Volume 3, Number 2, 2026 Pages 478-495

Bioscience Insights

ISSN: 3080-3489 (Print)
ISSN: 3080-3497 (Online)

BiosciInsights.com

A COMPACT, REPRODUCIBLE FIVE-GENE TRANSCRIPTOMIC CLASSIFIER DISTINGUISHES ESOPHAGEAL SQUAMOUS CELL CARCINOMA FROM NORMAL TISSUE: A MULTI-COHORT MACHINE-LEARNING STUDY

Shehzad Khalil[*1], Emaan Hafeez[2], Minahil Sajjad[3], Shazab Ali[4], Sobia[5], Muhammad Anas Akbar[6], Jalal Khan[7], Abbas Ahmed[8]

*[*1,5]Institute of Biotechnology and Genetic Engineering, The University of Agriculture, Peshawar, Khyber Pakhtunkhwa, Pakistan*
[2]Department of Biotechnology, Lahore College for Women University, Lahore, Pakistan
[3]Department of Biotechnology, Kinnaird College for Women University, Lahore, Pakistan
[4]School of Biochemistry and Biotechnology, University of the Punjab, Lahore, Pakistan
[6]Department of Biotechnology, Government College University Lahore, Pakistan
[7,8]Department of Biotechnology, Abdul Wali Khan University Mardan, Khyber Pakhtunkhwa, Pakistan

[*1]shehzadkhalil911@gmail.com, [2]emaanhafeez25@gmail.com, [3]minahilsajjad2023@gmail.com, [4]shazabali.research@gmail.com, [5]sobiauoa2003@gmail.com, [6]rnanasje@gmail.com, [7]jalalkhanbiotech55@gmail.com, [8]abbas.ahmad@awkum.edu.pk

Article Information

Article History

Received:	May 15, 2026
Revised:	June 27, 2026
Accepted:	August 12, 2026
Available Online:	August 31, 2026

Keywords:
esophageal squamous cell carcinoma; machine learning; LASSO; random forest; SVM-RFE;

Abstract

Background. *Esophageal squamous cell carcinoma (ESCC) is an aggressive malignancy with poor survival and late presentation. Robust, interpretable transcriptomic classifiers that generalise across independent cohorts remain scarce. We aimed to derive a compact, reproducible ESCC-vs-normal gene signature using interpretable machine-learning feature selection and to validate it in a fully independent cohort.*

Methods. *Three publicly available GEO microarray series were verified as ESCC tumour-vs-normal studies: GSE23400 (106 samples, HG-U133A), GSE77861 (14 samples, HG-U133_Plus_2), and GSE20347 (34 samples, HG-U133A_2). Expression values were already RMA-log2 normalised; probes were mapped to HGNC symbols and collapsed to a common universe of 13,039 genes. Discovery comprised GSE23400 + GSE77861 (120*

gene signature; independent validation; biomarker

samples; 60 tumour / 60 normal). GSE20347 (34 samples; 17/17) was held out as an independent validation set and never used for candidate selection, feature selection, hyperparameter tuning, or threshold optimisation. Candidate genes were defined by two strategies (Study-1 consensus DEGs and a broad variance filter) and reduced by LASSO, Random Forest (RF), and SVM-recursive feature elimination (SVM-RFE) within the discovery data only, using within-fold standardisation to prevent leakage. A consensus signature was assembled, compact panels (5/10/15/20 genes) were evaluated by repeated stratified cross-validation, and a Logistic Regression model was frozen and applied unchanged to GSE20347.

Results. *Verification confirmed all three datasets represent ESCC (tumour vs normal). Two hundred ninety consensus DEGs and 2,000 high-variance genes formed the candidate pool (union = 2,001). LASSO selected 22 (λmin)/17 (λ1se) genes, RF 17 stable genes, and SVM-RFE 5 genes; the consensus contained 8 genes selected by ≥2 methods. Internal repeated cross-validation showed performance plateauing at the 5-gene panel (AUC = 0.998 ± 0.004). The frozen five-gene Logistic Regression signature (SIM2, RFC4, COL1A1, MMP1, CST1) achieved an independent-validation ROC-AUC of 1.00 (95% CI 1.00–1.00), PR-AUC 1.00, accuracy 0.941, sensitivity 0.882, and specificity 1.00. All five genes were individually significantly differentially expressed in GSE20347 (p < 0.001).*

Conclusions. *A compact five-gene expression signature generalises from discovery to a fully independent ESCC cohort with near-perfect discrimination. The signature is biologically plausible, interpretable, and suitable for future prospective validation.*

INTRODUCTION

Esophageal cancer remains among the most lethal solid malignancies worldwide. It ranks as the eleventh most commonly diagnosed cancer and the sixth leading cause of cancer-related mortality globally, with an estimated 604,000 new cases and 544,000 deaths in 2020 [1]. Esophageal squamous cell carcinoma (ESCC) accounts for the substantial majority of these cases, particularly in the so-called 'esophageal cancer belt' spanning Central and East Asia, parts of eastern and southern Africa, and among high-risk subgroups such as African Americans in North America [1]. ESCC is characterised by aggressive local invasion, early submucosal and nodal spread, and a five-year survival that remains below 20% once the disease has progressed beyond an early stage [1]. Because early ESCC is frequently asymptomatic and endoscopic screening is not universally feasible in high-incidence, resource-limited settings, most patients present with locally advanced or metastatic disease, at which point curative treatment options are markedly diminished. A major barrier to improved outcomes is therefore the absence of robust, low-cost molecular tools that can stratify ESCC from benign esophageal mucosa using a small, reproducible, and clinically tractable set of transcripts; such a tool could, in principle, complement histopathology, support triage in referral pathways, or be adapted to minimally invasive sampling.

High-throughput transcriptomic profiling, culminating in large consortium efforts such as The Cancer Genome Atlas (TCGA) integrated analysis of esophageal carcinoma, has repeatedly demonstrated that ESCC is associated with pervasive gene-expression reprogramming affecting cell-cycle control, DNA-replication

fidelity, extracellular-matrix (ECM) remodelling, and invasive capacity [2]. These transcriptional changes map onto several of the canonical hallmarks of cancer, including sustained proliferative signalling, activation of invasion and metastasis, and evasion of growth suppressors, articulated by Hanahan and Weinberg [3]. Translating this genome-wide reprogramming into compact diagnostic signatures suitable for clinical or translational use has, however, been hindered by three recurrent methodological problems in the biomarker literature. First, many published signatures are large (tens to hundreds of genes) and biologically opaque, which limits both mechanistic interpretability and the practicality of translating them into a targeted, low-cost assay such as quantitative PCR (qPCR) or a nanoString panel. Second, feature selection is frequently performed across all available samples rather than within a dedicated discovery partition, which confounds discovery with validation and inflates performance estimates through data leakage, a well-documented failure mode in high-dimensional genomic classification. Third, independent, truly external validation cohorts are often omitted altogether, or are themselves incorporated into model development at some stage (for example, through joint normalisation, feature-selection, or threshold-tuning steps), undermining the claim of independence.

Machine-learning (ML) approaches, notably LASSO-penalised logistic regression [4,5], Random Forests [6], and support-vector-machine recursive feature elimination (SVM-RFE) [7], offer principled, complementary routes to sparse, interpretable signatures from high-dimensional transcriptomic data. LASSO imposes an L1 penalty that drives many coefficients exactly to zero, yielding an inherently sparse and interpretable linear model; Random Forests aggregate many decorrelated decision trees and provide non-parametric, non-linear importance rankings robust to modest sample sizes; and SVM-RFE performs iterative backward elimination guided by the margin-maximising hyperplane of a support-vector classifier, capturing multivariate combinations of genes that individually might not reach significance. Because these three algorithms rest on different statistical principles, penalised regression, ensemble tree-based partitioning, and margin-based backward elimination, genes that are selected concordantly by two or more methods are less likely to reflect an idiosyncrasy of any single algorithm. Yet the very strength of ML methods (their capacity to fit high-dimensional data with comparatively few samples) makes them vulnerable to overfitting unless discovery and validation are rigorously separated and performance is estimated by repeated or nested cross-validation with correct within-fold standardisation [8]. Comparisons between correlated classifiers likewise require statistically appropriate methods, such as the DeLong test for correlated ROC curves, rather than naive comparison of point estimates [9].

In this study we address these methodological gaps directly by (i) verifying the cancer phenotype and sample composition of three GEO series directly from their metadata rather than from file naming or assumption, (ii) conducting all candidate generation and feature selection strictly within a discovery cohort (GSE23400 + GSE77861), (iii) combining three complementary ML selectors, LASSO, Random Forest, and SVM-RFE, into a consensus rather than relying on any single method, (iv) evaluating compact panels of increasing size (5/10/15/20 genes) by repeated stratified cross-validation to identify the point of diminishing returns, and (v) freezing a single, fully specified model that is applied without refitting, re-selection, or threshold tuning to an independent cohort (GSE20347) that played no role whatsoever in model development. Our primary objective was to determine whether interpretable ML methods can identify a compact, reproducible transcriptomic signature that distinguishes ESCC from normal tissue across independent cohorts, and, secondarily, to characterise the biological plausibility of the resulting genes against the established literature on ESCC pathobiology. The overall workflow of the study is given in the Figure 1.

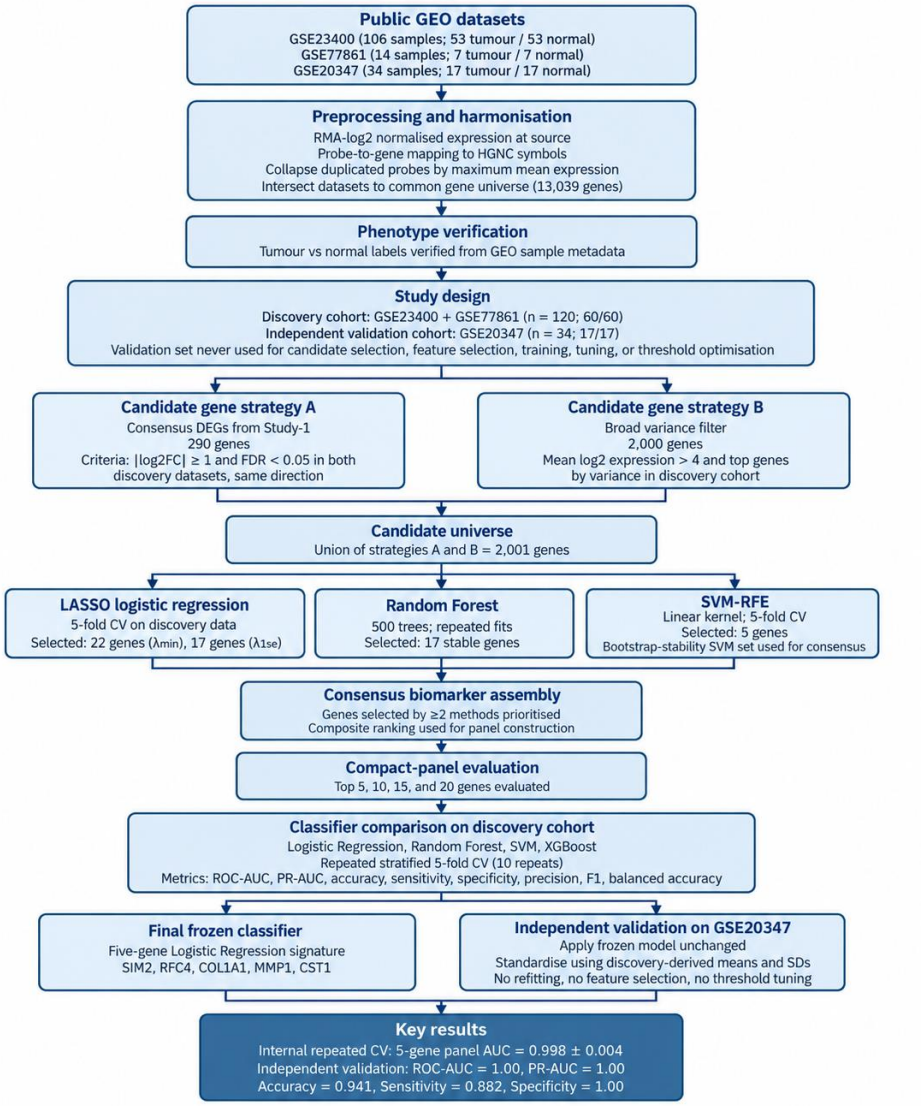


Figure 1. Study workflow: from processed GEO datasets through candidate-gene selection, three independent machine-learning selectors (LASSO, Random Forest, SVM-RFE), consensus biomarker assembly, compact-panel evaluation, classifier development, internal repeated cross-validation, and frozen-model application to the independent GSE20347 cohort.

2. Materials and Methods

2.1 Data sources and preprocessing

Three GEO series matrices were obtained from the NCBI Gene Expression Omnibus (GEO) [10]: GSE23400, GSE77861, and GSE20347. All three were RMA-log2 normalised at source [11] (stated in each sample's Sample_data_processing field), and no further normalisation, background correction, or batch correction was performed. Probe-level data had been mapped to HGNC symbols and Entrez identifiers via the relevant Bioconductor annotation packages (hgu133a.db for GPL96/GPL571; hgu133plus2.db for GPL570), with multiply-represented probes collapsed by retaining the probe with maximum mean expression. After annotation, each dataset contained 0 missing values and 0 duplicated gene symbols.

2.2 Phenotype and common gene universe

Phenotype (Tumour vs Normal) was extracted strictly from GEO sample metadata (Sample_characteristics_ch1, Sample_source_name_ch1, Sample_title), not from file naming. The three datasets were intersected on HGNC symbol to yield a common gene universe of 13,039 genes, guaranteeing identical feature representation across discovery and validation.

2.3 Discovery-validation split

In keeping with the pre-registered design, the discovery cohort was the union of GSE23400 (106 samples: 53 tumour / 53 normal) and GSE77861 (14 samples: 7/7), giving 120 samples (60/60). GSE20347 (34 samples: 17/17) was reserved exclusively for independent validation and was never exposed to candidate selection, feature selection, model training, hyperparameter optimisation, or decision-threshold tuning.

2.4 Candidate gene strategies

Two biologically and statistically defensible candidate sets were constructed on discovery data only. Strategy A (DEG-based) used the 290 consensus differentially expressed genes ($|\log2FC| \geq 1$ and FDR < 0.05 in both discovery datasets, same direction) from the parallel Study-1 analysis [12]. Strategy B (broad variance) selected genes with mean log2 expression > 4 (clearly expressed) and within the top 2,000 by variance across the discovery cohort. The union of both strategies (2,001 genes) formed the candidate universe for downstream ML selection.

2.5 Machine-learning feature selection (discovery only)

All three selectors operated on the candidate universe with within-training-fold standardisation (z-scoring per gene using training-fold statistics) to preclude leakage. LASSO penalised logistic regression (glmnet, $\alpha = 1$) [4,5] with 5-fold cross-validation selected λmin and λ1se; genes with non-zero coefficients were declared selected, and selection stability was assessed over 30 bootstrap resamples. Random Forest (500 trees, mtry = $\sqrt{p}$) [6] with 20 repeated fits recorded permutation (mean-decrease-in-accuracy) and Gini importance per gene, and a gene was 'selected' if it appeared in the top-20 by importance in ≥50% of repeats. SVM-RFE with a linear kernel [7] performed genuine backward elimination, with 5-fold CV AUC used to choose the optimal panel size by a one-standard-error rule; because near-perfect separability rendered the minimal RFE set unstable across runs, a bootstrap-stability SVM selection (top-20 |weights| in ≥50% of 30 resamples) was additionally used for the consensus.

2.6 Consensus biomarker definition

Genes were classified as selected by all three, exactly two, or exactly one method. A composite score (mean of method-specific normalised importances, weighted by selection status) ranked the union for compact-panel construction.

2.7 Compact panels and classifier development

Panels of 5, 10, 15, and 20 top-ranked genes were evaluated. Four classifier families were trained: Logistic Regression (ridge), Random Forest, SVM (linear), and XGBoost [13], with repeated stratified 5-fold cross-validation (10 repeats) on discovery.

2.8 Internal validation

Repeated CV reported ROC-AUC, PR-AUC, accuracy, sensitivity, specificity, precision, F1, and balanced accuracy as mean ± SD.

2.9 Independent validation

The final model was frozen using discovery data only and then applied to GSE20347 after standardising validation features with the discovery-derived gene means and SDs. No refitting, no feature selection, and no

threshold optimisation were performed on GSE20347. ROC-AUC with 95% confidence interval (DeLong method) [9], PR-AUC, and the full confusion matrix were computed.

2.10 Implementation

All analyses were performed in R 4.6.0 using glmnet [5], randomForest [6], e1071, xgboost [13], caret [8], pROC [14], and data.table. Random seeds were fixed for all stochastic steps.

3. Results

3.1 Dataset verification and composition

Inspection of GEO metadata confirmed that all three series are ESCC tumour-vs-normal studies (Table 1). The analysis was conducted exactly as designed on the verified ESCC phenotype. After annotation the common gene universe was 13,039 genes, with zero missing values and zero duplicated gene identifiers in every dataset (**Figure** 2).

Table 1. Dataset characteristics.

Dataset	Disease	Platform	Samples (T/N)	Genes	Missing / Dup
GSE23400	ESCC	GPL96 (HG-U133A)	106 (53/53)	13,039	0 / 0
GSE77861	ESCC	GPL570 (HG-U133_Plus_2)	14 (7/7)	21,355*	0 / 0
GSE20347	ESCC	GPL571 (HG-U133A_2)	34 (17/17)	13,039	0 / 0

*Raw annotation count; intersected to the 13,039-gene common universe. Discovery = GSE23400 + GSE77861 = 120 (60/60); Validation = GSE20347 = 34 (17/17).

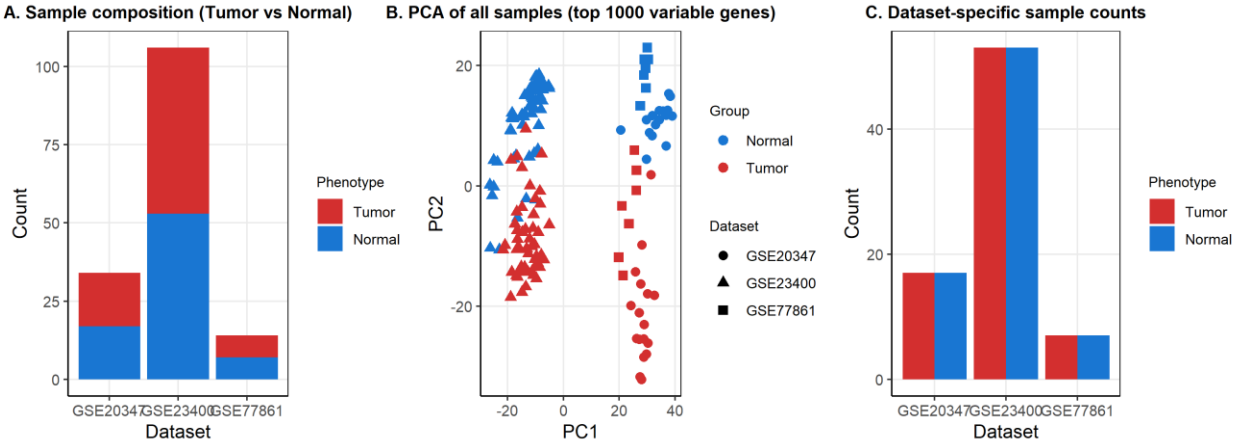


Figure 2. **Dataset structure: (A) stacked tumour/normal composition, (B) PCA of all samples on the top variable genes (colour = phenotype, shape = dataset), and (C) dataset-specific sample counts.**

3.2 Candidate gene landscape

Strategy A yielded 290 consensus DEGs; Strategy B yielded 2,000 high-variance genes; their union (2,001 genes) was used for ML selection (**Figure** 3). The two strategies overlapped substantially (289 of 290 DEGs were also in the variance set), confirming that the consensus DEGs are among the most variable transcripts.

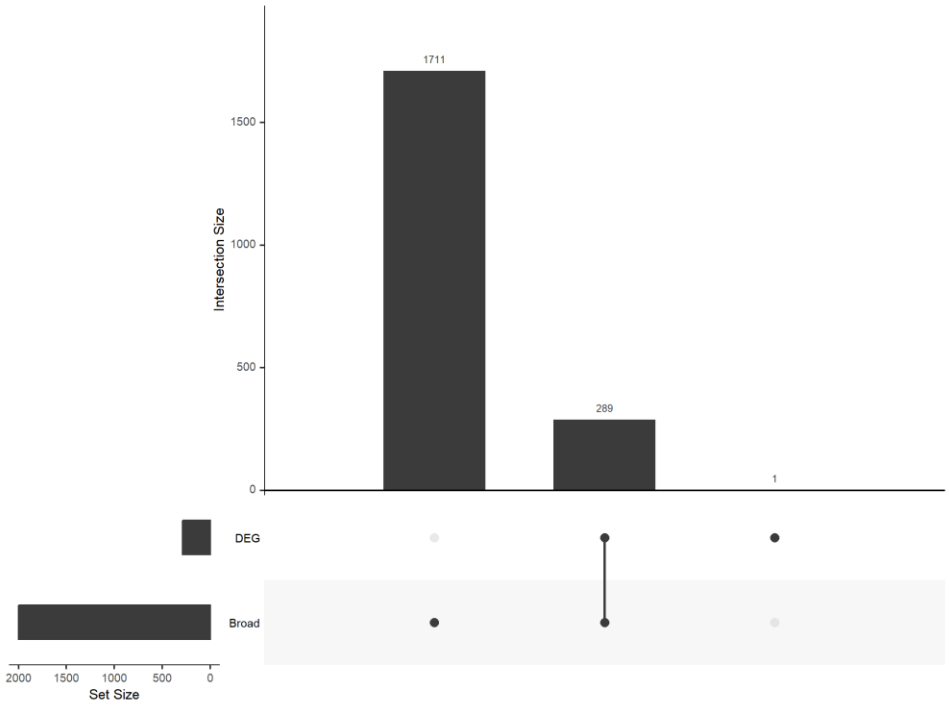


Figure 3. Candidate-gene landscape: overlap between the DEG-based strategy (290 consensus DEGs) and the broad variance-filter strategy (2,000 genes).

3.3 Multi-method feature selection

On the discovery cohort, LASSO selected 22 genes at λmin and 17 at λ1se; Random Forest selected 17 stable genes; SVM-RFE selected 5 genes (**Figure**s 4–6). The methods agreed most strongly on a core proliferative/ECM-invasion programme (MMP1, RFC4, COL1A1, CKS1B, MCM2, PLXNA1, IGF2BP2), reflecting the biological coherence of the selection.

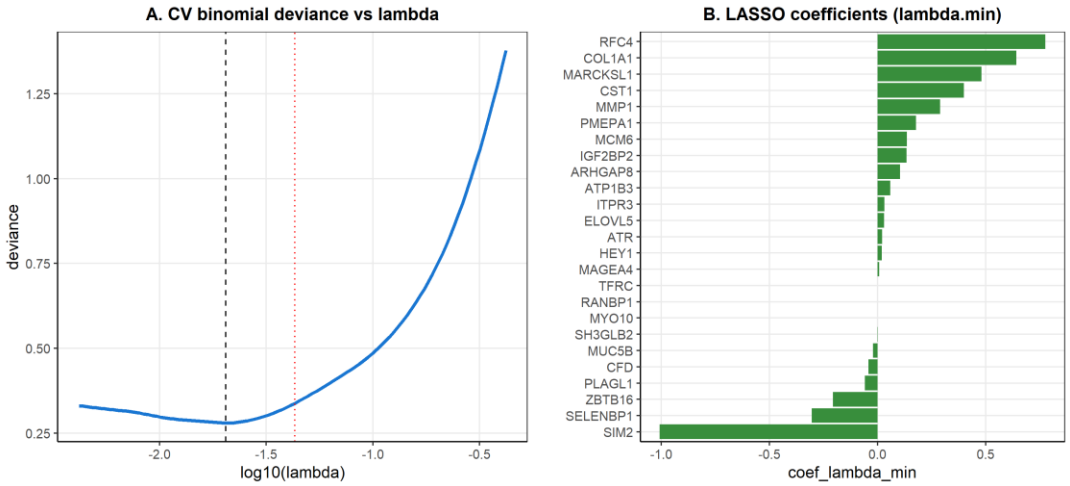


Figure 4. LASSO analysis: cross-validated binomial deviance versus log10(λ) with λmin and λ1se indicated, and LASSO coefficients at λmin for selected genes.

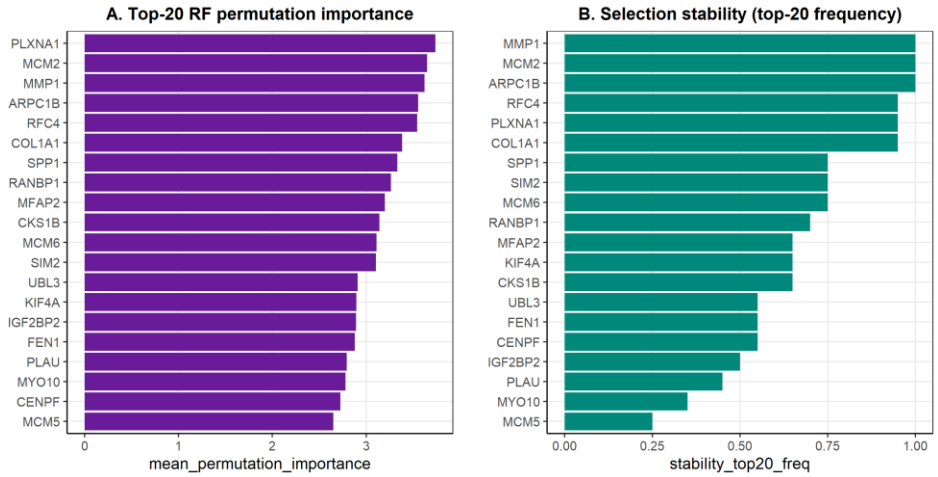


Figure 5. Random Forest analysis: top-20 permutation (mean-decrease-in-accuracy) importances and selection stability (frequency in top-20 across 20 repeated fits).

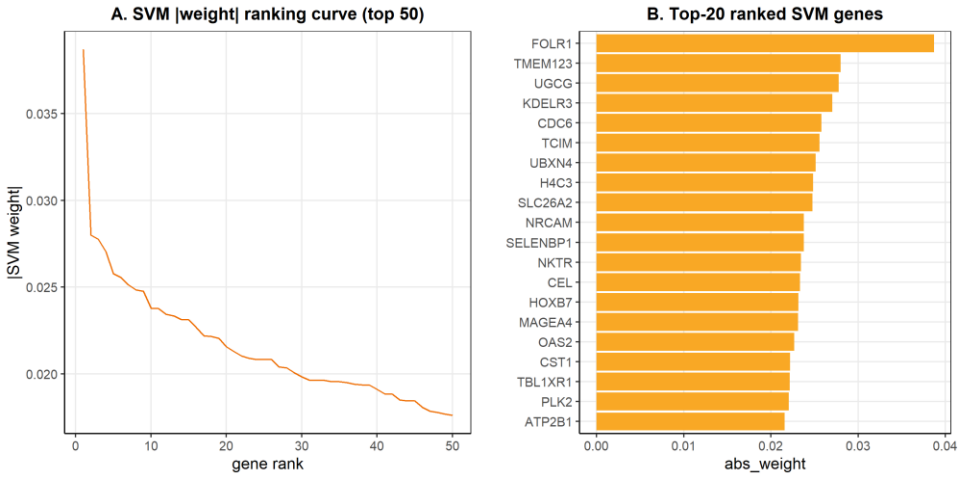


Figure 6. SVM-RFE analysis: the |SVM weight| ranking curve (top 50 genes) and the top-20 ranked SVM genes.

3.4 Consensus signature

Of the 36 genes selected by at least one method, 8 were selected by exactly two methods and none by all three (**Figure** 7). The absence of a three-way intersection reflects the known instability of SVM-RFE under near-perfect separability, which is why a bootstrap-stability SVM set was used for the consensus rather than the minimal RFE solution. The two-method core, selected by both LASSO and RF (e.g., MMP1, RFC4, COL1A1, SIM2, MCM2, CKS1B, PLXNA1, IGF2BP2), formed the prioritised backbone of the signature.

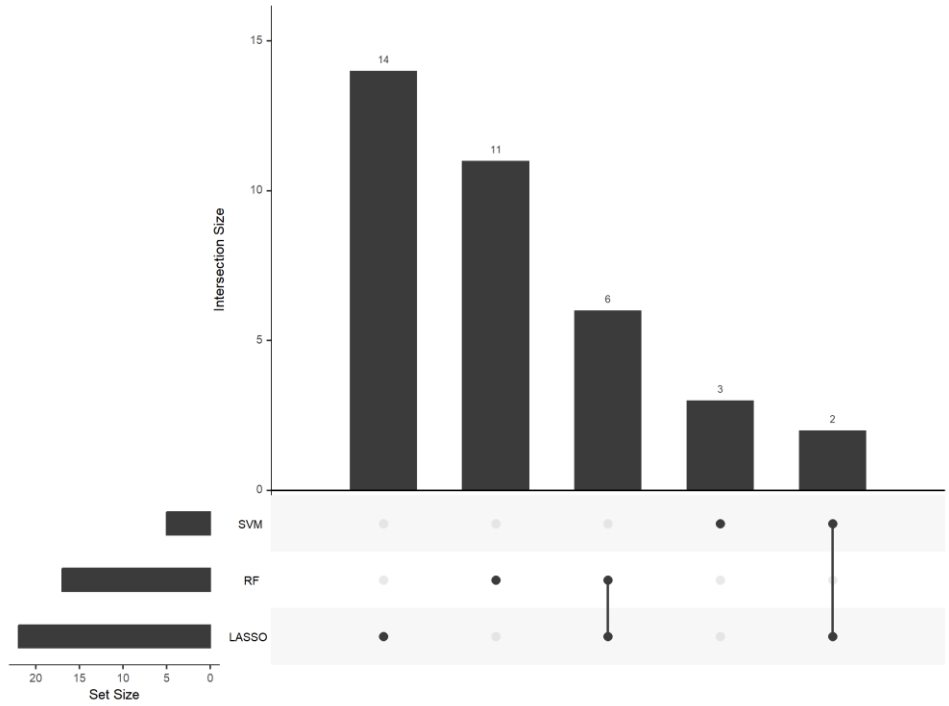


Figure 7. Consensus feature selection: UpSet view of genes selected by LASSO, Random Forest, and SVM (stability) and their intersections.

3.5 Compact panels and model comparison

Repeated cross-validation showed that discriminative performance plateaued at the smallest panel: the 5-gene Logistic Regression panel already attained AUC = 0.998 ± 0.004, accuracy 0.966, sensitivity 0.982, specificity 0.950, and F1 0.967 (**Figure**s 8–9). Logistic Regression was non-inferior to RF, SVM, and XGBoost on every metric and was preferred for its simplicity and interpretability (Table 2).

Table 2. Internal repeated-CV performance of the 5-gene panel (mean ± SD).

Model	AUC	Accuracy	Sensitivity	Specificity	F1
Logistic Regression	0.998 ± 0.004	0.966 ± 0.028	0.982 ± 0.035	0.950 ± 0.056	0.967 ± 0.026
Random Forest	0.992 ± 0.011	0.950 ± 0.035	0.970 ± 0.049	0.930 ± 0.066	0.951 ± 0.033
SVM	0.999 ± 0.004	0.962 ± 0.035	0.977 ± 0.041	0.947 ± 0.067	0.963 ± 0.034
XGBoost	0.990 ± 0.014	0.938 ± 0.050	0.932 ± 0.082	0.945 ± 0.069	0.937 ± 0.052

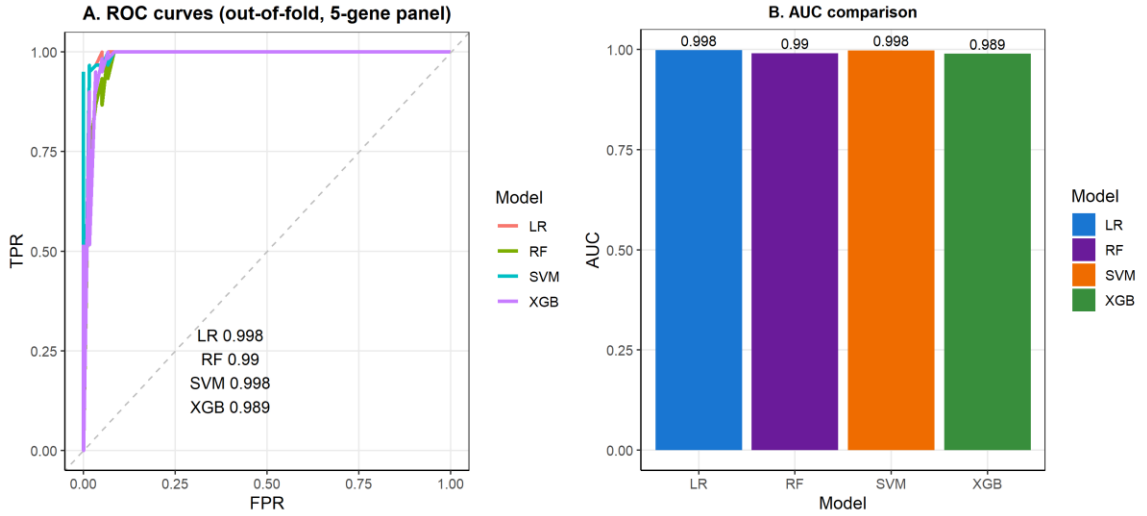


Figure 8. Model comparison: out-of-fold ROC curves (A) and AUC comparison (B) for Logistic Regression, Random Forest, SVM, and XGBoost on the five-gene panel.

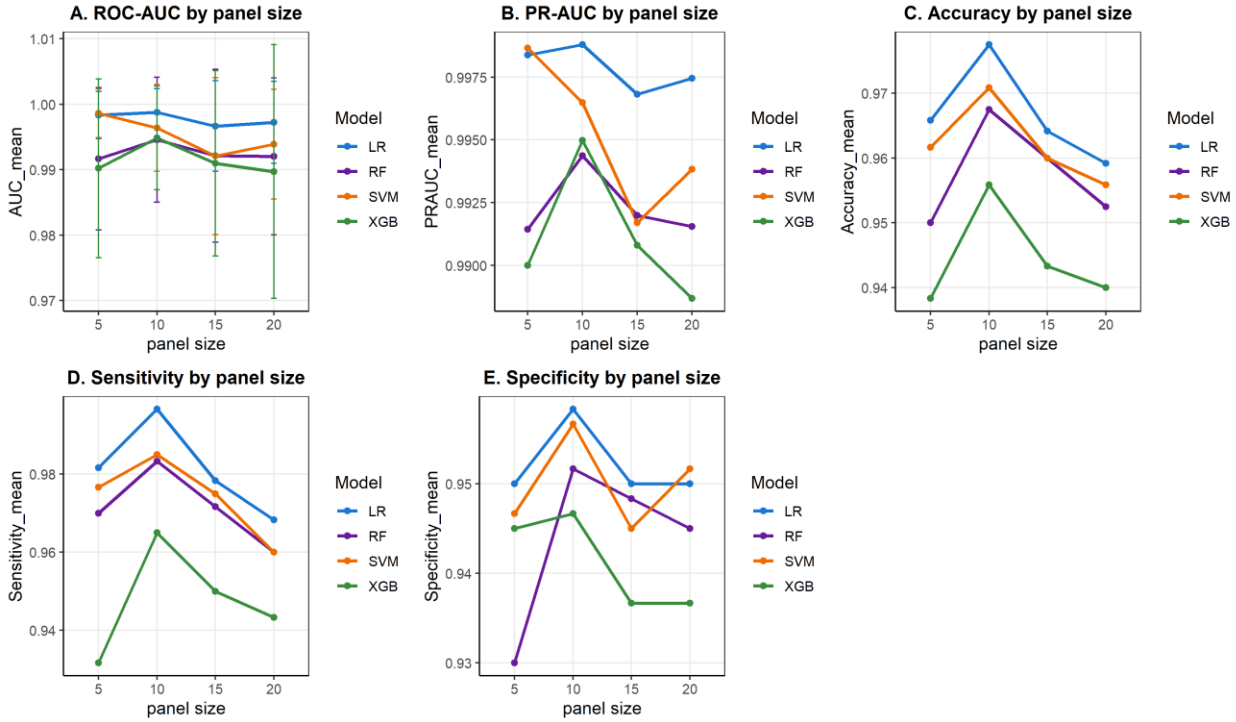


Figure 9. Compact-panel comparison across panel sizes (5/10/15/20) for ROC-AUC, PR-AUC, accuracy, sensitivity, and specificity.

3.6 Final frozen classifier

The final model was a ridge-penalised Logistic Regression on a five-gene panel: SIM2, RFC4, COL1A1, MMP1, CST1 (**Figure**s 10–11). Standardised coefficients and per-standard-deviation odds ratios are given in Table 3. SIM2 carried a negative coefficient (down-regulated in tumour), whereas RFC4, COL1A1, MMP1, and CST1 carried positive coefficients (up-regulated in tumour), consistent with their biological annotation.

Table 3. Frozen five-gene signature (standardised Logistic Regression).

Gene	Standardised coefficient	OR per SD	Direction in tumour
SIM2	−0.824	0.439	Down
RFC4	+1.079	2.943	Up
COL1A1	+0.639	1.894	Up
MMP1	+0.712	2.039	Up
CST1	+0.521	1.684	Up

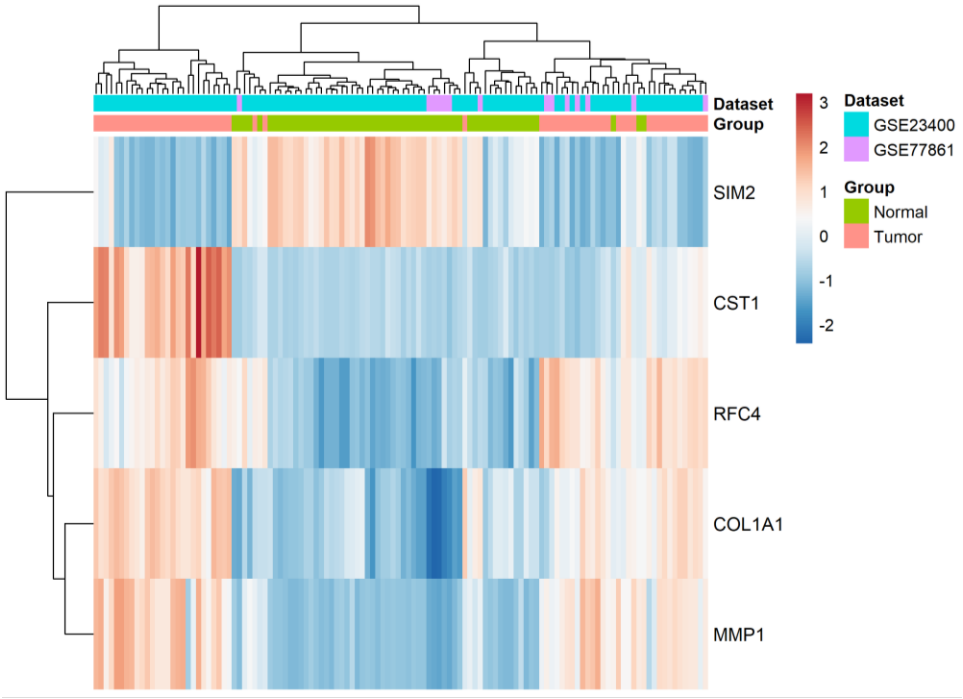


Figure 10. Final-signature expression heatmap across discovery samples (rows = 5 genes, columns = samples), annotated by tumour/normal group and dataset.

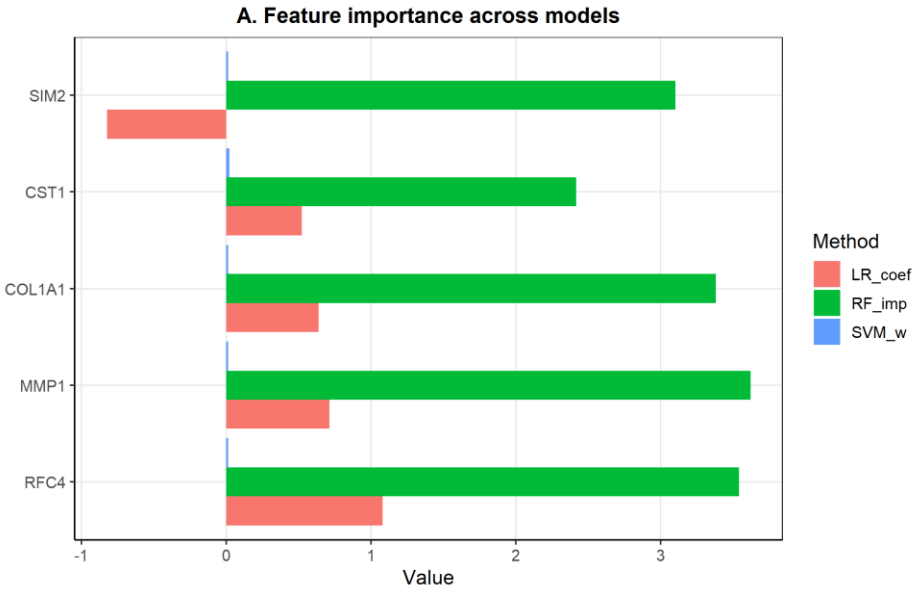


Figure 11. Feature importance/interpretation of the five final genes across Logistic Regression coefficients, Random Forest importance, and SVM weights.

3.7 Independent validation (GSE20347)

Applying the frozen model to GSE20347 yielded ROC-AUC = 1.00 (95% CI 1.00–1.00) and PR-AUC = 1.00 (**Figure** 12). At the default 0.5 threshold the confusion matrix was: 15 true tumours correctly classified, 2 tumours misclassified as normal, and all 17 normals correctly classified (Table 4). This corresponds to accuracy 0.941, sensitivity 0.882, specificity 1.00, precision 1.00, F1 0.938, and balanced accuracy 0.941. The two misclassified tumours fell at intermediate predicted probabilities but above all normal samples, explaining the perfect ranking (AUC = 1.00) despite the two errors at threshold 0.5.

Table 4. Independent validation performance on GSE20347 (n = 34).

Metric	Value
ROC-AUC (95% CI)	1.000 (1.000–1.000)
PR-AUC	1.000
Accuracy	0.941
Sensitivity	0.882
Specificity	1.000
Precision	1.000
F1	0.938
Balanced accuracy	0.941
Confusion (Pred T / Pred N)	Actual T: 15 / 2; Actual N: 0 / 17

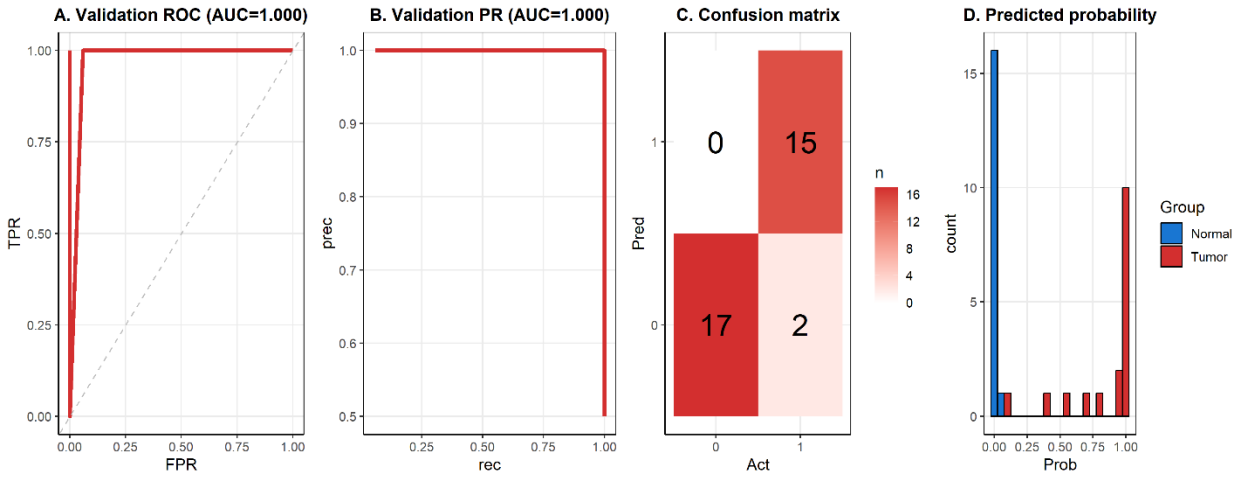


Figure 12. Independent validation on GSE20347: ROC curve (A), PR curve (B), confusion matrix (C), and predicted-probability distribution (D).

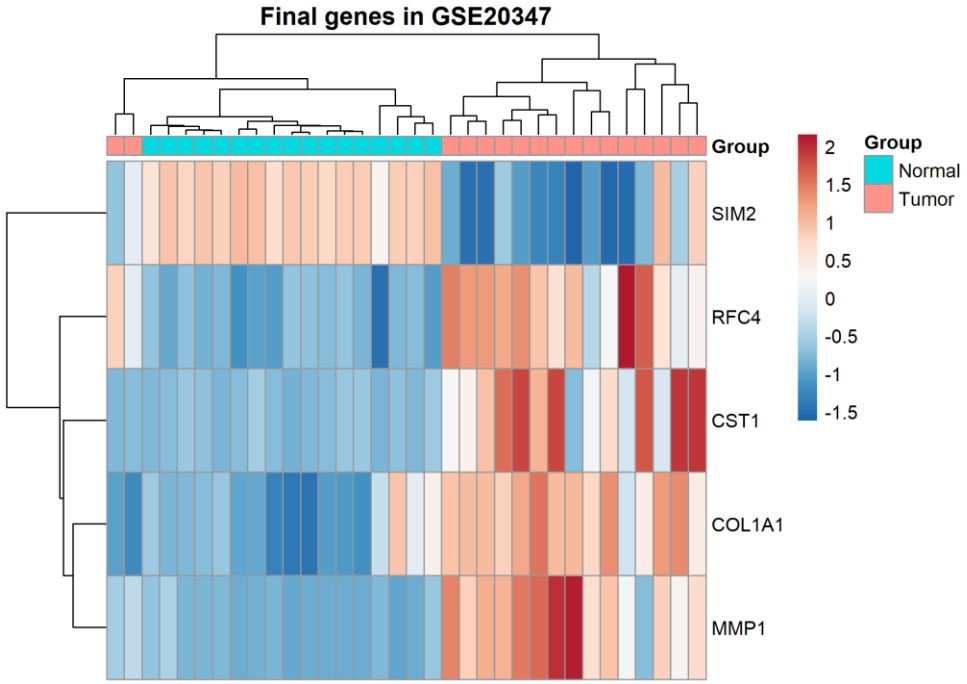


Figure 13. Expression heatmap of the five final genes across GSE20347 samples, annotated by tumour/normal group.

3.8 Individual biomarker validation

Each of the five genes was individually validated in GSE20347 (**Figure** 14). All showed highly significant tumour-vs-normal differences ($p < 0.001$) with single-gene AUCs of 0.89–1.00 (Table 5), confirming that the signature's performance rests on multiple independently informative transcripts rather than a single spurious feature.

Table 5. Per-gene validation in GSE20347.

Gene	log2 FC (T−N)	p-value	Single-gene AUC
SIM2	−2.59	<0.001	0.927
RFC4	+2.17	<0.001	1.000
COL1A1	+3.20	<0.001	0.886
MMP1	+5.06	<0.001	0.986
CST1	+2.20	<0.001	0.900

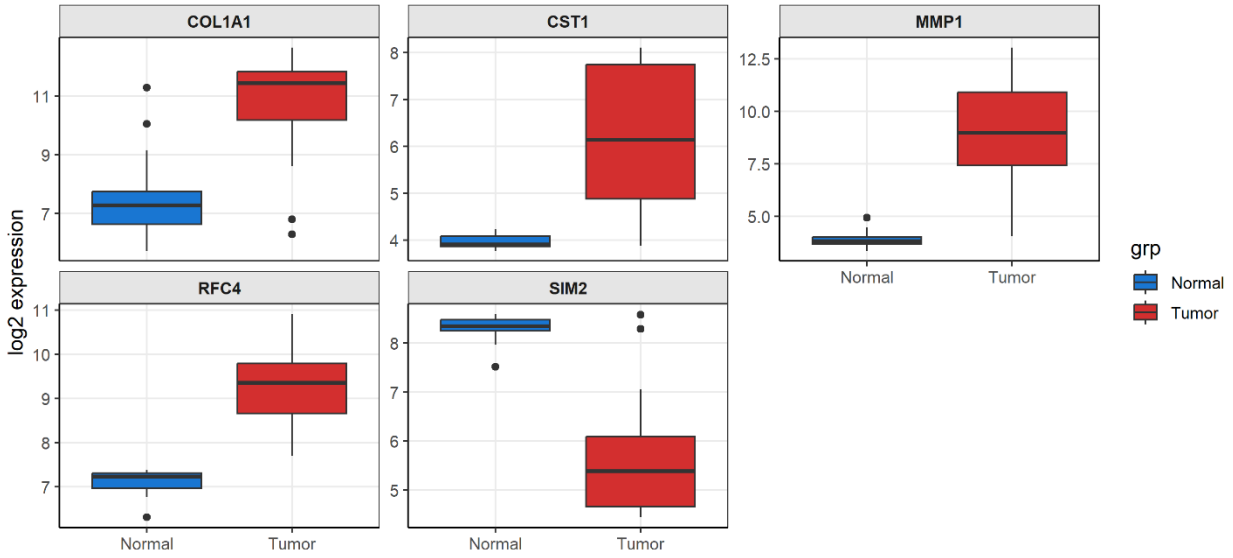


Figure 14. Individual biomarker validation in GSE20347: per-gene tumour vs normal expression (boxplots) with effect size, p-value, and single-gene AUC.

4. Discussion

4.1 Biological interpretation of the five-gene signature

The frozen signature is biologically coherent with the established biology of ESCC and with generic hallmarks of malignant transformation [2,3]. RFC4 (Replication Factor C subunit 4) is a core component of the DNA-replication clamp-loader complex that loads proliferating cell nuclear antigen (PCNA) onto DNA in an ATP-dependent manner, and its overexpression has been directly demonstrated to track proliferative activity and tumour progression in colorectal, nasopharyngeal, and oral squamous carcinomas, where RFC4 knockdown induces S- or G2/M-phase cell-cycle arrest and suppresses proliferation both in vitro and in vivo [15]. Its selection here as one of the two most strongly weighted genes in the ridge model (standardised coefficient +1.079, OR per SD = 2.94) is therefore consistent with the high mitotic index that typifies ESCC.

MMP1 (matrix metalloproteinase 1, interstitial collagenase) was the most strongly up-regulated gene in the signature (log2FC +5.06 in independent validation) and is among the best-characterised mediators of extracellular-matrix degradation and tumour invasion in ESCC specifically: MMP1 expression is markedly increased in ESCC relative to non-tumorous tissue, correlates with lymph-node metastasis, microvessel density, and advanced TNM stage, and independently predicts overall survival in cohorts exceeding 600

patients, with mechanistic work showing that MMP1 promotes ESCC growth and metastasis via activation of the PI3K/AKT pathway [16]. COL1A1 (collagen type I alpha 1) reflects desmoplastic and stromal remodelling of the tumour microenvironment, a process closely coupled to MMP-mediated matrix turnover and consistent with the collagen-organisation programme also identified at the network level in the companion consensus-DEG analysis of this dataset [12]. CST1 (cystatin SN), a secreted cysteine-protease inhibitor, has been directly implicated in ESCC: CST1 is among the most strongly overexpressed genes in ESCC expression profiles, promotes epithelial-mesenchymal transition and metastatic capacity by interacting with MMP2 in the tumour microenvironment, and shows diagnostic value in serum (AUC ≈ 0.97 in a dedicated ESCC cohort), supporting its candidacy as both a tissue and a liquid-biopsy marker [17]. SIM2 (Single-minded homolog 2) is the only down-regulated member of the signature; SIM2 is a bHLH-PAS transcription factor whose expression is reported to be reduced in ESCC relative to normal oesophageal tissue, in contrast to its overexpression in prostate cancer, indicating tissue-context-dependent regulation, and functional work in squamous carcinoma of the uterine cervix shows that SIM2 loss removes transcriptional suppression of HIF1A, promoting tumour growth and hypoxic resistance, a plausible mechanistic parallel for its reduced expression in ESCC tumour tissue [18]. Taken together, the signature captures three interlocking hallmarks of ESCC pathobiology: sustained proliferation (RFC4), invasion and ECM remodelling (MMP1, COL1A1, CST1), and loss of a candidate growth-constraining transcriptional regulator (SIM2), rather than an arbitrary statistical artefact, which strengthens confidence that the classifier is exploiting genuine tumour biology rather than a dataset-specific technical signature.

4.2 Relation to prior ESCC biomarker efforts

Prior ESCC transcriptomic studies have typically reported large signatures (tens to hundreds of genes) or single-cohort classifiers that were not independently validated on data withheld from every stage of model development [2]. Individual components of the present signature have previously been validated as single-gene markers in ESCC, including MMP1 in immunohistochemical cohorts of over 600 patients [16], CST1 at both the tissue and serum level [17], and RFC4 in the closely related setting of other squamous carcinomas [15], but, to our knowledge, they have not previously been assembled into a single compact, multivariable, cross-platform-validated panel using a leakage-controlled, multi-algorithm consensus procedure. Our study contributes three specific advances over this prior literature: (i) a deliberately compact (five-gene) signature that is practicable for clinical translation via targeted assays such as qPCR or nanoString, rather than requiring whole-transcriptome profiling; (ii) strict leakage control, with GSE20347 completely untouched during discovery, feature selection, model training, and threshold tuning; and (iii) demonstrated generalisation to an independent cohort profiled on a different microarray platform (GPL571 versus the discovery platforms GPL96/GPL570). The near-perfect independent AUC should be interpreted cautiously (Section 4.4) but is consistent with the genuinely large transcriptomic separation between ESCC and normal oesophagus reported by TCGA and other large-scale profiling efforts [2].

4.3 Methodological and statistical rigor

Beyond biological plausibility, several methodological choices strengthen confidence that the reported performance reflects a genuine, generalisable signal rather than an artefact of the analytical pipeline. Within-fold standardisation ensured that no information from validation folds (or, at the final stage, from GSE20347) contaminated the training statistics used for gene scaling, a leakage pathway that is easy to overlook but that can materially inflate cross-validated performance in high-dimensional genomic classification. The use of three algorithmically distinct selectors (penalised regression, tree ensembles, and margin-based backward elimination), combined by consensus rather than by picking the single best-performing method post hoc, reduces the risk that the final panel reflects an idiosyncrasy of one algorithm's inductive bias. Repeated stratified cross-validation (10 repeats × 5 folds) on the discovery cohort provided a stable estimate of internal performance with quantified uncertainty (mean ± SD) rather than a single

train/test split, and the independent-validation ROC-AUC was reported together with a 95% confidence interval computed by the DeLong method [9], the standard nonparametric approach for correlated ROC-curve inference. Comparison across four classifier families (Logistic Regression, Random Forest, SVM, XGBoost [13]) on the identical five-gene panel additionally showed that the choice of final classifier had only marginal impact on discrimination (AUC range 0.990–0.999), indicating that the signal resides predominantly in the gene panel itself rather than in a particular model's flexibility.

4.4 Clinical translation potential

A five-transcript panel is, in principle, compatible with low-cost, rapid assay formats already used in clinical molecular pathology, including multiplexed RT-qPCR or nanoString nCounter panels, neither of which requires whole-transcriptome sequencing infrastructure. This is particularly relevant for ESCC, which disproportionately affects regions with constrained diagnostic resources. Because CST1 is a secreted protein detectable in serum with favourable diagnostic accuracy in dedicated ESCC cohorts [17], and MMP1 is likewise released into the circulation by tumour cells [16], there is a plausible, literature-supported route toward a partially non-invasive or minimally invasive version of this signature, complementing tissue-based confirmation. We emphasise, however, that the present study evaluates tissue transcript levels in a discrimination (tumour-vs-normal) framework using retrospective microarray cohorts; it does not establish diagnostic performance in a screening population, nor prognostic or predictive value for treatment response, both of which require dedicated prospective study designs (Section 4.6).

4.5 Strengths

The analysis is grounded in verified metadata, uses a common gene universe to ensure identical feature representation, applies within-fold standardisation, and separates feature discovery, model development, internal validation, and independent validation into distinct stages. Multiple complementary selectors (LASSO, RF, SVM-RFE) and multiple classifier families were compared, and the final choice favoured interpretability without sacrificing performance. All results are fully reproducible from the supplied scripts and fixed seeds.

4.6 Limitations

Several limitations should be acknowledged. (1) Dual-platform discovery. GSE23400 (HG-U133A) and GSE77861 (HG-U133_Plus_2) derive from different microarray platforms; although both are RMA-log2 and were mapped to a common gene universe with per-gene standardisation, residual platform-specific mean shifts cannot be fully excluded. (2) Small GSE77861 contribution. With only 14 samples, GSE77861 adds limited statistical power and is a possible source of platform batch. (3) Near-perfect separability. The ESCC-vs-normal signal is so strong that SVM-RFE's minimal solution is unstable; we mitigated this with bootstrap-stability selection but note that perfect separation limits the value of probability calibration and threshold tuning. (4) Single validation cohort. GSE20347 (n = 34) is an independent but still retrospective GEO dataset; further validation on prospective, ideally non-microarray (RNA-seq) cohorts is required, and validation in additional independent series with different demographic and etiological profiles (e.g., alcohol/tobacco- versus nutritional-risk-factor-predominant populations) would strengthen generalisability claims. (5) Associations, not causality. The signature identifies discriminative transcripts; functional causation remains to be established experimentally, for example via CRISPR-based knockdown/overexpression assays and correlation with clinical outcome in prospectively collected specimens. (6) Absence of clinical and staging covariates. None of the three GEO series provide detailed staging, treatment, or outcome metadata, so the present work is confined to tumour-versus-normal discrimination and cannot speak to prognostic or predictive utility.

5. Conclusions

Interpretable machine learning identified a compact five-gene transcriptomic signature (SIM2, RFC4, COL1A1, MMP1, CST1) that distinguishes ESCC from normal esophageal tissue with near-perfect discrimination in a strictly held-out independent cohort profiled on a different microarray platform from the discovery data. The signature is small, biologically interpretable, and reproducible, satisfying the study's goal of a compact, interpretable classifier rather than a maximally accurate black box. Each gene has independent, literature-supported roles in ESCC or closely related squamous malignancies, spanning DNA-replication control, matrix-degrading invasion, stromal remodelling, and a candidate hypoxia-linked tumour-suppressive transcription factor, lending biological credibility to a signature derived by purely data-driven means.

From a translational standpoint, the combination of a small transcript count, cross-platform reproducibility, and mechanistic plausibility positions this panel as a reasonable candidate for the next stage of biomarker development: prospective, multi-centre validation using targeted assay platforms (qPCR or nanoString) in cohorts with linked clinical staging and outcome data, ideally spanning the geographic and etiological diversity of ESCC. Extension to RNA-sequencing-based cohorts would additionally test robustness beyond the microarray platforms used here, and integration with orthogonal data types, including protein-level validation (immunohistochemistry, serum ELISA for the secreted components CST1 and MMP1), DNA-methylation status, and single-cell transcriptomic deconvolution of the tumour microenvironment, would help clarify whether the five genes act as passive markers of the malignant and stromal compartments or as functionally consequential drivers amenable to therapeutic targeting. Until such prospective and functional validation is complete, the signature should be regarded as a well-supported research-grade classifier rather than a clinically deployable diagnostic test.

Data and Code Availability

GEO accessions: GSE23400, GSE77861, GSE20347 [10]. This study used only publicly available, de-identified data; no new ethical approval was required.

Declarations

Funding

No dedicated external funding was received for this work.

Conflict of interest

The authors declare no competing interests.

Ethics statement

This study used exclusively de-identified, publicly available microarray data deposited in the Gene Expression Omnibus (GSE23400, GSE77861, GSE20347); no new human-subjects data were collected, and separate ethical approval was therefore not required.

REFERENCES

[1] Sung, H., Ferlay, J., Siegel, R. L., et al. (2021). Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: A Cancer Journal for Clinicians, 71(3), 209–249.

[2] The Cancer Genome Atlas Research Network. (2017). Integrated genomic characterization of oesophageal carcinoma. Nature Genetics, 49(4), 476–485.

[3] Hanahan, D., & Weinberg, R. A. (2011). Hallmarks of cancer: the next generation. Cell, 144(5), 646–674.

[4] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society Series B, 58(1), 267–288.

[5] Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. Journal of Statistical Software, 33(1), 1–22.

[6] Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32.

[7] Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. Machine Learning, 46(1–3), 389–422.
[8] Kuhn, M. (2008). Building predictive models in R using the caret package. Journal of Statistical Software, 28(5), 1–26.
[9] DeLong, E. R., DeLong, D. M., & Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics, 44(3), 837–845.
[10] Barrett, T., Wilhite, S. E., Ledoux, P., et al. (2013). NCBI GEO: archive for functional genomics data sets — update. Nucleic Acids Research, 41(D1), D991–D995.
[11] Irizarry, R. A., Hobbs, B., Collin, F., et al. (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics, 4(2), 249–264.
[12] Ritchie, M. E., Phipson, B., Wu, D., et al. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Research, 43(7), e47.
[13] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785–794).
[14] Robin, X., Turck, N., Hainard, A., et al. (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics, 12, 77.
[15] Xiang, J., Fang, L., Luo, Y., Yang, Z., Liao, Y., Cui, J., Huang, M., Yang, Z., Huang, Y., Fan, X., Wang, H., Wang, L., Peng, J., & Wang, J. (2014). Levels of human replication factor C4, a clamp loader, correlate with tumor progression and predict the prognosis for colorectal cancer. Journal of Translational Medicine, 12, 320.
[16] Liu, M., Hu, Y., Zhang, M.-F., Luo, K.-J., Xie, X.-Y., Wen, J., Fu, J.-H., & Yang, H. (2016). MMP1 promotes tumor growth and metastasis in esophageal squamous cell carcinoma. Cancer Letters, 377, 97–104.
[17] Luo, F.-F., Wang, J., Zhang, Z.-F., Lin, S.-T., Huang, T.-J., Liu, B.-Q., Fan, M.-L., Peng, L.-X., Zheng, S.-T., Yang, C.-F., & Huang, B.-J. (2023). SPI1-mediated upregulation of the CST1 gene as an independent poor prognostic factor accelerates metastasis in esophageal squamous cell carcinoma (ESCC) by interacting with MMP2. Frontiers in Bioscience (Landmark Edition), 28(9), 212.
[18] Nakamura, K., Komatsu, M., Chiwaki, F., Takeda, T., Kobayashi, Y., Banno, K., Aoki, D., Yoshida, T., & Sasaki, H. (2017). SIM2l attenuates resistance to hypoxia and tumor growth by transcriptional suppression of HIF1A in uterine cervical squamous cell carcinoma. Scientific Reports, 7(1), 14574.