Uncertainty-Aware Logistic Regression with Gray-Zone Refinement for Predicting Response to Neoadjuvant Chemotherapy in Breast Cancer

Aixa X. Torres Fuertes* AIXA.TORRES@UTEC.EDU.PE
Department of Bioengineering and Chemical Engineering, Universidad de Ingeniería y Tecnología - UTEC, Peru

Fátima R. Jara Cuya* FATIMA.JARA@UTEC.EDU.PE
Department of Bioengineering and Chemical Engineering, Universidad de Ingeniería y Tecnología - UTEC, Peru

Rodrigo Romero Tello * RODRIGO.ROMERO@UTEC.EDU.PE
Department of Bioengineering and Chemical Engineering, Universidad de Ingeniería y Tecnología - UTEC, Peru

Jesús A. Sullón Silva* JESUS.SULLON@UTEC.EDU.PE
Department of Electrical and Mechatronics Engineering, Universidad de Ingeniería y Tecnología - UTEC, Peru

Ariana M. Villegas Suarez* AVILLEGAS@UTEC.EDU.PE
Department of Computer Science, Universidad de Ingeniería y Tecnología - UTEC, Peru

Abstract

Predicting response to neoadjuvant chemotherapy (NAC) in breast cancer remains a clinical challenge. We developed a machine learning framework combining bibliographically-weighted Elastic Net for dimensionality reduction with regularized Logistic Regression (LR) as the primary model, and a selective escalation strategy using a multilayer perceptron (MLP) for ambiguous predictions. From GSE205568 (n=2551), 730 robust genes were selected. LR achieved strong performance (nested-CV AUCPR = 0.82, ROC-AUC = 0.93), but uncertainty analysis identified a "gray zone" near the decision threshold, concentrating misclassifications. Routing these cases to an MLP and aggregating outputs via stacking with isotonic recalibration improved gray-zone AUCPR by +0.24 and yielded perfect calibration (ECE ≈ 0). External validation on GSE25065 (n=198) showed that while discrimination transferred (ROC-AUC = 0.94, AUCPR = 0.76), recalibration and local threshold adjustment were required to recover clinically useful performance (F1 = 0.74, Recall = 0.95) (de Hond et al., 2023). These findings support the use of LR as a reliable baseline, augmented by explicit uncertainty detection and selective complexity to improve robustness in clinical prediction.

Keywords: Breast cancer, neoadjuvant chemotherapy, machine learning, logistic regression, uncertainty quantification, transcriptomics

Data and Code Availability The data used in this study are publicly available in the Gene Expression Omnibus (GEO) database under accession ID **GSE205568** and ID **GSE25065** from the National Center for Biotechnology Information (NCBI), accessible at: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi.

The code used for data preprocessing, analysis and result generation is deposited in a GitHub repository, available at: https://anonymous.4open.science/r/chemo-response-prediction-5785/.

Institutional Review Board (IRB) This study was conducted using publicly available data and did not involve direct participation or intervention with human or animal subjects. Therefore, approval from an Institutional Review Board (IRB) was not required.

1. Introduction

Breast cancer accounts for 26.1% of cancer diagnoses in women, making it the most frequent type of cancer in this group (PAHO, 2025). In 2022, the worldwide incidence reached 105.4 and mortality 38.9 per 100,000 women (WHO, 2025). At the regional level in the Americas, during the same year, it was reported as the leading cause of cancer-related deaths

* These authors contributed equally

among women, representing 15.7%, with more than 200,000 new cases recorded (PAHO, 2025). Peru has contributed to this trend, showing a 3.97% increase in mortality between 2013 and 2022 (Terrel-Poccomo et al., 2025). These metrics highlight the urgent need to continue efforts aimed at reducing mortality. It is well established that early action against cancer is crucial. In this context, neoadjuvant chemotherapy (NAC) is one of the first treatments administered. Its goal is to reduce tumor size to enable breast-conserving surgery and achieve a pathological complete response (pCR). Treatment response is classified as either pCR or residual disease (RD), representing the absence or presence of invasive cancer cells in the breast and axillary lymph nodes after treatment, respectively (Chica-Parrado et al., 2020). However, not all patients achieve a pCR, and this outcome is associated with poorer survival rates as well as a higher risk of locoregional recurrence and metastasis (Chica-Parrado et al., 2020; Trabulus et al., 2024). Anticipating prognosis in NAC could therefore serve as a valuable tool for making more effective clinical decisions to combat cancer in a timely manner.

Machine learning models have been widely applied for such predictions, particularly due to their ability to extract, analyze, and identify linear and non-linear relationships that can only be revealed when analyzing large-scale omics data (Krasniqi et al., 2025). Although several projects have generated gene expression data linked to neoadjuvant chemotherapy response, these datasets are often not fully compatible, which complicates cross-study comparisons and limits their utility for training robust machine learning models. Batch effects and technical heterogeneity across large-scale omics studies remain a major barrier to reproducibility, further emphasizing the need for strategies that enable the effective integration of multiple datasets to maximize their value for predictive modeling (Yu et al., 2024).

In that framework, this work presents a predictive framework for neoadjuvant chemotherapy response in breast cancer that integrates bibliographically weighted gene selection with regularized logistic regression as an interpretable baseline. The approach incorporates explicit uncertainty detection through a "gray zone," where ambiguous cases are routed to a multilayer perceptron, thereby improving local discrimination. A meta-learning strategy with isotonic recalibration ensures reliable calibration and a balanced trade-off between precision and recall. External validation confirms the need for recalibration and threshold adjustment to achieve clinically useful performance.

2. Data Selection

The models were built using gene expression microarray data and clinical information. For the training and testing phases, data from the GSE205568 project were used. This data set corresponds to a curated and standardized collection of breast cancer gene expression studies. The curation process included the integration of information from the treatment regimen, clinic pathological variables, clinical outcome data, and gene expression profiles (GEO, 2024). For the validation phase, data from the GSE25065 project were used, which includes gene expression profiles along with clinical and outcome information from breast cancer patients treated with neoadjuvant chemotherapy (GEO, 2011).

3. Data Preprocessing

The GSE205568 dataset was constructed by integrating 28 GEO datasets, which were selected and merged based on overlapping genes. Prior to analysis, the merged expression matrix was subjected to a standardized preprocessing pipeline. Low-variance genes were removed through variance filtering to reduce noise, and expression values were log2-transformed. Subsequently, quantile normalization was applied to minimize batch effects across studies (Vinu Jose et al., 2025).

Additional preprocessing steps included the exclusion of rows with missing values and the selection of patients who had received neoadjuvant chemotherapy, based on clinical metadata. To ensure consistent gene-level annotation, identifiers were converted from NCBI_IDs to official HUGO gene symbols. Clinical outcome variables related to NAC were dichotomized: patients achieving pathological complete response (pCR) were assigned a value of 1, while those with non-complete pathological response were assigned a value of 0. The results of the preprocessing step are summarized in Table 1.

For external validation, the GSE25065 dataset was employed. As this dataset contains probe-level measurements derived from Affymetrix arrays, probes corresponding only to the final gene set (resulting from dimensionality reduction applied to the training data) were retained. Probe identifiers were mapped to HUGO gene symbols to ensure compatibility with

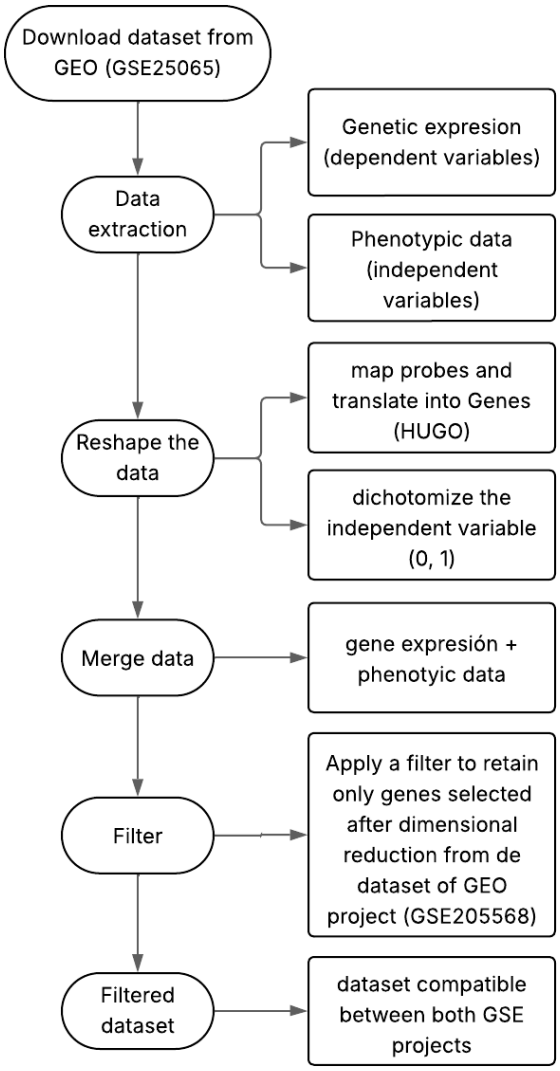

Figure 1: Workflow for harmonizing and integrating datasets from multiple projects of GEO

the processed GSE205568 dataset. Furthermore, clinical response data were used to dichotomize treatment outcomes in the validation cohort, where pCR was labeled 1 and residual disease (RD) was labeled 0. The protocol designed to harmonize both datasets used to make both datasets compatible is explained in Figure 1. This procedure can be used to make GEO datasets compatible, allowing them to serve as a validation cohort for machine learning models or to expand the training and testing datasets for model development.

4. Dimensionality Reduction

4.1. Elastic Net – Feature Selection

An Elastic Net with a stability selection scheme was applied (60 iterations with stratified resampling of rows and columns) using the SAGA solver (Algamal and Lee, 2015). A distinctive aspect of this procedure was the incorporation of a priority weight derived from evidence available in the scientific literature (National Library of Medicine, 2025). Specifically, a priority matrix was constructed from a file containing the number of PubMed hits associated with each gene; this file was computed *once* without labels and reused unchanged across folds to avoid leakage.

Publication counts were transformed into a logarithmic score (priority_log), standardized across genes (z-score), and mapped to a column-wise scaling factor applied *after* standardization of X:

$$z_i = \frac{\log(h_i) - \mu}{\sigma}, \qquad w_i = \mathrm{clip}(\exp(\gamma z_i),\, 1.0,\, 3.0)\,.$$

with slope $\gamma = 0.5$. The resulting adjusted value acted as a differential penalization, effectively emulating `penalty.factor` in `glmnet` through column rescaling prior to model fitting and used only during the stability-selection stage (the baseline classifier was re-fit without priority scaling).

Features were retained if selected in at least 60% of resamples (threshold 0.60), yielding a robust subset of 730 genes (Table 1). As a robustness check, we varied $\gamma \in \{0, 0.25, 0.5, 0.75, 1.0\}$ (and clipping limits in a secondary check) under the same nested-CV protocol, observing stable outer-fold performance.

Table 1: Dataset features before and after the preprocessing phase.

GSE project	Samples	Genes	Labels	Distribution
Before preprocessing				
GSE205568	3736	9185	pCR/npCR	–
GSE25065	198	22283	pCR/RD	–
After preprocessing				
GSE205568	2551	730	1/0	635/1916
GSE25065	198	730	1/0	42/156

pCR: pathological complete response; npCR: non–pathological complete response; RD: residual disease.

5. Mathematical Models Employed

Several supervised learning approaches were implemented for the binary classification of gene expression profiles. The models included Support Vector Machines (SVM) with linear kernel, Random Forest, XGBoost, and Naive Bayes (Gaussian and Bernoulli variants). All models were trained within a homogeneous preprocessing pipeline to ensure methodological consistency.

5.1. Logistic Regression (LR)

The primary focus of this study, however, was the Logistic Regression (LR) model with regularization. Logistic Regression estimates the probability of class membership through the sigmoid function:

$$P(y = 1|X) = \frac{1}{1 + e^{-z}} \tag{1}$$

$$z = \beta_0 + \sum_{j=1}^{p} \beta_j X_j \tag{2}$$

where the parameters β are estimated by maximizing the penalized likelihood. In this study, an L2 (ridge) regularization term was added:

$$\ell(\beta) = -\sum_{i=1}^{n} \Big[y_i \log P(y_i \mid X_i) + (1 - y_i) \log \big(1 - P(y_i \mid X_i)\big) \Big] + \lambda \|\beta\|_2^2. \tag{3}$$

This formulation is particularly well-suited for high-dimensional data ($p \gg n$), such as microarray experiments, since it controls overfitting while preserving predictive performance (Mahmood and Kadir, 2025; Liao and Chin, 2007).

Logistic Regression was prioritized in this study due to several methodological advantages: (i) it provides interpretable coefficients directly linked to gene-level effects, (ii) it scales efficiently to thousands of predictors when combined with regularization, (iii) it is robust under class imbalance when used with appropriate evaluation metrics, and (iv) it integrates naturally with prior feature selection steps, enabling the model to capture robust linear patterns from the reduced set of 730 genes.

5.2. Multi-Layer Perceptron (MLP)

The second model considered was a Multi-Layer Perceptron (MLP), implemented as a non-linear classifier. This model was specifically applied to the cases located in the *gray zone* defined by Logistic Regression, with the aim of refining predictions in scenarios of higher uncertainty. The architecture included an input layer for gene expression profiles, one or more hidden layers with non-linear activation functions (ReLU), and a sigmoid output layer for binary classification. Training was carried out using backpropagation with an adaptive optimizer.

Unlike Logistic Regression, the MLP can capture complex and non-linear relationships among genes, although it requires more careful hyperparameter tuning and is less interpretable. The application of the MLP in the gray zone is inspired by recent advances in selective classification or reject option in neural networks, which improve reliability by abstaining from predicting in ambiguous cases (Hasan et al.; Hendrickx et al., 2024). Unlike traditional approaches, where uncertain cases remain unclassified, our strategy leverages the MLP as a second-stage model to resolve ambiguity, thereby improving both coverage and predictive accuracy.

6. Results

6.1. Dataset summary and preprocessing

After curation and feature selection, the primary training dataset (GSE205568) comprised $n = 2551$ samples and a reduced feature set of $p = 730$ genes following Elastic Net stability selection with bibliographic prioritization. The external validation cohort (GSE25065) comprised $n = 198$ samples and was reduced to the same $p = 730$ features by mapping array probes to the selected gene set. Clinical outcomes were dichotomized as pathological complete response (pCR = 1) versus residual disease (RD / npCR = 0). All subsequent modeling was performed inside a homogeneous pipeline (variance filtering, log2-transformation, quantile normalization, feature selection) and evaluated using nested cross-validation where indicated.

6.2. Model comparison (nested cross-validation)

Table 2 summarizes the principal aggregated metrics obtained with a stratified nested CV protocol (outer

k=5, inner k=3). All models shared an identical preprocessing-and-training pipeline in which hyperparameters were tuned via GridSearchCV optimizing AUCPR; preprocessing steps were fit exclusively on the training split of each fold (no leakage), with fixed random seeds and class-weight balancing when applicable. Given the class imbalance, AUCPR was the primary optimization metric, and ROC-AUC, accuracy, F1, precision, and recall are reported for completeness. A regularized logistic regression (L2) provided the best trade-off between discrimination, calibration, and interpretability and was therefore selected as the principal baseline model.

6.3. Out-of-fold performance and calibration

Out-of-fold (OOF) predictions from the selected Logistic Regression model yielded: ROC-AUC = 0.9139, AUCPR = 0.7982, Brier score = 0.1190 and Expected Calibration Error (ECE) = 0.1022. The best F1 threshold on OOF predictions was $t_{\text{OOF}} = 0.6292$; a median operational threshold from nested folds was $t_{\text{median}} = 0.578$. These calibration metrics motivated a deeper analysis of uncertainty around the decision threshold (see next subsection).

6.4. Empirical "gray zone" of predictive uncertainty

Analysis of OOF prediction margins revealed a strong association between small distance to the decision threshold and increased classification error: Spearman correlation between absolute margin and error indicator was $\rho \approx -0.359$ (p-value $< 10^{-77}$). Binning of OOF probabilities identified a concentrated interval of increased error rates in the approximate probability range [0.55, 0.70]. An asymmetric empirical gray zone defined from OOF statistics was set to [0.553, 0.727]; approximately 2.7% of samples fell within this region. Because errors concentrate in this narrow band, we explored a selective strategy in which ambiguous cases are routed to a more expressive model.

We report absolute counts to contextualize impact given the small gray-zone fraction. In the development OOF set, the gray zone contained $N_{\text{gray}} = 69$ cases ($\approx 2.7\%$ of 2551). Within this subset, the LR baseline made 25 errors (FP=20, FN=5). The gray-boosted MLP reduced errors to 19 (FP=13, FN=6; $\Delta = -6$, -24%), and the calibrated stacking to 22 (FP=16, FN=6; $\Delta = -3$, -12%). Thus, the refinement step corrects a meaningful share of ambiguous cases despite the small coverage of the gray zone.

6.5. Selective complexity: gray-boosted MLP and stacking

A hierarchical pipeline was implemented in which a global LR classifier provided full-cohort predictions, while cases falling within the empirically defined gray zone were routed to a gray-boosted MLP. The final predictions for these ambiguous instances were then aggregated through a stacking meta-learner (meta-LR) (Wang et al., 2024) with isotonic recalibration. Out-of-fold evaluation showed that the gray-boosted MLP achieved a ROC-AUC of 0.9205 and an AUCPR of 0.8400. The stacking pipeline with isotonic recalibration obtained a ROC-AUC of 0.9204 and an AUCPR of 0.8124, while also achieving markedly improved calibration, as reflected by a Brier score of 0.0825 and an ECE of 0.0000. At the classification level, stacking yielded an F1 score of 0.7777, with precision and recall values of 0.7759 and 0.7795, respectively. For comparison, the baseline LR model evaluated on the same out-of-fold partition reached an F1 score of 0.7365, precision of 0.7111, and recall of 0.7638. These results indicate that selective escalation of model complexity in the gray zone not only improves discrimination but also achieves superior calibration and more balanced classification performance compared to the baseline alone.

6.5.1. Performance inside the gray zone

Table 3 reports performance restricted to the empirically defined gray-zone subset (OOF estimate).

The stacking pipeline substantially improves discrimination and positive predictive value within the gray zone, indicating that selective allocation of model complexity yields practical gains where the baseline is uncertain. Approximately 2.7% of samples were routed to the specialized branch under OOF estimation.

6.6. Uncertainty quantification and decision-curve analysis

Bootstrap (N=400) uncertainty bands were computed for ROC, PR and calibration curves for both LR and STACKING pipelines. These bands are presented in the figures recommended for inclusion (see Figure 2).

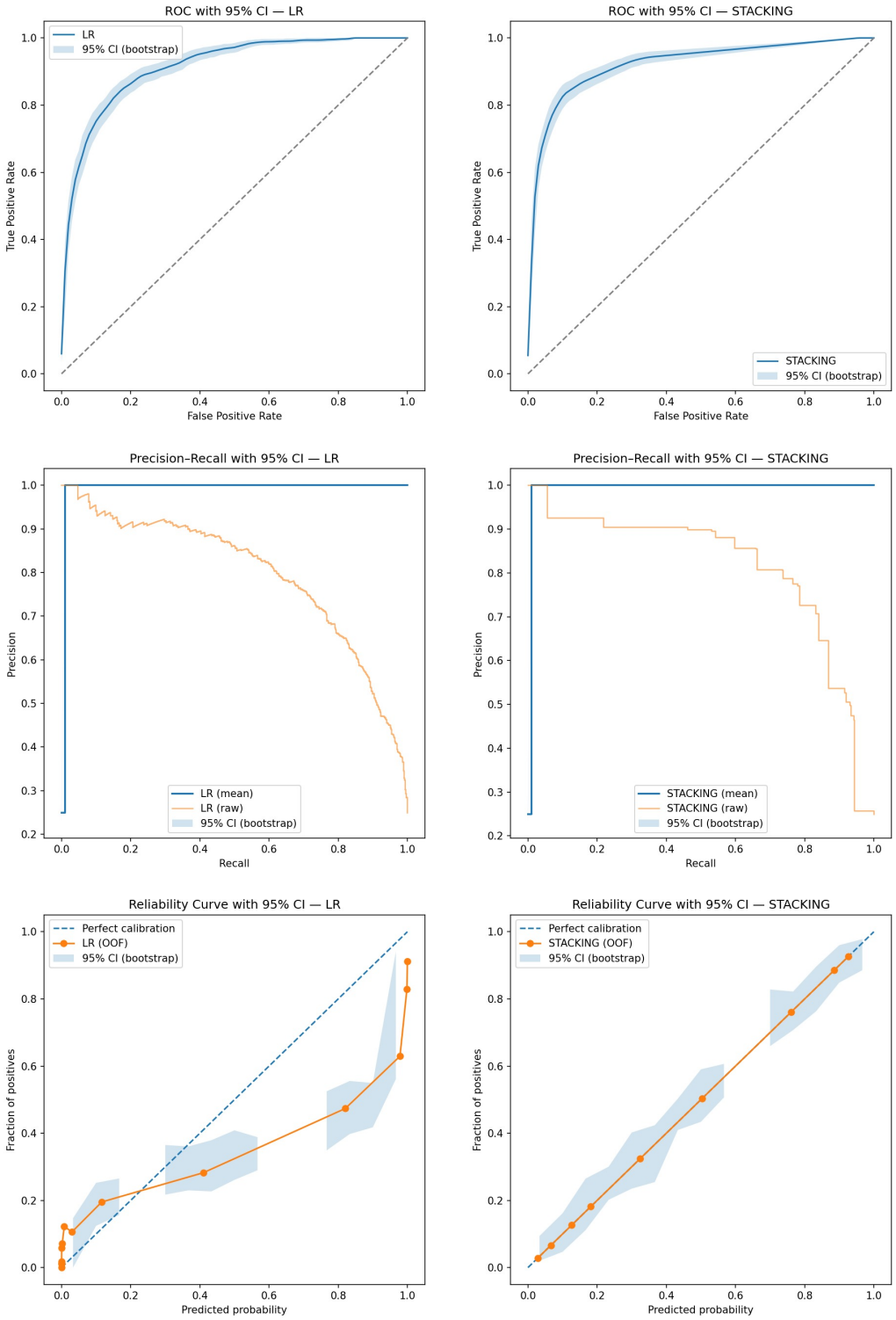

Figure 2: Bootstrap uncertainty bands for ROC, PR and calibration curves for Logistic Regression (left column) and STACKING (right column). Include these figures to document discrimination and calibration uncertainty.

Table 2: Nested cross-validation performance (outer 5-fold averages ± standard deviation). Primary optimization metric: AUCPR.

Model	AUCPR	ROC-AUC	Accuracy	F1	Precision	Recall
Logistic Regression (L2, C=0.1)	0.824 ± 0.034	0.926 ± 0.013	0.882 ± 0.013	0.765 ± 0.022	0.758 ± 0.038	0.775 ± 0.026
SVM (linear)	0.770 ± 0.030	0.900 ± 0.016	0.862 ± 0.010	0.726 ± 0.017	0.717 ± 0.033	0.739 ± 0.036
XGBoost	0.709 ± 0.026	0.859 ± 0.011	0.812 ± 0.019	0.650 ± 0.018	0.611 ± 0.047	0.701 ± 0.049
Random Forest	0.641 ± 0.026	0.827 ± 0.019	0.772 ± 0.035	0.601 ± 0.031	0.541 ± 0.058	0.685 ± 0.053
Gaussian Naive Bayes	0.476 ± 0.018	0.749 ± 0.010	0.694 ± 0.025	0.540 ± 0.020	0.433 ± 0.027	0.721 ± 0.045
Bernoulli Naive Bayes	0.434 ± 0.014	0.734 ± 0.015	0.655 ± 0.029	0.524 ± 0.007	0.401 ± 0.019	0.762 ± 0.062

Table 3: Out-of-fold (OOF) performance restricted to empirical gray-zone samples. Values shown to three decimal places.

Model	AUCPR	Precision	F1
LR (gray zone)	0.443	0.474	0.590
STACKING (gray zone)	0.681	0.600	0.625
Uplift (STACK – LR)	0.238	0.126	0.035

6.7. External validation on GSE25065

External performance was evaluated under two scenarios to illustrate the effect of threshold transfer and recalibration.

Scenario A — Fixed operational threshold We applied the model trained on GSE205568 with its out-of-fold (OOF) operating threshold $t_{\text{OOF}} = 0.6292$ directly to GSE25065. Discrimination was preserved (ROC-AUC = 0.9361, AUCPR = 0.7773), but the fixed threshold produced no positive predictions in the external cohort (Accuracy = 0.7879, F1 = 0.0000, Precision = 0.0000, Recall = 0.0000); the confusion matrix $\begin{bmatrix} 156 & 0 \\ 42 & 0 \end{bmatrix}$ illustrates the collapse. This reflects the non-transferability of decision thresholds across cohorts with different prevalence or measurement scales, not a failure of ranking.

Scenario B — Isotonic recalibration with held-out testing To adapt probabilities and the decision boundary without contaminating evaluation, we used a single stratified split of the full external cohort into a *calibration* subset and an independent *test* subset (StratifiedShuffleSplit, test_size = 0.50, random_state = 42). No model refitting or weight updates were performed at any stage.

Protocol. The trained model parameters remained fixed throughout.

1. Compute raw model probabilities $p = \Pr(\text{response} \mid x)$ on GSE25065.
2. On the *calibration* subset, fit a monotone isotonic regressor $m : [0,1] \to [0,1]$ (with out-of-bounds clipping) to obtain calibrated scores $\tilde{p} = m(p)$.
3. On the *calibration* subset only, select the operating threshold t_{calib} that maximizes F1 along the precision–recall curve of $\tilde{p}$.
4. Freeze $m(\cdot)$ and t_{calib}.
5. Apply the frozen $m(\cdot)$ and t_{calib} to the *held-out test* subset and compute final metrics and the confusion matrix.

No labels or distributional information from the test subset influenced calibration or threshold selection.

Table 4: Calibration and held-out test results.

Split	AUROC	AUPRC	Acc.	F1	P	R	CM
Calibration	0.9423	0.7923	0.9091	0.8085	0.7308	0.9048	$\begin{smallmatrix} 71 & 7 \\ 2 & 19 \end{smallmatrix}$
Test (held-out)	0.9377	0.7605	0.8586	0.7407	0.6061	0.9524	$\begin{smallmatrix} 65 & 13 \\ 1 & 20 \end{smallmatrix}$

Acc.: Accuracy; F1: F1-score; P: Precision; R: Recall; CM: Confusion Matrix.

Results. This shows that isotonic recalibration plus locally optimized thresholding restores clinically useful operating points in the external cohort while strictly avoiding information leakage.

6.8. Model interpretation (SHAP)

After fixing the 730-gene panel, we re-fit the final L2-regularized logistic regression on the full development set using the same pipeline (median imputation + standardization) and computed SHAP values with the linear explainer on the log-odds scale. Global importance was summarized as mean absolute SHAP across samples. Importantly, bibliographic priorities were *not* used at this stage; hence, importances reflect

the model's learned contributions within the selected panel rather than publication frequency.

The highest contributors included *PHIP*, *FMO2*, *PACS1*, *GALNT6* and *REST*, consistent with proliferation/survival, xenobiotic and glycosylation axes, and transcriptional regulation; alongside these, less-discussed genes in this context (e.g., *MSANTD2*, *ABTB2*, *TESMIN*) also emerged, suggesting that literature-informed selection helped focus the search space while the ranking itself is data-driven (Figure 3). For transparency at the ensemble level, we interpreted the stacking meta-learner (logistic regression) over the OOF feature matrix $Z = \{\texttt{oof\_lr}, \texttt{oof\_mlp}, |\texttt{oof\_lr} - t_{\mathrm{LR}}|, \texttt{oof\_mlp} - \texttt{oof\_lr}, \texttt{is\_gray}\}$ using the same linear SHAP framework. The meta-model relied primarily on calibrated base probabilities (`oof_mlp`, `oof_lr`), with smaller contributions from disagreement/uncertainty features, yielding a parsimonious and interpretable combination (Figure 4).[1]

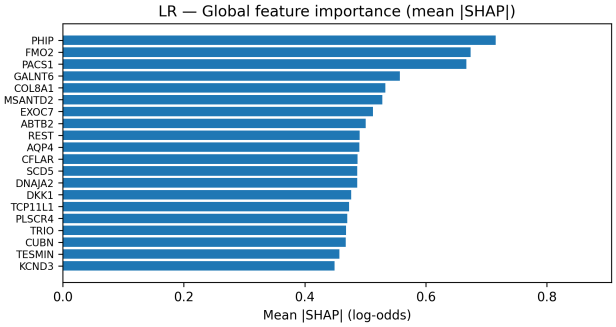

Figure 3: Global gene importance for LR (mean |SHAP|).

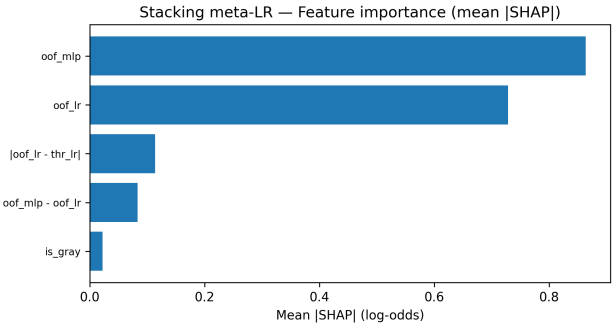

Figure 4: Stacking meta-learner feature importance (mean |SHAP|) over OOF meta-features. The meta-model weights `oof_mlp` and `oof_lr` most strongly, with modest contributions from disagreement/uncertainty terms.

7. Discussion

The results indicate that a bibliographically-weighted Elastic Net feature selection followed by a regularized logistic regression yields strong and stable discrimination for predicting pathological complete response from microarray data (nested-CV AUCPR ≈ 0.82; ROC-AUC ≈ 0.93), but that global performance metrics mask concentrated uncertainty near the decision threshold. An empirical gray zone derived from OOF statistics (approximately 2.7% of samples, interval $[0.553, 0.727]$) accumulated a disproportionate share of classification errors, and routing these ambiguous cases to a gray-boosted multilayer perceptron combined via stacking produced meaningful improvement within that subset (AUCPR uplift $\approx +0.238$) while preserving or slightly improving global calibration. External validation confirmed that discrimination largely transfers to an independent cohort but also demonstrated that an operating threshold selected in the training environment may fail in a new population; isotonic recalibration and local threshold selection were therefore necessary to recover clinically useful recall and precision. Practical limitations include the small absolute size of the gray-zone subgroup (which constrains population-level impact), the increased complexity and operational burden introduced by a stacked pipeline, and the modality-specific nature of the study (microarray preprocessing and probe-to-gene mapping), all of which warrant careful re-evaluation prior to deployment in different assay platforms or clinical settings. A useful comparison between similar computational projects is the twelve-gene signature of Wu et al. (2022), which offers good discrimination (AUC = 0.83) with high clinical feasibility but does not address calibration or cohort transferability. Our framework instead focuses on reliability through uncertainty detection and recalibration, underscoring complementary trade-offs between simplicity for deployment and robustness for safe generalization.

Overall, these findings support a pragmatic framework in which a robust, interpretable linear baseline is retained for the majority of cases, uncertainty is explicitly detected, and selective escalation to higher-

1. Concise biological references for the highlighted genes will be added in the revision.

capacity models is applied only for well-characterized ambiguous instances, with mandatory cohort-specific recalibration before clinical use.

7.1. Gene-level interpretation

After fixing the 730-gene panel, SHAP analysis of the final L2–logistic regression (priorities not used at this stage) showed that the strongest drivers were *PHIP*, *FMO2*, *PACS1*, *GALNT6* and *REST*, mapping to plausible biological axes for NAC response (growth/survival, xenobiotic metabolism, glycosylation/ECM remodeling and transcriptional regulation). In parallel, less-discussed genes in this setting (e.g., *MSANTD2*, *ABTB2*, *TESMIN*) also contributed non-trivially, indicating that while literature-informed selection narrowed the search space, the learned ranking is data-driven rather than a reflection of publication frequency. At the ensemble level, SHAP on the meta-learner confirmed that calibrated base probabilities (`oof_mlp`, `oof_lr`) dominate, with smaller gains from simple disagreement/uncertainty terms; this supports a efficient stacking strategy that remains interpretable while delivering the gray-zone improvements reported above.

8. Conclusion

This study demonstrates that a bibliographically-weighted Elastic Net feature selection followed by a regularized logistic regression yields a reliable and interpretable predictive baseline for neoadjuvant chemotherapy response from curated microarray data, attaining nested-CV AUCPR values near 0.82 and ROC-AUC values near 0.93. While global discrimination is high, a concentrated empirical gray zone around the decision threshold accumulates most classification uncertainty; selectively directing these ambiguous instances to a gray-boosted multilayer perceptron and aggregating predictions via a stacking meta-learner produces measurable improvements in precision and AUCPR within that subset, with a reported AUCPR uplift of approximately 0.238 inside the gray zone and modest improvements in overall F1 and calibration. Crucially, external validation reveals that although discrimination can generalize across cohorts, clinically useful binary decisions do not transfer without adjustment: isotonic recalibration and locally chosen operating thresholds were necessary to recover high recall and acceptable precision in the independent cohort. These results suggest that while the baseline logistic regression is usable, its raw probabilities are unreliable at extremes and near the decision boundary, underscoring the need for explicit calibration. In contrast, stacking with isotonic calibration achieves near-perfect reliability (Brier and ECE approaching zero) and, more importantly, corrects misclassifications precisely in the gray zone of clinical uncertainty—arguably the most critical region for safe decision-making. For translational deployment, we therefore recommend adopting a conservative, interpretable baseline model, implementing explicit uncertainty detection with selective escalation of model complexity, and institutionalizing cohort-specific probability recalibration and threshold selection as part of the deployment protocol. Future work should prospectively evaluate this framework in independent, prospectively collected cohorts and examine its applicability to other molecular modalities and clinical contexts to firmly establish generalizability and clinical utility.

References

Zakariya Yahya Algamal and Muhammad Hisyam Lee. Regularized logistic regression with adjusted adaptive elastic net for gene selection in high dimensional cancer classification. *Computers in Biology and Medicine*, 67:136–145, 2015. ISSN 0010-4825. doi: 10.1016/j.compbiomed.2015.10.008.

María Rosario Chica-Parrado, Ana Godoy-Ortiz, Begoña Jiménez, Nuria Ribelles, Alba Emilio Barragán, Isabel, et al. Resistance to neoadjuvant treatment in breast cancer: Clinicopathological and molecular predictors. *Cancers*, 2020. doi: 10.3390/cancers12082012.

Anne A. H. de Hond, Ilse M. J. Kant, Mattia Fornasa, Giovanni Cinà, Paul W. G. Elbers, Patrick J. Thoral, M. Sesmu Arbous, and Ewout W. Steyerberg. Predicting readmission or death after discharge from the ICU: External validation and retraining of a machine learning model. *Critical Care Medicine*, 51(2):291–300, February 2023. doi: 10.1097/CCM.0000000000005758.

GEO. Validation cohort for genomic predictor of response and survival following neoadjuvant taxane-anthracycline chemotherapy in breast cance. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=gse25065, 2011.

GEO. A pooled treatment-curated breast cancer gene-expression dataset. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE205568, 2024.

Mehedi Hasan, Moloud Abdar, Abbas Khosravi, Uwe Aickelin, Pietro Lio', Ibrahim Hossain, Ashikur Rahman, and Saeid Nahavandi. Survey on leveraging uncertainty estimation towards trustworthy deep neural networks: The case of reject option and post-training processing. doi: 10.48550/arXiv.2304.04906. URL https://doi.org/10.48550/arXiv.2304.04906.

K. Hendrickx, L. Perini, and D. Van der Plas et al. Machine learning with a reject option: a survey. *Machine Learning*, 2024. doi: 10.1007/s10994-024-06534-x. URL https://doi.org/10.1007/s10994-024-06534-x.

Eriseld Krasniqi et al. Multimodal deep learning for predicting neoadjuvant treatment outcomes in breast cancer: a systematic review. *Biology Direct*, 2025. doi: 10.1186/s13062-025-00661-8.

James G. Liao and Kui V. Chin. Logistic regression for disease classification using microarray data: model selection in a large p and small n case. *Bioinformatics*, 23(15):1945–1951, August 2007. doi: 10.1093/bioinformatics/btm287.

Nozad H. Mahmood and Dler H. Kadir. An elastic net approach to logistic regression for genetic selection in high-dimensional brain cancer data. *Cihan University-Erbil Scientific Journal (CUESJ)*, 9(1):14–23, 2025. doi: 10.24086/cuesj.v9n1y2025.pp14-23.

National Library of Medicine. Pubmed. https://pubmed.ncbi.nlm.nih.gov/, 2025. Accessed September 2025.

PAHO. Cancer. https://www.paho.org/en/topics/cancer, 2025. Accessed September 2025.

Melanie Kiara Terrel-Poccomo, Grecia Santillán-Romero, Carlos Quispe-Vicuña, Jorge Ybaseta-Medina, J. Smith Torres-Roman, et al. Trends in breast cancer mortality in peru and its geographical areas from 2013 to 2022 and prediction until 2027. *BMC Cancer*, 2025. doi: 10.1186/s12885-025-13872-z.

Fadime Didem Can Trabulus et al. Predictors of recurrence in breast cancer patients with pathological partial response. *Revista da Associação Médica Brasileira*, 2024. doi: 10.1590/1806-9282.20231215.

Laurence Buisseret Francoise Rothe Christos Sotiriou Vinu Jose, David Venet et al. Pooled breast cancer gene expression dataset with detailed treatment regimen characterization. https://osf.io/preprints/osf/jbwe4_v1, 2025.

M. Wang, Y. Qian, Y. Yang, H. Chen, and W. F. Rao. Improved stacking ensemble learning based on feature selection to accurately predict warfarin dose. *Frontiers in Cardiovascular Medicine*, 10:1320938, January 2024. doi: 10.3389/fcvm.2023.1320938.

WHO. Cancer today: Globocan heatmap. https://gco.iarc.fr/today/en/dataviz/maps-heatmap?mode=population&cancers=20, 2025.

Jia Wu, Wenfang Zhang, Shudong Chen, Yuanyuan Liu, Rui Bi, Hongnan Mo, Guosheng Ren, Zhaohui Ruan, Xiang Wang, Shusen Wang, Zhi-Ming Shao, and Yujie Sun. A novel twelve-gene signature to predict neoadjuvant chemotherapy response and prognosis in breast cancer. *Frontiers in Oncology*, 12:985001, 2022. doi: 10.3389/fonc.2022.985001. URL https://doi.org/10.3389/fonc.2022.985001.

Ying Yu, Yixin Chen, Xianglin Yuan, Yixuan Zhang, Xiaodong Fang, Yu Hou, Yaning Yang, Fang Bai, Xiaoxue Cui, Huanhuan Cui, Shuangshuang Zhang, Bo Wang, Shuzhe Ding, Yifan Feng, Zhimin Qi, Xiang Zhou, Xuegong Zhang, Cheng Li, et al. Comprehensive evaluation of batch effect correction methods for integrative omics data analysis. *Genome Biology*, 2024. doi: 10.1186/s13059-024-03401-9. URL https://genomebiology.biomedcentral.com/articles/10.1186/s13059-024-03401-9.