

External Transferability of TCGA-Trained Pan-Cancer Transcriptomic Classifiers Across ICGC and GEO Cohorts

Sungmin Yang*, Hong-Gee Kim**

*Postdoctoral Researcher, Dental Research Institute, Seoul National University, Seoul, Korea

**Professor, Biomedical Knowledge Engineering Laboratory, Seoul National University, Seoul, Korea

[Abstract]

In this paper, we propose a controlled external transferability evaluation framework for pan-cancer transcriptomic classifiers across independent cohorts and measurement platforms. The key idea of this framework is to distinguish internal predictive accuracy from feature stability, external generalization, and preservation of cancer-discriminative transcriptomic structure. We trained and selected models using only TCGA RNA-seq data for five cancer types and directly applied the locked models, without retraining or feature reselection, to independent ICGC RNA-seq and GSE2109 microarray cohorts. Also, we evaluated 48 model configurations and performed diagnostic analyses of calibration, cancer-centroid geometry, measurement resolution, perturbation sensitivity, and biological interpretation. Experimental results showed that models with nearly identical TCGA performance exhibited substantially different external transferability. In particular, the primary model retained high performance in ICGC but declined in GSE2109, whereas ANOVA500+XGBoost(XGB) showed more stable cross-platform performance. Post hoc analyses showed that reduced external performance was accompanied by weaker preservation of cancer-type-specific transcriptomic structure, a pattern consistent with the possibility that shifts in cross-cohort representation contributed to the loss of external generalizability. These findings demonstrate that independent cross-cohort and cross-platform validation is necessary for robust evaluation of transcriptomic cancer classifiers.

► **Key words:** pan-cancer classification, transcriptomic transferability, external validation, domain shift, cross-platform generalization

-
- First Author: Sungmin Yang, Corresponding Author: Hong-Gee Kim
 - *Sungmin Yang (syang90@snu.ac.kr), Dental Research Institute, Seoul National University
 - **Hong-Gee Kim (yj2945@snu.ac.kr), Biomedical Knowledge Engineering Laboratory, Seoul National University
 - Received: 2026. 09. 11, Revised: 2026. 09. 22, Accepted: 2026. 10. 11.

[요 약]

본 논문에서는 독립 코호트와 서로 다른 측정 플랫폼에서 범암 전사체 분류기의 외부 전이 가능성을 평가하기 위한 체계적 검증 프레임워크를 제안한다. 본 프레임워크의 핵심은 내부 예측 정확도와 특징 안정성, 외부 일반화 성능, 그리고 암종 구분에 필요한 전사체 구조의 보존성을 구분하여 평가하는 데 있다. 이를 위해 BRCA, KIRC, OV, PAAD, PRAD의 5개 암종을 대상으로 TCGA RNA-seq에서만 특징선택과 모델 학습을 수행하고, 학습된 모델을 재학습이나 특징 재선택 없이 독립적인 ICGC RNA-seq과 GSE2109 microarray 코호트에 적용하였다. 또한 총 48개 모델 구성을 비교하고, 확률 보정, 암종 중심점 기하구조, 측정 해상도, 교란 민감도 및 생물학적 해석을 포함한 추가 진단 분석을 수행하였다. 실험 결과, TCGA 내부에서 거의 동일한 성능을 보인 모델들이 외부 데이터에서는 뚜렷하게 다른 전이 성능을 나타냈다. 특히 주모델은 ICGC에서 높은 성능을 유지했으나 GSE2109에서 크게 감소한 반면, ANOVA500+XGBoost(XGB)는 보다 안정적인 플랫폼 간 성능을 보였다. 추가적인 사후 분석에서 외부 코호트의 성능 저하는 암종 간 전사체 구분 구조의 보존 감소와 함께 관찰되었으며, 이는 코호트 간 표현 구조의 변화가 외부 일반화 성능 저하에 일부 기여했을 가능성과 일치한다. 이러한 결과는 전사체 기반 암종 분류기의 강건성을 평가하기 위해 독립 코호트와 플랫폼 간 외부 검증이 필수적임을 보여준다.

▶ **주제어:** 범암 분류, 전사체 전이 가능성, 외부 검증, 데이터셋 시프트, 플랫폼 간 일반화

I. Introduction

Gene-expression profiles retain strong tissue-lineage and tumor-associated transcriptional signals and have therefore been widely used for cancer-type and tissue-of-origin classification. The Cancer Genome Atlas (TCGA) established a large and systematically characterized pan-cancer resource that enabled comparative molecular analyses across tumor types [1]. Subsequent studies demonstrated that transcriptome-wide expression profiles can distinguish multiple cancer types with very high accuracy, including primary and metastatic tumors [2,3]. More recent work has extended this direction toward interpretable tissue-of-origin classifiers evaluated in independent GEO datasets and clinical tumor samples [4], while contemporary reviews continue to document rapid expansion of machine-learning and deep-learning approaches for pan-cancer classification [5]. Collectively, these studies demonstrate that cancer transcriptomes contain strong discriminative information. As classification performance within large curated datasets approaches a practical

ceiling, however, further improvements in internal accuracy provide increasingly limited evidence about whether the learned decision rule will remain reliable outside the development environment.

The more difficult problem is therefore external transferability. Internal cross-validation is essential for model development, but training and validation folds remain embedded within largely the same data-generating environment. Cross-study validation provides a more stringent assessment because it evaluates prediction in studies that were not part of model development [6]. This distinction is particularly important for transcriptomic data, in which differences in assay technology, laboratory procedures, preprocessing, and expression representation can alter the numerical structure presented to a classifier. RNA sequencing itself is highly reproducible under controlled conditions [7], but direct comparison with microarray data remains challenging because the two technologies differ in dynamic range, background behavior, transcript and probe representation, and

preprocessing. Several approaches have therefore been developed for cross-platform normalization and integration of gene-expression data [8-11]. Recent systematic evaluation has further shown that normalization strategy can substantially affect machine-learning compatibility between RNA-seq and microarray datasets [11]. Within-sample relative-expression or rank-based representations provide an alternative strategy by reducing dependence on absolute measurement scale, but relative gene ordering can itself be affected by platform-specific probe behavior [12]. Thus, measurement of the same genes across datasets does not necessarily imply preservation of the multivariate expression relationships required by a trained classifier.

Feature selection and classifier architecture introduce an additional source of variation in model transfer. Feature selection can reduce dimensionality and redundancy in high-dimensional biological data [13], while approaches such as minimum-redundancy maximum-relevance (mRMR) [14], ElasticNet regularization [15], and tree-based feature importance [16] prioritize different statistical properties of the expression space. Likewise, linear classifiers and nonlinear ensemble methods such as XGBoost and CatBoost impose different decision structures [16-17]. Consequently, a configuration that is optimal under internal cross-validation is not necessarily the one whose feature relationships remain most useful after transport. Internal predictive accuracy, feature-selection stability, and external transferability should therefore be regarded as related but non-equivalent properties of model robustness.

These issues can be understood more broadly as forms of dataset shift, in which the predictor or outcome distributions encountered during model evaluation differ from those represented during development [18]. In high-throughput genomics, technical and biological batch effects can be sufficiently large to dominate relevant molecular variation [19]. Importantly, however, detecting

dataset identity does not by itself establish that a predictive model will fail. An external cohort may differ strongly from TCGA while still preserving the cancer-discriminative structure required for accurate classification. Conversely, relatively modest changes in specific decision-relevant relationships may substantially impair prediction. The central methodological question is therefore not simply whether two datasets differ, but whether the cancer-discriminative representation learned by a classifier is preserved after transfer.

Despite substantial progress in pan-cancer transcriptomic classification, this distinction has received less attention than predictive accuracy itself. Recent classifiers increasingly include external evaluation and biological interpretation [4,5], but external performance is commonly reported as an endpoint rather than systematically examined to determine why a locked model succeeds or fails after transfer. In particular, comparisons across independent RNA-seq cohorts and RNA-seq-to-microarray environments provide an opportunity to distinguish ordinary cross-cohort variation from the more substantial representation changes that may accompany measurement-platform differences. Examining model-independent cancer geometry, effective measurement resolution, feature stability, calibration, perturbation sensitivity, and classifier attribution can therefore provide information that a single external accuracy metric cannot.

Accordingly, this study was organized around one primary research question: to what extent do TCGA-trained five-cancer transcriptomic classifiers retain predictive performance when transferred without retraining to independent ICGC RNA-seq and GEO microarray cohorts, and which measurable changes in transcriptomic representation accompany any loss of transferability? All supervised feature selection and model fitting were confined to TCGA. Five feature-selection methods were evaluated at three feature-set sizes and combined with three classifier

architectures, producing 45 selected-feature configurations: three classifier-specific all-gene baselines yielded 48 configurations in total. After development, the fitted models and selected features were locked and transferred without retraining or feature reselection to non-TCGA ICGC RNA-seq projects and GSE2109 GPL570 microarray samples. The ICGC analysis therefore provided an independent cross-cohort RNA-seq evaluation, whereas GSE2109 was treated conservatively as a secondary cross-platform stress test.

The methodological contribution of this study is a controlled framework for evaluating transcriptomic transferability rather than a new classification algorithm. The framework combines TCGA-only supervised model development with locked evaluation in an independent RNA-seq cohort and a cross-platform microarray cohort, explicitly separates predictive performance from feature stability and representation-level diagnostics such as cancer-centroid geometry and effective measurement resolution, and keeps post hoc biological interpretation separate from model selection. This design extends conventional external validation beyond reporting external accuracy alone by characterizing which decision-relevant properties of the transcriptomic representation are preserved after transfer.

II. Related Work

1. Pan-cancer transcriptomic classification and external validation

Previous TCGA-based studies have demonstrated that whole-transcriptome expression profiles contain strong tissue- and cancer-type signals and can support highly accurate classification of both primary and metastatic tumors [2-3]. However, high predictive performance within a development cohort does not necessarily establish that the learned decision structure will remain stable in independent datasets. Cross-study validation

therefore provides an important framework for assessing model transportability because it evaluates performance outside the original development environment rather than repeatedly resampling the same cohort [6]. This distinction is particularly relevant in transcriptomic prediction, where changes in patient populations, specimen characteristics, laboratory procedures, sequencing centers, batch composition, measurement platforms, and preprocessing pipelines can produce substantial dataset shift [18-19].

2. Cross-platform representation and domain shift

Differences between RNA sequencing and microarray measurements represent an additional challenge for model transfer. Several cross-platform normalization strategies have been proposed to align expression distributions between RNA-seq and microarray datasets [8-11], while within-sample relative-expression or rank-based representations have been explored as a way to reduce dependence on absolute expression scale [10]. Such transformations may improve comparability, but they do not guarantee preservation of the multivariate relationships used by a trained classifier. In the present study, external cohort labels and outcomes were not used to estimate target-specific transformations or to reselect features after model development. Instead, TCGA-trained models were locked and directly evaluated in independent ICGC RNA-seq and GEO microarray cohorts. This design provides a stricter assessment of predictive transfer, while allowing residual cohort- and platform-associated distortion to be examined explicitly rather than removed retrospectively.

3. Explainable model interpretation and biological validity

Model interpretation provides a complementary means of determining whether transferable predictive signals correspond to recognizable biological structure. TreeSHAP enables feature-level

attribution for tree-based models [20], whereas gene set enrichment analysis (GSEA) evaluates coordinated biological programs across a ranked transcriptome [21]. These approaches address different levels of interpretation and must be separated from predictive validation. In particular, enrichment analysis based only on supervised-selected classifier features can be affected by the feature-selection process. Accordingly, the present study distinguishes over-representation analysis (ORA) of classifier-selected genes from transcriptome-wide preranked GSEA. Biological concordance is therefore used as a secondary validity check to determine whether discriminative signals correspond to established tissue- or cancer-associated transcriptional programs, rather than as evidence of independent predictive validation or novel causal mechanisms.

III. Materials and Methods

1. Study design, cohorts, and common gene universe

The analysis included breast invasive carcinoma (BRCA), kidney renal clear cell carcinoma (KIRC), ovarian carcinoma (OV), pancreatic adenocarcinoma (PAAD), and prostate adenocarcinoma (PRAD), which were represented with usable sample support in all three data sources. TCGA primary tumors were used for model development. TCGA analyses were restricted to primary-tumor samples (sample-type code 01). Participant identifiers were defined by the first 12 characters of the TCGA barcode. When multiple eligible primary-tumor profiles were available for the same participant, a single profile was retained deterministically by prioritizing the sample with the greatest number of finite gene measurements, with the sample identifier used as a tie-breaker. Thus, each TCGA participant contributed at most one expression profile to model development and cross-validation. ICGC Release 28 non-TCGA

projects were used for independent RNA-seq validation [22]. GSE2109, generated on Affymetrix Human Genome U133 Plus 2.0 (GPL570), served as a secondary cross-platform stress test [23]. GSE2109 was selected because it provided representation of all five target cancer types within a single GPL570 microarray cohort and therefore enabled evaluation under a measurement platform distinct from the RNA-seq development and ICGC cohorts. Given the limited support for some classes, particularly PAAD, it was treated as a secondary cross-platform stress test rather than an untouched confirmatory cohort. TCGA-origin ICGC projects were excluded. The non-TCGA ICGC evaluation comprised BRCA-KR for BRCA, RECA-EU for KIRC, OV-AU for OV, PACA-AU and PACA-CA for PAAD, and PRAD-CA and PRAD-FR for PRAD. Only primary-tumor specimens and the project-provided `normalized_read_count` expression field were retained. When multiple eligible molecular samples were available for the same donor, one profile was selected deterministically by prioritizing the sample with the largest number of non-missing gene measurements, with sample identifier used as a tie-breaker. This yielded one expression profile per donor. RECA-EU was further restricted to clear-cell renal carcinoma. GSE2109 samples were assigned to the five target classes using predefined phenotype terms, and samples not matching these rules were excluded. For GSE2109, samples containing “breast” together with carcinoma, adenocarcinoma, ductal, or lobular terms were mapped to BRCA; renal or kidney samples containing “clear cell” to KIRC; ovarian, ovary samples containing serous, carcinoma, or adenocarcinoma terms to OV; pancreatic adenocarcinoma to PAAD; and prostate carcinoma or adenocarcinoma to PRAD. External labels were not used for TCGA feature ranking or primary-model selection. Cohort counts are summarized in Table 1.

Gene mapping yielded 35,828 unique symbols in TCGA, 20,826 in GPL570, and 16,305 genes

measurable across every selected ICGC project. The pre-load intersection contained 15,014 genes; 14,949 remained completely observed after matrix construction and defined the common measurable universe. This universe was defined by assay representability, not by external labels or predictive performance.

Before supervised model development, external assay availability was used to define the common measurable feature space, and external phenotype information was used to construct the prespecified external evaluation cohorts. External labels or performance were not used for TCGA feature ranking, model fitting, hyperparameter selection, or primary-configuration selection. External labels, predictive outcomes, or external performance were not used for feature ranking, model selection, or hyperparameter selection. Here, 'locked transfer' denotes that, after this common measurable universe was fixed and TCGA-only model development was completed, the selected features, fitted model parameters, and preprocessing rule were frozen and applied to the external cohorts without target-specific retraining, feature reselection, or distributional adaptation. The 14,949 gene universe is specific to the assays evaluated in this study. A future external platform may yield a different measurable gene intersection. Direct application of the present locked model therefore requires that its selected genes be measurable, whereas a materially different feature space would require a new development-and-locking cycle rather than target-specific adaptation of the existing model.

Table 1. Cohort composition used for model development and external evaluation.

Dataset	BRCA	KIRC	OV	PAAD	PRAD	Total
TCGA	1,095	533	426	178	497	2,729
ICGC non-TCGA	50	91	82	275	169	667
GSE2109 GPL570	365	211	256	19	85	936

TCGA counts represent one primary-tumor expression profile per participant after participant-level deduplication. ICGC counts represent one primary-tumor expression profile per donor from BRCA-KR, RECA-EU, OV-AU, PACA-AU, PACA-CA, PRAD-CA, and PRAD-FR.

2. Gene harmonization and within-sample rank representation

Identifiers were harmonized to HGNC symbols in R/Bioconductor [24]. Ensembl version suffixes were removed, ambiguous one-to-many mappings were discarded, and multiple valid measurements mapping to the same gene were collapsed by the median. TCGA transcripts per million (TPM) values were transformed as $\log_2(\text{TPM}+1)$; GSE2109 was obtained from the processed GEO Series Matrix for the GPL570 platform, and the supplied expression values were retained without additional between-sample normalization. GPL570 probe identifiers were mapped to HGNC gene symbols using the Bioconductor `hgu133plus2.db` annotation; probes with ambiguous one-to-many gene mappings were excluded, and multiple valid measurements mapping to the same gene were collapsed by the median. After gene harmonization, each sample was independently transformed to percentile ranks across the common 14,949 gene universe. No target-specific normalization or cross-cohort batch correction was applied before external prediction. Each sample was then independently converted to percentile ranks across the identical 14,949 gene universe using $(\text{rank}-1)/(\text{p}-1)$. This transformation does not estimate external-cohort normalization parameters and was chosen to reduce dependence on absolute expression scale while preserving within-sample ordering. No cross-cohort batch adjustment such as ComBat was applied, because the objective was to evaluate transfer into the external domains without adapting their distributions to TCGA [25].

3. TCGA-only model development and locked external validation

Model development used five-fold stratified cross-validation with seed 42. Every supervised feature-selection step was repeated inside each TCGA training fold. Within each TCGA training fold, the 8,000 genes with the largest training-set variance were retained before supervised feature

ranking. ANOVA ranking used the scikit-learn `f_classif` statistic, whereas mutual-information ranking used `mutual_info_classif`. For the feature-class quotient variant of minimum-redundancy maximum-relevance (mRMR-FCQ) [14], the 2,000 genes with the highest ANOVA F statistics within the variance-filtered set were used as candidates. Relevance was defined by the multiclass ANOVA F statistic, redundancy by the mean absolute Pearson correlation with already selected genes, and genes were selected greedily according to the ratio of relevance to redundancy. ElasticNet ranking [15] used standardized training data and logistic regression with `l1_ratio=0.5`, `C=0.1`, the saga solver, balanced class weights, a maximum of 2,500 iterations; gene importance was defined as the mean absolute coefficient across classes. XGBoost-based ranking [16] used 250 trees, maximum depth 5, learning rate 0.05, subsample 0.8, `colsample_bytree=0.8`, balanced sample weights, with genes ranked using the model `feature_importances_`. All feature-selection procedures were fitted independently within each TCGA training fold. The top 250, 500, and 1,000 genes from each of the five selectors defined 15 selected-feature settings. Pairing these settings with logistic regression, XGBoost, and CatBoost produced 45 models; three additional all-gene baselines, one per classifier using all 14,949 common genes, yielded 48 configurations in total. For the reported final analysis, classifier hyperparameters were fixed in the analysis pipeline before cross-validation and external evaluation; no hyperparameter optimization was performed within the reported cross-validation procedure or using ICGC or GSE2109 performance. Logistic regression used standardized features, `C=1.0`, balanced class weights, the `lbfgs` solver, a maximum of 3,000 iterations. XGBoost used 350 trees, maximum depth 5, learning rate 0.05, subsample 0.8, `colsample_bytree=0.8`, the multi:softprob objective, `mlogloss`, histogram tree construction, balanced sample weights. CatBoost [17] used 400 iterations,

depth 7, learning rate 0.05, `MultiClassloss`, automatic balanced class weights. Core modeling used NumPy 2.4.6, pandas 3.0.5, SciPy 1.17.1, scikit-learn 1.9.0 [26], joblib 1.6.0, XGBoost 3.2.0, CatBoost 1.2.10, and Matplotlib 3.11.1; pathway analysis used GSEapy 1.3.1 [27].

The primary configuration was selected from TCGA cross-validation only using mean Macro-F1, followed by Matthews correlation coefficient (MCC), smaller feature count, and deterministic model order. Neither ICGC nor GSE2109 performance was used for feature selection, hyperparameter determination or primary-configuration selection. Mean pairwise Jaccard similarity among the five fold-specific selected-gene sets quantified feature-selection stability. Each configuration was then refitted on the full TCGA cohort, locked, and applied to ICGC and GSE2109 without retraining. Macro-F1, MCC, balanced accuracy, class-specific precision/recall/F1, AUROC, Brier score, and expected calibration error (ECE) were retained. External 95% confidence intervals for Macro-F1 and MCC were estimated from 1,000 stratified bootstrap replicates with replacement. For ICGC, resampling was performed independently within project-code strata, with each project belonging to one target cancer; for GSE2109, resampling was performed independently within cancer-type strata. Each bootstrap replicate therefore preserved the original project/class sample counts. The 95% confidence interval was defined by the 2.5th and 97.5th percentiles of the bootstrap metric distribution.

4. Post hoc diagnostics of external performance loss

All diagnostic analyses were conducted after the main external results were known and did not alter the prespecified primary configuration. To test sample-support effects, ICGC and GSE2109 were repeatedly subsampled to matched cancer-specific counts for 5,000 repeats, with a second equal-support analysis using 19 samples per

cancer. A PAAD-excluded sensitivity analysis removed true PAAD samples from evaluation while retaining predictions into PAAD as errors for the remaining cancer classes, thereby assessing the extent to which the smallest GSE2109 class contributed to the observed performance pattern. Probability behavior was summarized by Brier score [28], log loss, normalized entropy, and 10-bin ECE. Calibration metrics were interpreted descriptively because class prevalence differed across datasets.

Residual domain identity was assessed using within-cancer TCGA-versus-external logistic-regression classifiers and a separate post hoc principal component analysis (PCA). For PCA, the same 14,949 gene within-sample percentile-rank representation used in the main analysis was reconstructed. To limit visual imbalance, up to 200 samples were randomly selected from each dataset-by-cancer stratum. Samples from TCGA, ICGC, and GSE2109 were then pooled, the 2,000 genes with the largest variance in this pooled diagnostic subset were retained, and a two-component PCA was jointly fitted using randomized SVD. Thus, PCA was not fitted on TCGA alone and projected onto the external cohorts; it was used solely as an unsupervised post hoc visualization of dataset and cancer-type structure and did not contribute to model fitting or selection. Cancer-class geometry was then examined independently of the trained classifier by comparing TCGA and external cancer centroids using Pearson correlation, cosine similarity, RMSE, mean rank shift, and standardized mean difference(SMD), followed by nearest-TCGA-centroid classification. Effective measurement resolution was summarized by the fraction of unique expression values per sample. Additional controls included Gaussian rank noise, random feature permutation, alternative within-sample representations, and GPL570 multi-probe concordance. These controls were used to reject simple explanations rather than to

optimize performance on the external cohorts.

5. Post hoc biological interpretation of one externally robust candidate

Biological interpretation was intentionally restricted to the ANOVA500+XGBoost configuration rather than repeated across all models. This model was selected after external performance inspection because it combined high feature stability, 500 features, and strong performance in both external cohorts. It therefore remains a post hoc candidate; the TCGA-only mRMR-FCQ250+logistic-regression(LR) model remains the prespecified primary configuration. Native XGBoost TreeSHAP contributions were computed in raw-margin space using the locked model. For each cancer and dataset, class-specific support for each of the 500 genes was summarized as the difference between mean SHAP contribution in samples of that cancer and mean contribution in all other samples. The 50 largest positive TCGA differences were defined as class-supporting features. Preservation of the complete 500-gene attribution profile was quantified by Spearman correlation between TCGA and each external cohort.

ORA was performed on all 500 ANOVA-selected classifier genes and separately on the top 50 class-supporting TCGA TreeSHAP genes. The 14,949 common measurable genes were used as the background, and Benjamini-Hochberg correction was applied within cancer, analysis, and gene-set library. For conventional GSEA, all 14,949 TCGA genes were ranked separately for each cancer using a one-versus-rest Welch t statistic on percentile-rank expression. GSEA used Hallmark and Reactome collections with a weighted preranked statistic, 10,000 permutations, gene-set sizes 10-500 [21,27,29-30]. Because these contrasts compare one tumor type against the other four tumor types, positive and negative normalized enrichment score(NES) values indicate relative differences within this five-cancer panel and must not be interpreted as tumor-versus-normal changes.

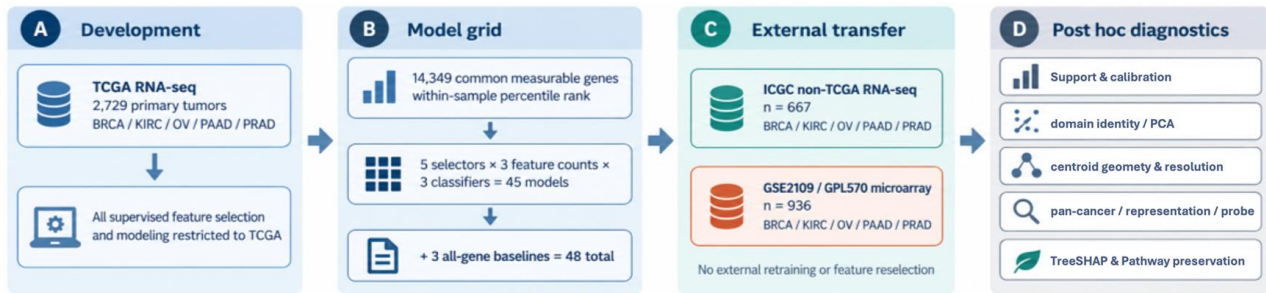


Fig. 1. Study design for external transferability evaluation. (A) TCGA RNA-seq was used exclusively for supervised feature selection and model fitting. (B) Five selectors at three feature counts generated 15 feature settings; three classifiers produced 45 selected-feature models, and three all-gene baselines yielded 48 total configurations. (C) Locked models were transferred without retraining to non-TCGA ICGC RNA-seq and GSE2109/GPL570. (D) Post hoc diagnostics were used to explain transfer behavior and did not redefine the prespecified primary model.

For TCGA-significant pathways, leading-edge genes were fixed in TCGA first. Sample-level pathway scores were then computed in TCGA, ICGC, and GSE2109 without pathway re-selection. For each cancer-specific pathway score, 95% AUROC confidence intervals were estimated from 1,000 stratified bootstrap replicates with replacement, resampling positive and negative samples separately within each dataset. Intervals were defined by the 2.5th and 97.5th percentiles of the bootstrap AUROC distribution. External labels were used only for this diagnostic assessment. Empirical GSEA FDR values recorded as 0.000 were reported as <0.001 in the manuscript to avoid literal-zero interpretation. The complete development-transfer-diagnostic workflow is summarized in Fig. 1.

IV. Results

1. Near-ceiling TCGA performance separated only under external transfer

Across all 48 configurations, TCGA cross-validation Macro-F1 ranged from 0.9963 to 0.9991. The prespecified primary model was mRMR-FCQ250+logistic regression(LR) model (Macro-F1 0.9991; Jaccard stability 0.911). It retained high performance in ICGC (Macro-F1 0.989) but fell to 0.813 in GSE2109. The difference between near-identical internal scores and substantially

different external behavior demonstrates that TCGA cross-validation alone did not identify the most transferable configuration. For consistency, configurations are denoted as feature selector, number of selected genes, and classifier; for example, mRMR-FCQ250+Logistic regression model, ANOVA500+XGB, and XGB-FS1000+XGBoost, where LR denotes logistic regression and XGB-FS denotes XGBoost-based feature selection.

Classifier architecture mattered differently by external environment. Across the 15 matched selected-feature settings, XGB showed a higher observed GSE2109 Macro-F1 than logistic regression in all 15 comparisons, with a mean difference of 0.051. In ICGC, LR showed a higher observed Macro-F1 in 14 of the 15 matched settings, with a mean difference of 0.006. These comparisons are descriptive. XGB-FS1000+XGB produced the highest GSE2109 Macro-F1 (0.906), but its fold feature Jaccard was only 0.314. ANOVA500+XGB achieved nearly the same GSE2109 result (0.905), slightly higher ICGC performance (0.983), and much greater feature stability (0.944). For this reason it was treated as the parsimonious externally robust candidate for interpretation, not as a replacement for the prespecified primary model. Representative configurations are summarized in Table 2 and Fig. 2A. Complete results for all 48 configurations, including external uncertainty estimates, are provided in Supplementary Table S1.

Table 2. Representative internal and external performance. External-performance-selected models are explicitly post hoc.

Configuration	Genes	TCGA CV F1	Jaccard	ICGC F1	GSE2109 F1
mRMR-FCQ250+LR (primary)	250	0.999	0.911	0.989	0.813
ANOVA500+XGB (post hoc)	500	0.998	0.944	0.983	0.905
XGB-FS1000+XGB (post hoc)	1,000	0.998	0.314	0.981	0.906
AllGenes+LR	14,949	0.997	1.000	0.917	0.786
AllGenes+XGB	14,949	0.998	1.000	0.983	0.868

2. Sample support and PAAD contributed to the GSE2109 loss

Matching ICGC and GSE2109 to identical class supports for 5,000 repeats left a positive ICGC-minus-GSE2109 Macro-F1 difference of 0.120 for the primary model (95% resampling interval 0.090-0.150), 0.081 for ANOVA500+XGB (0.061-0.102), and 0.081 for XGB-FS1000+XGB (0.060-0.101). Thus, the lower GSE2109 performance was not explained by unequal class counts alone.

PAAD was the clearest class-specific failure of the primary model. In GSE2109, PAAD recall was 0.947 but precision was 0.180, producing F1 0.303 and 100 PAAD predictions for only 19 true PAAD samples. ANOVA500+XGB improved PAAD precision/recall/F1 to 0.765/0.684/0.722. Removing true PAAD samples increased primary-model GSE2109 Macro-F1 to 0.941, showing that PAAD accounted for a large share of the loss but not the entire cross-platform deficit. As a descriptive probability diagnostic, the multiclass Brier score was 0.001 in TCGA out-of-fold predictions, 0.016 in ICGC, and 0.174 in GSE2109. These support, error-pattern, and calibration findings are shown in Fig. 2B-D.

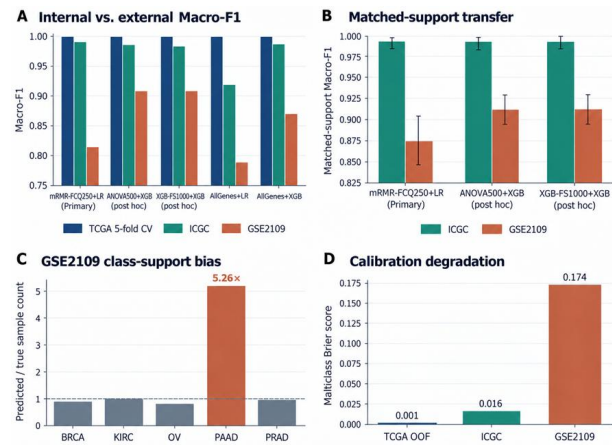


Fig. 2. External predictive transfer and primary-model error diagnostics. (A) Representative configurations had near-ceiling TCGA cross-validation performance but separated after transfer, especially in GSE2109. (B) Matching cancer-specific sample support across ICGC and GSE2109 did not remove the external gap. (C) The primary model markedly overcalled PAAD in GSE2109. (D) Multiclass Brier score increased from TCGA out-of-fold predictions to ICGC and most strongly to GSE2109.

3. Cancer-discriminative geometry showed weaker preservation in GSE2109

Residual dataset identity was readily detectable in both external cohorts despite their different predictive performance. Within each cancer, domain classifiers distinguished TCGA from ICGC and TCGA from GSE2109 with AUROC 1.000, even though the primary classifier transferred well to ICGC and less well to GSE2109. PCA likewise showed strong dataset separation while cancer-type structure remained visible across domains (Fig. 3A-B).

Cancer-class geometry also showed differential preservation across the two external cohorts. For the primary 250-gene set, mean cancer-specific centroid correlation with TCGA was 0.935 in ICGC but 0.815 in GSE2109. Mean absolute rank shift increased from 0.062 to 0.115 and mean absolute SMD from 0.666 to 1.165. A model-independent nearest-TCGA-centroid classifier achieved Macro-F1 0.975 in ICGC and 0.879 in GSE2109. The median fraction of unique expression values per sample was 0.892 in TCGA, 0.874 in ICGC, and 0.658 in GSE2109, consistent with greater rank tying and lower effective measurement resolution in the microarray cohort. These geometry and resolution findings are summarized in Fig. 3C-D and Table 3.

Table 3. Diagnostic evidence used to interpret the external performance gap

Hypothesis	Key observation	Conservative interpretation
Unequal sample support	Matched resampling retained gaps of 0.081-0.120	Not the main explanation
PAAD alone	PAAD F1 0.303; PAAD-excluded GSE F1 rose to 0.941	Major contributor, not sole cause
Domain identity	Within-cancer TCGA-vs-external AUROC = 1.000 for both cohorts	Dataset shift exists in both; presence alone is insufficient
Class geometry	Centroid correlation 0.935 ICGC vs 0.815 GSE; nearest-centroid F1 0.975 vs 0.879	Supports weaker preservation of discriminative geometry in GSE
Measurement resolution	Unique-value fraction 0.892 TCGA, 0.874 ICGC, 0.658 GSE	Plausible contributor to rank compression
Generic noise/scaling/probes	Perturbation, alternative scaling, and probe concordance did not reproduce model ranking	No single simple technical factor explains the result

Additional diagnostic controls did not reproduce the observed pattern of external model performance. Configuration-level mean feature shift showed no significant association with external model ranking. Strong Gaussian rank noise and random feature permutation degraded XGBoost models more than the primary logistic-regression model, indicating that the perturbation response did not mirror the observed GSE2109 model ranking for XGBoost. Rank-Gaussian and within-sample z-score representations did not remove the cross-platform gap, and the XGB-FS1000 feature set showed the lowest multi-probe concordance. The combined diagnostic evidence is summarized in Table 3, and these negative controls are shown in Fig. 4.

4. Biological interpretation of the ANOVA500+XGB candidate showed coherent but heterogeneous class programs

Global XGBoost importance and class-specific TreeSHAP indicated that the 500-gene model relied on distinct feature subsets for each cancer. The three largest TCGA class-supporting TreeSHAP differences were TCF21, TRPS1, and GATA3 for BRCA; POU3F3, DMGDH, and SLC17A3 for KIRC; SOX17, FAM181A, and KLHL14 for OV; PDX1, INS, and TM4SF4 for PAAD; and KLK2, HOXB13, and KLK3 for PRAD. These features are model attributions rather than univariate differential-expression effects and should not be interpreted independently of the correlated 500-gene model.

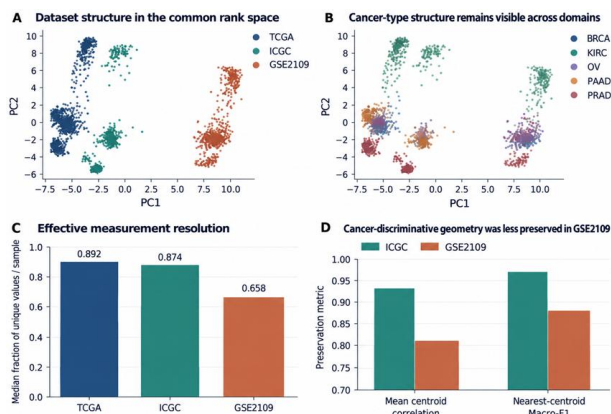


Fig. 3. Domain structure, measurement resolution, and class-geometry preservation. (A) PCA colored by dataset shows strong residual domain structure. (B) The same coordinates colored by cancer type show that class structure remains detectable. (C) GSE2109 has lower effective measurement resolution, measured by the median fraction of unique expression values per sample. (D) For the primary 250-gene feature set, mean cancer-centroid correlation with TCGA and model-independent nearest-TCGA-centroid Macro-F1 were both lower in GSE2109 than in ICGC. PCA was jointly fitted to a pooled diagnostic subset of the three cohorts and was not used for prediction.

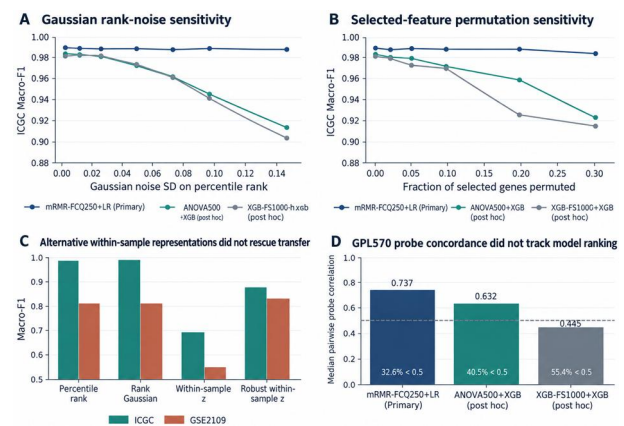


Fig. 4. Diagnostic controls against simple technical explanations. (A) Gaussian rank noise and (B) selected-feature permutation degraded the XGBoost configurations more than the primary LR model at stronger perturbation levels, showing that perturbation sensitivity did not reproduce the observed GSE2109 model ranking for GSE2109 performance. (C) Alternative within-sample representations did not consistently improve transfer. (D) GPL570 multi-probe concordance was lowest for the XGB-FS1000 feature set despite its strong GSE2109 performance, so probe discordance did not track model ranking.

ORA of all 500 classifier features identified three Hallmark collections after FDR correction: pancreas beta cells (10 overlapping genes; fold enrichment 7.47; FDR 2.37×10^{-5}), androgen response (11 genes; fold enrichment 3.57; FDR 0.0053), and KRAS signaling down (15 genes; fold enrichment 2.46; FDR 0.017). When the analysis was restricted to the top 50 class-supporting TreeSHAP genes, Hallmark ORA remained significant for PAAD pancreas-beta-cell genes (FOXA2, GCG, INS, ISL1, PDX1, SLC2A2; fold enrichment 44.85; FDR 7.95×10^{-8}) and PRAD androgen-response genes (ABCC4, HOMER2, KLK2, KLK3, NKX3-1; fold enrichment 16.25; FDR 9.43×10^{-5}). No class-specific Hallmark ORA term survived FDR 0.05 for BRCA, KIRC, or OV, a negative result that argues against presenting every class as equally pathway-resolved from the selected 50-gene lists.

Full-universe TCGA GSEA provided broader transcriptome-level context. BRCA showed strong positive enrichment for E2F targets (NES 3.19), G2/M checkpoint (2.96), and estrogen-response programs (early 2.41; late 2.27). KIRC was enriched for hypoxia (2.67), interferon-gamma response (2.51), and inflammatory signaling while E2F targets were relatively depleted (NES -2.71). OV showed strong E2F (3.01) and G2/M (2.66) enrichment. PAAD showed TNF/NF- κ B signaling (3.45), epithelial-mesenchymal transition (3.43), KRAS signaling up (3.01), and pancreas beta cells (2.84). PRAD showed a dominant androgen-response signal (3.07) with relative depletion of E2F (-2.83) and interferon-gamma response (-2.55). All values cited here had FDR <0.001 unless otherwise stated. Because the reference group is the other four cancers, these terms describe relative five-class biology rather than tumor-versus-normal dysregulation.

A representative pathway from each cancer was then scored without reselecting genes in the external cohorts. BRCA estrogen-response-late AUROC was 0.994 in TCGA, 0.981 in ICGC, and 0.946 in GSE2109; KIRC hypoxia was

0.992/0.990/0.969; OV E2F targets 0.881/0.974/0.808; PAAD KRAS signaling up 0.967/0.958/0.948; and PRAD androgen response 1.000/0.996/0.982. Thus, several biologically coherent programs were highly transportable even across the microarray cohort, whereas proliferation-related programs were less consistently preserved. Across all 25 selected Hallmark pathways, median pathway AUROC was 0.921 in TCGA, 0.958 in ICGC, and 0.887 in GSE2109.

Feature-attribution preservation was also heterogeneous. Mean Spearman correlation of the 500-feature class-specific TreeSHAP profile was 0.737 between TCGA and ICGC and 0.590 between TCGA and GSE2109. GSE2109 correlations were particularly low for OV (0.403) and PAAD (0.390), but remained high for KIRC (0.836) and PRAD (0.769). This pattern is consistent with heterogeneous, cancer-dependent differences in feature-attribution preservation across the external cohorts. The biological results are summarized in Table 4 and Fig. 5.

Table 4. Conservative biological summary of the post hoc ANOVA500+XGB candidate. The representative Hallmark terms shown here had TCGA GSEA FDR <0.001. External AUROCs quantify preservation of TCGA-defined pathway scores and were not used for pathway selection

Cancer	Top 3 TCGA TreeSHAP genes	Representative Hallmark program	NES	ICGC AUC	GSE2109 AUC
BRCA	TCF21, TRPS1, GATA3	Estrogen response (late)	2.27	0.981	0.946
KIRC	POU3F3, DMGDH, SLC17A3	Hypoxia	2.67	0.990	0.969
OV	SOX17, FAM181A, KLHL14	E2F targets	3.01	0.974	0.808
PAAD	PDX1, INS, TM4SF4	KRAS signaling up	3.01	0.958	0.948
PRAD	KLK2, H0XB13, KLK3	Androgen response	3.07	0.996	0.982

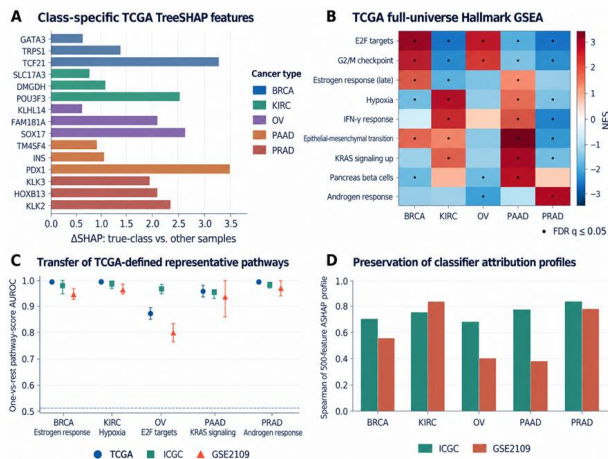


Fig. 5. Post hoc biological interpretation of the ANOVA500+XGB candidate. (A) Top class-supporting TCGA TreeSHAP features. (B) Selected Hallmark GSEA programs from complete 14,949 gene TCGA one-versus-rest rankings; dots indicate FDR $q \leq 0.05$. (C) Transfer of one representative TCGA-defined pathway per cancer; error bars indicate 95% percentile confidence intervals from 1,000 one-versus-rest stratified bootstrap replicates. (D) Spearman preservation of the complete 500-feature class-specific TreeSHAP profile. Pathway discovery and feature attribution were defined in TCGA; external cohorts were used only for post hoc preservation diagnostics.

V. Discussion

The central finding of this study is methodological rather than algorithmic: near-perfect internal performance in TCGA did not identify which transcriptomic classifier would remain most robust after external transfer. The prespecified mRMR-FCQ250+LR model(primary) generalized with little loss to independent ICGC RNA-seq cohorts but showed a substantially larger decline in GSE2109. In contrast, XGBoost-based configurations transferred more effectively to the microarray cohort while showing no consistent advantage in ICGC. These results do not indicate that one classifier family is universally more robust. Rather, they suggest that predictive transfer depends on the interaction between classifier architecture and the specific form of external distributional shift.

The diagnostic analyses provide further insight into this difference. Dataset identity was readily detectable in both ICGC and GSE2109, indicating

that the presence of domain shift alone was insufficient to explain predictive degradation. A clearer distinction emerged from preservation of cancer-discriminative expression geometry. Compared with ICGC, GSE2109 showed lower effective rank resolution, weaker correspondence of cancer-specific centroids with TCGA, and poorer model-independent nearest-centroid transfer. In contrast, Gaussian perturbation, feature permutation, alternative within-sample transformations, and GPL570 probe discordance did not reproduce the observed ranking of classifier performance. Together, these findings are more consistent with structured and class-dependent deformation of multivariate transcriptomic relationships than with a single global batch-effect magnitude or generic noise sensitivity.

This distinction is important because external dataset differences are not equivalent to predictive failure. Both external cohorts were readily separable from TCGA at the dataset level, yet the primary classifier retained almost all of its performance in ICGC while degrading substantially in GSE2109. Thus, the practical question is not simply whether an external dataset differs from the development cohort, but whether the expression relationships required by the classifier remain sufficiently preserved. The centroid and nearest-centroid analyses support this interpretation independently of the trained classifier, providing evidence that preservation of cancer-discriminative geometry is more informative than dataset identity alone.

The biological analyses provide a complementary interpretation of the transferred signal. The post hoc ANOVA500+XGB candidate relied on recognizable lineage-associated features, while full-transcriptome GSEA identified coherent cancer-associated programs. BRCA showed estrogen-response signals consistent with established breast cancer biology [31]; KIRC showed hypoxia-related enrichment consistent with VHL and oxygen-sensing dysregulation [32]; OV showed

prominent cell-cycle programs consistent with the proliferative genomic characteristics of high-grade serous ovarian carcinoma [33]; PAAD showed KRAS-, EMT-, and pancreatic-lineage-associated programs compatible with established pancreatic ductal adenocarcinoma biology [34]; and PRAD exhibited strong androgen-response signals consistent with the molecular taxonomy of prostate cancer [35]. These findings support the biological plausibility of the learned representation, but they do not demonstrate that the classifier discovered novel tumor mechanisms.

Interpretation of these biological signals also requires caution. The five classes originate from different organs, and the classifier may therefore exploit tissue-lineage, sex-associated, stromal, and cellular-composition signals in addition to tumor-specific molecular alterations. Furthermore, the GSEA contrasts compare each cancer type with the other four cancers rather than with matched normal tissue. Positive or negative enrichment therefore describes relative positioning within the five-cancer classification space rather than tumor-versus-normal dysregulation. The observation that class-specific Hallmark ORA reached clear significance only for PAAD and PRAD further argues against assuming uniform pathway-level specificity across all five cancer types.

Taken together, the results indicate that internal accuracy, feature stability, external predictive transfer, and preservation of decision-relevant expression structure represent distinct dimensions of model robustness. Near-ceiling internal validation alone was insufficient to reveal these differences. External evaluation across independent cohorts and measurement environments therefore provides information that cannot be recovered solely from additional within-cohort cross-validation, particularly when models are intended to operate beyond the data environment in which they were developed.

VI. Limitations and Future Research

This study has several limitations. The five-cancer panel was defined by availability across TCGA, ICGC, and GSE2109 and therefore does not represent all tumor types or within-organ subtypes. GSE2109 included only 19 PAAD samples and had been examined during earlier exploratory analyses, so it should be considered a secondary cross-platform stress test rather than an untouched confirmatory cohort. In addition, cohort and platform differences occur simultaneously between RNA-seq and GSE2109, preventing causal attribution of the observed performance loss specifically to assay technology. The common 14,949-gene space was also restricted to genes measurable across all datasets, and detailed clinical covariates such as tumor purity, stromal content, and sex were not uniformly harmonized. Because cancer-type prevalence differed across cohorts and class balancing was used for some classifiers, absolute Brier-score and ECE values are not prevalence-invariant measures. We therefore interpret these calibration metrics descriptively within each evaluation cohort rather than as prevalence-adjusted evidence of superior or inferior calibration across cohorts.

Future research should therefore focus on locked validation in a completely independent cohort, preferably with matched biological samples profiled across multiple platforms. Such studies would allow the post hoc ANOVA500+XGB candidate to be evaluated without further model selection and could determine whether preservation of cancer-discriminative geometry consistently predicts external transferability. Incorporating harmonized clinical and pathological covariates would further clarify how much of the observed shift reflects measurement technology versus biological cohort composition.

VII. Conclusions

This study showed that near-ceiling internal performance does not necessarily predict the external transferability of pan-cancer transcriptomic classifiers. Models that showed almost indistinguishable performance within TCGA exhibited substantially different behavior when transferred without retraining to independent ICGC RNA-seq and GSE2109 microarray cohorts. The prespecified mRMR-FCQ250+LR model(primary) retained high performance in ICGC but showed a marked decline in GSE2109, whereas the post hoc ANOVA500+XGB candidate achieved more balanced performance across both external cohorts with high feature-selection stability.

Diagnostic analyses further suggested that the observed cross-platform performance loss did not appear to be fully accounted for by sample imbalance, generic noise, alternative expression scaling, or probe discordance alone. Reduced preservation of

cancer-discriminative transcriptomic geometry was also observed in the external cohorts, a pattern consistent with a potential contribution of representation-level shifts to impaired transfer. TreeSHAP and transcriptome-wide GSEA additionally supported the biological coherence of the learned signals, while their heterogeneous preservation across cancer types highlighted the importance of evaluating biological interpretability and predictive transferability separately.

Overall, the main contribution of this study is a controlled framework that distinguishes internal accuracy, feature stability, external predictive performance, and preservation of decision-relevant transcriptomic structure as complementary dimensions of model robustness. These findings support the use of independent cross-cohort and cross-platform validation, together with representation-level diagnostics, when evaluating transcriptomic classifiers intended for use beyond their original development environment. Further

Supplementary Table S1. Cohort construction and sample-selection audit for TCGA, ICGC, and GSE2109

Panel A. Primary transferability metrics					Panel B. Secondary external metrics				
Configuration	TCGA CV Macro-F1 mean ± SD	Jaccard	ICGC Macro-F1 [95% CI]	GSE2109 Macro-F1 [95% CI]	Configuration	ICGC MCC [95% CI]	ICGC Bal. Acc.	GSE2109 MCC [95% CI]	GSE2109 Bal. Acc.
ANOVA250+LR	0.998 ± 0.003	0.898	0.979 [0.966-0.990]	0.847 [0.825-0.874]	ANOVA250+LR	0.976 [0.960-0.988]	0.988	0.890 [0.866-0.913]	0.933
ANOVA250+XGB	0.997 ± 0.002	0.898	0.980 [0.967-0.992]	0.867 [0.826-0.904]	ANOVA250+XGB	0.982 [0.970-0.992]	0.99	0.891 [0.876-0.921]	0.867
ANOVA250+CatBoost	0.997 ± 0.002	0.898	0.982 [0.967-0.993]	0.858 [0.835-0.883]	ANOVA250+CatBoost	0.984 [0.970-0.994]	0.991	0.905 [0.884-0.927]	0.938
ANOVA500+LR	0.998 ± 0.003	0.944	0.987 [0.977-0.994]	0.853 [0.830-0.880]	ANOVA500+LR	0.986 [0.974-0.994]	0.992	0.883 [0.859-0.905]	0.928
ANOVA500+XGB	0.998 ± 0.002	0.944	0.983 [0.971-0.994]	0.905 [0.864-0.937]	ANOVA500+XGB	0.984 [0.972-0.994]	0.991	0.920 [0.898-0.940]	0.897
ANOVA500+CatBoost	0.998 ± 0.002	0.944	0.990 [0.978-0.998]	0.869 [0.845-0.895]	ANOVA500+CatBoost	0.992 [0.982-0.998]	0.994	0.903 [0.881-0.924]	0.936
ANOVA1000+LR	0.999 ± 0.002	0.936	0.986 [0.975-0.995]	0.820 [0.797-0.843]	ANOVA1000+LR	0.984 [0.972-0.994]	0.991	0.858 [0.835-0.882]	0.916
ANOVA1000+XGB	0.998 ± 0.002	0.936	0.983 [0.971-0.994]	0.892 [0.851-0.925]	ANOVA1000+XGB	0.984 [0.974-0.994]	0.991	0.901 [0.878-0.924]	0.889
ANOVA1000+CatBoost	0.998 ± 0.003	0.936	0.984 [0.972-0.994]	0.871 [0.846-0.897]	ANOVA1000+CatBoost	0.986 [0.974-0.994]	0.992	0.904 [0.883-0.925]	0.938
Mi250+LR	0.998 ± 0.003	0.851	0.987 [0.977-0.997]	0.841 [0.819-0.865]	Mi250+LR	0.988 [0.978-0.996]	0.928	0.886 [0.864-0.908]	0.928
Mi250+XGB	0.996 ± 0.002	0.851	0.980 [0.968-0.991]	0.848 [0.820-0.879]	Mi250+XGB	0.978 [0.964-0.990]	0.988	0.894 [0.872-0.917]	0.9
Mi250+CatBoost	0.997 ± 0.002	0.851	0.990 [0.978-0.998]	0.851 [0.829-0.876]	Mi250+CatBoost	0.992 [0.984-0.998]	0.994	0.895 [0.874-0.916]	0.932
Mi500+LR	0.997 ± 0.002	0.869	0.992 [0.983-0.999]	0.856 [0.832-0.882]	Mi500+LR	0.992 [0.984-0.998]	0.994	0.892 [0.871-0.915]	0.933
Mi500+XGB	0.998 ± 0.002	0.869	0.977 [0.965-0.989]	0.867 [0.833-0.903]	Mi500+XGB	0.978 [0.965-0.990]	0.987	0.908 [0.887-0.929]	0.891
Mi500+CatBoost	0.998 ± 0.002	0.869	0.990 [0.978-0.998]	0.864 [0.839-0.895]	Mi500+CatBoost	0.992 [0.982-0.998]	0.994	0.907 [0.884-0.927]	0.928
Mi1000+LR	0.998 ± 0.002	0.885	0.985 [0.974-0.995]	0.847 [0.821-0.873]	Mi1000+LR	0.984 [0.972-0.994]	0.991	0.880 [0.856-0.902]	0.916
Mi1000+XGB	0.998 ± 0.002	0.885	0.982 [0.969-0.993]	0.877 [0.850-0.907]	Mi1000+XGB	0.984 [0.972-0.994]	0.991	0.912 [0.892-0.932]	0.921
Mi1000+CatBoost	0.999 ± 0.002	0.885	0.982 [0.968-0.993]	0.865 [0.841-0.893]	Mi1000+CatBoost	0.984 [0.971-0.994]	0.991	0.907 [0.887-0.929]	0.928
mRMR-FCQ250+LR	0.999 ± 0.001	0.911	0.989 [0.979-0.997]	0.813 [0.795-0.831]	mRMR-FCQ250+LR	0.988 [0.978-0.996]	0.992	0.861 [0.837-0.883]	0.916
mRMR-FCQ250+XGB	0.997 ± 0.002	0.911	0.982 [0.970-0.993]	0.888 [0.847-0.920]	mRMR-FCQ250+XGB	0.984 [0.972-0.994]	0.991	0.901 [0.876-0.921]	0.888
mRMR-FCQ250+CatBoost	0.998 ± 0.003	0.911	0.985 [0.972-0.995]	0.870 [0.845-0.895]	mRMR-FCQ250+CatBoost	0.988 [0.977-0.996]	0.992	0.906 [0.884-0.925]	0.939
mRMR-FCQ500+LR	0.998 ± 0.003	0.914	0.992 [0.982-0.999]	0.807 [0.788-0.827]	mRMR-FCQ500+LR	0.992 [0.984-0.998]	0.994	0.850 [0.824-0.874]	0.913
mRMR-FCQ500+XGB	0.998 ± 0.002	0.914	0.983 [0.971-0.994]	0.896 [0.860-0.930]	mRMR-FCQ500+XGB	0.984 [0.972-0.994]	0.991	0.908 [0.886-0.929]	0.902
mRMR-FCQ500+CatBoost	0.998 ± 0.003	0.914	0.986 [0.975-0.996]	0.860 [0.837-0.886]	mRMR-FCQ500+CatBoost	0.988 [0.978-0.996]	0.992	0.903 [0.883-0.925]	0.936
mRMR-FCQ1000+LR	0.999 ± 0.002	0.935	0.989 [0.978-0.997]	0.796 [0.777-0.817]	mRMR-FCQ1000+LR	0.988 [0.978-0.996]	0.992	0.831 [0.806-0.855]	0.901
mRMR-FCQ1000+XGB	0.998 ± 0.002	0.935	0.981 [0.967-0.992]	0.901 [0.861-0.934]	mRMR-FCQ1000+XGB	0.982 [0.968-0.992]	0.99	0.912 [0.891-0.933]	0.894
mRMR-FCQ1000+CatBoost	0.999 ± 0.002	0.935	0.988 [0.977-0.998]	0.863 [0.838-0.893]	mRMR-FCQ1000+CatBoost	0.990 [0.982-0.998]	0.993	0.895 [0.872-0.916]	0.923
ElasticNet250+LR	0.997 ± 0.004	0.7	0.992 [0.982-0.999]	0.831 [0.810-0.853]	ElasticNet250+LR	0.992 [0.984-0.998]	0.994	0.877 [0.855-0.898]	0.922
ElasticNet250+XGB	0.998 ± 0.002	0.7	0.985 [0.974-0.995]	0.887 [0.850-0.924]	ElasticNet250+XGB	0.986 [0.974-0.996]	0.992	0.904 [0.883-0.926]	0.891
ElasticNet250+CatBoost	0.999 ± 0.002	0.7	0.986 [0.975-0.996]	0.873 [0.846-0.901]	ElasticNet250+CatBoost	0.988 [0.978-0.996]	0.992	0.907 [0.885-0.929]	0.932
ElasticNet500+LR	0.998 ± 0.002	0.551	0.990 [0.980-0.999]	0.851 [0.826-0.877]	ElasticNet500+LR	0.990 [0.980-0.998]	0.993	0.893 [0.870-0.915]	0.93
ElasticNet500+XGB	0.998 ± 0.002	0.551	0.983 [0.971-0.994]	0.900 [0.863-0.935]	ElasticNet500+XGB	0.984 [0.972-0.994]	0.991	0.908 [0.886-0.930]	0.893
ElasticNet500+CatBoost	0.998 ± 0.002	0.551	0.989 [0.977-0.998]	0.878 [0.855-0.905]	ElasticNet500+CatBoost	0.990 [0.980-0.998]	0.993	0.914 [0.894-0.934]	0.943
ElasticNet1000+LR	0.998 ± 0.003	0.668	0.990 [0.981-0.999]	0.852 [0.828-0.879]	ElasticNet1000+LR	0.990 [0.982-0.998]	0.993	0.888 [0.866-0.910]	0.929
ElasticNet1000+XGB	0.997 ± 0.003	0.668	0.984 [0.972-0.994]	0.895 [0.852-0.930]	ElasticNet1000+XGB	0.984 [0.973-0.994]	0.991	0.912 [0.892-0.933]	0.885
ElasticNet1000+CatBoost	0.999 ± 0.002	0.668	0.988 [0.974-0.998]	0.864 [0.835-0.891]	ElasticNet1000+CatBoost	0.990 [0.979-0.998]	0.993	0.898 [0.875-0.920]	0.926
XGB-FS250+LR	0.999 ± 0.002	0.278	0.989 [0.979-0.997]	0.869 [0.845-0.898]	XGB-FS250+LR	0.988 [0.978-0.996]	0.992	0.890 [0.860-0.913]	0.931
XGB-FS250+XGB	0.998 ± 0.002	0.278	0.982 [0.970-0.994]	0.900 [0.863-0.932]	XGB-FS250+XGB	0.984 [0.972-0.994]	0.991	0.918 [0.877-0.939]	0.896
XGB-FS250+CatBoost	0.999 ± 0.001	0.278	0.988 [0.976-0.998]	0.873 [0.847-0.901]	XGB-FS250+CatBoost	0.990 [0.980-0.998]	0.993	0.908 [0.886-0.929]	0.93
XGB-FS500+LR	0.999 ± 0.002	0.229	0.989 [0.979-0.996]	0.848 [0.824-0.872]	XGB-FS500+LR	0.988 [0.978-0.996]	0.992	0.872 [0.852-0.892]	0.909
XGB-FS500+XGB	0.998 ± 0.002	0.229	0.984 [0.973-0.994]	0.898 [0.861-0.933]	XGB-FS500+XGB	0.984 [0.972-0.994]	0.991	0.912 [0.892-0.933]	0.884
XGB-FS500+CatBoost	0.998 ± 0.002	0.229	0.990 [0.980-0.998]	0.869 [0.842-0.892]	XGB-FS500+CatBoost	0.992 [0.984-0.998]	0.994	0.911 [0.890-0.930]	0.931
XGB-FS1000+LR	0.998 ± 0.003	0.314	0.987 [0.975-0.995]	0.835 [0.808-0.862]	XGB-FS1000+LR	0.986 [0.974-0.994]	0.992	0.871 [0.848-0.896]	0.909
XGB-FS1000+XGB	0.998 ± 0.002	0.314	0.981 [0.967-0.992]	0.906 [0.869-0.941]	XGB-FS1000+XGB	0.982 [0.968-0.992]	0.99	0.922 [0.902-0.943]	0.898
XGB-FS1000+CatBoost	0.998 ± 0.002	0.314	0.986 [0.974-0.996]	0.864 [0.840-0.891]	XGB-FS1000+CatBoost	0.988 [0.978-0.996]	0.992	0.903 [0.882-0.924]	0.937
AllGenes+LR	0.995 ± 0.002	N/A	0.917 [0.897-0.936]	0.786 [0.751-0.822]	AllGenes+LR	0.894 [0.871-0.920]	0.94	0.789 [0.763-0.815]	0.81
AllGenes+XGB	0.998 ± 0.002	N/A	0.983 [0.971-0.994]	0.868 [0.814-0.910]	AllGenes+XGB	0.984 [0.972-0.994]	0.991	0.899 [0.877-0.919]	0.849
AllGenes+CatBoost	0.999 ± 0.002	N/A	0.984 [0.973-0.994]	0.869 [0.842-0.899]	AllGenes+CatBoost	0.986 [0.975-0.994]	0.992	0.897 [0.875-0.919]	0.925

confirmation in an untouched external cohort will be required before the post hoc ANOVA500+XGB configuration can be regarded as a validated final model.

VIII. Data and Code Availability

The original TCGA, ICGC Release 28, and GSE2109 expression data are available from their respective public repositories. The ICGC projects analyzed were BRCA-KR, RECA-EU, OV-AU, PACA-AU, PACA-CA, PRAD-CA, and PRAD-FR; GSE2109 is available from GEO under accession GSE2109. Study-specific analysis scripts, sample manifests, processed data, harmonized gene lists, and derived result tables are available from the corresponding author upon reasonable request.

ACKNOWLEDGEMENT

This research was supported by the Korea-EU Joint Research Program through the National Research Foundation of Korea (NRF), funded by the Ministry of Science and ICT (MSIT), Republic of Korea (Grant No. RS-2023-00302123).

REFERENCES

- [1] The Cancer Genome Atlas Research Network, J. N. Weinstein, E. A. Collisson, G. B. Mills, et al., "The Cancer Genome Atlas Pan-Cancer analysis project," *Nature Genetics*, 45(10), 1113–1120, 2013. DOI: 10.1038/ng.2764.
- [2] Y. Li, K. Kang, J. M. Krahn, N. Croutwater, K. Lee, D. M. Umbach, and L. Li, "A comprehensive genomic pan-cancer classification using The Cancer Genome Atlas gene expression data," *BMC Genomics*, 18, 508, 2017. DOI: 10.1186/s12864-017-3906-0.
- [3] J. K. Grewal, B. Tessier-Cloutier, M. Jones, et al., "Application of a Neural Network Whole Transcriptome-Based Pan-Cancer Method for Diagnosis of Primary and Metastatic Cancers," *JAMA Network Open*, 2(4), e192597, 2019. DOI: 10.1001/jamanetworkopen.2019.2597.
- [4] Y. Hao, H. Huang, D. Huang, J. Ruan, X. Liu, and J. Zhang, "OncoTrace-TOO: Interpretable Machine Learning Framework for Cancer Tissue-of-Origin Identification Using Transcriptomic Signatures," *Cancer Reports*, 8(8), e70311, 2025. DOI: 10.1002/cnr2.70311.
- [5] J. Wang, J. Zhang, X. Dai, C. Yan, and C. Fang, "Computational models for pan-cancer classification based on multi-omics data," *Frontiers in Genetics*, 16, 1667325, 2025. DOI: 10.3389/fgene.2025.1667325.
- [6] C. Bernau, M. Riester, A.-L. Boulesteix, G. Parmigiani, C. Huttenhower, L. Waldron, and L. Trippa, "Cross-study validation for the assessment of prediction algorithms," *Bioinformatics*, 30(12), i105–i112, 2014. DOI: 10.1093/bioinformatics/btu279.
- [7] SEQC/MAQC-III Consortium, "A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium," *Nature Biotechnology*, 32, 903–914, 2014. DOI: 10.1038/nbt.2957.
- [8] A. A. Shabalín, H. Tjelmeland, C. Fan, C. M. Perou, and A. B. Nobel, "Merging two gene-expression studies via cross-platform normalization," *Bioinformatics*, 24(9), 1154–1160, 2008. DOI: 10.1093/bioinformatics/btn083.
- [9] J. A. Thompson, J. Tan, and C. S. Greene, "Cross-platform normalization of microarray and RNA-seq data for machine learning applications," *PeerJ*, 4, e1621, 2016. DOI: 10.7717/peerj.1621.
- [10] J. M. Franks, G. Cai, and M. L. Whitfield, "Feature specific quantile normalization enables cross-platform classification of molecular subtypes using gene expression data," *Bioinformatics*, 34(11), 1868–1874, 2018. DOI: 10.1093/bioinformatics/bty026.
- [11] S. M. Foltz, C. S. Greene, and J. N. Taroni, "Cross-platform normalization enables machine learning model training on microarray and RNA-seq data simultaneously," *Communications Biology*, 6, 222, 2023. DOI: 10.1038/s42003-023-04588-6.
- [12] Q. Guan, R. Chen, H. Yan, et al., "Differential expression analysis for individual cancer samples based on robust within-sample relative gene expression orderings across multiple profiling platforms," *Oncotarget*, 7(42), 68909–68920, 2016. DOI: 10.18632/oncotarget.11996.
- [13] I. Guyon and A. Elisseeff, "An Introduction to Variable and Feature Selection," *Journal of Machine Learning Research*, 3, 1157–1182, 2003.
- [14] C. Ding and H. Peng, "Minimum redundancy feature selection from microarray gene expression data," *Journal of Bioinformatics and Computational Biology*, 3(2), 185–205, 2005. DOI: 10.1142/S0219720005001004.
- [15] H. Zou and T. Hastie, "Regularization and Variable Selection Via the Elastic Net," *Journal of the Royal Statistical Society: Series B*, 67(2), 301–320, 2005. DOI: 10.1111/j.1467-9868.2005.00503.x.
- [16] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting

- System," Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794, 2016. DOI: 10.1145/2939672.2939785.
- [17] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "CatBoost: unbiased boosting with categorical features," *Advances in Neural Information Processing Systems*, 31, 2018.
- [18] J. G. Moreno-Torres, T. Raeder, R. Alaiz-Rodriguez, N. V. Chawla, and F. Herrera, "A unifying view on dataset shift in classification," *Pattern Recognition*, 45(1), 521–530, 2012. DOI: 10.1016/j.patcog.2011.06.019.
- [19] J. T. Leek, R. B. Scharpf, H. C. Bravo, et al., "Tackling the widespread and critical impact of batch effects in high-throughput data," *Nature Reviews Genetics*, 11, 733–739, 2010. DOI: 10.1038/nrg2825.
- [20] S. M. Lundberg, G. Erion, H. Chen, et al., "From local explanations to global understanding with explainable AI for trees," *Nature Machine Intelligence*, 2, 56–67, 2020. DOI: 10.1038/s42256-019-0138-9.
- [21] A. Subramanian, P. Tamayo, V. K. Mootha, et al., "Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles," *Proceedings of the National Academy of Sciences USA*, 102(43), 15545–15550, 2005. DOI: 10.1073/pnas.0506580102.
- [22] J. Zhang, R. Bajari, D. Andric, et al., "The International Cancer Genome Consortium Data Portal," *Nature Biotechnology*, 37(4), 367–369, 2019. DOI: 10.1038/s41587-019-0055-9.
- [23] R. Edgar, M. Domrachev, and A. E. Lash, "Gene Expression Omnibus: NCBI gene expression and hybridization array data repository," *Nucleic Acids Research*, 30(1), 207–210, 2002. DOI: 10.1093/nar/30.1.207.
- [24] R. C. Gentleman, V. J. Carey, D. M. Bates, et al., "Bioconductor: open software development for computational biology and bioinformatics," *Genome Biology*, 5, R80, 2004. DOI: 10.1186/gb-2004-5-10-r80.
- [25] W. E. Johnson, C. Li, and A. Rabinovic, "Adjusting batch effects in microarray expression data using empirical Bayes methods," *Biostatistics*, 8(1), 118–127, 2007. DOI: 10.1093/biostatistics/kxj037.
- [26] F. Pedregosa, G. Varoquaux, A. Gramfort, et al., "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, 12, 2825–2830, 2011.
- [27] Z. Fang, X. Liu, and G. Peltz, "GSEAPy: a comprehensive package for performing gene set enrichment analysis in Python," *Bioinformatics*, 39(1), btac757, 2023. DOI: 10.1093/bioinformatics/btac757.
- [28] G. W. Brier, "Verification of Forecasts Expressed in Terms of Probability," *Monthly Weather Review*, 78(1), 1–3, 1950. DOI: 10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2.
- [29] A. Liberzon, C. Birger, H. Thorvaldsdottir, M. Ghandi, J. P. Mesirov, and P. Tamayo, "The Molecular Signatures Database Hallmark gene set collection," *Cell Systems*, 1(6), 417–425, 2015. DOI: 10.1016/j.cels.2015.12.004.
- [30] M. Milacic, D. Beavers, P. Conley, et al., "The Reactome Pathway Knowledgebase 2024," *Nucleic Acids Research*, 52(D1), D672–D678, 2024. DOI: 10.1093/nar/gkad1025.
- [31] The Cancer Genome Atlas Network, "Comprehensive molecular portraits of human breast tumours," *Nature*, 490, 61–70, 2012. DOI: 10.1038/nature11412.
- [32] The Cancer Genome Atlas Research Network, "Comprehensive molecular characterization of clear cell renal cell carcinoma," *Nature*, 499, 43–49, 2013. DOI: 10.1038/nature12222.
- [33] The Cancer Genome Atlas Research Network, "Integrated genomic analyses of ovarian carcinoma," *Nature*, 474, 609–615, 2011. DOI: 10.1038/nature10166.
- [34] The Cancer Genome Atlas Research Network, "Integrated Genomic Characterization of Pancreatic Ductal Adenocarcinoma," *Cancer Cell*, 32(2), 185–203.e13, 2017. DOI: 10.1016/j.ccell.2017.07.007.
- [35] The Cancer Genome Atlas Research Network, "The Molecular Taxonomy of Primary Prostate Cancer," *Cell*, 163(4), 1011–1025, 2015. DOI: 10.1016/j.cell.2015.10.025.

Authors



Sungmin Yang received the B.S. degree in Computer Engineering from University, and the M.S. and Ph.D degrees from the Seoul National University, Seoul, Korea. Dr. Yang is currently a Postdoctoral Researcher at

Seoul National University. His research interests include bioinformatics, artificial intelligence, machine learning, deep learning, cancer genomics, single-cell and multi-omics data analysis, computational biology, and biomedical data mining.



Hong-Gee Kim received the B.S degree from Korea University, Seoul, Korea, in 1985, and the M.S. and Ph.D degrees from the University of Georgia, USA, in 1993 and 1996, respectively.

Dr. Kim joined the faculty of the School of Dentistry at Seoul National University, Seoul, Korea, in 2005. He is currently a Professor of Biomedical Informatics and the Director of the Biomedical Knowledge Engineering Laboratory at Seoul National University. His research interests include biomedical informatics, ontology engineering, biomedical knowledge graphs, semantic web technologies, artificial intelligence, machine learning, and multi-omics data integration.