AUC [15]. In our study, the AUC value for the Random Forest algorithm was also calculated as %91.67, supporting the findings of Montazeri et al.

In a study conducted by Zhou et al. (2024) on the Wisconsin breast cancer dataset, various machine learning algorithms were employed for predicting benign and malignant tumors, with the best modeling achieved through a combination of AdaBoost and logistic regression, resulting in an accuracy rate of %99.12 [16]. In our study, both AdaBoost and Logistic Regression exhibited

similar evaluation metrics. These findings suggest that future research could explore the AdaBoost-logistic regression combination proposed by Zhou et al.

Additionally, Ramakrishna et al. (2023) achieved the highest performance in breast cancer diagnosis by combining AdaBoost and Random Forest algorithms, resulting in an accuracy rate of %97.95. These types of hybrid algorithms could be considered for further analyses in light of the results obtained in our study. The use of hybrid algorithms is believed to provide higher accuracy and generalization capabilities [17]. In our study, AdaBoost stood out, particularly with high precision and AUC values, but lower results were obtained for recall. Therefore, it may be beneficial for future studies to investigate the hybrid modeling strategies proposed by Ramakrishna et al. to improve recall values.

Conversely, the %78.57 recall value of AdaBoost indicates the model's potential to miss some true positive cases. This situation suggests the negative impact of undiagnosed cases on clinical outcomes in cancer screenings. Notably, the study by He and Garcia (2009) highlights that low recall values can pose significant issues in imbalanced datasets and are critical for the early detection of diseases [18]. In this context, the low recall rate suggests that AdaBoost may overlook certain cases, which could affect patient prognosis.

The SVM and Random Forest algorithms have exhibited lower AUC values compared to AdaBoost. While SVM demonstrates balanced performance with a %75 accuracy and %70 precision, it falls short in classification capacity with a %75 AUC value. This finding is consistent with the studies conducted by Hsu and Lin (2002), which reported that SVM tends to perform inadequately, particularly in imbalanced datasets [19].

Logistic Regression has achieved an exceptionally high precision rate of %99.99; however, its %75 recall value indicates that the model is inadequate in capturing true positive cases. Studies by Menard (2002) have noted that the low recall values of Logistic Regression can lead to classification errors, particularly pronounced in imbalanced datasets [20]. The imbalance observed in this study raises the risk that Logistic Regression may miss some cases. These findings highlight the potential negative implications of false negative results on clinical outcomes. SHAP analysis indicates that tumor size, age, and metastasis status are the most significant factors in breast cancer diagnosis. The weights of these factors in the model are similarly emphasized in clinical literature. For instance, Smith et al. (2019) highlight the critical importance of tumor size on cancer prognosis and demonstrate that delays in cancer diagnosis associated with increasing age can adversely affect clinical outcomes. In our study, the model's predictive ability increased with age, although a decline was observed for individuals over 60. This finding suggests that while age increases cancer risk, other clinical factors also come into play in older populations [21]. The prominence of tumor size as one of the most important factors in the model, as indicated by SHAP analysis, aligns with findings in the literature. Notably, the study by Harris et al. (2016) reveals a direct relationship between increasing tumor size and metastasis, establishing this factor as a significant determinant in breast cancer diagnosis [22]. In our study, it was observed that tumor sizes exceeding 5 cm significantly enhanced the model's predictions. The comparison of sample sizes in other studies with the sample size in this study reveals the existence of not only studies with a similar number of participants but also those conducted with significantly larger datasets. Despite the small size of our dataset, high evaluation metrics have been achieved; however, working with larger datasets presents important opportunities to enhance the generalizability of the results and increase statistical power.

4. Conclusions

In conclusion, the AdaBoost model has emerged as the most successful classification algorithm for breast cancer diagnosis in this study. Its high precision rate provides a significant advantage in minimizing false positive cases, which is crucial for clinical applications. However, the lower recall value indicates potential risks associated with missed cases, which should not be overlooked. Therefore, it is recommended that hybrid models combining different algorithms or conducting model optimizations be implemented in clinical settings to achieve more balanced results. Furthermore, the results of the SHAP analysis have clearly identified the model's key determining factors, confirming the critical roles of tumor size and age in breast cancer diagnosis.

Future studies should assess the performance of algorithms more comprehensively, particularly using larger datasets and various data sources.

Author Statements:

- **Ethical approval:** The conducted research is not related to either human or animal use.

- **Conflict of interest:** The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper
- **Acknowledgement:** I sincerely thank Ahmet Öznacar for his invaluable contributions, which were crucial to this study's successful completion.
- **Author contributions:** The authors declare that they have equal right on this paper.
- **Funding information:** The authors declare that there is no funding to be acknowledged.
- **Data availability statement:** The dataset is available at https://www.kaggle.com/datasets/abdelrahman16/breast-cancer-prediction

References

[1] Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., & Jemal, A. (2018). Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, 68(6), 394-424. https://doi.org/10.3322/caac.21492

[2] Duffy, M. J., Patnick, J., & Vaughan, T. L. (2017). Screening for breast cancer. *British Medical Journal*, 358, j4242. https://doi.org/10.1136/bmj.j4242

[3] Lehmann, R., et al. (2020). AI in breast cancer: Review of current clinical applications and future perspectives. *Journal of Clinical Medicine*, 9(8), 2453. https://doi.org/10.3390/jcm9082453

[4] Yala, A., et al. (2019). A deep learning model to triage breast cancer patients. *Nature*, 573, 170-174. https://doi.org/10.1038/s41586-019-1456-4

[5] Gurcan, M. N., Boucher, G., Can, A., & Madabhushi, A. (2009). Histopathological image analysis: A review. *IEEE Transactions on Biomedical Engineering*, 56(2), 292-306. https://doi.org/10.1109/TBME.2009.2014043

[6] McKinney, S. M., et al. (2020). International evaluation of an AI system for breast cancer screening. *Nature*, 577, 89-94. https://doi.org/10.1038/s41586-019-1799-6

[7] Tavakoli, S., et al. (2021). The impact of artificial intelligence on breast cancer diagnosis and treatment: A comprehensive review. *Cancer Control*, 28(1), 10732748211009885. https://doi.org/10.1177/10732748211009885

[8] Freund, Y., & Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1), 119-139. https://doi.org/10.1006/jcss.1997.1504

[9] Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3), 273-297. https://doi.org/10.1007/BF00994018

[10] Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5-32. https://doi.org/10.1023/A:1010933404324

[11] Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (3rd ed.). Wiley. https://doi.org/10.1007/s10916-019-1272-1

[12] Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS 2017)* (pp. 4765-4774). Curran Associates Inc.

[13] Zuo, D., Yang, L., Jin, Y., et al. (2023). Machine learning-based models for the prediction of breast cancer recurrence risk. *BMC Medical Informatics and Decision Making*, 23(276). https://doi.org/10.1186/s12911-023-02377-z

[14] Chai, Y., Zhaoxiong, Y., Zhu, W., & Wang, H. (2024). Association between sleep duration, depression and breast cancer in the United States: A national health and nutrition examination survey analysis 2009–2018. *Annals of Medicine*, 56(1). https://doi.org/10.1080/07853890.2024.2314235

[15] Montazeri, M., Montazeri, M., Montazeri, M., & Beigzadeh, A. (2016). Machine learning models in breast cancer survival prediction. *Technology and Health Care*, 24(1), 31-42. https://doi.org/10.3233/THC-151064

[16] Zhou, S., Hu, C., Wei, S., & Yan, X. (2024). Breast cancer prediction based on multiple machine learning algorithms. *Technology in Cancer Research & Treatment*, 23. https://doi.org/10.1177/15330338241234791

[17] Ramakrishna, M. T., Venkatesan, V. K., Izonin, I., Havryliuk, M., & Bhat, C. R. (2023). Homogeneous Adaboost ensemble machine learning algorithms with reduced entropy on balanced data. *Entropy (Basel)*, 25(2), 245. https://doi.org/10.3390/e25020245

[18] He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9), 1263-1284. https://doi.org/10.1109/TKDE.2008.239

[19] Hsu, C., & Lin, C. (2002). A comparison of methods for multiclass support vector machines. *IEEE Transactions on Neural Networks*, 13(2), 415-425. https://doi.org/10.1109/72.991427

[20] Menard, S. (2002). *Applied logistic regression analysis* (2nd ed.). Sage Publications.

[21] Smith, R. A., Andrews, K. S., Brooks, D., Fedewa, S. A., Manassaram-Baptiste, D., Saslow, D., & Wender, R. C. (2019). Cancer screening in the United States, 2019: A review of current American Cancer Society guidelines and current issues in cancer screening. *CA: A Cancer Journal for Clinicians*, 69(3), 184-210. https://doi.org/10.3322/caac.21557

[22] Harris, L., Fritsche, H., Mennel, R., Norton, L., Ravdin, P., Taube, S., & Winchester, D. (2016). American Society of Clinical Oncology 2007 update of recommendations for the use of tumor markers in breast cancer. *Journal of Clinical Oncology*, 25(33), 5287-5312. https://doi.org/10.1200/JCO.2007.14.2364