Supplementary Information

Pathway-based machine learning for breast cancer risk stratification: an interpretable framework validated in two independent cohorts

[Author information removed for double-blind review]

Contents

1 Supplementary Methods

1.1 Cohort inclusion and exclusion

For TCGA-BRCA, patients were included if they had: (1) invasive breast adenocarcinoma histology; (2) available tumor RNA-seq expression data; (3) available overall survival annotation; and (4) PAM50 subtype assignment. Patients with missing RNA-seq data or incomplete survival data were excluded.

For GSE96058/SCAN-B, all patients with available RNA-seq expression data, overall survival data, and PAM50 subtype assignment were included per the published SCAN-B cohort [1, 2]. Patients censored before 60 months were kept in survival analyses but removed from binary classification analyses ($n = 103$ removed for binary classification).

1.2 Pathway gene set curation

Genes for each pathway were selected based on three criteria: (1) a well-established functional role in the pathway, supported by the MSigDB Hallmark gene set collection [3]; (2) consistent detection across both cohort RNA-seq platforms (RSEM for TCGA; FPKM for SCAN-B); and (3) prior use as a prognostic marker in the breast cancer literature [4, 5]. Gene sets were finalized before any analysis and were not changed based on results. *PGAP3* was noted before analysis as a potential missingness risk in SCAN-B given platform differences. Its absence from GSE96058 was handled by removing it from the HER2 signaling score in that cohort only, reducing the gene count for that pathway from 5 to 4.

1.3 Hyperparameter rationale

Hyperparameters were fixed before any analysis based on the following reasoning and were not changed based on observed results:

- **Elastic net:** An l_1-ratio of 0.5 gives balanced L1 and L2 penalty, appropriate for the correlated pathway features used here. $C = 1.0$ is the scikit-learn default and was kept without tuning.
- **Random forest:** 500 trees is well established as sufficient for stabilizing out-of-bag error in cohorts of this size. Maximum depth of 8 limits overfitting on the 213-patient TCGA training set.

- **Gradient boosting:** Learning rate of 0.03 and maximum depth of 3 follow standard guidance for regularized gradient boosting on small feature sets.

1.4 Cross-validation procedure

Stratified five-fold cross-validation on GSE96058 was carried out as follows. The dataset was split into five folds stratified by five-year outcome status, so that event rates were approximately equal across folds. For each fold: (1) the StandardScaler was fitted on the four training folds and applied to the held-out fold; (2) the classifier was trained on the four training folds; (3) predicted probabilities were generated for the held-out fold. Out-of-fold predictions were combined across all five folds to compute AUC-ROC. The same five-fold assignment was used for survival analysis, with stratification by event status and follow-up tertile.

1.5 ssGSEA implementation

The ssGSEA algorithm was implemented following Barbie et al. [6]. For each sample, genes were ranked by expression value in descending order. For each pathway gene set, the enrichment score was computed as the normalised sum of the cumulative rank fraction for genes inside the set minus the fraction for genes outside the set, weighted by the absolute difference of the two cumulative distribution functions. Scores were normalised across samples to unit variance within each pathway. The resulting per-sample, per-pathway scores were used as inputs to the same cross-validation pipeline as the mean z-score features.

1.6 Net reclassification index

The category-free (continuous) net reclassification index (NRI) was computed by comparing predicted probabilities from the combined model against those from the subtype-only model for the same patients. The NRI was split into event and non-event components. Bootstrap confidence intervals were computed with 1,000 iterations of stratified resampling (random seed 42).

1.7 Decision curve analysis

Net benefit was computed at threshold probabilities from 0.01 to 0.99 in increments of 0.01, using the formula:

$$\text{Net benefit} = \frac{\text{TP}}{n} - \frac{\text{FP}}{n} \cdot \frac{p_t}{1 - p_t}$$

where TP and FP are true and false positives at threshold p_t and n is the total sample size. Reference strategies were treat-all and treat-none (net benefit = 0).

1.8 Software and computational environment

All analyses were run in Python 3.10. Key packages used: scikit-learn 1.3.0 (classifiers, cross-validation, StandardScaler); lifelines 0.27.8 (Cox model, Kaplan-Meier); shap 0.42.1 (SHAP values via TreeExplainer); statsmodels 0.14.0 (Schoenfeld residuals); scipy 1.11.0 (DeLong test); matplotlib 3.7.1 and seaborn 0.12.2 (figures). Total runtime for the full cross-validation pipeline was approximately 45 minutes on a standard laptop.

2 Supplementary Results

2.1 Hazard ratio estimates from the combined Cox model

Table S1 shows hazard ratio estimates with 95% confidence intervals from the combined Cox proportional hazards model fitted on the full GSE96058 cohort (not cross-validated). A hazard ratio greater than 1 indicates increased hazard per unit increase in the standardised feature.

Table S1: Supplementary Table S1. Hazard ratios from the combined Cox model (full GSE96058 cohort). HR = hazard ratio; CI = 95% confidence interval. All features are standardised before model fitting.

Feature	**HR**	**95% CI**	p
Clinical features			
Age at diagnosis	1.31	[1.22–1.41]	<0.001
Tumor size	1.18	[1.10–1.27]	<0.001
Node positive	1.62	[1.34–1.96]	<0.001
ER negative	1.41	[1.13–1.76]	0.002
Ki-67 high	1.53	[1.27–1.85]	<0.001
NHG high	1.29	[1.07–1.56]	0.008
Pathway scores			
Proliferation	1.14	[1.03–1.26]	0.011
Estrogen response	0.83	[0.74–0.93]	0.001
Immune infiltration	0.91	[0.82–1.01]	0.072
Apoptosis	0.92	[0.84–1.01]	0.084
EMT	1.08	[0.98–1.19]	0.121
HER2 signaling	1.06	[0.97–1.16]	0.198
Angiogenesis	1.05	[0.96–1.15]	0.307
Prolif./Apoptosis ratio	1.11	[1.01–1.22]	0.031

2.2 Luminal B exploratory analysis

Owing to sample size constraints, formal within-subtype Cox modelling was not performed for Luminal B ($n = 344$), HER2-enriched ($n = 127$), or Basal-like ($n = 249$) subgroups. A descriptive Kaplan-Meier analysis of risk groups within Luminal B (defined by median combined Cox prediction in the full cohort) showed directionally consistent separation (log-rank $p = 0.041$; C-index 0.604), though confidence intervals were much wider. These results are exploratory and should not be interpreted as confirmatory.

2.3 Full AUC comparison including gradient boosting

Table S2 shows the complete AUC results including gradient boosting, which was omitted from Table 3 in the main text for space.

Table S2: Supplementary Table S2. Full binary classification AUC-ROC (stratified five-fold CV, GSE96058). Bootstrap 95% CI from 1,000 iterations.

Feature set	**Elastic net**	**RF**	**GB**
Subtype only (5)	0.613	0.613	0.613
Pathway only (8)	0.641 [0.612–0.670]	0.645 [0.616–0.674]	0.633 [0.604–0.662]
Clinical only (9)	0.855 [0.832–0.878]	0.856 [0.833–0.879]	0.842 [0.818–0.866]
Combined (17)	0.855 [0.832–0.878]	**0.856** [0.833–0.879]	0.827 [0.801–0.853]

Supplementary References

References

[1] Saal, L. H. et al. The Sweden Cancerome Analysis Network – Breast (SCAN-B) initiative. *Genome Med.* **7**, 20 (2015).

[2] Brueffer, C. et al. Clinical value of RNA sequencing-based classifiers for prediction of the five conventional breast cancer biomarkers. *JCO Precis. Oncol.* **2**, 1–18 (2018).

[3] Liberzon, A. et al. The Molecular Signatures Database (MSigDB) Hallmark gene set collection. *Cell Syst.* **1**, 417–425 (2015).

[4] Hanahan, D. & Weinberg, R. A. Hallmarks of cancer: the next generation. *Cell* **144**, 646–674 (2011).

[5] Salgado, R. et al. The evaluation of tumor-infiltrating lymphocytes (TILs) in breast cancer. *Ann. Oncol.* **26**, 259–271 (2015).

[6] Barbie, D. A. et al. Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. *Nature* **462**, 108–112 (2009).