RESEARCH ARTICLE

Editorial Process: Submission:05/19/2024 Acceptance:10/13/2024

Predicting HER2 Status Associated with Breast Cancer Aggressiveness Using Four Machine Learning Models

Amel Boulmaiz[1,2]*, Hajira Berredjem[1]*, Khadidja Cheikchouk[3], Aicha Boulkrah[1], Hayet Aouras[4], Hanene Djedi[3]

Abstract

Objective: Breast cancer (BC) is a heterogeneous disease with various biological and clinical subtypes. HER2 status (human epidermal growth factor receptor 2) is a crucial biomarker, associated with aggressive tumor behavior and poor prognosis. Advanced algorithmic models can aid in predicting cancer growth and metastasis, serving as valuable clinical tools for classification and treatment. Effective treatment strategies in oncology rely on accurate decision-making and early identification of factors associated with positive outcomes. Breast cancer (BC) presents challenges in understanding its contributing factors and establishing precise diagnostic methods. Our research introduces a novel method utilizing machine learning (ML) techniques to explore the relationship between various clinical and molecular variables focusing on predicting the status of the human epidermal growth factor receptor 2 (HER2), a key aggressiveness biomarker in BC. This objective aligns with leveraging artificial intelligence (AI) to support decision-making and address diagnostic considerations during treatment. **Methods:** Four ML models, namely logistic regression, random forest, LightGBM, and CatBoost, were implemented and evaluated using Python. The dataset was compiled by extracting medical records of BC patients, covering the period from 2018 to 2020. The model's predictive performance was evaluated using accuracy, precision, recall, and F1-score as performance metrics. **Result:** The models achieved varying accuracies between 86.36 and 95.45%. The logistic regression model achieved an accuracy of 90.90% while the random forest and LightGBM models achieved an accuracy of 86.36%. The CatBoost model outperformed others with a greater accuracy of 95.45%, indicating its superior predictive capabilities for HER2 status. The ML models demonstrated potential in predicting HER2 status, enabling early detection and facilitating personalized treatment strategies. **Conclusion:** Our findings emphasize the significance of AI and ML techniques in improving BC outcomes and guiding decision-making. Further research is required to explore the broader applications of ML in predicting comprehensive BC outcomes in diverse healthcare settings and among heterogeneous populations.

Keywords: Breast cancer- Clinical decision- HER2- Machine Learning- prediction accuracy

Asian Pac J Cancer Prev, **25 (10)**, 3609-3618

Introduction

Breast cancer (BC), characterized by uncontrolled cell proliferation, presents a considerable challenge in the realm of global health, significantly contributing to morbidity and mortality among women [1, 2]. It is recognized as one of the most prevalent malignancies affecting the female population, with the World Health Organization reporting a notable incidence of approximately 2.3 million new cases in 2020 [3]. A pivotal factor in the diagnosis and treatment of BC is the Human Epidermal Growth Factor Receptor 2 (*HER2*) gene, known to be overexpressed in roughly 20% of BC cases [4, 5]. This overexpression is associated with aggressive molecular subtypes of lobular and ductal carcinoma, underscoring the urgency for precise prediction of HER2 status [6, 7]. Accurate determination of HER2 status is crucial, as it not only influences prognosis but also guides the use of targeted therapies, highlighting the transition towards personalized treatment paradigms in oncology [8, 9]. Despite significant strides in the treatment of BC, accurately and reliably predicting HER2 status from clinical and pathological data remains a substantial gap in current practices. In light of the critical role of HER2 in BC prognosis and treatment, several studies have aspired to enhance the prediction accuracy of HER2 status. Machine learning (ML) approaches have emerged as a promising avenue, leveraging computational algorithms to refine diagnostic

*[1]Laboratory of Applied Biochemistry and Microbiology, Department of Biochemistry, Faculty of Sciences, University of Badji Mokhtar, 23000 Annaba, Algeria. [2]Environmental Biosurveillance Laboratory, Department of Biology, Faculty of Sciences, University of Badji Mokhtar, 23000 Annaba, Algeria. [3]Medical Oncology Service, Cancer Research Centre (CLCC), 23000 Annaba, Algeria. [4]Abdallah Nouaouiria Health Hospital Establishment - El Bouni, 23000 Annaba, Algeria.*For Correspondence: h_berjem@yahoo.fr, hajira.berredjem@univ-annaba.dz*

accuracy and patient outcomes. The diagnostic process across various types of cancers can be time-consuming for experts, underscoring the essential role of computer-assisted systems in BC diagnosis [10, 11]. Many studies have proved the important role of computer-aided diagnostics (CAD) in the early detection and classification of BC [12, 13]. This study distinguishes itself from prior research by focusing on HER2 status prediction using ML algorithms in BC diagnosis, while previous studies have highlighted the significance of HER2 as a biomarker and the efficacy of targeted therapies, with one or two models and a small dataset [14, 15]. These advancements in data mining and ML algorithms have significantly enhanced diagnostic precision. Umer et al. [16] have identified the challenges presented by the vast and heterogeneous nature of healthcare databases, highlighting the potential of Artificial Intelligence (AI) and ML to address these complexities through accurate and tailored diagnostic systems. Ming Ni et al. [17] utilized a Fisher discriminant analysis model based on radiomic features from MRI to predict clinicopathological subtypes of breast cancer with high accuracy. Researchers also developed an XGBoost machine learning model using radiomic fusion features from PET/CT images to predict *HER2* expression status in breast cancer patients, demonstrating improved predictive performance compared to other models [18]. Additionally, a CNN approach was developed to predict HER2 status and response to trastuzumab therapy using hematoxylin & eosin (H&E) stained tumor samples, achieving high accuracy in predicting both slide-level HER2 status and treatment response [19]. Given the pivotal role of HER2 status in guiding BC treatment and the existing gaps in its precise prediction, this study hypothesizes that advanced ML models can significantly improve the prediction accuracy of HER2 status from clinical and pathological data. This hypothesis is based on the idea that computer algorithms can find complex patterns in healthcare data that are hard to see with traditional diagnostic methods. The objective of this work is to evaluate the efficacy of various classification models, including Logistic Regression (LR), Random Forest (RF), LightGBM (LGBM), and CatBoost (CB), in predicting the HER2 status in BC diagnosis. This study aims to bridge the gap in accurate HER2 status prediction, thereby facilitating the development of personalized treatment strategies. Effective treatment of BC depends on several clinical parameters; therefore, accurate assessment of HER2 presence in BC is crucial for therapeutic decision-making [20]. Although targeted therapies such as trastuzumab and pertuzumab have significantly improved the prognosis for HER2-positive BC, these drugs are designed to target the overproduction of HER2, which can slow or stop tumour growth and improve patient outcomes [21, 22]. A significant contribution of this research lies in its comprehensive evaluation of ML models using a robust methodological framework that integrates a comprehensive evaluation of our models. This approach not only enhances the precision and reliability of predictions but also offers insights into the comparative performance of different algorithms, underscoring the strengths and weaknesses of each model, by leveraging a detailed analysis of clinical and pathological factors associated with BC outcomes. This study will refine HER2 prediction, improving diagnostic tools and treatment planning in personalized oncology care.

Materials and Methods

Study design and data collection

This prospective cohort study aims to investigate the determinants of BC outcomes. The study enrolled 109 BC patients, and data collection occurred from 2018 to 2020 at EHS Abdellah Nouaouria, El-Bouni-Annaba Hospital in Algeria, and the Anti-Cancer Center (CAC), University Hospital of Annaba, Algeria. The inclusion criteria consisted of histologically confirmed diagnosis of BC, availability of complete medical data including diagnostic test results, administered treatments, and follow-up outcomes, as well as sufficient information on clinical and pathological characteristics such as disease stage, histological type, hormone receptor (ER/PR) and HER2 status. Patients diagnosed with any other types of cancer in addition to BC, those with incomplete medical data or missing records, individuals with a history of prior illnesses or treatments that could potentially affect the study outcomes, and patients with significant co-morbidities or underlying health conditions were excluded from the study. Data collection was performed using a standardized data collection form, extracting demographic information and clinical characteristics including age, tumour size, number of nodes, metastasis, histologic type, stage, hormone receptor status, and HER2 status from medical records.

Ethics approval and consent to participate

The present study prospectively analyzes data retrieved from medical records. It adheres to the principles outlined in the Declaration of Helsinki and received approval from the local ethics committee at CHU Ibn Roched (No. 1726). Informed consent was obtained from all participants prior to their involvement in the study.

Machine Learning analysis

Four ML models, namely LR, RF, LGBM and CB, were implemented and evaluated using Python and supervised learning models were used. Data pre-processing was conducted by removing records containing missing values before training the models, ensuring the quality and consistency of the training data.

Baseline characteristic of BC cases according to HER2 status

To describe the *HER2* status in our dataset, we first summarized the study population using descriptive statistics. Categorical variables were presented as frequencies and percentages, providing an overview of the distribution of HER2 status. To further investigate the relationship between HER2 status, age, and hormonal subtype, we employed boxplots. The associations between the predictor variables (age, hormonal subtype) and the dependent variable (HER2 status) were assessed using LR. We examined the coefficients (odds ratios) and their

associated p-values to determine the significance of these associations. In addition to LR, we utilized an RF model to identify different features associated with HER2 status. The importance score of each feature was calculated using the trained RF model, representing its contribution to the prediction accuracy.

Model Implementation

To create predictive models for HER2 status classification, we used four machine learning (ML) algorithms: Logistic Regression (LR), Random Forest (RF), LightGBM (LGBM), and CatBoost (CB). These models were chosen based on their effectiveness and established usage in similar tasks. For more details about RF, LGBM, and CB, readers are referred to Saber M et al. [23]. The different steps of this experimentation process are graphically represented in the flowchart (Figure 1). All metrics were performed with the same split test train, ensuring a more objective and reproducible approach.

Model performance evaluation

The model performance evaluation focused primarily on the accuracy metric, which provides a comprehensive measure of the overall precision of the models predictions compared to the true labels.

$$\text{Accuracy} = TP + TN\,TP + FP + FN + TN \times 100 \quad (1)$$

By training the models on a comprehensive dataset, the study leveraged the capabilities of these algorithms to achieve accurate classifications and predictions. Additional metrics, including precision, recall, F1-score, sensibility and sensitivity were also computed to gain a deeper understanding of the model performance, considering true positives, false positives, and false negatives.

$$\text{Recall} = TP / (TP + FN) \times 100 \quad (2)$$

$$\text{Precision} = TP / (TP + FP) \times 100 \quad (3)$$

$$\text{F-measure} = 2 \times \text{Precision} \times \text{Recall} / (\text{Precision} + \text{Recall}) \quad (4)$$

The confusion matrix was calculated to further assess the model's performance.

$$ER = (FN + FP) / (TP + FP + FN + TN) = 1 - \text{Accuracy} \quad (5)$$

Sensitivity, also referred to as the true positive rate, is a crucial performance metric in binary classification models. It quantifies the model's ability to correctly identify positive examples.

$$\text{Sensitivity} = \text{True Positives} / (\text{True Positives} + \text{False Negatives}) \quad (6)$$

Specificity, also known as the true negative rate, is another important measure in binary classification. It assesses the model's ability to correctly identify negative examples.

$$\text{Specificity} = \text{True Negatives} / (\text{True Negatives} + \text{False Positives}) \quad (7)$$

At least, we employed two essential metrics to evaluate the performance of our classification models: AUC-ROC (Area Under the Receiver Operating Characteristic Curve) and loss function. AUC-ROC is a commonly used measure to assess the discriminatory ability of a model by quantifying its capability to distinguish between positive and negative classes. It ranges from 0.5 to 1.0, with a value of 1.0 indicating perfect performance. Also, loss function quantifies the discrepancy between model predictions and actual values, with a lower loss value indicating a better alignment between predictions and ground truth [24].

The common approach to solve supervised learning tasks is to minimize the loss function L:

$$L(f(x), y) = \Sigma[wi \times l(f(xi), yi)] + J(f) \quad (8)$$

Where:

l(f(x), y) is the value of the loss function at the point (x, y).

wi is the weight of the i-th object.

J(f) is the regularization term.

The loss function L measures the discrepancy between the predicted output f(x) and the true label y for a given input x. It is computed by summing the weighted individual losses l(f(xi), yi) over all training examples, where each example's weight wi reflects its relative importance [25, 26].

Results

Baseline characteristic of BC cases according to HER2 status

Table 1 showed that the median age at diagnosis was 47 (31-75) years. Overall, 77 (77.46%) patientswere aged 50 years or below, while 32 (29.36%) patients were above 50 years. Upon HER2 biopsie's evaluation, 85 (77.98%) revealed positive expression, while 24 (22.01%) revealed no *HER2* expression. Regarding the tumor grade, frequencies were 71.56%, 3.67% and 24.77% for grade II, grade I and grade III, respectively. The histological results showed that ductal carcinoma was the most common histological subtype (87.16%), while lobular carcinoma was less frequent (12.84%).

Regarding hormonal subtype, luminal A subtype was the most frequent (59.63%), followed by luminal B (21.10%), *HER2* overexpression (8.26%), and basal-like subtype (11.01%). We explored the relationship between HER2 status and the variables of interest; we used boxplots to visualize the distribution of HER2+ and HER2- cases and hormonal subtype across age (Figure 2). The median age for both HER2+ and HER2- groups was found to be 47 years. The interquartile range (IQR) for the HER2+ and HER2- groups ranges from 43 to 50 and from 42 to 54, respectively. In contrast, the HER2- group exhibits a wider spread of ages, suggesting greater variability in age within this group. In the other hand, the median age for each subtype is as follows: 47 for Luminal A, 45 for Luminal B, 49 for HER2 overexpression, and 49.5 for

Table 1. Patient's Distribution According to *HER2* Status.

	All patients (n = 109) n (%)	*HER2*+ (n = 85) n (%)	*HER2*- (n = 24) n (%)
Median Age (years)	47 (31-75)	-	-
≤ 50	77 (77.64)	58 (68.23)	19 (79.16)
> 50	32 (29.36)	27 (31.76)	5 (20.83)
NTGM Grade			
I	4 (3.67)	3 (3.52)	1 (4.16)
II	78 (71.56)	63 (74.11)	15 (62.50)
III	27 (24.77)	19 (22.35)	8 (33.33)
Histological subtype			
Ductal carcinoma	95 (87.16)	72 (84.70)	23 (95.83)
Lobular carcinoma	14 (12.84)	10 (11.76)	4 (16.66)
Hormonal subtype			
Luminal A	65 (59.63)	61 (71.76)	4 (16.66)
Luminal B	23 (21.10)	12 (14.11)	11 (45.83)
HER2 overexpression	9 (8.26)	0 (0.00)	9 (37.50)
Basal Like	12 (11.01)	12 (14.11)	0 (0.00)

Triple Negative. The IQR represents the range between the 25th and 75th percentiles of the age distribution. The IQR for Luminal A, Luminal B, and both HER2 overexpression and Triple Negative subtypes, indicating that the age distribution in Luminal A has a wider spread compared to the other subtypes. Additionally, we performed LR analysis to evaluate the influence of these variables on HER2 status (Table 2).

Furthermore, our analysis revealed that the variable "BMI" exhibits a significant negative association with HER2. Each unit increase in BMI is associated with a decrease of 69.17% in the odds of the dependent variable. The low p-value (p= 0.0098) confirms the significance of this association. The variable "Hormonal Subtype" exhibits a positive association with the dependent variable, but the p-value was > 0.05. The odds ratio of 2.0935 suggests an increase in the odds of the dependent variable.

Based on the feature importance scores obtained for the RF model (Figure 3), the hormonal subtype is the most significant feature and has the highest significance score of 0.30. The age and the tumour size (T) are also important features with scores of 0.20 and 0.09, respectively. This results suggests that the RF model's prediction is significantly influenced by both age and tumour size. Other features such as Lymph Node (N), Progesterone Receptor (PR), and Estrogen Receptor (ER), as well as BMI, also have some importance with scores of 0.05 to 0.07. However, Grade NGHT, Ki67, Metastasis (M), and Histological Type features have lower importance scores with values under 0.04

Machine learning models

The results presented in Figure 4 are based on evaluating the models. The performance of each model in predicting the HER2 status in BC is as follows: the LR model achieved an accuracy of 90.90% with a precision

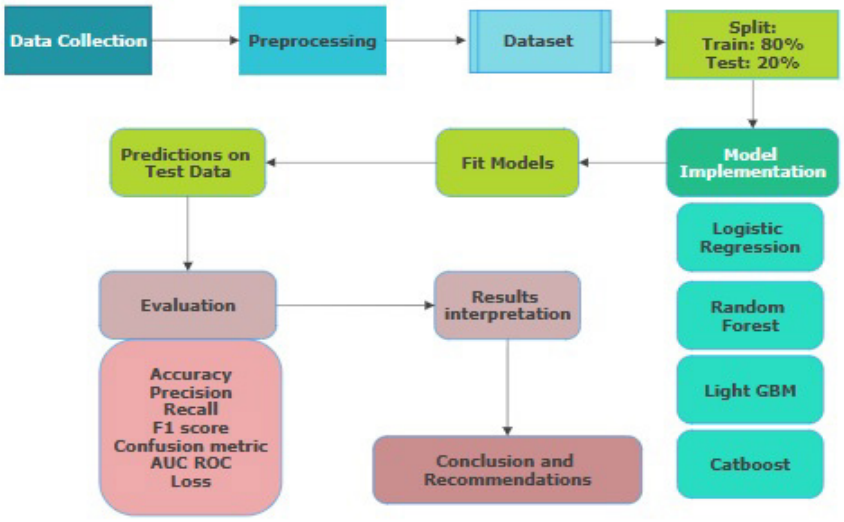
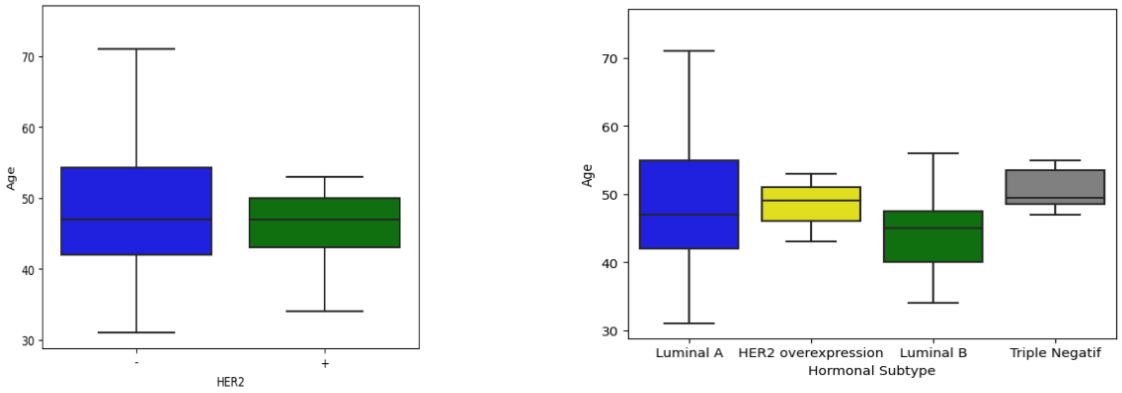


Figure 1. Methodology and Performance Evaluation of Predictive Models for *HER2* Classification in Breast Cancer



(a) (b)

Figure 2. (a) *HER2* dependent and (b) Hormonal subtype-dependent variation in age distribution.

Table 2. Relationship between the Predictor Variable and the Binary Outcome Variable Using LR

Variables	Coefficient	p-value	OR	(IC95%)
Age	-0.0343	0.2617	0.9661	[0.9334 - 1.0000]
BMI	-1.1764	0.0098	0.3083	[0.0956 - 0.9998]
Grade NGHT	0.2294	0.8371	1.2578	[0.9999 - 1.5835]
Hormonal Subtype	0.7388	0.2330	2.0935	[1.0004 - 4.3847]
T	-0.2308	0.5368	0.7938	[0.6312 - 0.9999]
N	-0.1923	0.6263	0.8250	[0.6811 - 0.9999]
M	-0.0680	0.8884	0.9341	[0.8738 - 1.0000]
Histological Type	-0.1359	0.6833	0.8728	[0.7629 - 1.0000]
Ki67	0.1082	0.8236	1.1142	[0.9329 - 0.9999]
PR	0.3119	0.2986	1.3660	[1.0004 - 1.8653]
ER	-0.4549	0.5622	0.6344	[0.4023 - 0.9992]

of 77.77%, recall of 100%, and F1-score of 87.5%. The RF and LGBM models achieved an accuracy of 86.36% with a precision of 75.0%, recall of 60.0%, and F1-score of 66.67%. The CB model outperformed the others with an accuracy of 95.45%, precision of 100%, recall of 80.0%, and F1-score of 88.89%. These results highlight the superior performance of CB in predicting the HER2 status in BC.

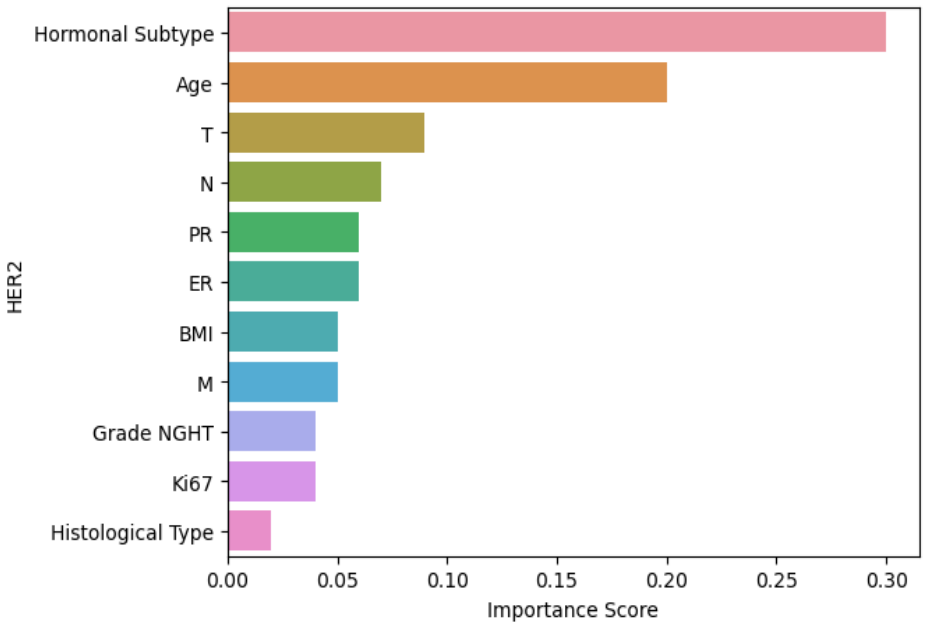
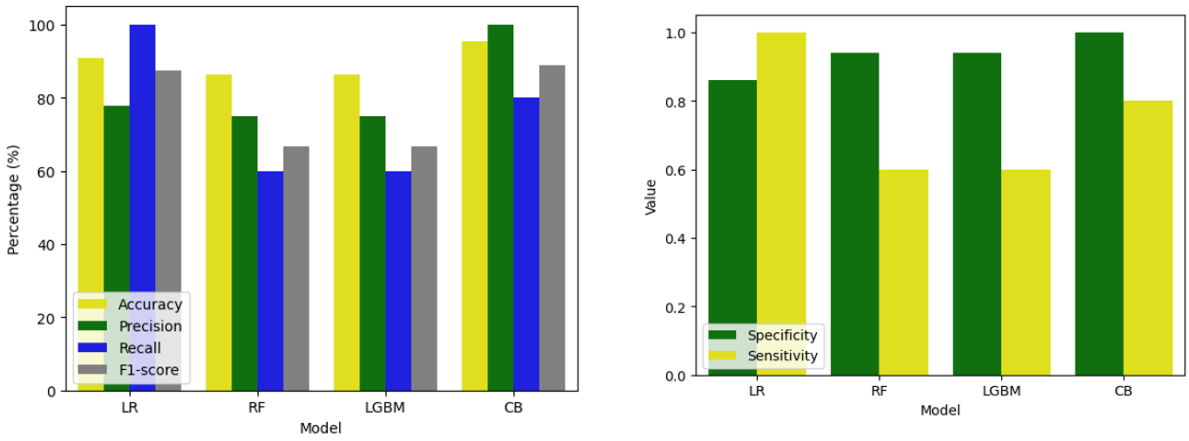


Figure 3. Feature Importance Plot for Random Forest



(a) (b)

Figure 4. The Performance Indicator (a) Accuracy, Recall, Precision and F1-Score; (b) Specificity and Sensitivity of the proposed models.

The specificity and sensitivity values for each model represent specific outcomes obtained during the evaluation of classification model performance. For the LR model, it indicates that the model has a specificity of 86% and a sensitivity of 100%. This means that the LR model has a relatively high ability to correctly identify negative and positive examples.

For the RF model and LGBM model, the specificity and the sensitivity were 94% and 60%, respectively. Finally, CB showed high values of specificity and sensitivity (100% and 80%), indicating a good ability to identify both positive and negative examples. The LR model achieved an accuracy of 90.90%, indicating that it correctly predicted the HER2 status of BC with a

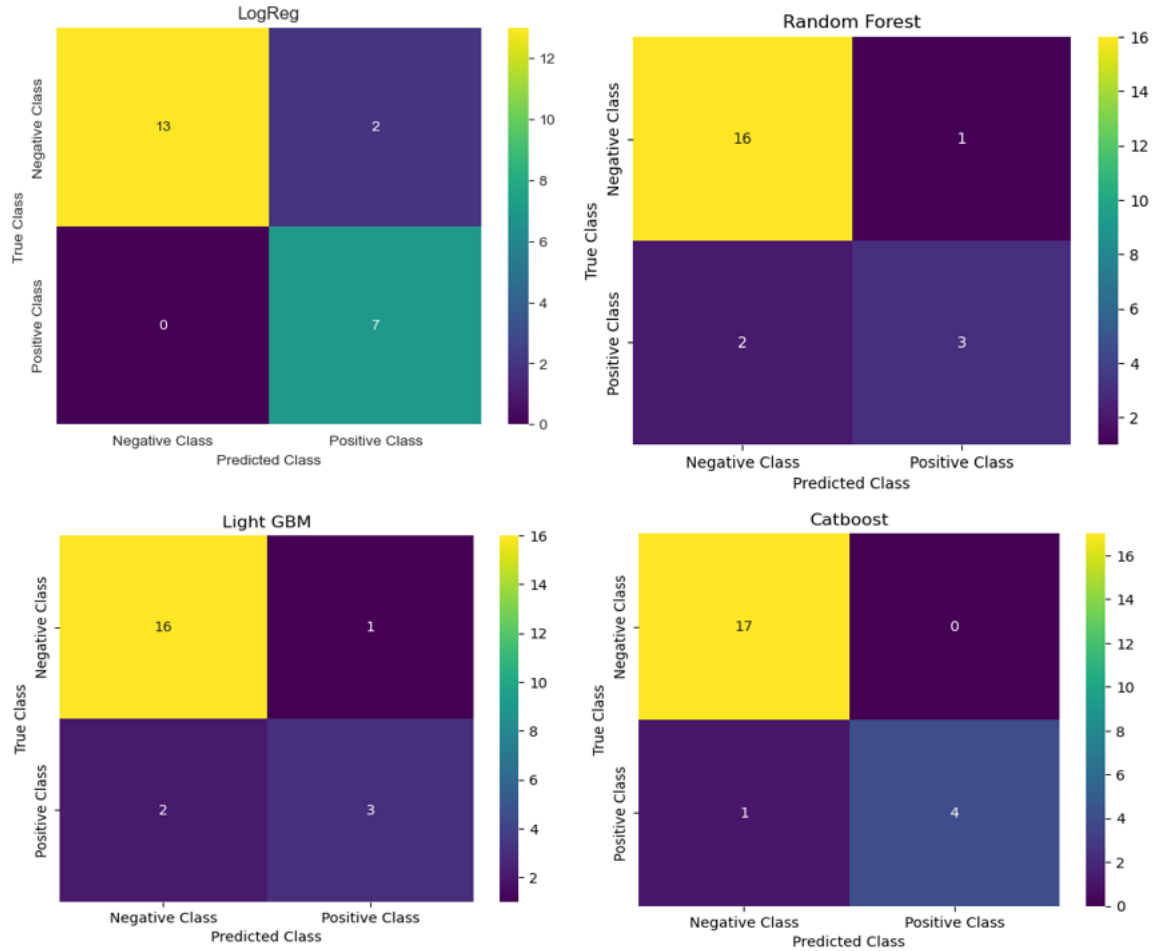
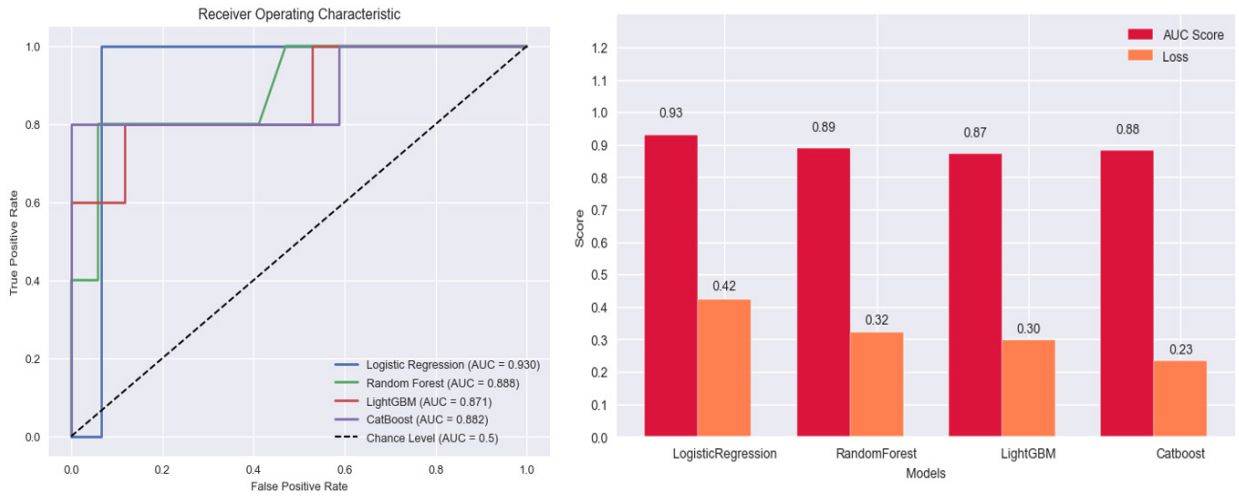


Figure 5. Confusion Metrics of Logistic Regression (LR), Random Forest (RF), LightGBM (LGBM) and CatBoost (CB).



(a) (b)

Figure 6. Area Under the Receiver Score and Loss for each model (a) ROC with AUC scores; (b) Loss for the implemented models

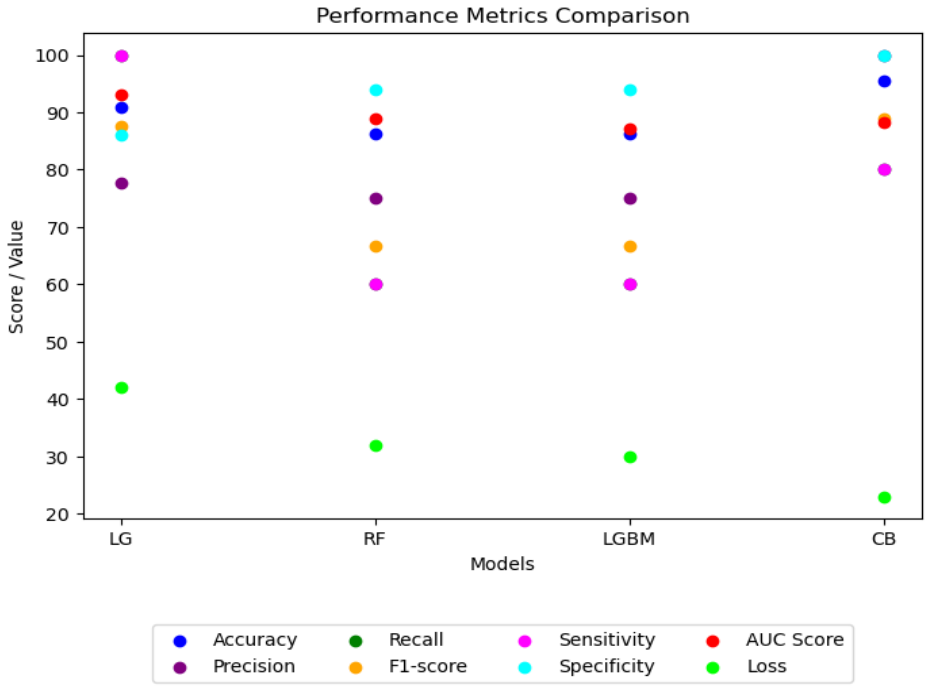


Figure 7. The Cross-Validation Score/Value According to the Number of Features in each Iteration. LG, Logistic Regression; RF, Random Forest; LGBM, LightGBM; CB, CatBoost

relatively high level of accuracy. As it is demonstrated in Figure 5, the confusion matrix for this model shows a true negative count of 13 true positive count, 7 false positives and 2 false negatives. The RF model and LGBM achieved an accuracy of 86.36%. The confusion matrix reveals 16 true negatives, 3 true positives, 0 false positive and 1 false negative. This indicates that the model correctly identified a significant portion of negative cases but had some misclassifications for positive cases. The CB model stands out with an impressive accuracy of 95.45%. The confusion matrix demonstrates its capability, with 17 true negatives, 4 true positives, 1 false positive and 0 false negative. This suggests a high level of correct predictions for both negative and positive cases.

Based on the obtained results, the used classification models have demonstrated good performance in terms of AUC-ROC for HER2 status prediction. Among the evaluated models, LR achieved an AUC-ROC of 0.930 indicating a very good ability to discriminate HER2 status. Additionally, RF, LGBM and CB models exhibited good performance with AUC-ROC values of 0.89, 0.88 and 0.87, respectively. The results of the loss function for the different classification models were also informative. LR presented a loss of 0.422, indicating some divergence between the predicted values and the actual values of HER2 status. The RF model displayed a slightly higher

loss with a value of 0.321, suggesting a slightly higher dispersion of predictions compared to the actual values. On the other hand, the LGBM and CB models exhibited remarkable performance with losses of 0.297 and 0.235, respectively. These low loss values indicate a high agreement between the predictions and the actual values of HER2 status, demonstrating the accuracy and quality of these models in predicting HER2 status (Figure 6).

Overall, the results suggest that CB stands out as the top-performing model for predicting HER2 status in BC. It demonstrates exceptional performance with high accuracy, precision, recall, AUC-ROC, and minimal loss. LR also performs well, especially in terms of accuracy and AUC-ROC. LGBM excels in AUC-ROC and demonstrates low loss, while RF, slightly less accurate, shows higher loss values. These results provide a distinct ranking of model performance, with CB and logistic regression emerging as the leading performers (Figure 7).

Discussion

HER2 positive cancers are associated with faster growth and are likely to respond to targeted therapy, while HER2 negative cancers grow more slowly and are less likely to recur or spread. About 20% of patients with metastatic BC have overexpression of HER2 gene [27]. The test is conducted via immunohistochemistry (IHC) to quantify the HER2 protein on cancer cells or fluorescence in situ hybridization (FISH) to detect additional copies of HER2 genes [28]. In metastatic BC, overexpression of the HER2 protein is associated with severe histology features, high mortality, a high risk of breast cancer recurrence, and a poor response to therapeutic interventions [29-31]. Recent research has centered on creating ML driven predictive frameworks for evaluating HER2 status in BC, with the aim of enhancing the comprehensibility and dependability of these models for healthcare practitioners. AI and ML techniques are transforming the field of oncology, particularly in the diagnosis and treatment of BC [32-33]. Our study conducted a comprehensive comparative analysis of classification models, including LR, RF, LGBM, and CB to predict the HER2 status in BC diagnosis. The findings indicated that the CB model outperformed others regarding precision, accuracy, and overall performance. This superiority can be attributed to CB's advanced handling of categorical features. Its efficient gradient boosting mechanism significantly reduced overfitting, a common challenge in ML models [34]. Notably, the CB model demonstrated a remarkable balance between precision and recall, crucial for minimizing false negatives in clinical diagnoses, thereby, ensuring patients receive appropriate treatment. Precise assessment of HER2 is also essential for identifying patients who may benefit from HER2 targeted therapy, either alone or in combination with chemotherapy, to ensure precision chemotherapy [35-36]. The efficacy of CB in our study corroborates with existing literature that highlights the potential of gradient-boosting models in medical diagnostics. By leveraging a complex ensemble of decision trees, CB effectively captures intricate clinical and pathological data patterns, which more simplistic models often miss [37-38]. This detailed data parsing is particularly vital in oncology, where subtle variations in gene expression can drastically alter treatment pathways. Moreover, our application of multiple metrics for model evaluation, including accuracy, F1 score, and AUC-ROC, aligns with best practices in ML, ensuring a holistic assessment of model performance. Our study identified several factors influencing the prediction of HER2 status in BC patients, including BMI, age, hormonal subtype, nodal involvement, and tumor size. Specifically, higher BMI levels were significantly associated with a negative HER2 status. Obesity and overweight status have been linked to reduced chances of achieving a pathological complete response (pCR) in HER2-positive luminal BC, particularly in the context of neoadjuvant therapy [39]. Additionally, meta-analyses have confirmed the negative impact of BMI on pCR rates in patients undergoing anti-HER2-based neoadjuvant therapy. Leveraging comprehensive clinical and pathological data, clinicians can make more informed and personalized treatment decisions, significantly impacting the management of BC [40]. The main subtypes of BC, including Luminal A, Luminal B, HER2-positive, and Triple-negative BC (TNBC), exhibit distinct characteristics based on hormone receptor and HER2 status, influencing treatment strategies and patient outcomes. Furthermore, age, tumour size, and lymph node involvement play critical roles in determining the stage of BC and guiding treatment decisions [41] by analyzing a combination of these factors, predictive models can aid healthcare providers in making informed decisions regarding treatment strategies and patient care. In responding to our research hypothesis, the findings unequivocally demonstrate that advanced ML models can significantly improve the prediction accuracy of HER2 status from clinical and pathological data. This study addresses the gap in precise HER2 status prediction and sets a precedent for applying advanced computational algorithms in personalizing BC treatment strategies. In pursuit of predicting HER2 status in BC and improving the clarity and reliability of these models for healthcare practitioners, our study demonstrates significant methodological rigor. Indeed, a ML model capable of better defining predictors associated with therapy response may influence therapeutic decisions for these patients in the near future. However, it is essential to acknowledge limitations, such as the reliance on the evaluation of model performance on a specific dataset. Further studies and validations are necessary to confirm and generalize these findings.

Conclusion and Future Perspective

The CB model exhibited the highest accuracy of 95.45% among all the tested models. It demonstrated a strong performance in correctly classifying both negative and positive cases, with a low misclassification rate. These findings suggest that the CB model may be the most reliable choice for predicting hormonal subtypes of BC. However, it is essential to compare this model with other established or emerging models to determine the optimal approach for predicting HER2 status. Future research should focus on identifying additional prognostic

factors such as genetic information, sequencing data, or patient lifestyle data, to enrich and enhance the accuracy of existing models.

The use of AI paves the way for more precise personalized medicine, where therapeutic decisions can be guided by reliable diagnostic predictions. Comparisons with other existing or emerging models are also crucial to determine the best approaches for predicting HER2 status. In addition, larger and more diverse patient cohorts are imperative to account for variations in genetic, environmental and lifestyle factors that can influence BC outcomes. Finally, exploring the potential of multi-omics models could open new avenues for understanding the underlying biology of BC and improving patient outcomes.

Author Contribution Statement

HB and AmelB: Conceptualization, methodology and design of the work. AmelB and AichaB: Investigation. AmelB: Software and Validation. HB, HD and HA: Project administration. KC and HD: Resources. HB and AmelB: wrote the initial draft. All authors reviewed and edited the manuscript.

Acknowledgements

This study was supported by the Algerian Ministry of High Education and Scientific Research, and the ATRSSV Algerian National Agency.

Data Availability

All the relevant data is presented in the manuscript.

Ethical Declaration

The present study adheres to the principles outlined in the Declaration of Helsinki and received approval from the local ethics committee at CHU Ibn Rochd (No. 1726 of October 2022). Informed consent was obtained from all participants prior to their involvement in the study.

Conflict of Interest

The authors declare that they have no conflict of interest.

References

1. Chtihrakkannan R, Kavitha P, Mangayarkarasi T, Karthikeyan R. Breast cancer detection using machine learning. Int J Innov Technol Explor Eng. 2019;8:3123-6. https://doi.org/10.35940/ijitee.K2498.0981119
2. Botlagunta M, Botlagunta MD, Myneni MB, Lakshmi D, Nayyar A, SaiGullapalli J, et al. Classification and diagnostic prediction of breast cancer metastasis on clinical data using machine learning algorithms. Sci Rep. 2023;13:485-502. https://doi.org/10.1038/s41598-023-27548-w
3. Anderson BO, Ilbawi AM, Fidarova E, Weiderpass E, Stevens L, Abdel-Wahab M, et al. The Global Breast Cancer Initiative: a strategic collaboration to strengthen health care for non-communicable diseases. Lancet Oncol. 2023;22:578-81. https://doi.org/10.1016/S1470-2045(21)00071-1
4. Bhushan A, Gonsalves A, Menon JU. Current State of Breast Cancer Diagnosis, Treatment, and Theranostics. Pharmaceutics. 2021;13:723-47. https://doi.org/10.3390/pharmaceutics13050723
5. Cardoso F, Kyriakides S, Ohno S, Penault-Llorca F, Poortmans P, Rubio IT, et al. Early breast cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. Ann Oncol. 2019;30:1194-220. https://doi.org/10.1093/annonc/mdz173
6. Schettini F, Chic N, Brasó-Maristany F, Paré L, Pascual T, Conte B, et al. Clinical, pathological, and PAM50 gene expression features of HER2-low breast cancer. NPJ Breast Cancer. 2021 ;7:1-12. https://doi.org/10.1038/s41523-020-00208-2
7. Mutai R, Barkan T, Moore A, Sarfaty M, Shochat T, Yerushalmi R, et al. Prognostic impact of HER2-low expression in hormone receptor positive early breast cancer. Breast. 2021;60:62–9. https://doi.org/10.1016/j.breast.2021.08.016
8. Bergeron A, Bertaut A, Beltjens F, Charon-Barra C, Amet A, Jankowski C, et al. Anticipating changes in the *HER2* status of breast tumours with disease progression towards better treatment decisions in the new era of HER2-low breast cancers. Br J Cancer. 2023;129:122–34.https://doi.org/10.1038/s41416-023-02287-x
9. Goddard KA, Weinmann S, Richert-Boe K, Chen C, Bulkley J, Wax C . *HER2* evaluation and its impact on breast cancer treatment decisions. Public Health Genomics. 2012 ;15 :1–10. https://doi.org/10.1159/000325746
10. Guo Z, Xie J, Wan Y, Zhang M, Qiao L, Yu J, Chen S, Li B, Yao Y. A review of the current state of the computer-aided diagnosis (CAD) systems for breast cancer diagnosis. Open Life Sci. 2022;17:1600–11. https://doi.org/10.1515/biol-2022-0517
11. Loizidou K, Skouroumouni G, Nikolaou C, Pitris C. A review of computer-aided breast cancer diagnosis using sequential mammograms. Tomography. 2022;8:2874-92. https://doi.org/10.3390/tomography8060241
12. Liew XY, Hameed N, Clos J. A review of computer-aided expert systems for breast cancer diagnosis. Cancers. 2021;13:2764. https://doi.org/10.3390/cancers13112764
13. Mirbabaie M, Stieglitz S, Frick NRJ. Artificial intelligence in disease diagnostics: A critical review and classification on the current state of research guiding future direction. Health Technol. 2021;11:693-731. https://doi.org/10.1007/s12553-021-00555-5
14. Fanizzi A, Latorre A, Bavaro DA, Bove S, Comes MC, Di Benedetto EF, et al. Prognostic power assessment of clinical parameters to predict neoadjuvant response therapy in HER2-positive breast cancer patients: A machine learning approach. Cancer Med. 2023;12:20663-9. https://doi.org/10.1002/cam4.6512
15. Amrane M, Oukid S, Gagaoua I, Ensari T. Breast cancer classification using machine learning. In2018 electric electronics, computer science, biomedical engineerings' meeting (EBBT) 2018 Apr 18 (pp. 1-4). IEEE.
16. Umer M, Naveed M, Alrowais F, Ishaq A, Hejaili AA, Alsubai S, et al. Breast cancer detection using convoluted features and ensemble machine learning algorithm. Cancers. 2022;14:6015. https://doi.org/10.3390/cancers14236015
17. Ni M, Zhou X, Liu J, Yu H, Gao Y, Zhang X, et al. Prediction of the clinicopathological subtypes of breast cancer using a Fisher discriminant analysis model based on radiomic features of diffusion-weighted MRI. BMC Cancer. 2020;20:1073. https://doi.org/10.1186/s12885-020-07557-y
18. Chen Y, Wang Z, Yin G, Sui C, Liu Z, Li X, et al. Prediction of *HER2* expression in breast cancer by combining PET/CT radiomic analysis and machine learning. Ann Nucl Med. 2022;36:172-82. https://doi.org/10.1007/s12149-

021-01688-3
19. Farahmand S, Fernandez AI, Ahmed FS, Rimm DL, Chuang JH, Reisenbichler E, et al. Deep learning trained on hematoxylin and eosin tumor region of interest predicts *HER2* status and trastuzumab treatment response in HER2+ breast cancer. Mod Pathol. 2022;35:44-51. https://doi.org/10.1038/s41379-021-00911-w
20. Goddard KA, Weinmann S, Richert-Boe K, Chen C, Bulkley J, Wax C. *HER2* evaluation and its impact on breast cancer treatment decisions. Public Health Genomics. 2012;15:1–10. https://doi.org/10.1159/000325746
21. Wolff AC, Hammond MEH, Allison KH, Harvey BE, Mangu PB, Bartlett JMS, et al. Human Epidermal Growth Factor Receptor 2 Testing in Breast Cancer: American Society of Clinical Oncology/College of American Pathologists Clinical Practice Guideline Focused Update. J Clin Oncol. 2018;36:2105-22. https://doi.org/10.1200/JCO.2018.77.8738
22. Swain SM, Shastry M, Hamilton E. Targeting HER2-positive breast cancer: advances and future directions. Nat Rev Drug Discov. 2023;22:101-26. https://doi.org/10.1038/s41573-022-00579-0
23. Saber M, Boulmaiz T, Guermoui M, Abdrabo KI, Kantoush SA, Sumi T, et al. Enhancing flood risk assessment through integration of ensemble learning approaches and physical-based hydrological modeling. Geomatics Nat Haz Risk. 2023 ;14 :2203798. https://doi.org/10.1080/19475705.2023.2203798
24. Fawcett T. An introduction to ROC analysis. Patt Recog Letters. 2006 ;27:861-74. https://doi.org/10.1016/j.patrec.2005.10.010
25. Heaton J, Goodfellow I, Bengio Y, Courville A. Deep learning. Genet Program Evolvable Mach. 2018;19:305-7. https://doi.org/10.1007/s10710-017-9314-z
26. Christmann A, Steinwart I. Loss Functions and Their Risks. In: Support Vector Machines. Information Science and Statistics. New York, NY: Springer. 2008. p. 21-47. https://doi.org/10.1007/978-0-387-77242-4_2
27. Horimoto Y, Ishizuka Y, Ueki Y, Higuchi T, Arakawa A, Saito M. Comparison of tumors with *HER2* overexpression versus *HER2* amplification in HER2-positive breast cancer patients. BMC Cancer. 2022;22:242. https://doi.org/10.1186/s12885-022-09351-4
28. Che Y, Ren F, Zhang X, Cui L, Wu H, Zhao Z. Immunohistochemical *HER2* Recognition and Analysis of Breast Cancer Based on Deep Learning. Diagnostics. 2023;13:263. https://doi.org/10.3390/diagnostics13020263
29. Slamon DJ, Clark GM, Wong SG, Levin WJ, Ullrich A, McGuire WL. Human breast cancer: correlation of relapse and survival with amplification of the HER-2/neu oncogene. Science. 1987;235:177-82. https://doi.org/10.1126/science.3798106
30. Tandon AK, Clark GM, Chamness GC, Ullrich A, McGuire WL. HER-2/neu oncogene protein and prognosis in breast cancer. J Clin Oncol. 1989;7:1120-8. https://doi.org/10.1200/JCO.1989.7.8.1120
31. Moasser MM. The oncogene HER2: its signaling and transforming functions and its role in human cancer pathogenesis. Oncogene. 2007;26:6469-87. https://doi.org/10.1038/sj.onc.1210477
32. Cordova C, Muñoz R, Olivares R, Minonzio J, Lozano C, Gonzalez P, et al. *HER2* classification in breast cancer cells: A new explainable machine learning application for immunohistochemistry. Oncology Lett. 2023 ;25 :1-9. https://doi.org/10.3892/ol.2022.13630
33. Conde-Sousa E, Vale J, Feng M, Xu K, Wang Y, Della Mea V, et al. HEROHE Challenge: Predicting *HER2* Status in Breast Cancer from Hematoxylin–Eosin Whole-Slide Imaging. J Imaging. 2022;8:213. https://doi.org/10.3390/jimaging8080213
34. Hancock JT, Khoshgoftaar TM. Cat Boost for big data: an interdisciplinary review. J Big Data. 2020;7:1-45. https://doi.org/10.1186/s40537-020-00369-8
35. Liu ZH, Wang K, Lin DY, Xu J, Chen J, Long XY, et al. Impact of the updated 2018 ASCO/CAP guidelines on *HER2* FISH testing in invasive breast cancer: a retrospective study of *HER2* FISH results of 2233 cases. Breast Cancer Res Treat. 2019;175:51-7. https://doi.org/10.1007/s10549-019-05148-5
36. Slamon DJ, Leyland-Jones B, Shak S, Fuchs H, Paton V, Bajamonde A, et al. Use of chemotherapy plus a monoclonal antibody against *HER2* for metastatic breast cancer that overexpresses HER2. N Engl J Med. 2001;344:783-92. https://doi.org/10.1056/NEJM200103153441101
37. Zhang Z, Zhao Y, Canes A, Steinberg D, Lyashevska O. written on behalf of AME Big-Data Clinical Trial Collaborative Group. Predictive analytics with gradient boosting in clinical medicine. Ann Transl Med. 2019;7:1-7. https://doi.org/10.21037/atm.2019.03.29
38. Park DJ, Park MW, Lee H, Kim YJ, Kim Y, Park YH. Development of machine learning model for diagnostic disease prediction based on laboratory tests. Sci Rep. 2021;11:1-11. https://doi.org/10.1038/s41598-021-87171-5
39. Ebinç S, Oruç Z, Sezgin Y, Karhan O, Bilen E, Yerlikaya H, et al. Factors influencing *HER2* discordance in nonmetastatic breast cancer and the role of neoadjuvant therapy. Future Oncol. 2022;18:3399-408. https://doi.org/10.2217/fon-2022-0198
40. Ligorio F, Zambelli L, Fucà G, Lobefaro R, Santamaria M, Zattarin E, et al. Prognostic impact of body mass index (BMI) in HER2+ breast cancer treated with anti-*HER2* therapies: from preclinical rationale to clinical implications. Ther Adv Med Oncol. 2022 ;14 :1-18. https://doi.org/10.1177/17588359221079123
41. O'Shaughnessy J, Gradishar W, O'Regan R, Gadi V. Risk of recurrence in patients with HER2+ early-stage breast cancer: literature analysis of patient and disease characteristics. Clin Breast Cancer. 2023;23:350-62. https://doi.org/10.1016/j.clbc.2023.03.007