



Machine Learning Classification of Breast Cancer Molecular Subtypes and Clinical Phenotypes from GEO Expression Cohorts, with a Leakage-Controlled External Validation Protocol on GSE1456

Anushka Choudaha, Dr. Elamathi Natarajan*

Department of Bioinformatics, Biotechnika Info Labs, Bengaluru, India.

*Corresponding author: Dr. Elamathi Natarajan, E-mail: elamathi@biotechnika.org

ABSTRACT

Background. Breast cancer comprises biologically distinct intrinsic subtypes with different outcomes. Public GEO cohorts allow reproducible computational classification, but small test sets, heterogeneous labels and information leakage often inflate reported performance. This study classified breast cancer subtypes and clinical phenotypes from expression data and tested whether subtype classification transfers to an independent cohort.

Methods. Four Affymetrix cohorts (GSE20685, GSE45827, GSE2034, GSE7390) were processed through a common pipeline: probe-to-gene annotation, quality control, differential-expression (DEG) feature selection on the training split only, and Logistic Regression, SVM and Random Forest classification (80/20 stratified hold-out), each against its own available label. A model trained on GSE45827 with nested cross-validation (CV) was then tested once on GSE1456 as an independent external cohort.

Results. Hold-out accuracy was 90.9–92.4% for six molecular clusters (GSE20685), 96.2–100% for four subtypes (GSE45827), 70.7–79.3% for bone relapse (GSE2034) and 80.0–90.0% for ER status (GSE7390); in GSE2034 no model clearly exceeded the 75.9% majority-class baseline. On GSE1456 (102 labelled tumours), the SVM selected by nested CV (97.4% internal accuracy) reached only 48.0% accuracy (bootstrap 95% CI 38–58; majority baseline 38.2%). Basal (recall 0.88) and HER2 (0.80) tumours partly transferred; luminal A (0.21) and luminal B (0.30) did not.

Conclusion. Within a cohort, expression-based classification is accurate for subtype and ER labels but not for relapse. Subtype classification did not transfer reliably to an independent cohort: external accuracy (48.0%) was far below the internal estimate (97.4%).

Keywords: Breast cancer; GEO; Differential expression; Feature selection; External validation; PAM50

1. INTRODUCTION

Breast cancer is heterogeneous. Expression profiling resolved intrinsic subtypes [1,2] and the PAM50 assay [3] turned them into a clinical-grade classifier. GEO [4] provides many annotated cohorts, yet merged-cohort studies must handle platform and batch heterogeneity [5], and high-dimensional feature selection performed before cross-validation is a recognised source of optimistic bias [6].

The five GEO cohorts used here have been reused extensively, mostly for prognosis rather than subtype classification (Table 1). The PAM50 framework underlying our labels was introduced by Parker et al. [3].



GSE20685 was generated for microarray-based subtype and outcome analysis [7] and later served for DEG- and pathway-based classification [8], and for training or validating prognostic signatures, including 4-, 5-, 9- and 18-gene panels and aging-, hypoxia- and microenvironment-related models [9–15]. GSE45827 contains subtype-defined tumours [16] and has been re-analysed for DEG, enrichment, PPI and hub-gene discovery, and lncRNA expression [17–19]. GSE2034 [20] and GSE7390 [21] underpin relapse and 76-gene-signature validation work, and were used for stem-cell-derived, phosphatase and PTP4A2 signature or biomarker studies [22–24]. Cross-cohort studies examined merged-dataset survival prediction [5], PAM50-stratified gene modules [25] and correlation-based variable selection with independent validation [6]. GSE1456, a population-based Swedish cohort [26], has been used for BRD4, subpathway, insulin-pathway, RDGN, ME1 and serotonin-receptor signature studies and pan-signature comparisons [27–33].

Table 1. Role of the selected GEO datasets in the cited literature

Dataset	Prior uses (references)
GSE20685	Subtype characterization [7]; DEG/pathway classification [8]; signature training/validation [9–15]
GSE45827	Subtype expression [16]; DEG, PPI, hub genes, lncRNA [17–19]
GSE2034	Relapse and signature work [6,20]; merging [5]; biomarkers/pathways [23,24]; signature comparison [29]
GSE7390	76-gene validation [21]; PAM50 modules [25]; selection/validation [6]; signatures and biomarkers [9,22–24,32]
GSE1456	Cohort [26]; merging [5]; signatures, pathways, predictive models [27–33]

Few of these studies combine multi-cohort integration, fold-internal feature selection, leave-one-cohort-out testing and a strictly untouched external cohort for **subtype classification**. This work targets that gap. We therefore (i) built a uniform DEG-then-classifier pipeline across four cohorts, (ii) reported performance against majority-class baselines with confidence intervals, and (iii) performed an external validation on GSE1456 in which no test-cohort labels were used for feature selection, tuning or model choice.

2. MATERIALS AND METHODS

2.1. Cohorts and labels

Expression and phenotype data were downloaded from GEO (GSE20685, GSE45827: GPL570; GSE2034, GSE7390: GPL96). The cohorts do not share a label scheme, so each was analysed with the label it actually provides (Table 2). No pooled model across cohorts was trained, because the cohorts have no common ground-truth subtype. In GSE45827, 25 samples without a subtype (non-tumour/other) were excluded from the within-cohort analyses, leaving 130 tumours; they were kept as extra training classes in the external-validation run (Sections 2.4 and 3.5).

2.2. Preprocessing

Sample IDs were matched between expression and metadata (all samples matched). Missing values were checked (none were found). Expression was inspected by boxplots, density plots and PCA; values were already log2-scale. Probes were mapped to gene symbols using the GPL570/GPL96 annotation and multiple probes per gene were averaged (or the highest-mean probe retained in GSE45827), giving 13.5k–23.5k genes per cohort.

2.3. DEG feature selection and models

Data were split 80/20 (stratified, seed 42). Features were chosen by univariate DEG screening on the training split only, with Benjamini–Hochberg FDR control (GSE20685: 500 genes; GSE45827: 174 genes from the top 50 per subtype; GSE2034 and GSE7390: top 50 genes). Features were standardized with training-fitted

scalers. Logistic Regression, SVM and Random Forest (500 trees, balanced class weights) were trained and evaluated on the hold-out set by accuracy, precision, recall, F1 and confusion matrices; accuracy confidence intervals are Wilson 95% intervals.

2.4. External validation protocol (GSE1456)

The notebook trains on GSE45827 and evaluates once on the U133A samples of GSE1456 (labels mapped Basal, ERBB2→HER2, Luminal A, Luminal B; Normal Like and No Subtype samples were excluded from scoring). Genes were matched by official symbol (10,525 shared genes) and no missing gene was zero-filled. Gene selection (ANOVA F-test, $k = 50, 100$ or 200 ; called "DEG-selected" below) and scaling were part of each model pipeline, and hyper-parameters were tuned by grid search on macro-F1. Logistic Regression, SVM and Random Forest were compared by nested cross-validation (5 outer, 3 inner folds) on GSE45827 only; the model with the highest internal macro-F1 was selected, refitted on the whole training cohort (5-fold grid search) and applied once to GSE1456. Two external normalizations were evaluated, per-cohort z-scoring and strict scaling with training-cohort means and SDs, together with a PAM50-gene baseline and 1,000-sample bootstrap intervals (seed 42). All 155 GSE45827 samples were used for training, including the 25 without a tumour subtype, which were kept as two additional classes (see Section 3.5).

3. RESULTS AND DISCUSSION

3.1. Cohort summary and hold-out performance

Hold-out results for the four within-cohort analyses are summarised in Table 2 and Figures 1–3.

Table 2. Test accuracy (Wilson 95% CI) and weighted F1 on the 20% hold-out set of each cohort.

Cohort	Label	n	Test n	Majority baseline	Features	Logistic Regression	SVM	Random Forest
GSE20685	Kao molecular clusters (type I–VI)	327	66	28.4%	500	92.4% (83–97) F1 0.924	90.9% (82–96) F1 0.908	92.4% (83–97) F1 0.924
GSE45827	PAM-type subtypes: Basal, HER2, Luminal A, Luminal B	130	26	31.5%	174	96.2% (81–99) F1 0.961	100.0% (87–100) F1 1.000	100.0% (87–100) F1 1.000
GSE2034	Bone relapse (yes/no)	286	58	75.9%	50	70.7% (58–81) F1 0.710	72.4% (60–82) F1 0.724	79.3% (67–88) F1 0.746
GSE7390	ER status (coded 1/0)	198	40	67.7%	50	80.0% (65–90) F1 0.789	90.0% (77–96) F1 0.898	87.5% (74–95) F1 0.870

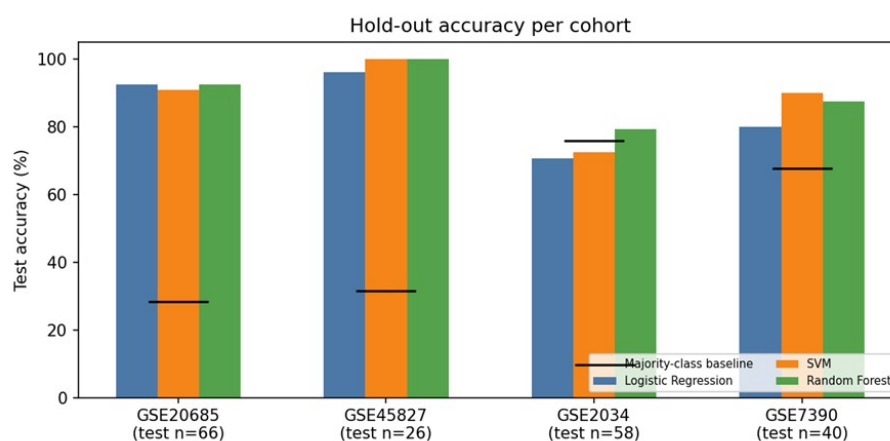


Figure 1. Hold-out accuracy by cohort and model; black bars are majority-class baselines.

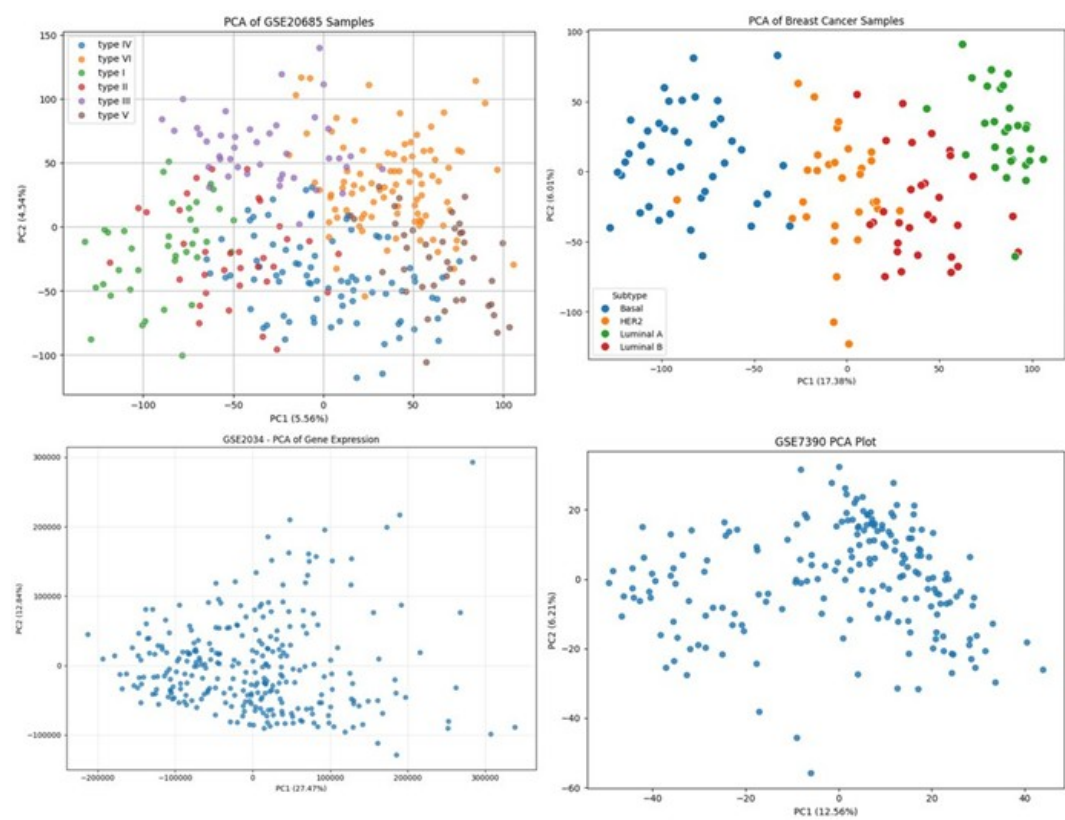


Figure 2. PCA of GSE20685, GSE45827, GSE2034 and GSE7390 (left to right, top to bottom).

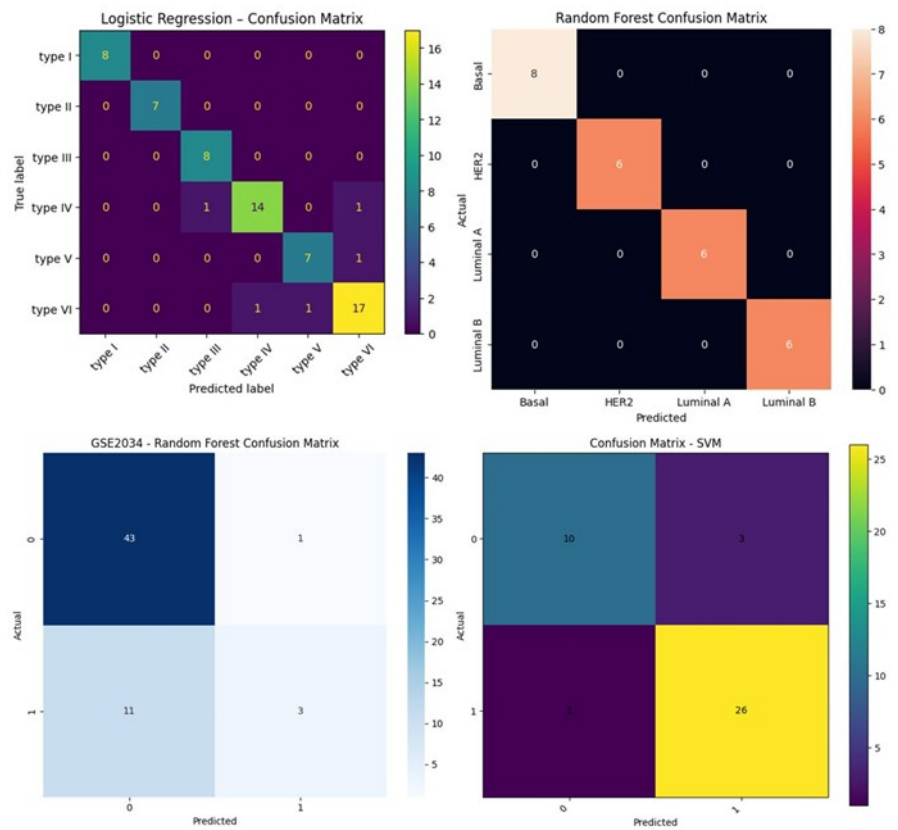


Figure 3. Confusion matrices: GSE20685 (Logistic Regression), GSE45827 (Random Forest), GSE2034 (Random Forest), GSE7390 (SVM).

3.2. Per-cohort findings

GSE20685. Logistic Regression and Random Forest reached 92.4% (61/66) and SVM 90.9%; types I–III were classified almost perfectly, while types IV–VI (F1 0.80–0.94) were confused with one another. Selected genes included SLC39A6, GATA3, XBP1, MLPH and ESR1, consistent with luminal biology.

GSE45827. SVM and Random Forest classified all 26 test tumours correctly and Logistic Regression 25/26. Top Random Forest genes were ESR1, MIEN1, PGAP3, MLPH, GRB7 and ERBB2 (Figure 4), matching the luminal/HER2 axis. With only 26 test samples, a perfect score has a lower 95% bound of about 87%.

GSE2034. Random Forest was the best model (79.3%), but this is only 3.4 points above the 75.9% majority baseline, and its relapse recall was 0.21 (precision 0.75); Logistic Regression and SVM were below baseline (Figures 4 and 5). Top DEGs were dominated by immunoglobulin and lymphocyte genes (IGKC, IGLC1, CD79A, CD27), indicating a tumour-infiltrate signal rather than a relapse biology that generalises.

GSE7390. SVM reached 90.0% (36/40) versus a 67.5% baseline; ER-related genes (ESR1, GATA3, TFF1, CA12, MAPT) led the DEG list, which is expected because the label is ER status.

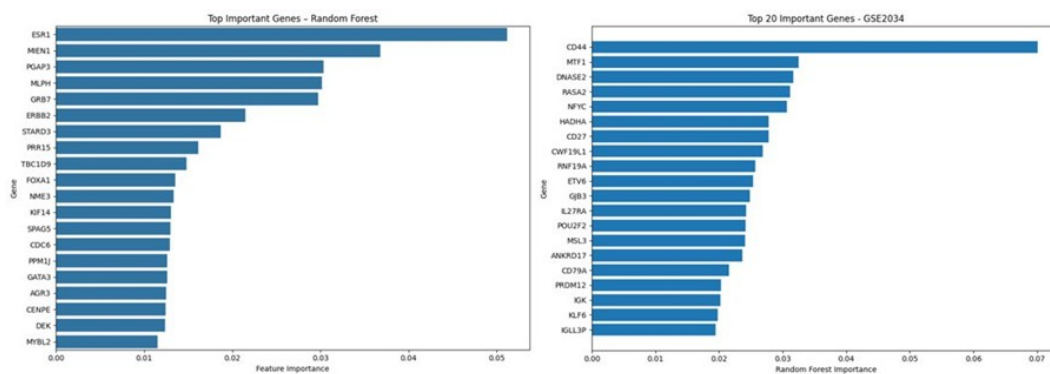


Figure 4. Random Forest feature importance for GSE45827 (left) and GSE2034 (right).

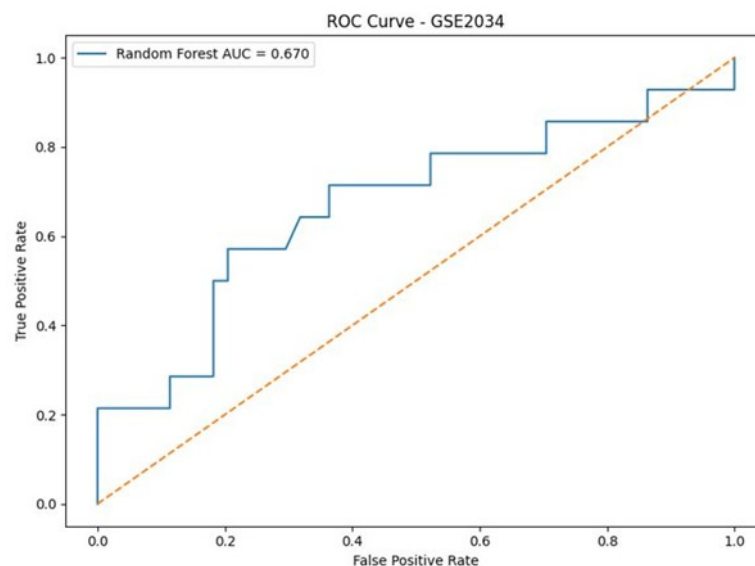


Figure 5. ROC curve, GSE2034.

3.3. External validation on GSE1456

The external analysis applied the model trained on GSE45827 to GSE1456 without any refitting. As run, the training set contained all 155 GSE45827 samples: basal 41, HER2 30, luminal A 29 and luminal B 30, plus the 25

samples without a tumour subtype, which were retained as two extra classes ("normal", n=11; "unknown", n=14) rather than excluded as described in Section 2.1 (see Section 3.5). The cohorts share 10,525 gene symbols (GSE45827: 14,725 genes; GSE1456 U133A: 13,237 genes, 159 samples). Of the 159 GSE1456 samples, 102 carried a Basal, HER2, luminal A or luminal B label (basal 25, HER2 15, luminal A 39, luminal B 23) and were scored; 57 Normal-like or No-subtype samples were not scored. Always predicting luminal A would give 38.2% accuracy, and random guessing among four classes 25%.

Internal nested cross-validation within GSE45827 was high for every model (Table 5). With DEG-selected genes, SVM reached 97.4% accuracy (macro-F1 0.976), Logistic Regression 96.1% (0.967) and Random Forest 93.5% (0.945); PAM50-only genes gave 91.0–96.8%. SVM had the highest internal macro-F1 and was therefore designated by the pre-specified rule, not by its external score, as the model to report. The grid search retained k = 100 genes for every model (the same 100 genes, 17 of them PAM50 genes).

Table 3. External validation on GSE1456 (n=102 labelled tumours; majority baseline 38.2%) of the SVM selected by internal nested CV in each configuration. Accuracy intervals are bootstrap percentile intervals (1,000 resamples); macro-F1 is averaged over the four evaluable subtypes and its interval is given for the primary analysis only. Macro AUC is the mean of the four one-vs-rest AUCs read from Figure 6b and was not computed for the other rows. Row 1 is the primary analysis. Source: GSE1456_external_metrics.csv and notebook output.

Features	Normalization	Model	Accuracy (95% CI)	Balanced acc.	Macro-F1 (95% CI)	Macro AUC
DEG-selected	Cohort z-score	SVM	48.0% (38–58)	0.547	0.50 (0.41–0.58)	≈0.77
DEG-selected	Training scaler	SVM	22.5% (15–31)	0.250	0.09	—
PAM50 genes	Cohort z-score	SVM	43.1% (33–53)	0.512	0.46	—

On GSE1456 performance fell sharply (Table 3). The selected SVM with DEG-selected genes and per-cohort z-scoring classified 49 of 102 tumours correctly (accuracy 48.0%, bootstrap 95% CI 38–58; balanced accuracy 0.547; macro-F1 0.50, CI 0.41–0.58; weighted F1 0.46), compared with 97.4% in internal cross-validation. This is 9.8 points above the 38.2% majority-class rate, but the lower bound of the interval sits at that rate, so the data do not demonstrate a reliable gain over always predicting the commonest class. Logistic Regression and Random Forest gave similar accuracy (49.0% and 47.1%; Table 5).

Table 4. Per-subtype performance of the selected SVM on GSE1456 (DEG-selected genes, cohort z-score). Confusion matrix, reference (rows) by predicted (columns: basal, HER2, luminal A, luminal B): basal 22/1/2/0; HER2 1/12/0/2; luminal A 1/0/8/30; luminal B 11/5/0/7.

Subtype	n	Precision	Recall	F1	One-vs-rest AUC
Basal	25	0.63	0.88	0.73	0.90
HER2	15	0.67	0.80	0.73	0.92
Luminal A	39	0.80	0.21	0.33	0.77
Luminal B	23	0.18	0.30	0.23	0.47
Macro average	102	0.57	0.55	0.50	≈0.77

The errors were systematic rather than diffuse (Table 4, Figure 6a). Basal tumours (recall 0.88, AUC 0.90) and HER2 tumours (recall 0.80, AUC 0.92) were recognised reasonably well, although 11 of 23 reference luminal B tumours were called basal and 5 HER2. Luminal A was largely missed: only 8 of 39 were recognised (recall 0.21) and 30 were called luminal B. Luminal B recall was 0.30 and its one-vs-rest AUC was 0.47, no better than chance (Figure 6b). The luminal A/B boundary therefore did not transfer, whereas the more separable basal and HER2 classes partly did. The gene panel selected in the training cohort was dominated by proliferation and cell-cycle genes (for example AURKA, BUB1, CCNB1, CDK1, CDC20 and KIF2C) together with ESR1, ERBB2, GRB7 and FOXA1. Whether the luminal shift reflects platform and normalization differences, different label

definitions in the two cohorts (GSE1456 labels were taken from GEO sample metadata and their derivation was not verified here), or both cannot be separated with these data.

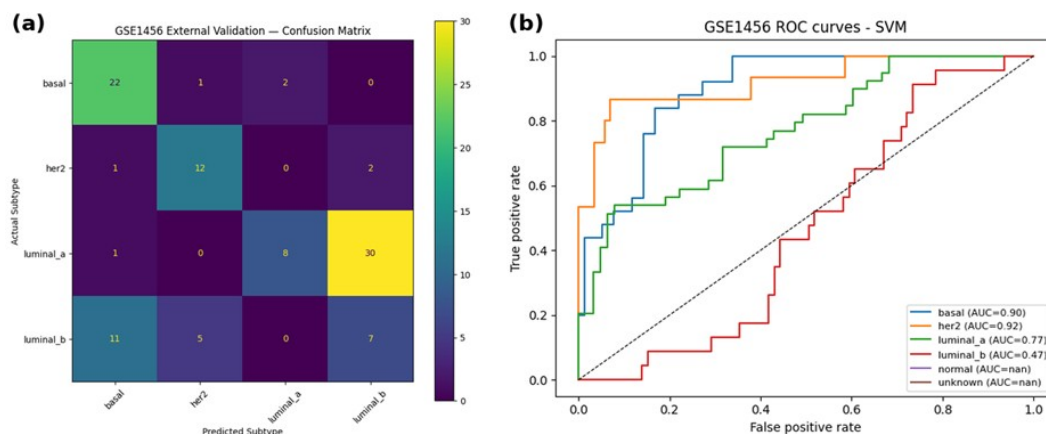


Figure 6. GSE1456 external validation of the selected SVM: (a) confusion matrix (rows, reference; columns, predicted); (b) one-vs-rest ROC curves. "normal" and "unknown" are training-only classes with no external samples, so their AUC is undefined.

Normalization mattered more than the choice of model or gene set (Table 5). With strictly leakage-free scaling (training-cohort means and SDs applied to GSE1456), accuracy with DEG-selected genes collapsed to 5.9% (Logistic Regression), 22.5% (SVM) and 28.4% (Random Forest); the first two are below the 25% expected by chance. PCA of the raw shared genes separated the two cohorts completely along PC1 (43.5% of variance; Figure 7a), although mean expression levels were similar (6.88 versus 6.62), indicating a large platform- or batch-related offset. Per-cohort z-scoring removed this separation (Figure 7b), but it standardises each cohort with its own gene means and SDs; it uses no test labels, yet it needs a batch of external samples and is therefore not a single-patient prediction rule.

Table 5. Accuracy (macro-F1) for every model and feature set: internal nested CV in GSE45827 and external testing in GSE1456 (n=102) under two normalizations. External macro-F1 is averaged over the four evaluable subtypes. Bootstrap 95% intervals for external accuracy are about ± 10 points (33–61% across the z-scored configurations). The bold row is the primary analysis.

Features	Model	Internal nested CV (GSE45827)	External, cohort z-score	External, training scaler
DEG-selected	Logistic Regression	96.1% (0.967)	49.0% (0.51)	5.9% (0.14)
DEG-selected	SVM	97.4% (0.976)	48.0% (0.50)	22.5% (0.09)
DEG-selected	Random Forest	93.5% (0.945)	47.1% (0.50)	28.4% (0.25)
PAM50 genes	Logistic Regression	94.2% (0.938)	47.1% (0.51)	44.1% (0.43)
PAM50 genes	SVM	96.8% (0.967)	43.1% (0.46)	50.0% (0.52)
PAM50 genes	Random Forest	91.0% (0.906)	51.0% (0.54)	37.3% (0.29)

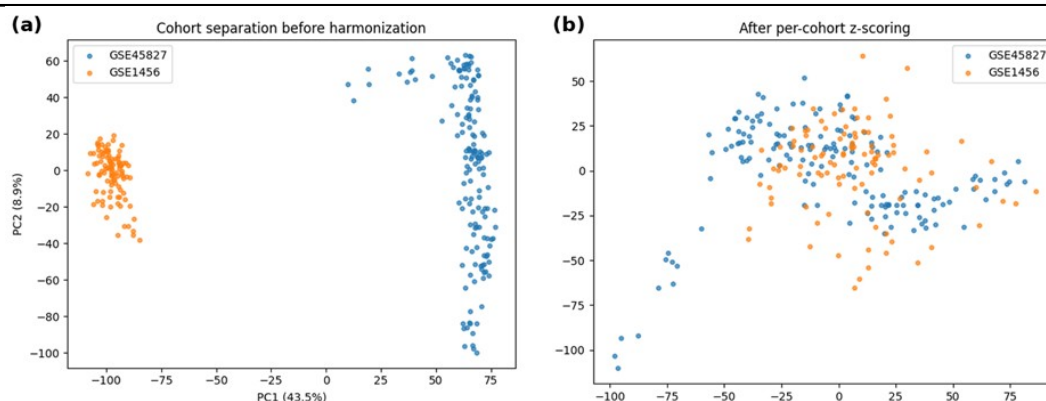


Figure 7. PCA of GSE45827 and GSE1456 on the shared genes: (a) before and (b) after per-cohort z-scoring.

A baseline using only the 43 PAM50 genes present in both cohorts was no better than the DEG-selected panel: 43.1–51.0% with cohort z-scoring and 37.3–50.0% with training-cohort scaling. The spread among models and feature sets under z-scoring (43–51%) is well inside the bootstrap intervals of roughly ± 10 points, so none of these configurations can be ranked reliably. The macro-AUC exported by the notebook was not computable because the two training-only classes have no external samples; Table 3 instead gives the mean of the four evaluable one-vs-rest AUCs (about 0.77), read from Figure 6b.

4. Discussion

Labels reflecting tumour biology that is strongly encoded in expression (subtype clusters, ER status) were classified with high accuracy, whereas bone-relapse prediction in GSE2034 barely exceeded the majority baseline. The small hold-out sets (26–66 samples) make the point estimates imprecise; repeated or nested cross-validation is preferable. Because label definitions differ by cohort, the numbers are not comparable across cohorts and do not constitute a pooled multi-cohort subtype classifier.

The external result qualifies the within-cohort figures. Near-perfect hold-out and nested-CV accuracy in GSE45827 (97.4%) did not carry over to GSE1456 (48.0%), although both are Affymetrix breast-tumour cohorts with the same subtype vocabulary. Basal and HER2 tumours, which have strong and distinctive expression signals, partly transferred; the luminal A/B distinction, which rests on a continuous proliferation gradient and on how each cohort sets its cut-off, did not. Subtype classifiers should therefore be evaluated on independent cohorts with harmonised labels and a pre-specified normalization, and within-cohort accuracy should not be read as expected performance elsewhere. Because the external set has 102 tumours, the confidence intervals span about 20 points, and the training run included two non-tumour classes (see Section 3.5), these numbers are a first external estimate rather than a definitive one.

4.1. Limitations

(i) Single 80/20 splits. (ii) DEG screening used univariate tests, and the number of retained genes differs by cohort (50–500). (iii) A PAM50 centroid-scoring step in the project's master notebook used randomly generated placeholder centroids, so its subtype calls were not used here; real PAM50 centroids are required before PAM50 labels can be assigned to GSE2034 or GSE7390. (iv) In the master notebook the GSE2034 matrix had values in 0–1, suggesting that detection p-values were read instead of intensities, so only the dataset-specific notebook results are reported. (v) Only the U133A samples of GSE1456 have data matched to GPL96. (vi) In the external run, the 25 GSE45827 samples without a tumour subtype were kept as two extra classes ("normal", "unknown") instead of being removed, so the training set (155 samples) differs from the 130 tumours of Section 2.1 and the model could assign samples to classes absent from the external labels; the run should be repeated on the 130 subtyped tumours. (vii) Per-cohort z-scoring uses the external cohort's own gene means and SDs (not its labels), and strict training-parameter scaling failed because of a strong cohort offset. (viii) GSE1456 subtype labels were parsed from GEO sample text and were not verified against the original publication; they may have been derived differently from the GSE45827 labels. (ix) The notebook's exported macro-F1 averaged over six class labels, two of which have no external samples; values in this paper are rescaled to the four evaluable classes, and the notebook did not export a usable macro-AUC.

5. CONCLUSION

Expression-based models classify subtype and ER labels well on held-out data from the same cohort, but not relapse. In the independent GSE1456 validation, four-class subtype accuracy fell to 48.0% (95% CI 38–58) against a 97.4% internal estimate: basal and HER2 tumours partly transferred, luminal A and B did not, and results depended strongly on normalization. Cross-cohort subtype classification therefore needs harmonised labels and normalization before any claim of generalisability.

ACKNOWLEDGEMENTS

None.

CONFLICT OF INTEREST

The authors declare that they have no conflict of interest.

ETHICAL STATEMENTS

This study is a secondary analysis of publicly available, de-identified gene-expression datasets from the NCBI Gene Expression Omnibus. No new human participants, samples or animals were involved, and no additional ethical approval or informed consent was required.

DATA AVAILABILITY

All datasets are public in the NCBI Gene Expression Omnibus (GSE20685, GSE45827, GSE2034, GSE7390, GSE1456; <https://www.ncbi.nlm.nih.gov/geo/>).

CODE AVAILABILITY

All analyses were run in Google Colab (Python 3). Each notebook downloads its GEO data directly and can be re-run end to end:

BREAST_CANCER_GSE20685.ipynb — GSE20685 hold-out analysis (Sections 3.1–3.2). Colab: <https://colab.research.google.com/drive/1VT1ejckdMYD8AV3vW6oYbrZTWY99hntO?usp=sharing>

BREAST_CANCER_GSE_45827.ipynb — GSE45827 hold-out analysis (Sections 3.1–3.2). Colab: https://colab.research.google.com/drive/1NbD0_Kek97IErY0DaVbXna-OC9KM2j3L?usp=sharing

BREAST_CANCER_GSE_ID_2034.ipynb — GSE2034 bone-relapse analysis (Sections 3.1–3.2). Colab: https://colab.research.google.com/drive/1FxacbJPcyntzXJ6F-ICc_SC-2FVd9oOV?usp=sharing

BREAST_CANCER_GSE_ID_7390.ipynb — GSE7390 ER-status analysis (Sections 3.1–3.2). Colab: https://colab.research.google.com/drive/1_hYDyau1z4-EEbTJzuthWjmfCj6QsaE1?usp=sharing

FINAL_EXTERNAL_VALIDATION_1456.ipynb — GSE45827→GSE1456 external validation (Section 3.3). Colab: <https://colab.research.google.com/drive/1CQH71BQFIClt9He39PE85VW4VuRUgnLf?usp=sharing>

Master_pipeline_breast_cancer_dataset_use_gseid__1_.ipynb — Master pipeline notebook (see Section 3.5, items iii–iv). Colab: <https://colab.research.google.com/drive/1CQH71BQFIClt9He39PE85VW4VuRUgnLf?usp=sharing>

References

- [1] Perou CM, Sørlie T, Eisen MB, et al. Molecular portraits of human breast tumours. *Nature*. 2000;406:747-52. doi:10.1038/35021093
- [2] Sørlie T, Perou CM, Tibshirani R, et al. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc Natl Acad Sci U S A*. 2001;98:10869-74. doi:10.1073/pnas.191367098
- [3] Parker JS, Mullins M, Cheang MCU, et al. Supervised risk predictor of breast cancer based on intrinsic subtypes. *J Clin Oncol*. 2009;27:1160-7. PMID: 19204204
- [4] Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res*. 2002;30:207-10. doi:10.1093/nar/30.1.207
- [5] Yasrebi H, Sperisen P, Praz V, Bucher P. Can survival prediction be improved by merging gene expression data sets? *PLoS One*. 2009;4:e7431. PMID: 19851466
- [6] Hikichi S, Sugimoto M, Tomita M. Correlation-centred variable selection of a gene expression signature to predict breast cancer metastasis. *Sci Rep*. 2020;10:7923. PMID: 32404965
- [7] Kao KJ, Chang KM, Hsu HC, Huang AT. Correlation of microarray-based breast cancer molecular subtypes and clinical outcomes: implications for treatment optimization. *BMC Cancer*. 2011;11:143. PMID: 21501481
- [8] [Authors]. A pathway-based classification of breast cancer integrating data on differentially expressed genes, copy number variations and microRNA target genes. [Journal. Year;Vol:Pages]. PMID: PMC3887768
- [9] [Authors]. Identification of a 9-gene prognostic signature for breast cancer. *Cancer Med*. 2020;9(24):9471-84. doi:10.1002/cam4.3523. PMID: 33090721
- [10] [Authors]. A four-gene signature in the tumor microenvironment that significantly associates with the prognosis of patients with breast cancer. [Journal. Year;Vol:Pages]. PMID: 32791092
- [11] [Authors]. An aging-related gene signature-based model for risk stratification and prognosis prediction in breast cancer. [Journal. Year;Vol:Pages]. PMID: 34785957
- [12] [Authors]. Development and validation of a hypoxia-related prognostic signature for breast cancer. [Journal. Year;Vol:Pages]. PMID: 32724434
- [13] [Authors]. Construction and identification of a novel 5-gene signature for predicting the prognosis in breast cancer. [Journal. Year;Vol:Pages]. PMID: [unavailable]
- [14] [Authors]. Identification of a 4-mRNA metastasis-related prognostic signature for patients with breast cancer. [Journal. Year;Vol:Pages]. PMID: [unavailable]
- [15] [Authors]. Validation of the 18-gene classifier as a prognostic biomarker of distant metastasis in breast cancer. [Journal. Year;Vol:Pages]. PMID: 28886126
- [16] GSE45827. Expression data from breast cancer subtypes. NCBI Gene Expression Omnibus [Internet]. Available from: <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE45827>
- [17] [Authors]. Expression analysis of long non-coding RNAs related with FOXM1, GATA3, FOXA1 and ESR1 in breast tissues. [Journal. Year;Vol:Pages]. PMID: 34094972
- [18] Lin Y, et al. Identification of potential key genes for HER-2 positive breast cancer based on bioinformatics analysis. *Medicine (Baltimore)*. 2020;99:e18445. PMID: 31895772
- [19] Wang S, Shang P, Yao G, Ye C, Chen L, Hu X. A genomic and transcriptomic study toward breast cancer. *Front Genet*. 2022;13:989565. PMID: 36313438
- [20] Wang Y, Klijn JG, Zhang Y, et al. Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *Lancet*. 2005;365:671-9. doi:10.1016/S0140-6736(05)17947-1
- [21] Desmedt C, Pictet F, Loi S, et al. Strong time dependence of the 76-gene prognostic signature for node-negative breast cancer patients in the TRANSBIG multicenter independent validation series. *Clin Cancer Res*. 2007;13:3207-14. PMID: 17545524
- [22] Yin Z, et al. A 41-gene signature derived from breast cancer stem cells as a predictor of survival. *J Exp Clin Cancer Res*. 2014;33:49. PMID: 24906694
- [23] [Authors]. The prognostic significance of protein tyrosine phosphatase 4A2 in breast cancer. [Journal. Year;Vol:Pages]. PMID: 26203261

- [24] [Authors]. Microarray phosphatome profiling of breast cancer patients unveils a complex phosphatase regulatory role of the MAPK and PI3K pathways in estrogen receptor-negative breast cancers. [Journal. Year;Vol:Pages. PMID]
- [25] Liu R, Zhang W, Liu ZQ, Zhou HH. Gene modules associated with breast cancer distant metastasis-free survival in the PAM50 molecular subtypes. *Oncotarget*. 2016;7:21686-98. PMID: 26934123
- [26] Pawitan Y, Bjöhle J, Amler L, et al. Gene expression profiling spares early breast cancer patients from adjuvant therapy: derived and validated in two population-based cohorts. *Breast Cancer Res*. 2005;7:R953-64. doi:10.1186/bcr1325
- [27] Crawford NPS, et al. Bromodomain 4 activation predicts breast cancer survival. *Proc Natl Acad Sci U S A*. 2008;105:6380-5. PMID: 18427120
- [28] [Authors]. Inference of patient-specific subpathway activities reveals a functional signature associated with the prognosis of patients with breast cancer. [Journal. Year;Vol:Pages]. PMID: 29971923
- [29] Huang S, Murphy L, Xu W. Genes and functions from breast cancer signatures. *BMC Cancer*. 2018;18:473. PMID: 29699511
- [30] [Authors – please verify]. Identification of a prognostic signature based on the expression of genes related to the insulin pathway in early breast cancer. *Breast Care*. 2021;[Vol:Pages]. PMID: 34248472
- [31] [Authors]. RDGN-based predictive model for the prognosis of breast cancer. [Journal. Year;Vol:Pages. PMID]
- [32] [Authors]. ME1 promotes basal-like breast cancer progression and associates with poor prognosis. *Sci Rep*. 2018; [Vol:Pages]. PMID: 30425310
- [33] [Authors]. Integrative dissection of 5-hydroxytryptamine receptors-related signature in the prognosis and immune microenvironment of breast cancer. [Journal. Year;Vol:Pages]. PMID: 37795441