ESMO GOOD SCIENCE BETTER MEDICINE BEST PRACTICE

ESMO OPEN SCIENCE FOR OPTIMAL CANCER CARE

ORIGINAL ARTICLE

Development of two machine learning models to predict conversion from primary HER2-0 breast cancer to HER2-low metastases: a proof-of-concept study

F. Miglietta[1,2†], A. Collesei[3†], C. Vernieri[4,5], T. Giarratano[1], C. A. Giorgi[1], F. Girardi[1], G. Griguolo[1,2], M. Cacciatore[6], A. Botticelli[7], A. Vingiani[5,8], G. Fotia[4,5], F. Piacentini[9], D. Massa[1,2], F. Zanghì[1,2], M. Marino[1,2], G. Pruneri[4,8], M. Fassan[10,11], A. P. Dei Tos[11], M. V. Dieci[1,2*] & V. Guarneri[1,2]

[1]Oncology Unit 2, Istituto Oncologico Veneto (IOV) - IRCCS, Padova; [2]Department of Surgery, Oncology and Gastroenterology, University of Padova, Padova; [3]Bioinformatics - Clinical Research Unit, Istituto Oncologico Veneto, IOV - IRCCS, Padova; [4]Medical Oncology Department, Fondazione IRCCS Istituto Nazionale dei Tumori (INT), Milan; [5]Oncology and Hemato-Oncology, Department University of Milan, Milan; [6]Pathology Unit, ULSS 9 - Treviso-Azienda ULSS 2 Marca Trevigiana, Treviso; [7]Department of Radiological, Oncological and Pathological Science, Policlinico Umberto I, "Sapienza" University of Rome, Rome; [8]Department of Advanced Diagnostics, Fondazione IRCCS Istituto Nazionale dei Tumori, Milan; [9]Department of Medical and Surgical Sciences for Children and Adults, University Hospital of Modena, Modena; [10]Istituto Oncologico Veneto (IOV) - IRCCS, Padova; [11]Pathology Unit, Azienda Universitaria Ospedaliera di Padova, Padova, Italy

Available online xxx

Background: HER2-low expression has gained clinical relevance in breast cancer (BC) due to the availability of anti-HER2 antibody—drug conjugates for patients with HER2-low metastatic BC. The well-reported instability of HER2-low status during disease evolution highlights the need to identify patients with HER2-0 primary BC who may develop a HER2-low phenotype at relapse. In response to the urgency of maximizing treatment access, we utilized artificial intelligence to predict this occurrence.

Patients and methods: We included a large multicentric retrospective cohort of patients with BC who underwent tissue resampling at relapse. The dataset was preprocessed to address relevant issues such as missing data, feature abundance, and target class imbalance. We then trained two models: one focused on explainability [Extreme Gradient Boosting (XGBoost)] and another aimed at performance (an ensemble of XGBoost and support vector machine).

Results: A total of 1200 patients were included in this study. Among 386 patients with HER2-0 primary BC and matched HER2 status at relapse, 42.5% ($n = 157$) converted to a HER2-low phenotype. The explainable model achieved a balanced accuracy of 58%, with a sensitivity of 53% and a specificity of 64%. The most important variables for this model were primary BC phenotype [mean Shapley value (SHAP) 0.540], primary BC histological type (SHAP 0.101), grade (SHAP 0.182), and sites of relapse (SHAP 0.008-0.213). The ensemble model had a balanced accuracy of 64%, with a sensitivity of 75% and a specificity of 53%.

Conclusions: This work represents one of the first proof-of-concept applications of machine learning models to predict a highly relevant phenomenon for drug access in modern BC oncology. Starting with an explainable model and subsequently integrating it with an ensemble approach enabled us to enhance performance while maintaining transparency, explainability, and intelligibility.

Key words: machine learning, explainability, breast cancer, HER2

INTRODUCTION

Breast cancer (BC) is the most frequently diagnosed solid tumor in women worldwide.[1] It represents a highly heterogeneous disease, characterized by diverse biological and clinicopathological features, thus accounting for different clinical behaviors in terms of prognosis and treatment sensitivity.

In clinical practice, the classification driving the treatment decision process in terms of prognostic stratification and drug access is mostly based on hormone receptor (HR) expression and HER2 status by immunohistochemistry (IHC) and *in situ* hybridization (ISH) analyses.[2,3] This allows the identification of three major BC subtypes: HR-positive/HER2-negative (HR+/HER2− BC), HER2-positive (HER2+ BC), and triple-negative (TNBC) BC.

*Correspondence to: Prof. Maria Vittoria Dieci, Department of Surgery, Oncology and Gastroenterology—University of Padova, Division of Oncology 2, Istituto Oncologico Veneto—IRCCS, Via Gattamelata 64, 35128 Padova, Italy. Tel.: +39-0498215931

E-mail: mariavittoria.dieci@unipd.it (M. V. Dieci).

†These authors contributed equally to this work.

In the past years, a finer classification based on the assessment of these traditional biomarkers has been put forward. Particularly, the dichotomic stratification in HER2+/− has now evolved as a three-tier classification, characterized by a further subdivision of HER2− BC into HER2-0 and HER2-low phenotypes, with the latter further classified based on IHC score 1+ or 2+ in the absence of gene amplification by ISH.[4] The recognition of HER2-low as a self-standing entity builds on the availability of the novel anti-HER2 antibody–drug conjugate trastuzumab deruxtecan for the treatment of patients with advanced HER2-low BC.[5] This scenario is set to evolve further in the near future due to the expansion of the target population for trastuzumab deruxtecan, which has been demonstrated to be effective even in patients with HER2-ultra low advanced disease (HER2 IHC >0 and <1+).[6] We and others reported that BC can undergo dynamic evolution throughout its natural history, and this can be captured as a phenotypic switch from the primary disease to recurrence.[7-9] This phenomenon has also been confirmed in the context of the HER2-low phenotype,[10,11] with a substantial proportion of patients with HER2-0 phenotype on primary tumors evolving toward HER2-low BC at disease relapse, with clinical implication in terms of enrichment of their therapeutic armamentarium. Based on these notions, international guidelines currently emphasize the clinical value of resampling the metastatic disease, also from the perspective of re-evaluating HER2 expression in case of HER2-0-only status throughout the disease history.[2,4] Within this framework, the development of tools capable of assisting in the identification of patients for whom a relapse/metastasis biopsy can provide substantially impactful clinical information at many levels is of great interest and, in this context, artificial intelligence (AI) may represent the ideal approach to address the complexity of this need. Indeed, although AI promises to integrate into many aspects of BC management, the main areas of application are restricted, so far, to BC early detection, prediction of BC development in higher-risk populations, and computational pathology. A broader and less niche use of AI-based tools is highly prioritized in BC research and requires the identification of relevant clinical questions to be addressed. Based on these premises we conducted a proof of principle study, developing two machine learning-based models, each addressing a different need, explainability, and performance, to predict the phenomenon of HER2-low phenotype gain from primary BC to relapse.

METHODS

Population

Patients (female or male aged ≥18 years) diagnosed with BC and undergoing resampling of relapse at five different Italian Institutions (Istituto Oncologico Veneto - IRCCS, Padova; Treviso Hospital, Italy; Fondazione IRCCS Istituto Nazionale Tumori, Milan; Oncologia Medica del Policlinico Umberto I, Rome; Division of Medical Oncology, Department of Medical and Surgical Sciences for Children and Adults, University Hospital of Modena) between January 1999 and December 2022 were included. Patients experiencing contralateral BC in the absence of other sites of recurrence were excluded. Estrogen receptor (ER) expression and HER2 status of primary and recurrent BC were retrieved from the original report and HER2− cases were reclassified as HER2-0 (IHC score = 0) versus HER2-low (IHC score = 1+/2+ in the absence of HER2 gene amplification by ISH). Based on ER expression and HER2 status, BC phenotype was stratified as follows: TN (ER = 0% and HER2-0/low), ER-low (ER = 1%-9% and HER2-0/low), ER+/HER2− (ER = 10%-100% and HER2-0/low), and HER2+ (any ER, HER2+). Clinicopathological, treatment, and follow-up data were also collected.

Machine learning model design

Intuitively, the modelization we present in this work aims at predicting patients likely to switch from '0' to 'low' HER2 status between primary tumor and recurrence. When achieved successfully, this task allows preselecting candidate patients likely to switch, and send them to a dedicated and personalized treatment path. Because of limitations in the clinical dataset, model training is not a simple operation for several reasons.

First and foremost, given the relatively restricted number of patients (1200 observations) and well-known cancer heterogeneity, the model is likely to overly adapt to the training set, thus causing overfitting, a recurring issue that must not be overlooked.[12] Therefore particular effort should be put into the training of a model able to generalize well with respect to the population and limit bias, eventually at the cost of reducing the overall performance.

Second, data missingness is a curse that, if poorly treated, may represent a source of bias in the model. Gaps in the observations must be tested for randomness to choose whether it is the best option to eliminate, by working only with complete cases, or imputing an entry in the dataset.

The third reason is represented by the strong imbalance within the target variable: as the switch between HER2-0 in primary malignancy and HER2-low in recurrence is a rare event, constituting just 18% of all cases, a naive model training on such an imbalanced dataset would certainly lead to distorted results.

As all aforementioned challenges were tackled, we introduce two models, each focusing on two aspects, explainability and performance. The impact of variables in the final model should be measurable and understandable, on the one hand, to avoid black boxes without practical meaning, and on the other to enhance visibility and potentially reduce sources of bias by enabling informed supervision on biological topics. Furthermore, an applied model must aim at optimal performance to enhance translation into clinical practice. Ultimately, we compared these two models with a simpler one, generalized linear model (GLM) to answer whether it is worth and beneficial to deploy nonlinear models in such a scenario. The models'

specifications and design will be explained in the following paragraphs.

Data missingness and feature selection. An immediate consideration to account for is data missingness. If admissible, data are commonly assumed to be missing completely at random, therefore allowing to simply remove observations or entire variables that are cursed by missingness over a certain defined threshold. In our case, this assumption cannot hold: the package naniar in R (RRID:SCR_001905) offers an implementation of Little's missing completely at random test,[13] which returned a significant *P* value ($\leq$0.05) in the dataset at hand, suggesting that features' missing values correlate over observations (i.e. missing not at random). Supplementary Figure S1, available at https://doi.org/10.1016/j.esmoop.2024.104087, provides a visualization of this phenomenon, and Supplementary Table S1, available at https://doi.org/10.1016/j.esmoop.2024.104087, details the worst cases of missingness. Therefore we are forced to assume patterns of missingness and impute the dataset as it is, ensuring not to introduce bias by removing relevant pieces of information. As the majority of variables is categorical, and the continuous ones were factorized and one-hot encoded through informed cut-offs (i.e. age, 50 years as threshold), for the imputation task we selected the multiple correspondence analysis (MCA) method,[14] implemented in package missMDA in R (R Foundation, Vienna, Austria), which is well suited for categorical data. We carried out the imputation procedure before applying any feature selection step, to exploit the relationship and redundancy between variables to better reconstruct an approximation of the 'real' value. However, to make the model focus on a reduced set of variables and increase its applicability to clinical practice, it is necessary to carry out feature selection: we chose to apply the Boruta algorithm[15] through the omonymous R package, which highlighted 10 features (primary BC's histology, grade and phenotype-related variables, plus recurrence site) as important for subsequent training.

Class imbalance. As we have already made clear, the target to be predicted is the switch characteristic from HER2-0 level to HER2-low level. This feature is obtained by combining the levels of HER2 at primary malignancy (401/1200 at the HER2-0 level) and at recurrence (382/1200 at the HER2-low level). Thus these two particular measurements are singularly occurring in approximately one-third of the population at hand, while their combination (target switch) is even rarer, present in 217 out of 1200 patients (18%). This scenario complicates the development of a functional model: as the prediction class is extremely unbalanced, the training step would cause the classification to focus on the majority class. This would eventually lead to high overall accuracy, as the majority class would be well detected, and the inability to correctly predict the minority class.[16] Therefore we randomly subdivided the dataset into training and test sets (70-30 split) by exploiting the balanced partition method available in package groupdata2. Then, we deployed a sampling technique to balance the target representation in the training data alone. To choose whether it was more beneficial to oversample, downsample, or apply a mixture of both, we generated sampled sets from the original training set, one for each combination of method and sample size, and trained a naive linear model to compare performances. At the end of this process, the optimal combination was a mixture of oversampling and downsampling, with a final sample size of 650. This corresponds approximately to downsampling the majority class by half and oversampling the minority class by double. The starting and resulting sample sizes for each of the two instances of the target class are shown in Supplementary Material, Figure S2, available at https://doi.org/10.1016/j.esmoop.2024.104087.

XGBoost model.

Model Framework and Shapley Values. Our main argument driving the decision process of the model was explainability. Therefore we first tried to fit a GLM, unfortunately not leading to sufficiently satisfactory results, due to the lack of tunable hyperparameters. We then opted to use tree-based models, selecting Extreme Gradient Boosting (XGBoost) for its scalability,[17] parallelization capabilities, and compatibility with 'shapr', the R implementation of the suite of algorithms designed to calculate Shapley values. After training a model, the Shapley value can be calculated as the average marginal contribution to the prediction of each single feature value, across all possible combinations of features. This metric enables the estimation of both the impact that each variable has on the target of interest and the direction of the said impact.[18] In our scenario, Shapley values suggest relevant baseline clinical features with respect to switching patients.

Overfitting. When training a classification model, one must avoid developing an algorithm that lacks generalization ability. This problem is commonly known as overfitting, and it is usually taken care of by validating the model on an external cohort. Unfortunately, given the scarce sample size of our dataset (particularly in terms of relative shortage of switching patients), we chose to apply bootstrap as a resampling technique, during the phase of training of the model, to ensure internal validation before moving to the test set.

SuperLearner: ensemble models. The explainable model has been tuned and trained to achieve, to our best effort, the highest performances in terms of balanced accuracy between the two classes. This is justified by the objective of the explainable model, which is characterizing the general relationships between relevant variables and the target. Ultimately, we felt the need to address a more contingent issue, which is the need to obtain the best overall performance, at the cost of explainability. One way to achieve this is by deploying the so-called ensemble models. Ensemble models constitute a machine learning approach able to combine multiple models in the learning

process: they suit our situation well, with a dataset characterized by high variance, and a single model with suboptimal accuracy.[19] Therefore we tried to combine the model we just presented, XGBoost, and Support Vector Machines, a flexible classifier able to perform well in nonlinear scenarios. The ensemble model, deployed thanks to the SuperLearner package in R, allowed us to overcome the limited generalization capability of a single learner, defined 'weak', combining the strength of two promising ones. After selecting the two weak models, we fine-tuned their hyperparameters to maximize their overall performance.

Ethical statement

The study was conducted in accordance with Good Clinical Practice (GCP) principles and received approval from the Institutional Ethical Committee (Institutional Review Board). Informed consent was obtained from all participating patients.

RESULTS

Population and HER2 dynamics

A total of 1200 patients were included. Primary BC phenotype data were available for 975 patients, with the following distribution: 199 HER2+ cases (20.4%) and 783 HER2− cases (79.6%). Among the HER2− group, 13.9% were TNBC (n = 135), 3.6% were ER-low (n = 35), and 62.1% were ER+/HER2− (n = 606). The main clinicopathological features of the HER2− population (n = 783) are detailed in Supplementary Table S2, available at https://doi.org/10.1016/j.esmoop.2024.104087. In brief, the vast majority of patients were female; 51.2% were HER2-0, and 48.8% were HER2-low. The distribution of BC phenotype across HER2 categories is presented in Table 1. In particular, among HER2-low cases, 13.7% were TN, 4.5% were ER-low, and 81.8% were ER+, respectively. Notably, TNBC was significantly enriched in the HER2-0 cases, while ER+ BC was significantly enriched in the HER-low cases (Table 1). The dynamics of HER2 status from primary BC to metastases are reported in Figure 1. In particular, among patients with HER2-0 primary BC who had matched HER2 status at relapse (n = 735), 42.5% converted to the HER2-low phenotype.

Explainable model performances

The model achieved a balanced accuracy of 58%, with a sensitivity of 53% and a specificity of 64%.

The plot of the Shapley values and the relative importance of the model variables are illustrated in Figure 2. The importance plot (Figure 2, top) highlights the overall contribution of each variable. The primary BC phenotype was of the greatest importance, followed by variables related to the site of relapse, primary BC histological type, and tumor grade. In particular, non-ER+/HER2− primary BC phenotype; no special type primary BC histology; high-grade primary tumors; usual metastatic localizations; and visceral, liver, nonlung, nonsoft tissue/skin metastases predicted the probability of switching from HER2-0 primary to HER2-low recurrent disease. Figure 2 (bottom) provides a cloud-like representation in the form of a Beeswarm plot, showing the contribution of each patient to the classification outcome, pushing predictions toward higher or lower probabilities within the trained model. In our scenario, the outcome is defined by the presence or absence of a switch from HER2-0 to HER2-low. Additional visual representation is presented in the Supplementary Material, specifically the Force plot in Figure S3, available at https://doi.org/10.1016/j.esmoop.2024.104087.

Table 1. Association between primary BC phenotype and HER2 status

Association	Total, n (%)	HER2-0, n (%)	HER2-low, n (%)	P value
TNBC	134 (100)	82 (61.2)	52 (38.8)	0.025
ER-low	34 (100)	17 (50.0)	17 (50.0)	
ER+	599 (100)	289 (48.2)	310 (51.8)	

BC, breast cancer; ER, estrogen receptor; TNBC, triple-negative breast cancer.

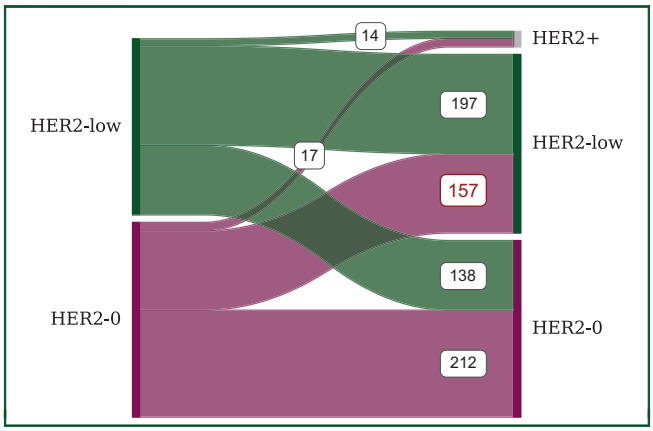

Figure 1. Sankey diagram illustrating the dynamics of HER2 level distribution between primary breast cancer (BC) and relapse. The target of interest is represented by patients who switched from HER2-0 to HER2-low, comprising 157 observations.

Ensemble model performances

The model achieved a balanced accuracy of 64%, with a sensitivity of 75% and a specificity of 53%. Given the study's focus on maximizing the likelihood of detecting conversion from HER2-0 primary BC to HER2-low recurrence, the model was trained to prioritize sensitivity at the expense of specificity. This approach ultimately improved balanced accuracy, increasing it by 6 percentage points. As shown in Table 2, the contributions of the two models were 69.1% for XGBoost and 30.9% for support vector machines.

Performance comparison with simpler models

Explainable and performance-related solutions, such as XGBoost and Ensemble models, involve nonlinear frameworks that require significant computation power and optimization efforts. As a result, one might consider using linear models, such as logistic regression, which are relatively easy to deploy and can be explained through odds ratio analysis. Therefore we compared our two main results

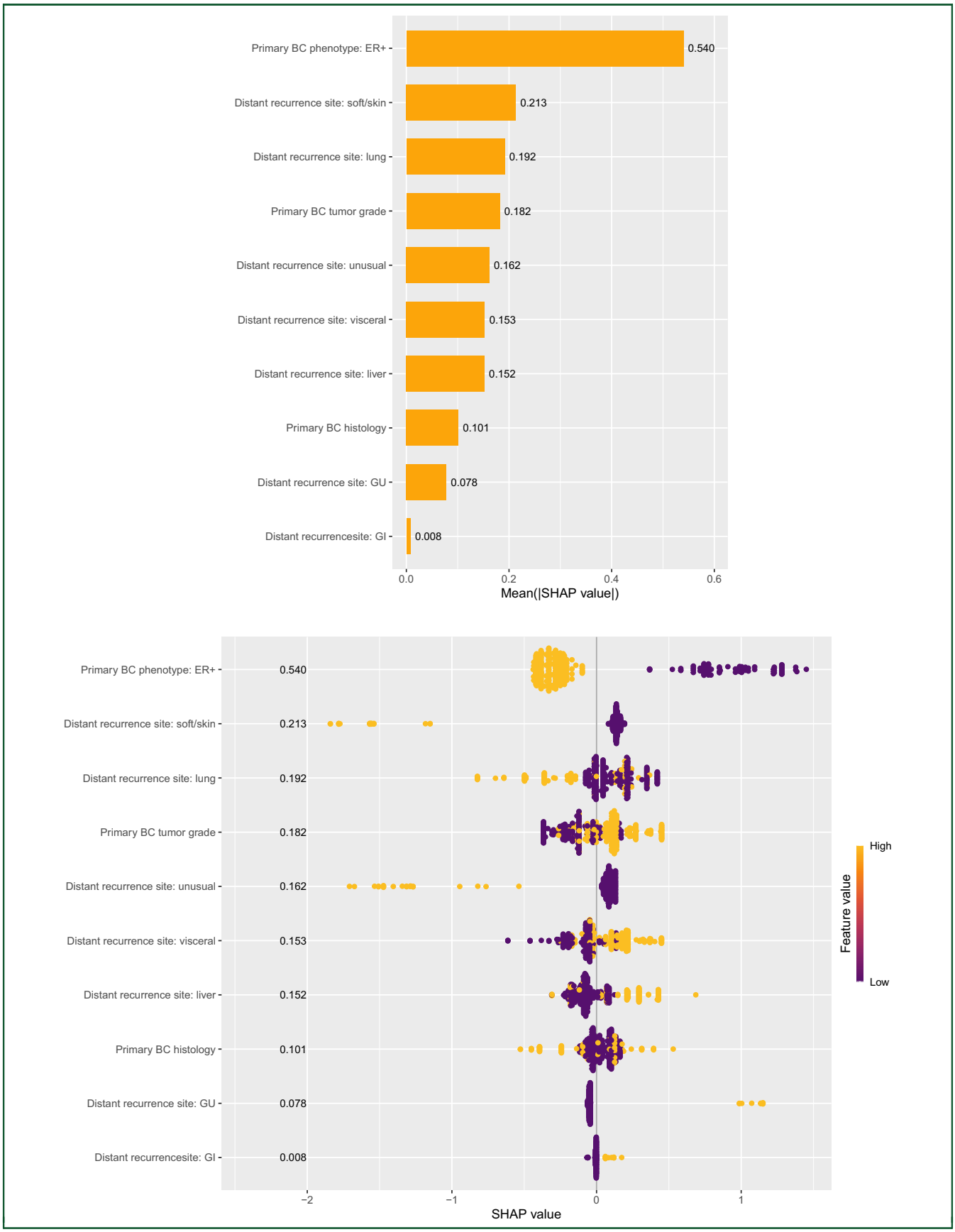

Figure 2. Elaboration of Shapley (SHAP) values into visual insights. Both graphs illustrate the overall contribution of each feature to the prediction: the Importance plot (top) aggregates the absolute SHAP values, providing a compact and additive explanation, while the Beeswarm plot (bottom) offers a detailed overview of each observation in a cloud-like visualization.
BC, breast cancer; ER, estrogen receptor; GI, gastrointestinal; GU, genitourinary.

https://doi.org/10.1016/j.esmoop.2024.104087

Table 2. The contribution of the two weak learners within the ensemble model

Learners	Risk[a]	Weight[a]
Extreme Gradient Boosting	0.201	0.691
Support vector machine	0.213	0.309

[a]Risk is a measure of accuracy that should be minimized for optimal performance, whereas weight indicates the importance of each component within the overall ensemble.

with a simple GLM. Figure 3 illustrates the performance of the logistic regression model using key metrics: balanced accuracy, sensitivity, and specificity. When focusing on explaining the biological relationships between features and the target, it is essential to prioritize balanced accuracy. Instead, when maximizing performance for predicting patients switching from HER2-0 to HER2-low, which is the positive value within the class of interest, sensitivity must be taken into account. Although the GLM and XGBoost models show similar performance, XGBoost demonstrates a slight advantage in balanced accuracy over the simpler model. This little advantage might be crucial in representing the biological context effectively. The coherence of Shapley values is influenced by overall model performance, particularly when dataset limitations prevent achieving high accuracy; for this reason, we recommend deploying XGBoost. On the other hand, the Ensemble model performs extremely well in predicting the positive class, achieving a sensitivity of up to 75%. This means the model can identify three out of four patients likely to switch in the following months at baseline. For completeness, we also report the specificity metric, even though that is beyond the main scope. It is evident that the Ensemble model prioritizes the positive class, leading to a trade-off with the less relevant class.

DISCUSSION

This study included 1200 patients with BC undergoing resampling of relapse/metastases from five centers and developed a nine-variable model based on classical clinicopathological features. The explainable model showed promising accuracy in predicting the acquisition of HER2-low phenotype at relapse in the case of HER2-0 primary tumor. In addition, the assessment of the Shapley values offered the opportunity to understand the importance that each variable retained for the model, thus allowing us to capture the potential clinicopathological drivers of the final observation (HER2-low gain). In particular, we observed that a non-ER+/HER2− primary BC phenotype was the most important feature associated with our target event, indirectly suggesting that ER-low/HER2− or TN phenotype at BC diagnosis was capable of predicting, with the highest importance, the acquisition of HER2-low phenotype on metastases. A possible explanation for such finding is that ER+/HER2− tumors are inherently at a higher likelihood of showing HER2-low phenotype at diagnosis,[20] thus potentially downsizing the relative impact of the acquisition of low HER2 expression levels at relapse/on metastases in this BC subtype. On the same ground may lie the observation that other variables that typically proceed in tandem with the total or subtotal absence of ER expression, namely, high-grade and no special type histology, were similarly important—albeit to a lesser extent—for the model. The other cluster of variables emerging as highly important for the model is those reflecting the site of metastases. In particular, the model isolated specific patterns of relapse/metastatic localization characterizing patients with HER2-0 primary BC gaining HER2-low phenotype at disease resampling. Lacking HER2-low phenotype and inherent biological significance, it is unlikely that this finding may have a solid biological driver. A more convincing explanation might be that certain metastatic patterns may be more technically prone to yield a HER2-low result. Indeed, the available evidence is scattered and inconsistent regarding the intrametastatic heterogeneity in terms of the prevalence of HER2-low expression and tendency of HER2 status instability,[10,21,22] thus being limited in scope in terms of the benchmark. Having met the need to support the observations related to the phenomenon of interest (HER2-low gain) with explainable drivers, we shifted our focus toward the possibility of enhancing performance by combining the XGBoost model with support vector machines, a flexible classifier. By doing so, we were able to reach a 75% sensitivity, which, in our view, albeit improvable, is remarkable. Indeed, we proposed a model sufficiently powered to overall meet the clinical principle of accountability for reliability, thus laying the ground for its external clinical validation. However, when evaluating the potential of the models, one needs to account for a foundational aspect driving the translation from the statistical to the clinical point of view: feature importance refers to the weight that the model attributes to each variable when computing the classification decision. One may be tempted to attribute causal links between features and the target, only based on these correlational relationships. It is worth reminding that more complex and assumption-demanding algorithms need to be deployed in order to draw such conclusions. In fact, we have been rather cautious when depicting the biological landscape, labeling our findings as potential indicators. We instead focused on the reliability and influence of our models in clinical practice, emphasizing on their predictive impact. With this clarification, we established a straightforward, clinically relevant, and prioritized objective in the current contemporary landscape of advanced BC: predicting the gain of the HER2-low phenotype. We demonstrate that applying an AI-based approach to this question, using easily obtainable traditional clinicopathological features, is both feasible and reliable. Ideally, this model fits within the specific setting of a patient with advanced, pretreated HER2-0 BC, who may be a potential candidate for trastuzumab deruxtecan in the presence of evidence indicating a gain of the HER2-low phenotype, for whom the decision to carry out a rebiopsy of a metastatic site is entirely guided by this consideration. However, our primary intent was to provide a proof of concept, and we do not expect our model to have an immediate clinical impact.

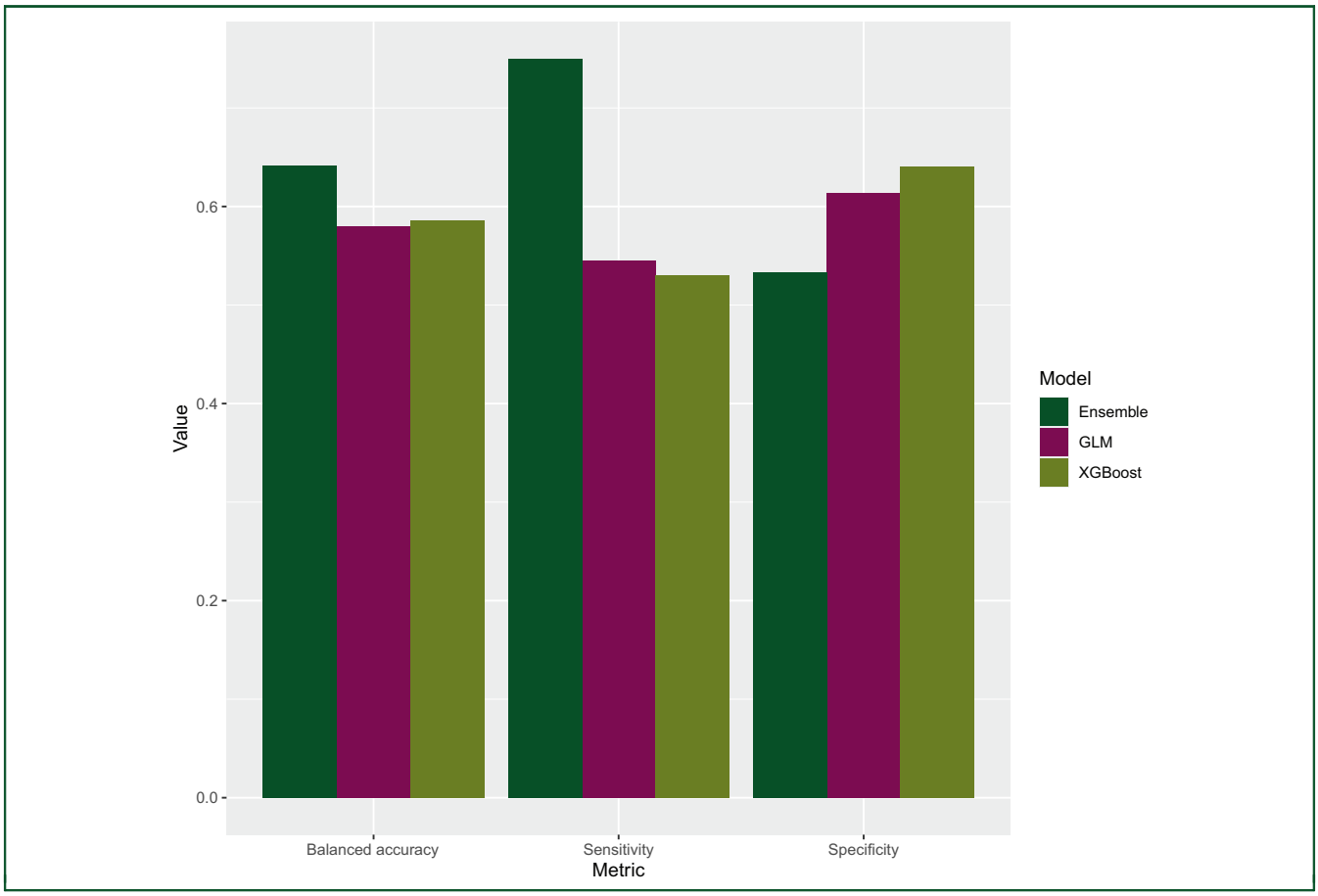

Figure 3. Performance comparison between models in terms of balanced accuracy, specificity, and sensitivity. The simplest model, the generalized linear model (GLM), performs reasonably well but shows lower performance compared with the Ensemble and Extreme Gradient Boosting (XGBoost) models.

It is indeed imperative to envisage a further evolution and expansion of this model, to cover the various levels of complexity of the information that can be obtained from a disease rebiopsy. Currently, the decision on whether to carry out a metastasis resampling on a patient with BC is guided by various clinical needs, sometimes overlapping, such as confirming the diagnosis of metastatic disease, evaluating the dynamics of the overall receptor/phenotypic status, and searching for diverse druggable targets. Furthermore, within the specific context of our study, enhancing the model with the ability to predict HER2-low gain in patients whose initial relapse biopsy revealed a HER2-0 status would be particularly valuable and make the model more acceptable. Such capability could inform the decision to carry out an additional biopsy of a metastatic site with the specific aim of capturing a druggable target, such as HER2-low. Based on that, before integrating such AI-based models within BC decision-making algorithms, they must be demonstrated to be capable of recapitulating all these aspects. Furthermore, efforts should be directed toward ensuring that such AI models are sufficiently flexible and capable of adjusting and adapting to the constantly evolving landscape of advanced BC treatment. In this context, the recent emergence of the HER2-ultralow entity as druggable in terms of trastuzumab deruxtecan access [6] serves as an emblematic example, imposing the need to generate specific data regarding the possible dynamic behavior also of this novel category.

Major strengths of this study are (i) the vast sample size, making it, to the best of our knowledge, the largest modeling study on the topic of HER2-low status instability during disease evolution; (ii) the twofold development of models serves a dual objective: on one side, a focus on explainability and the unraveling of biological mechanisms; on the other, the pursuit of high performance; and (iii) the reduction to essential variables through Boruta feature selection, with the aim of pinpointing the attention toward less, but more informative features. Some limitations should be acknowledged as well: (i) the model builds on a retrospective clinical platform, inherently subjected to the risk of selection, information, and confusion bias; (ii) the model has been developed and trained to set the sights on sensitivity while slightly sacrificing the overall accuracy. This choice was based on the clear and deliberate intent of maximizing the likelihood of capturing HER2-low phenotype acquisition, therefore judging the relatively low positive predictive value as acceptable; (iii) this study lacks external validation, raising concerns about the generalizability of the findings. While a simpler pipeline might have delivered better performances, we decided to prioritize reliability

over optimizing metrics. Additionally, we used resampling rather than an external cohort to cross-validate our framework. At the same time, we ensured the development of the best practices to avoid overfitting, thereby safeguarding the generalization capabilities of the model itself. Finally, it should be noted that optimizing the model by incorporating additional capabilities, particularly the ability to predict HER2-low gain in patients with HER2-0 status in both the primary tumor and the first metastatic sample, would be highly beneficial. Future efforts will undoubtedly focus on achieving these objectives.

In conclusion, we believe that this work represents the first proof of concept for applying a machine learning model to predict a highly relevant clinical phenomenon in the field of modern oncology: the acquisition of a druggable target. By starting with an explainable model and subsequently integrating it within an ensemble approach, we are able to enhance performance while maintaining transparency, explainability, and intelligibility.

FUNDING

This work was supported by the Italian Ministry of Health - Ricerca Corrente to VG [grant number L03P11].

DATA AVAILABILITY

Code availability: https://github.com/antoniocollesei/HER2-conversion-BRCAmet (RRID:SCR_002630). Data supporting the findings of this study are not publicly available due to privacy or ethical restrictions, but are accessible upon reasonable request. We encourage investigators interested in data access and collaboration to contact the corresponding author.

DISCLOSURE

FM reports, outside the submitted work, the following: personal fees from Roche, Novartis, Gilead, Menarini, Seagen/Pfizer, MSD, AstraZeneca, Daiichi-Sankyo. CV reports, outside the submitted work, the following: consultancy/advisory board for Novartis, Pfizer, Eli Lilly, Menarini, and Daiichi Sankyo; honoraria as a speaker from Novartis, Istituto Gentili, Accademia di Medicina, Eli Lilly; research grants from Roche (to the institution). TG reports, outside the submitted work, the following: personal fees from Eli lilly, iqvia, genetic. FG reports, outside the submitted work, the following: Travel support: Eli Lilly, Gilead, Novartis, Honoraria for lectures: AstraZeneca, Eli Lilly, Gilead G.Griguolo reports, outside the submitted work, the following: personal fees for consultancy/advisory role from Gilead, Seagen, Menarini; honoraria as a speaker from Eli Lilly, Novartis, MSD; travel support from Gilead. AV reports, outside the submitted work, the following: speaker honoraria from Roche and Lilly. FP reports, outside the submitted work, the following: consultancy/advisory board for Daichii SankyoNovartis, Pfizer, Roche and Daiichi Sankyo; honoraria as a speaker from Novartis, MSD, Pfizery. DM reports, outside the submitted work, the following: travel grants: Eli Lilly. MF reports, outside the submitted work, the following: Amgen, Astellas, Astra Zeneca, BMS, Diapath, Eli Lilly, GSK, Incyte, IQvia, Janssen Pharma, MSD, Novartis, Pierre Fabre, Pfizer, Sanofi, Roche. APDT reports, outside the submitted work, the following: MVD reports, outside the submitted work, the following: personal fees for consultancy/advisory role from: Eli Lilly, Pfizer, Novartis, Seagen, Gilead, MSD, Exact Sciences, AstraZeneca, Roche, Daiichi Sankyo, Roche. VG reports, outside the submitted work, the following: personal fees for advisory board membership for AstraZeneca, Daiichi Sankyo, Eisai, Eli Lilly, Exact Sciences, Gilead, Merck Serono, MSD, Novartis, Pfizer, Olema Oncology, Pierre Fabre; personal fees as an invited speaker for AstraZeneca, Daiichi Sankyo, Eli Lilly, Exact Sciences, Gilead, GSK, Novartis, Roche, and Zentiva; personal fees for expert testimony for Eli Lilly. The other authors have declared no conflicts of interest.

REFERENCES

1. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin*. 2018;68:394-424.
2. Gennari A, André F, Barrios CH, et al. ESMO Clinical Practice Guideline for the diagnosis, staging and treatment of patients with metastatic breast cancer. *Ann Oncol*. 2021;32:1475-1495.
3. Loibl S, André F, Bachelot T, et al. Early breast cancer: ESMO Clinical Practice Guideline for diagnosis, treatment and follow-up. *Ann Oncol*. 2023;35:159-182.
4. Tarantino P, Viale G, Press MF, et al. ESMO Expert Consensus Statements (ECS) on the definition, diagnosis, and management of HER2-low breast cancer. *Ann Oncol*. 2023;34:645-659.
5. Modi S, Jacot W, Yamashita T, et al. Trastuzumab deruxtecan in previously treated HER2-low advanced breast cancer. *N Engl J Med*. 2022;387:9-20.
6. Curigliano G, Hu X, Dent RA, et al. Trastuzumab deruxtecan (T-DXd) vs physician's choice of chemotherapy (TPC) in patients (pts) with hormone receptor-positive (HR+), human epidermal growth factor receptor 2 (HER2)-low or HER2-ultralow metastatic breast cancer (mBC) with prior endocrine therapy (ET): primary results from DESTINY-Breast06 (DB-06). *J Clin Oncol*. 2024;42:LBA1000-0.
7. Dieci MV, Barbieri E, Piacentini F, et al. Discordance in receptor status between primary and recurrent breast cancer has a prognostic impact: a single-institution analysis. *Ann Oncol*. 2013;24:101-108.
8. Guarneri V, Giovannelli S, Ficarra G, et al. Comparison of HER-2 and hormone receptor expression in primary breast cancers and asynchronous paired metastases: impact on patient management. *Oncologist*. 2008;13:838-844.
9. Grinda T, Joyon N, Lusque A, et al. Phenotypic discordance between primary and metastatic breast cancer in the large-scale real-life multicenter French ESME cohort. *NPJ Breast Cancer*. 2021;7:41.
10. Miglietta F, Dieci MV, Griguolo G, Guarneri V. Neoadjuvant approach as a platform for treatment personalization: focus on HER2-positive and triple-negative breast cancer. *Cancer Treat Rev*. 2021;98:102222.
11. Tarantino P, Gandini S, Nicolò E, et al. Evolution of low HER2 expression between early and advanced-stage breast cancer. *Eur J Cancer*. 2022;163:35-43.
12. Hawkins DM. The problem of overfitting. *J Chem Inf Comput Sci*. 2004;44(1):1-12.
13. Little RJ. A test of missing completely at random for multivariate data with missing values. *J Am Stat Assoc*. 1988;83(404):1198-1202.
14. Abdi H, Valentin D. Multiple correspondence analysis. *Encycl Meas Stat*. 2007;2(4):651-657.
15. Kursa MB, Rudnicki WR. Feature selection with the Boruta package. *J Stat Softw*. 2010;36:1-13.

16. Menardi G, Torelli N. Training and assessing classification rules with imbalanced data. *Data Min Knowl Discov*. 2014;28:92-122.
17. Chen T, He T, Benesty M, et al. XGBoost: eXtreme gradient boosting. R package version 04-2. The 22nd ACM SIGKDD. *International Conference*. 2015;1(4):1-4.
18. Merrick L, Taly A. The explanation game: explaining machine learning models using Shapley values. In: Proceedings of the Fourth Machine Learning and Knowledge Extraction: 4th IFIP TC 5, TC 12, WG 8.4, WG 8.9, WG 12.9 International Cross-Domain Conference, CD-MAKE 2020*, Dublin, Ireland, August 25—28, 2020*. Springer; 2020. p. 17-38.
19. Naimi AI, Balzer LB. Stacked generalization: an introduction to super learning. *Eur J Epidemiol*. 2018;33:459-464.
20. Schettini F, Prat A. Dissecting the biological heterogeneity of HER2-positive breast cancer. *Breast*. 2021;59:339-350.
21. Lin M, Luo T, Jin Y, et al. HER2-low heterogeneity between primary and p aired recurrent/metastatic breast cancer: implications in treatment and prognosis. *Cancer*. 2024;130:851-862.
22. Almstedt K, Krauthauser L, Kappenberg F, et al. Discordance of HER2-low between primary tumors and matched distant metastases in breast cancer. *Cancers*. 2023;15:1413.