

Cross-Cohort Robustness of Machine-Learning Gene-Expression Biomarkers for Lung Adenocarcinoma and Squamous Cell Carcinoma

Areen Jain

Gretchen Whitney High School

Abstract

Gene-expression classifiers can separate lung adenocarcinoma (LUAD) from lung squamous cell carcinoma (LUSC), but high apparent accuracy does not guarantee reproducible biomarkers. We developed a locked, cross-cohort analysis that combined nested cross-validation, Elastic-Net feature selection, resampling-based stability analysis, compact-panel evaluation, and a processing-to-platform shift ladder. The discovery cohort was GSE37745 (106 LUAD and 66 LUSC samples); the prespecified primary external cohort was GSE50081 (127 LUAD and 42 LUSC samples). A frozen Elastic-Net model achieved an internal out-of-fold ROC-AUC of 0.985 and an external ROC-AUC of 0.909 (DeLong 95% CI, 0.849–0.970). The external AUC degradation of 0.076 exceeded the prespecified 0.05 tolerance, showing that strong discrimination did not fully transfer without loss. Of 83 model-selected genes, 58 were selected in at least 80% of 100 stratified subsamples. Selection frequency was positively associated with external effect size (Spearman $\rho = 0.385$, $p < 0.001$), and 10- and 20-gene panels retained 98.0% and 98.5% of the full-model external AUC, respectively. Performance remained high across several microarray cohorts but fell to near chance in TCGA RNA-seq data (AUC 0.511), emphasizing platform shift as a major boundary of generalization. These results support stability and external validation as essential complements to internal accuracy in transcriptomic biomarker studies.

Keywords: lung adenocarcinoma; lung squamous cell carcinoma; gene expression; Elastic Net; biomarker stability; external validation; cross-cohort robustness

1 Introduction

1.1 Background

Lung cancer is biologically heterogeneous, and histologic subtype is clinically relevant for diagnosis, treatment selection, and interpretation of molecular studies. Lung adenocarcinoma (LUAD) and lung squamous cell carcinoma (LUSC) have distinct transcriptional programs, making them a useful test case for evaluating whether machine-learning biomarkers capture reproducible biology or merely exploit cohort-specific structure. Public gene-expression datasets make this question tractable, but they also create a high-dimensional, small-sample setting in which thousands of measured genes compete for relatively few labeled samples.

The central methodological risk is optimistic performance. If feature selection or preprocessing uses the complete dataset before cross-validation, information from held-out samples can leak into the model and inflate accuracy [1,2]. The problem is especially serious for gene lists: many nearly equivalent signatures can predict the same endpoint, while the particular genes selected can change substantially with the sampled patients [3,4]. Therefore, a useful biomarker study should report not only discrimination but also locked external performance, selection stability, calibration, and sensitivity to distributional shift.

1.2 Literature Review

Prior transcriptomic studies have demonstrated that LUAD and LUSC can be separated with high accuracy, including analyses that transferred histology signatures between array cohorts and TCGA [5,6]. However, the broader biomarker literature shows that internal performance is often overstated when selection is performed outside the resampling loop [1,2]. Reviews and empirical studies have also found that published signatures may fail to reproduce across resplits or independent datasets, and that random or proliferation-related gene sets can appear competitive without representing a stable disease mechanism [7,8].

Cross-cohort transfer is further complicated by batch and platform effects. Expression studies may differ in sample processing, normalization, patient population, geography, array platform, and sequencing technology. Joint correction can itself leak information when validation samples influence the transformation, so independent-cohort normalization and explicit shift diagnostics are safer for a locked evaluation [9–11]. For high-dimensional classification, simple regularized linear methods remain strong baselines when evaluated fairly with nested cross-validation [12,13].

Stability selection provides a complementary view of reproducibility by measuring how

often features are selected under perturbations of the training data [14,15]. The present study extends that idea by asking whether stable genes have larger external effects and whether stability-ranked panels preserve external performance. This design follows the reporting logic of TRIPOD and TRIPOD+AI, emphasizing prespecification, locking, external validation, and transparent limitations [16,17].

2 Data Description

The discovery dataset was GSE37745, an Affymetrix GPL570 cohort with 172 usable samples: 106 LUAD and 66 LUSC. GSE50081, also GPL570, was reserved as the primary external cohort and contained 169 usable samples: 127 LUAD and 42 LUSC. Secondary cohorts were GSE43580 ($n = 150$), GSE8894 ($n = 138$), GSE30219 ($n = 146$), and GSE41271 ($n = 263$). TCGA lung RNA-seq data were evaluated as the final, cross-platform step of the shift ladder.

Expression matrices were obtained from GEO or the TCGA Xena resource using accession-specific files. The TCGA expression representation and preprocessing were recorded separately from the microarray preprocessing. RNA-seq values were transformed and aligned to the frozen discovery-gene list before gene-wise standardization. Genes absent from TCGA were assigned a standardized value of zero after alignment. Because this procedure involves a different measurement platform and substantial missing-gene overlap, TCGA was treated as a cross-platform stress test rather than as directly equivalent validation of the microarray model. Probe sets were mapped to HGNC symbols using platform annotations; multi-mapping probes were removed and duplicate symbols were collapsed by maximum mean expression. A cohort was $\log_2$ transformed when its median raw intensity exceeded 50. No samples were excluded for outlier status by default. The final discovery matrix contained 21,655 harmonized genes; GSE41271 contained 16,664 discovery genes after platform intersection.

3 Theoretical Framework

3.1 Standardization and the prediction function

For gene g and sample i , within-cohort standardization was defined as

$$z_{ig} = \frac{x_{ig} - \mu_g}{\sigma_g + \epsilon}, \quad (1)$$

Table 1: Cohort composition used in the analysis.

Cohort	Usable samples	LUAD	LUSC
GSE37745 (discovery)	172	106	66
GSE50081 (primary external)	169	127	42
GSE43580	150	77	73
GSE8894	138	63	75
GSE30219	146	85	61
GSE41271	263	183	80
TCGA RNA-seq	1,017	502	515

where μ_g and σ_g were estimated without labels in the relevant training or scoring cohort and ϵ prevents division by zero. The classifier maps a standardized expression vector $\mathbf{z}_i$ to a probability

$$p_i = \Pr(y_i = 1 \mid \mathbf{z}_i) = \sigma(\beta_0 + \mathbf{z}_i^\top \boldsymbol{\beta}), \quad \sigma(t) = \frac{1}{1 + e^{-t}}, \quad (2)$$

where $y_i = 1$ denotes LUSC and $y_i = 0$ denotes LUAD.

3.2 Elastic-Net regularization

The primary model estimates the intercept and coefficient vector by minimizing the penalized logistic loss:

$$(\hat{\beta}_0, \hat{\boldsymbol{\beta}}) = \arg \min_{\beta_0, \boldsymbol{\beta}} \left\{ -\frac{1}{n} \sum_{i=1}^n [y_i \log p_i + (1 - y_i) \log(1 - p_i)] + \lambda \left[\alpha \|\boldsymbol{\beta}\|_1 + \frac{1 - \alpha}{2} \|\boldsymbol{\beta}\|_2^2 \right] \right\}, \quad (3)$$

The ℓ_1 term encourages sparsity, while the ℓ_2 term stabilizes correlated gene groups. Hyperparameters were selected inside inner cross-validation, so the outer validation folds did not influence feature selection or tuning.

3.3 Discrimination, calibration, and stability

The ROC-AUC is the probability that a randomly selected LUSC sample receives a higher score than a randomly selected LUAD sample, with half-credit assigned to ties:

$$\text{AUC} = \Pr(p^+ > p^-) + \frac{1}{2} \Pr(p^+ = p^-). \quad (4)$$

At the locked Youden threshold τ , sensitivity and specificity were computed from the confusion matrix. Calibration was summarized by the slope and intercept from a logistic

calibration model. For feature g , selection frequency over B perturbations was

$$\pi_g = \frac{1}{B} \sum_{b=1}^B I(g \in S_b), \quad (3)$$

where S_b is the selected set in subsample b . Pairwise set agreement was measured by the Jaccard index

$$J(S_a, S_b) = \frac{|S_a \cap S_b|}{|S_a \cup S_b|}. \quad (4)$$

Finally, pathway over-representation used the hypergeometric tail probability for observing at least k selected genes in a pathway of size K within a measured background of size N :

$$P(X \geq k) = \sum_{j=k}^{\min(K,n)} \frac{\binom{K}{j} \binom{N-K}{n-j}}{\binom{N}{n}}, \quad (5)$$

with Benjamini–Hochberg false-discovery-rate correction.

4 Methodology

4.1 Locked analysis plan and preprocessing

The analysis plan was written before model freezing and assigned GSE37745 as discovery and GSE50081 as primary external validation. All data-dependent steps were placed inside resampling folds: top-variance filtering, gene-wise scaling, hyperparameter selection, and model fitting. The discovery analysis used 25 outer folds generated by five repeats of five-fold stratified cross-validation and four inner folds for tuning. For the stability and compact-panel analyses, the candidate feature universe consisted of the 5,000 genes selected according to the prespecified variance-filtering procedure. The filtering rule did not use outcome labels. The origin of this candidate set is distinct from the full 21,655-gene measured background used for pathway analysis. Models included L2-regularized logistic regression, Elastic Net, random forest, and XGBoost; Elastic Net was the prespecified primary model because it combines interpretability and regularization.

After nested cross-validation, the consensus Elastic-Net hyperparameters were $C = 0.1$ and $\alpha = 0.5$. The final model selected 83 genes. The classification threshold was locked using Youden’s J statistic on discovery out-of-fold predictions ($\tau = 0.3376$). The model and threshold were then frozen before scoring GSE50081 or any shift-ladder cohort.

4.2 External validation and shift ladder

External scoring used label-free, per-cohort z-scoring. Missing genes were set to zero after scaling; this occurred for 646 genes in GSE41271 and 636 genes in TCGA. The primary endpoint was ROC-AUC on GSE50081 with a DeLong confidence interval [18]. Secondary metrics were PR-AUC, balanced accuracy, sensitivity, specificity, F1 score, Brier score, and calibration slope/intercept; PR-AUC was included because it is informative under class imbalance [19]. The shift ladder was evaluated without threshold or hyperparameter retuning. External cohorts were standardized using cohort-specific gene means and standard deviations calculated without using outcome labels. Therefore, the external results represent transfer under cohort-level standardization rather than deployment to an isolated individual sample without a reference batch.

4.3 Stability and compact panels

Feature stability was evaluated using 100 stratified subsamples, each containing 80% of the discovery cohort. Elastic Net was refit on each subsample. Genes with selection frequency at least 0.80 were labeled stable. Stability-ranked panels of 3, 5, 10, 20, 30, 50, 83, 100, and 5,000 genes were compared with size-matched random panels. The compact-panel criterion was prespecified as retention of at least 95% of full-model external AUC.

4.4 Pathway Analysis

The 58 stable genes were tested against the 21,655 measured discovery genes using Enrichr with the MSigDB Hallmark 2020, GO Biological Process 2021, KEGG 2021 Human, and Reactome 2022 libraries. Over-representation analysis used the measured discovery genes as the background, and p -values were adjusted using the Benjamini–Hochberg procedure. Terms with adjusted $p < 0.05$ were considered significant.

Feature stability was evaluated using 100 stratified subsamples containing 80% of the discovery cohort. Genes selected in at least 80% of subsamples were classified as stable. The analysis used random seed 20260921. Full analysis plans, model artifacts, tables, figures, and code are provided as supplemental materials.

5 Results

5.1 Internal model performance

All four candidate models performed strongly in discovery cross-validation (Table 2). Elastic Net achieved mean ROC-AUC 0.983 across outer folds, with mean PR-AUC 0.982 and balanced accuracy 0.980. Random forest had the highest mean ROC-AUC at 0.988, but the prespecified Elastic-Net model was retained because the study prioritized a sparse, interpretable frozen signature. The internal performance comparison is shown in Figure 1. Table 2 reports mean ROC-AUC across the 25 outer validation folds, whereas Figure 1 reports pooled out-of-fold ROC-AUC obtained by combining predictions across all outer folds. These values can differ because mean-fold AUC weights each fold equally, while pooled out-of-fold AUC weights each observation equally. The internal-to-external AUC degradation was calculated using the pooled out-of-fold value. All numerical results and figure labels have been made consistent with these definitions.

Table 2: Nested cross-validation performance in GSE37745. Values are means across outer folds; intervals are the 2.5th and 97.5th percentiles.

Model	ROC-AUC	PR-AUC	Balanced accuracy
L2 logistic regression	0.987 (0.947–1.000)	0.985	0.979
Elastic Net	0.983 (0.929–1.000)	0.982	0.980
Random forest	0.988 (0.959–1.000)	0.982	0.980
XGBoost	0.983 (0.943–1.000)	0.976	0.971

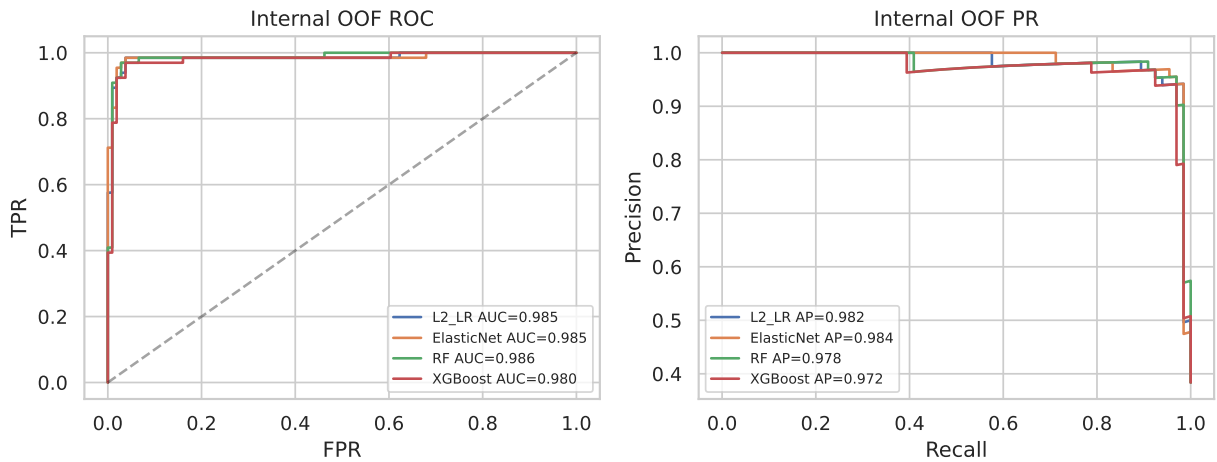


Figure 1: Internal performance of candidate classifiers in nested cross-validation.

5.2 Primary external validation

The frozen Elastic Net achieved ROC-AUC 0.909 in GSE50081 (95% CI, 0.849–0.970), PR-AUC 0.810, balanced accuracy 0.874, sensitivity 0.881, and specificity 0.866. The Brier score was 0.091 and the calibration slope was 0.764, indicating useful discrimination with some departure from ideal calibration. Internal out-of-fold ROC-AUC was 0.985, so the observed degradation was 0.076. The prespecified AUC threshold of 0.90 was met, but the prespecified degradation tolerance below 0.05 was not. External ROC curves and GSE50081 calibration are shown in Figure 2.

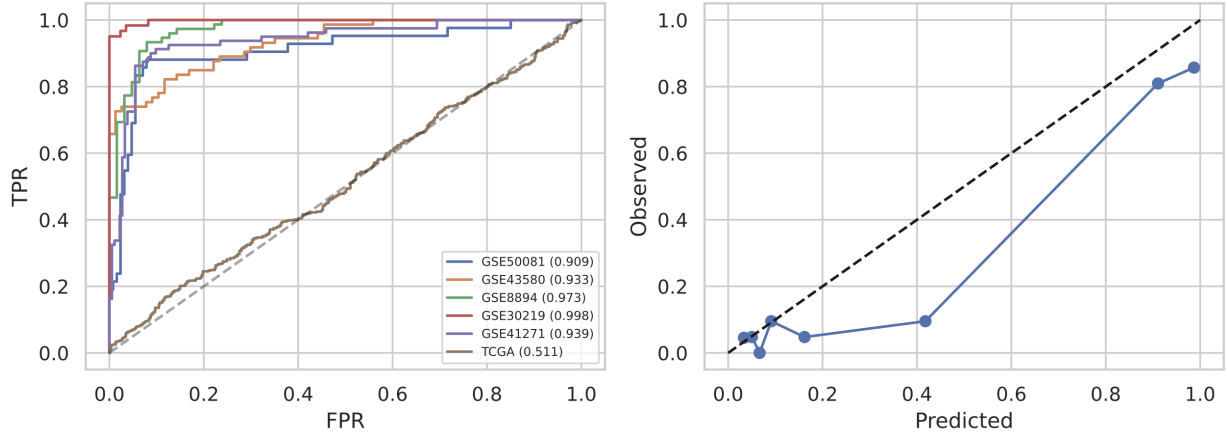


Figure 2: External ROC curves across validation cohorts and calibration in GSE50081. The RNA-seq TCGA curve is close to the diagonal.

5.3 Selection stability and external effects

Across 100 subsamples, 58 genes met the stable threshold of $\pi_g \geq 0.80$. Mean pairwise Jaccard agreement was 0.493, compared with an approximate chance value of 0.014, yielding a chance-corrected stability statistic of 0.486. Stable genes had larger mean absolute external Cohen’s d than unstable genes (1.501 versus 0.482). Selection frequency was positively associated with absolute external effect size ($\rho = 0.385$, $p = 7.76 \times 10^{-177}$), supporting the hypothesis that stability contains information about transferability.

5.4 Compact panels and shift ladder

The full frozen model scored AUC 0.910 in the panel-size analysis. The 10-gene and 20-gene stability-ranked panels achieved AUC 0.891 and 0.896, retaining 98.0% and 98.5% of full-model external AUC. Both met the prespecified 95% retention criterion. Random panels

also performed surprisingly well in this biologically separable task, with mean AUC 0.831 for 10 genes and 0.866 for 20 genes, so compact-panel superiority should not be inferred from discrimination alone. Panel-size performance is shown in Figure 3.

Across the microarray shift ladder, AUCs were 0.933 (GSE43580), 0.973 (GSE8894), 0.998 (GSE30219), and 0.939 (GSE41271). In contrast, TCGA RNA-seq performance was near chance (AUC 0.511; 95% CI, 0.475–0.546), with 636 missing model genes after alignment. This final result marks a sharp platform boundary rather than a smooth loss of accuracy. The relationship between feature stability and external effect size, together with performance across the shift ladder, is shown in Figure 4.

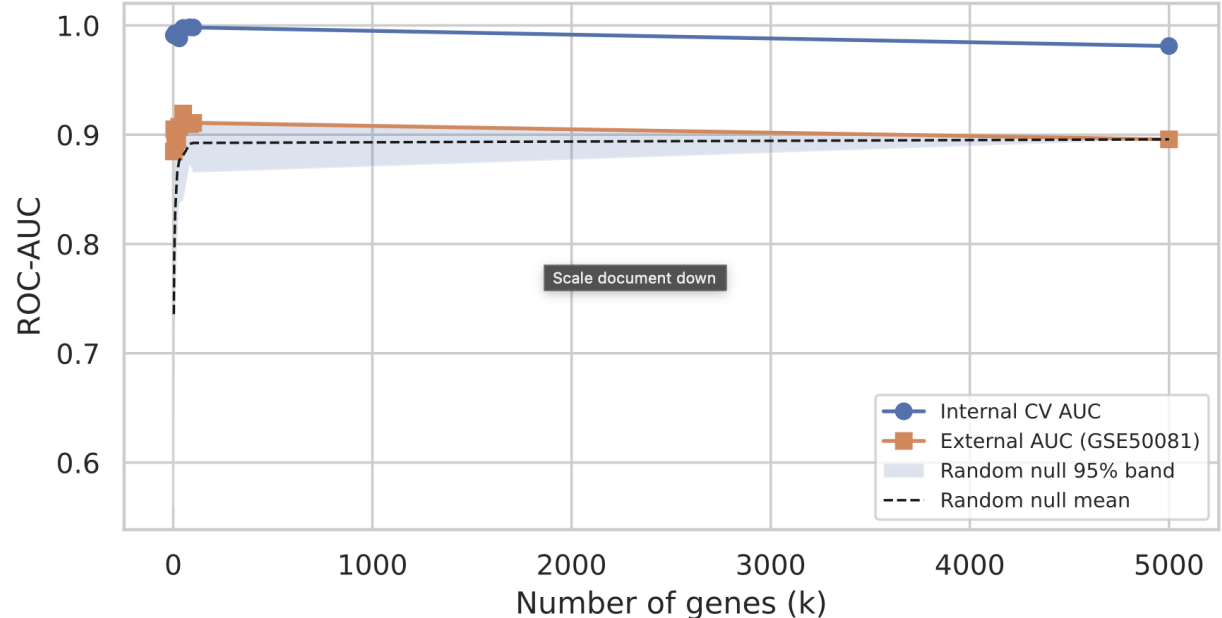


Figure 3: Internal and external AUC as a function of stability-ranked panel size. The shaded region shows the size-matched random-panel null band.

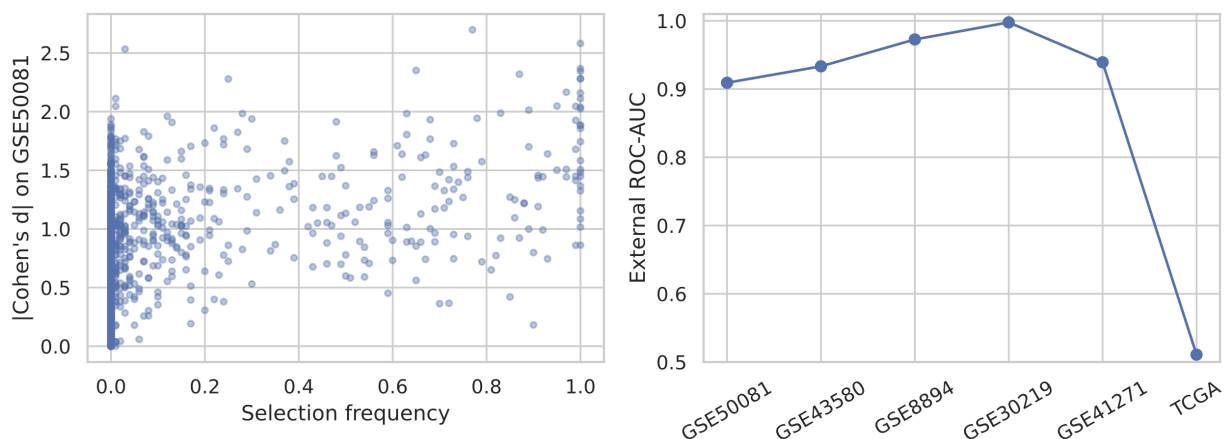


Figure 4: Relationship between selection stability and external effect size, and generalization across the shift ladder. Stability is associated with stronger external effects, while RNA-seq transfer fails.

5.5 Pathway enrichment

Five pathway terms were significant after correction. The strongest were estrogen response late (adjusted $p = 2.58 \times 10^{-4}$; odds ratio 12.93) and keratinization (adjusted $p = 0.0036$; odds ratio 17.17). Additional terms included p53 signaling, estrogen response early, and glycolysis. These results are exploratory because the selected genes were derived from a histology classification task and the measured-gene background is cohort-specific.

6 Discussion

6.1 Findings and Interpretations

This study evaluated transcriptomic biomarker robustness rather than internal accuracy alone. The frozen Elastic-Net model achieved strong discrimination in the primary external cohort, with AUC declining from 0.985 internally to 0.909 externally. Although the prespecified external AUC target was met, the degradation exceeded the tolerance threshold, indicating successful but imperfect transfer. Feature stability provided additional evidence of reproducibility: 58 of 5,000 candidate genes were selected in at least 80% of subsamples, and higher selection frequency was associated with larger external effects [3, 4, 15]. Stability-ranked 10- and 20-gene panels retained most of the full-model AUC, but the strong performance of random panels shows that compact prediction alone does not establish biological specificity. The model generalized across several independent microarray cohorts

but fell to near-chance performance in TCGA RNA-seq data, demonstrating that platform differences and missing genes can substantially limit transfer. The enrichment results involving keratinization, p53 signaling, glycolysis, and estrogen-response pathways were biologically plausible but exploratory because pathway significance does not establish causality.

6.2 Limitations

This study used retrospective public cohorts with histology labels inherited from their original studies rather than centralized pathology review. LUAD-versus-LUSC classification may also be easier than prognosis or treatment-response prediction, and prognostic signatures require different validation standards [23]. The microarray cohorts were not fully exchangeable, while the RNA-seq analysis involved substantial missing-gene overlap after harmonization. Stability was evaluated using 100 rather than 200 subsamples, as determined by the executed notebook. In addition, the analysis does not establish prospective clinical utility, performance in routine pathology workflows, or improvement beyond standard clinical and pathological variables.

6.3 Future Research Direction

Future research should evaluate whether the stable genes provide incremental value beyond pathology, smoking history, tumor purity, stage, and other clinical variables. Cross-platform validation should be treated as a separate modeling problem and should include representative RNA-seq cohorts, platform-aware transformations, and independent locked testing. Prospective studies could also assess calibration, clinical utility, and robustness in smaller biopsy specimens. Finally, biological validation of the stable genes and enriched pathways would help distinguish reproducible histology-associated signals from correlated but noncausal markers.

7 Supplemental Materials

The complete Python code, R Code, dataset, and instructions on how to run the codes are available in the project’s Github Repository for reproducibility and further exploration. The link is: <https://github.com/areenjain09/lung-biomarkers>

8 Acknowledgments

This work was made possible through the guidance of Dr. Olga Korosteleva, Professor of Statistics at California State University, Long Beach. Her mentorship and expertise greatly influenced both the direction of this project and my development as a researcher. I am also grateful to the teachers and mentors who have supported my academic growth over the years. Finally, I sincerely thank my family for their constant encouragement, support, and strength throughout this journey.

References

- [1] Ambroise C, McLachlan GJ. Selection bias in gene extraction on the basis of microarray gene-expression data. *Proc Natl Acad Sci USA*. 2002;99:6562–6566.
- [2] Varma S, Simon R. Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics*. 2006;7:91.
- [3] Ein-Dor L, Kela I, Getz G, et al. Outcome signature genes in breast cancer are not due to chance. *Bioinformatics*. 2005;21:171–177.
- [4] Ein-Dor L, Zuk O, Domany E. Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer. *Proc Natl Acad Sci USA*. 2006;103:5923–5928.
- [5] Girard L, Rodriguez-Canales J, Behrens C, et al. An expression signature as an aid to the histologic classification of non-small cell lung cancer. *Clin Cancer Res*. 2016;22:4880–4889.
- [6] Hou J, Aerts J, den Hamer B, et al. Gene expression-based classification of non-small cell lung carcinomas and survival prediction. *PLoS One*. 2010;5:e10312.
- [7] Michiels S, Koscielny S, Hill C. Prediction of cancer outcome with microarrays: a multiple random validation strategy. *Lancet*. 2005;365:488–492.
- [8] Venet D, Dumont JE, Detours V. Most random gene expression signatures are significantly associated with breast cancer outcome. *PLoS Comput Biol*. 2011;7:e1002240.
- [9] Leek JT, Scharpf RB, Bravo HC, et al. Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat Rev Genet*. 2010;11:733–739.

- [10] McCall MN, Murakami PN, Lukk M, et al. Frozen robust multiarray analysis (fRMA). *Biostatistics*. 2010;11:242–253.
- [11] Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*. 2007;8:118–127.
- [12] Dudoit S, Fridlyand J, Speed TP. Comparison of discrimination methods for the classification of tumors using gene expression data. *J Am Stat Assoc*. 2002;97:77–87.
- [13] Statnikov A, Wang L, Aliferis CF. A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification. *BMC Bioinformatics*. 2008;9:319.
- [14] Meinshausen N, Bühlmann P. Stability selection. *J R Stat Soc Series B*. 2010;72:417–473.
- [15] Nogueira S, Sechidis K, Brown G. On the stability of feature selection algorithms. *J Mach Learn Res*. 2018;18:1–54.
- [16] Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD). *Ann Intern Med*. 2015;162:55–63.
- [17] Collins GS, Dhiman P, Andorfer C, et al. TRIPOD+AI statement. *BMJ*. 2024;385:e078378.
- [18] DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves. *Biometrics*. 1988;44:837–845.
- [19] Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS One*. 2015;10:e0118432.
- [20] Liberzon A, Birger C, Thorvaldsdóttir H, et al. The Molecular Signatures Database hallmark gene set collection. *Cell Syst*. 2015;1:417–425.
- [21] Reimand J, Isik A, Nair S, et al. Pathway enrichment analysis and visualization of omics data using g:Profiler. *Nat Protoc*. 2019;14:482–517.
- [22] Subramanian A, Tamayo P, Mootha VK, et al. Gene set enrichment analysis. *Proc Natl Acad Sci USA*. 2005;102:15545–15550.

- [23] Shedden K, Marchetti F, Knudsen S, et al. Gene expression-based survival prediction in lung adenocarcinoma. *Nat Med*. 2008;14:822–827.