A Conserved Fibrinogen and Immune Evasion Gene Signature Predicts Mortality Across Lung Cancer Histological Subtypes: An Interpretable Machine Learning Discovery Study

Simbarashe G. Magwenzi
NYNOSK LLP, 71-75 Shelton Street, Covent Garden, London, United Kingdom, WC2H 9JQ
Corresponding author: simbarashe.magwenzi@nynosk.com

Running title: Fibrinogen and Immune Evasion Predict Lung Cancer Mortality
Keywords: lung cancer prognosis; fibrinogen; kynurenine pathway; KYNU; machine learning; SHAP; pan-lung cancer; immune evasion; XGBoost; external validation

Abstract

Background
TNM staging systematically misclassifies patients with favourable outcomes as high-risk, with direct consequences for treatment decisions. Whether the molecular drivers of lung cancer mortality are conserved across histological subtypes remains largely uncharacterised.

Methods
An XGBoost classifier was trained on 104 features (100 gene expression probes and four clinical variables) in 440 lung adenocarcinoma (LUAD) patients (GSE68465). A three-model ablation study quantified the independent contribution of gene expression over clinical staging alone. The trained model was applied without modification to an independent cohort (GSE30219, n=287), stratified into adenocarcinoma (n=85) and mixed-histology (n=202) subsets. SHapley Additive exPlanations (SHAP), Kaplan-Meier survival analysis, and decision curve analysis assessed biological and clinical utility.

Results
Clinical staging alone misclassified 49% of surviving patients as high-risk (specificity=0.51). Adding gene expression reduced this to 29% (specificity=0.71), corresponding to approximately 93 fewer incorrect high-risk designations per 1,000 patients. The full model achieved AUC=0.73 in the training test set, with stable external validation in the LUAD (AUC=0.71) and mixed-histology (AUC=0.69) subsets. When applied to the mixed-histology subset, the LUAD-trained model maintained near-identical biological feature hierarchies, with 12 of 14 top SHAP features shared across all three evaluation cohorts. The dominant signal in both external validation cohorts was the fibrinogen chain gene pair *FGG* and *FGA*, outranking pathological nodal stage and implicating coagulation-mediated tumour immune exclusion as a conserved mortality mechanism. Kynureninase (*KYNU*), a key mediator of IDO/TDO-mediated immune evasion, was independently recovered as a top predictor. Risk stratification significantly separated disease-free survival in both validation subsets (LUAD: log-rank p=0.044; mixed-histology: log-rank p=0.011).

Conclusion
Gene expression substantially improves survivor identification over clinical staging alone and reveals a conserved molecular signature of lung cancer mortality transcending histological boundaries. The dominance of fibrinogen and kynurenine pathway components across multiple histotypes suggests histotype-independent mechanisms of aggressiveness with direct implications for prognostic stratification and therapeutic targeting.

Introduction

Lung cancer kills more patients globally than any other malignancy, accounting for approximately 1.8 million deaths annually.[1] For the majority of patients with resected or localised disease, the primary prognostic instrument remains TNM pathological staging, an anatomical classification system that, despite successive refinements, captures only part of the biological information that determines whether a patient survives.[2] Within any given stage group, outcome varies substantially, depending on molecular features that staging does not measure.[2] The practical consequence of this imprecision is systematic misclassification: patients with molecularly favourable tumours may be labelled high-risk on the basis of anatomical staging and receive adjuvant treatment they do not require, whilst patients with molecularly aggressive tumours in early pathological stages may be under-monitored.

This staging gap creates a direct clinical need for molecular prognostic information that complements TNM classification. Gene expression profiling of lung tumours has produced numerous candidate signatures, yet most published work shares a common limitation: signatures are typically derived from and validated within a single histological subtype, most often adenocarcinoma. This leaves open a fundamental biological question that has rarely been directly addressed: whether the molecular mechanisms that drive lung cancer mortality are specific to each histological subtype, or whether they reflect conserved oncological processes shared across the lung cancer spectrum.

This question is clinically important. Lung cancer encompasses histologically diverse tumours, including adenocarcinoma, squamous cell carcinoma, large cell neuroendocrine carcinoma, basaloid carcinoma, small cell carcinoma, and others, each treated as a distinct disease with distinct therapeutic approaches. Yet the biological processes most likely to drive mortality, such as coagulation-mediated metastatic facilitation, immune evasion, metabolic reprogramming, and chromosomal instability, are not inherently histotype-specific. If a common molecular signature underlies poor outcomes across histological boundaries, its identification will have broad prognostic and therapeutic implications that no single-histotype study could reveal.

An XGBoost classifier[3] was therefore trained on transcriptomic and clinical data from 440 LUAD patients and applied, without modification or retraining, to an independent cohort of 287 patients spanning adenocarcinoma and multiple additional lung cancer histotypes. SHapley Additive exPlanations (SHAP)[4,5] served as an analytical instrument to identify which molecular signals the model captured and whether those signals were consistent across histological boundaries. The classifier was used as a tool for interrogating the transcriptomic basis of lung cancer mortality rather than an end in itself. The degree of cross-histology concordance that emerged, and the biological identity of the conserved signal, form the central contribution of this study.

Materials and Methods

Datasets

Gene expression data were accessed from the NCBI Gene Expression Omnibus (GEO).[6]

Training cohort (GSE68465): This publicly available dataset derives from the Director's Challenge Consortium for the Molecular Classification of Lung Adenocarcinoma,[7,8] a large, prospectively designed, multi-institutional study of gene expression-based survival prediction in

LUAD, widely regarded as a reference standard dataset for this indication. Microarray profiling was performed on the Affymetrix Human Genome U133A platform. From 462 samples, 19 non-tumoral normal lung controls were excluded. Two further samples were excluded for missing pathological stage data (pT or pN) and one for unconfirmed nodal stage (pX), yielding a final training cohort of 440 LUAD patients.

Validation cohort (GSE30219): This dataset derives from a study of epigenetically regulated gene expression in lung cancer [9,10] and comprises 307 patients profiled on the Affymetrix Human Genome U133 Plus 2.0 platform at a French centre. Following exclusion of 14 non-tumoral lung (NTL) samples and six samples with unconfirmed pathological staging (TX and/or NX), 287 patients were retained. To evaluate both histological concordance and cross-histology generalisability, the validation cohort was stratified into an adenocarcinoma subset (n=85) and a mixed-histology subset (n=202) comprising squamous cell carcinoma (SQC), large cell neuroendocrine carcinoma (LCNE), basaloid carcinoma (BAS), carcinoid (CARCI), small cell carcinoma (SCC), large cell carcinoma (LCC), and other types.

Preprocessing

Raw microarray data were processed using the GEOquery and limma packages in R.[11,12] Data were $\log_2$-transformed and Z-score normalised across samples within each dataset independently. The training cohort (GSE68465) was processed on the Affymetrix U133A platform and the validation cohort (GSE30219) on the Affymetrix U133 Plus 2.0 platform; both platforms share substantial probe overlap, and Z-score normalisation reduces systematic inter-platform differences at the gene expression level. No additional batch correction was applied. Any residual platform-specific technical variation would be expected to attenuate cross-cohort discrimination rather than inflate it, meaning that the observed generalisation performance represents a conservative estimate of the model's true prognostic capacity. Platform difference is acknowledged as a limitation of this study.

Differential Expression Analysis and Feature Selection

Differential expression between deceased and surviving patients in the training cohort was assessed using the limma empirical Bayes framework,[12] with Benjamini-Hochberg correction for multiple testing. Affymetrix internal control probes (AFFX- prefix) and probes without annotated gene symbols were excluded prior to feature selection, as these represent hybridisation quality controls rather than gene expression measurements. From the remaining 1,830 annotated, statistically significant probes (Padj < 0.05), the 50 with the largest positive $\log_2$ fold-change (upregulated in deceased patients) and the 50 with the largest negative $\log_2$ fold-change (downregulated in deceased patients) were selected as gene expression features. Combined with four clinical variables (age, sex, pN stage, pT stage), this yielded 104 model input features.

Model Development and Ablation

An XGBoost classifier[3] was trained with optimised hyperparameters that are available from the corresponding author upon reasonable request. Data were stratified and scaled prior to a fixed 80:20 stratified train-test split (n=352 training, n=88 held-out test). To quantify the independent prognostic contribution of gene expression relative to clinical staging, two ablation models were trained identically: a clinical-only model (age, sex, pN stage, pT stage) and a gene expression-only model (100 probes). All three configurations were evaluated on the same held-out test set. Model stability was assessed by stratified five-fold cross-validation on the training partition only (n=352). A classification threshold of 0.40 was established on training data following grid search optimisation.

External Validation

The trained model was applied without retraining to the GSE30219 cohort. Feature columns matching the 104 training features were extracted from the pre-processed validation data. Predictions were generated independently for the LUAD and mixed-histology subsets.

Performance Evaluation

Threshold-independent discrimination was assessed by the area under the receiver operating characteristic curve (AUC) and the area under the precision-recall curve (AUPRC), with 95% confidence intervals from 1,000 stratified bootstrap iterations (seed=42). Threshold-dependent metrics at 0.40 included accuracy, precision (positive predictive value), sensitivity, specificity, and F1 score. Calibration was assessed graphically. Clinical utility was quantified by decision curve analysis,[13] assessing net benefit relative to treat-all and treat-none strategies across probability thresholds of 0.05 to 0.60.

Interpretability

SHAP beeswarm plots were generated for the full training cohort (n=440) and both validation subsets (LUAD n=85, non-LUAD n=202) using the Python SHAP library.[4,5] Pairwise Spearman rank correlations between the top SHAP-identified gene features were computed for all three cohorts and visualised as annotated heatmaps with significance indicated at three levels ($p<0.001$, $p<0.01$, $p<0.05$).

Survival Analysis

Kaplan-Meier survival functions[14] were estimated using the Python lifelines library.[15] Patients were stratified into high-risk (predicted probability ≥0.40) and low-risk (< 0.40) groups based on the XGBoost model score. Survival analysis was performed separately in the LUAD and mixed-histology validation subsets and results were not pooled. Between-group differences were assessed by the log-rank test.[16] At-risk and event tables were generated at 50-month intervals.

A note on endpoint concordance: the training cohort provided overall survival (vital status: deceased or surviving) as the binary classification endpoint. Kaplan-Meier analysis in the validation cohort used disease-free survival (time to first relapse). In resected lung cancer, time to first progression and disease-free survival are considered functionally concordant endpoints, as most deaths in this population are preceded by documented disease recurrence. Cohort-specific survival curves are presented and interpreted independently.

Statistical Analysis and Software

All analyses were performed in Python 3.13 using XGBoost,[3] SHAP,[4,5] scikit-learn,[17] lifelines,[15] pandas, NumPy, matplotlib, and seaborn. Differential expression and preprocessing used R with limma[12] and GEOquery.[11] Reporting followed the TRIPOD guidelines for prediction model studies.[18]

Results

Patient Characteristics

Full patient characteristics for all three cohorts are presented in Table 1. The training cohort (GSE68465, n=440 LUAD) was approximately sex-balanced (221 male, 219 female), with a mean age of 64.44 ± 10.11 years. Pathological staging was predominantly early stage, with 149 patients at pT1 (33.9%) and 251 at pT2 (57.0%), and nodal spread in 141 patients (pN1 to pN2, 32.0%). Mortality was 53.4% (235 deceased, 205 surviving).

The LUAD validation subset (GSE30219, n=85) was predominantly male (66/19), with a mean age of 61.49 ± 9.28 years. Its staging distribution was substantially more favourable than the training cohort, with 71 of 85 patients at pT1 (83.5%) compared with 33.9% in training, and 82 of 85 at pN0 (96.5%) compared with 68.0% in training. This difference is clinically relevant: the limited nodal stage variability in the LUAD validation subset reduces the discriminative contribution of pN stage in that cohort and likely explains why the gene expression features, particularly *FGG*, rise to rank first in SHAP importance during validation. Mortality was 52.9% (45 deceased, 40 surviving).

The mixed-histology validation subset (n=202) comprised 179 male and 23 female patients, with a mean age of 61.53 ± 12.35 years, a broader stage distribution (pT1 to pT4, pN0 to pN3), and a higher mortality rate of 73.8% (149 deceased, 53 surviving). The pronounced sex imbalance in both GSE30219 subsets reflects the demographic composition of lung cancer patients at the French recruiting centre during the study period.

Differential Expression and Feature Selection

Differential expression analysis identified 1,880 probes at Padj < 0.05 from 22,283 tested. More probes were downregulated in deceased patients (1,128) than upregulated (752), suggesting that loss of gene expression, whether reflecting reduced transcriptional activity, downregulation of differentiation-associated programmes, or loss of tumour-protective gene expression, is a more prevalent feature of fatal LUAD than upregulation of oncogenic programmes. Following exclusion of Affymetrix control probes and unannotated probes, the 50 most upregulated and 50 most downregulated annotated probes by $\log_2$ fold-change were selected for model training (Figure 2).

Gene Expression Substantially Reduces Misclassification of Survivors: The Clinical Staging Gap

The results of the three-model ablation study are presented in Table 2. The clinical-only model (pT stage, pN stage, age, sex) achieved a specificity of 0.51 in the held-out test set, indicating that 49% of patients who survived were incorrectly classified as high-risk using staging information alone. In practical terms, for every 1,000 LUAD patients evaluated, clinical staging would incorrectly label approximately 228 survivors as high-risk. The gene expression-only model (specificity 0.54) partially addressed this gap, confirming that transcriptomic data carry independent prognostic information not captured by staging variables.

The full multimodal model, combining gene expression with clinical staging, achieved a specificity of 0.71, reducing the false-positive rate from 49% to 29%. For every 1,000 LUAD patients, this corresponds to approximately 93 fewer survivors incorrectly classified as high-risk compared with staging alone. Improvement was observed across all performance metrics: AUC increased from 0.65 (clinical-only) to 0.73 (full model), and F1 score from 0.65 to 0.72. Five-fold cross-validation confirmed model stability, with a mean AUC of 0.707 (SD 0.051). Cross-validation mean AUC for the gene expression-only model (0.682) exceeded that of the clinical-only model (0.675), with the minimum fold AUC higher for the gene expression-only model (0.629) than for the clinical-only model (0.618), indicating that gene expression features are somewhat less susceptible to performance variability across individual data partitions.

Cross-Histology Generalisation: A Conserved Mortality Signal Across Multiple Lung Cancer Subtypes

The degree of cross-histology concordance that emerged when the LUAD-trained model was applied to the mixed-histology validation subset exceeded initial expectations. The model achieved an AUC of 0.69 (95% CI: 0.60 to 0.78), an AUPRC of 0.82 against a baseline prevalence of 0.74, precision of 0.83, and significantly stratified disease-free survival (log-rank p=0.011),

with high-risk patients experiencing a 52.2% cumulative relapse rate compared with 32.9% in the low-risk group (Figure 7B). The AUC is unaffected by class prevalence and confirms genuine discriminative capacity independent of event rate.

In the LUAD-specific validation subset, performance was comparable: AUC 0.71 (95% CI: 0.58 to 0.81), with survival stratification demonstrating a two-fold difference in cumulative relapse rate between high-risk (42.9%) and low-risk (20.9%) groups (log-rank p=0.044; Figure 7A). The consistency of AUC across training (0.73), LUAD validation (0.71), and mixed-histology validation (0.69), a decline of only 0.04 across three independent evaluations spanning two independent cohorts, multiple histotypes, and two countries, is consistent with the model capturing biologically conserved signals rather than cohort-specific technical artefacts. Full performance metrics for both validation cohorts are presented in Table 3.

SHAP Analysis Identifies the Conserved Signal: The Fibrinogen Axis and Immune Evasion.

SHAP beeswarm analysis, performed independently on the full training cohort (n=440, Figure 3) and both validation subsets (n=85 LUAD and n=202 mixed-histology, Figure 4), revealed the molecular identity of the signal driving cross-histology generalisation. Twelve of the fourteen top SHAP-contributing features were shared across all three independent evaluations. Two clinical variables, pN stage and age, appeared consistently among the top contributors in the training cohort, confirming that the model correctly weights established prognostic determinants. Among the gene expression features, an identical set of eleven probes appeared in the top SHAP contributors across training, LUAD validation, and mixed-histology validation cohorts: *FGG*, *LAMA2*, *KYNU*, *NUP62CL*, *SLC15A1*, *CPS1*, *MFAP4*, *LY6D*, *KRT6C*/*KRT6B*/*KRT6A*, *SLC2A5*, and *NDC80*. *FGA* appeared additionally in both validation cohorts.

The fibrinogen axis

The most salient finding within the SHAP signature was the dominance of *FGG* (fibrinogen gamma chain) and *FGA* (fibrinogen alpha chain) in both external validation cohorts. In the LUAD and mixed-histology validation subsets alike, *FGG* ranked first among all features, outranking pathological nodal stage. Both genes encode subunits of fibrinogen, the hepatic precursor of fibrin that is central to haemostasis, thrombosis, and tumour microenvironment biology. Elevated circulating fibrinogen has been associated with worse outcomes in lung and other solid tumours across multiple observational studies.[19,20] Fibrin(ogen) deposition in tumours has been shown to facilitate metastasis by impairing natural killer cell-mediated elimination of circulating tumour cells[21] and by creating a pro-inflammatory, immunosuppressive stromal niche. The prognostic dominance of *FGG* and *FGA* expression in the primary tumour, observed across adenocarcinoma and multiple additional histotypes in an independent cohort to which the model was not exposed during training, supports the interpretation that intratumoural fibrinogen biology is a conserved determinant of lung cancer mortality. Fibrinogen is a routinely measured clinical laboratory marker; this finding provides specific mechanistic justification for prospective evaluation of tumour fibrinogen expression as a prognostic and potentially therapeutic target.

Kynurenine pathway immune evasion

KYNU (kynureninase) ranked consistently among the top five SHAP contributors across all three cohorts and was upregulated in deceased patients throughout. KYNU encodes the enzyme that catabolises kynurenine within the tryptophan degradation pathway, regulated by indoleamine 2,3-dioxygenase (IDO) and tryptophan 2,3-dioxygenase (TDO). This pathway is an established mechanism of tumour immune evasion: kynurenine accumulation in the tumour microenvironment suppresses cytotoxic T-lymphocyte function and promotes immunosuppressive regulatory T-cell differentiation.[22] The consistent identification of *KYNU* as a top mortality predictor by an unbiased transcriptomic analysis, replicated across

adenocarcinoma and mixed histotypes in two independent countries, provides convergent observational evidence for the prognostic importance of this pathway and for the biological rationale of targeting it therapeutically.

Additional signature component

LAMA2 (laminin subunit alpha-2), consistently downregulated in deceased patients, is compatible with basement membrane disruption facilitating invasion and epithelial-to-mesenchymal transition. *NDC80*, upregulated in deceased patients, is a kinetochore complex protein whose overexpression marks chromosomal instability, a hallmark of aggressive tumour biology. The metabolic transporters *SLC15A1* (PEPT1) and *SLC2A5* (GLUT5) reflect altered nutrient uptake consistent with metabolic reprogramming in aggressive tumours. *NUP62CL*, a nucleoporin, may reflect nuclear transport dysregulation in proliferating, molecularly unstable tumours. *MFAP4* (microfibrillar-associated protein 4), consistently downregulated in deceased patients, represents extracellular matrix disruption. Taken together, the biological portrait is coherent: the features driving poor prognosis are not specific to any single histological architecture but represent fundamental mechanisms of tumour aggressiveness operating across the lung cancer spectrum.

Pairwise Gene Co-expression Confirms Biological Coherence of the Signature

Pairwise Spearman correlations between the top SHAP-identified genes (Figure 5) were generally modest ($|r| < 0.30$ for most pairs), confirming that the top features are largely independent biological contributors. Two associations of biological interest emerged consistently. The *FGG-FGA* correlation was the strongest off-diagonal association in both training ($r=0.67$, $p<0.01$) and LUAD validation ($r=0.80$, $p<0.001$) cohorts, consistent with co-regulation of fibrinogen alpha and gamma chains; despite their correlation, both contribute independently to SHAP-ranked predictions. A consistent negative correlation between *MFAP4* and *NDC80* (training: $r=-0.44$, $p<0.001$; LUAD validation: $r=-0.53$, $p<0.01$) suggests a biological trade-off between extracellular matrix integrity and proliferative chromosomal instability. In the mixed-histology cohort, *LY6D* and *KRT6C*/*KRT6B*/*KRT6A* displayed strong co-expression ($r=0.72$, $p<0.001$), consistent with enrichment for squamous epithelial gene programmes in that cohort.

Risk Score Stratifies Disease-Free Survival Independently of Histotype

Kaplan-Meier analysis confirmed that the XGBoost risk score stratifies long-term clinical outcomes in both validation subsets independently (Figure 7). In the LUAD subset (n=85), high-risk patients (n=42) experienced a cumulative relapse rate of 42.9% (18 events) compared with 20.9% (9 events) in the low-risk group (n=43; log-rank $p=0.044$). Confidence intervals were appropriately wide given the available sample size and this result should be interpreted as exploratory. In the mixed-histology subset (n=202), high-risk patients (n=115) experienced a cumulative relapse rate of 52.2% (60 events) compared with 32.9% (24 events) in the low-risk group (n=73; log-rank $p=0.011$). The capacity of a classifier trained exclusively on adenocarcinoma to achieve significant survival stratification in a mixed-histology cohort to which it was not exposed during training is the most direct demonstration available in this study that the identified molecular signal reflects conserved rather than histotype-specific biology.

Clinical Utility Assessment

Decision curve analysis (Figure 8) evaluated net benefit across probability thresholds. In the LUAD validation subset (Figure 8A), the model provided net benefit above the treat-none strategy across all assessed thresholds. At threshold 0.40, model net benefit was approximately equivalent to treat-all, interpretable as clinically neutral at this prevalence (52.9% mortality). At thresholds

above 0.40, where the clinical cost of false positives increases, the model provided clear net benefit above treat-all.

In the mixed-histology validation subset (Figure 8B), the model provided net benefit above treat-none throughout. At a mortality prevalence of 73.8%, the treat-all strategy carries an inherently high net benefit that is mathematically difficult for a selective classifier to exceed; this reflects a property of decision curve analysis in high-prevalence cohorts rather than a limitation of model discrimination.

Discussion

A Conserved Molecular Signature of Lung Cancer Mortality

The degree of cross-histology concordance observed in this study was the most informative finding to emerge from the analysis. A classifier trained exclusively on adenocarcinoma, when applied to multiple additional histologically distinct lung cancer subtypes, identified near-identical biological features as the dominant predictors of mortality, with twelve of fourteen top SHAP features shared across all three independent evaluation cohorts spanning two countries, two platforms, and multiple histological architectures. This is not a performance result. It is a biological observation: it indicates that the molecular processes underlying lung cancer mortality are, in substantial part, conserved across histological subtypes rather than determined by the transcriptional programmes specific to each cell of origin. The convergence of those features on fibrinogen biology and kynurenine pathway signalling, replicated across training and two validation cohorts, provides a degree of confidence in these observations that no single-cohort or single-histotype study could achieve.

The Fibrinogen Axis: A Pan-Lung-Cancer Mortality Mechanism with Clinical Implications

FGG and *FGA*, encoding the gamma and alpha chains of fibrinogen respectively, ranked as the single most important predictive feature in both external validation cohorts, outranking pathological nodal stage. Fibrinogen is an acute-phase plasma protein synthesised by the liver and well-characterised in the context of coagulation. Its role in tumour biology extends beyond thrombosis: fibrin(ogen) deposition in the tumour microenvironment creates a scaffold that impairs natural killer cell-mediated cytotoxicity,[21] promotes angiogenesis, facilitates immune exclusion, and generates a pro-inflammatory stromal milieu that supports metastatic dissemination. Elevated plasma fibrinogen has been associated with worse outcomes in lung and other solid tumours.[19,20]

The present findings extend this literature in an important direction: it is not solely circulating fibrinogen levels but the expression of fibrinogen chain genes within the tumour itself that carries prognostic weight, and this intratumoural fibrinogen signal is conserved across adenocarcinoma, squamous cell carcinoma, neuroendocrine tumours, and other subtypes. The strong co-expression of *FGG* and *FGA* (r=0.67 to 0.80 across cohorts), alongside their independent SHAP contributions, indicates that both chains provide non-redundant prognostic information and that fibrinogen biology within the tumour is organised as a coordinated programme. Fibrinogen is a routine clinical laboratory measurement; this finding provides mechanistic justification for prospective evaluation of intratumoural fibrinogen expression as a prognostic biomarker. Whether anti-coagulant or anti-fibrinogen interventions may have therapeutic value in fibrinogen-high tumours is a hypothesis that these findings motivate but cannot confirm; prospective clinical evaluation would be required before any such conclusion could be drawn.

KYNU and the Kynurenine Pathway: Convergent Evidence for a Therapeutic Target

KYNU (kynureninase) appeared among the top five SHAP contributors in all three cohorts and was consistently upregulated in patients who died. The IDO1/TDO2-kynurenine-aryl hydrocarbon receptor axis is an established mechanism of adaptive immune evasion in cancer. [22] Kynurenine generated by tumour and stromal cells suppresses CD8+ T-lymphocyte function, promotes regulatory T-cell differentiation, and creates an immunosuppressive microenvironment. The IDO1 pathway has been investigated as a therapeutic target in non-small cell lung cancer and other tumour types; whilst clinical development has faced challenges, the mechanistic rationale for targeting kynurenine-mediated immune suppression remains under active investigation in the context of combination immunotherapy strategies.

The identification of *KYNU* as a top mortality predictor by an unbiased machine learning analysis, replicated independently across training and two external validation cohorts in two countries, constitutes convergent observational evidence for the prognostic importance of this pathway in lung cancer. That this signal is conserved across histotypes further supports the interpretation that IDO/TDO-mediated immune evasion is a broadly relevant feature of aggressive lung tumour biology. In the context of efforts to identify which patients are most likely to respond to, or resist, checkpoint immunotherapy, a transcriptomic predictor of kynurenine pathway activation may have utility beyond prognosis alone.

The Clinical Staging Gap and Its Implications for Treatment Decision-Making

The ablation study reveals a clinically important limitation of staging-based prognostication. The clinical-only model, comprising pT stage, pN stage, age, and sex, misclassified 49% of surviving patients as high-risk. For every 1,000 LUAD patients assessed, approximately 228 survivors would receive a high-risk designation based on this information alone. The multimodal model incorporating gene expression reduces this number to approximately 135, sparing approximately 93 patients per 1,000 from an incorrect high-risk classification.

The clinical consequences of high-risk misclassification are not trivial. In resected LUAD, a high-risk designation may influence recommendations for adjuvant chemotherapy, the intensity of post-operative surveillance, and clinical trial eligibility. Adjuvant chemotherapy in LUAD is associated with clinically meaningful toxicity, with absolute survival benefit that is modest and concentrated in patients with higher-risk disease.[23] Exposing patients with molecularly favourable tumours to these trade-offs based on staging misclassification represents an avoidable clinical cost. These data do not support replacement of clinical staging with molecular profiling but rather demonstrate that gene expression provides complementary information that meaningfully improves the identification of patients whose tumour biology is more favourable than their pathological stage implies.

What Cross-Histology Generalisation Reveals About Lung Cancer Biology

The observation that a LUAD-trained model maintains a near-identical feature hierarchy across multiple additional histotypes raises a question beyond prognosis: why do tumours of different cellular origins, arising through different mutational pathways, share the same molecular signature of poor outcomes? The most parsimonious explanation is that the features identified, including fibrinogen expression, kynurenine immune evasion, basement membrane disruption (*LAMA2*), chromosomal instability (*NDC80*), and metabolic reprogramming (*SLC15A1*, *SLC2A5*), are not part of the tumour's histotype-specific transcriptional programme, but are components of a shared biological response to tumour progression that transcends cell of origin. Aggressive tumour biology may converge on common mechanisms for immune escape, metabolic adaptation, and microenvironmental remodelling, regardless of histotype.

The consistent negative correlation between the extracellular matrix protein *MFAP4* and the chromosomal instability marker *NDC80* across cohorts reinforces this interpretation: tumours appear to polarise between an ECM-rich stromal programme and a proliferatively unstable, chromosomally disrupted programme, with the latter consistently associated with mortality. These observations are hypothesis-generating. Functional validation, including cell line and in vivo experimental work addressing the mechanistic roles of fibrinogen chain expression and KYNU activity in lung cancer aggressiveness across histotypes, will be required to confirm these inferences.

Limitations

The following limitations should be considered when interpreting these findings. First, both cohorts utilised Affymetrix microarray platforms; the absence of batch correction between the U133A and U133 Plus 2.0 arrays represents a methodological limitation, though any residual platform-specific technical variation would be expected to attenuate cross-cohort discrimination, making the observed generalisation a conservative estimate. Second, generalisation to RNA-sequencing-based clinical workflows has not been established. Third, the LUAD-specific external validation subset was small (n=85), and the Kaplan-Meier result in this subset (p=0.044) should be interpreted as exploratory. Fourth, the staging distribution of the LUAD validation subset (pT1: 83.5%, pN0: 96.5%) was substantially more favourable than the training cohort (pT1: 33.9%, pN0: 68.0%), which may have affected the relative contributions of clinical and gene expression features and limits direct comparison of feature hierarchies between training and LUAD validation SHAP analyses. Fifth, gene feature selection was data-driven rather than biologically pre-specified, and the identified features should be regarded as hypothesis-generating rather than confirmed therapeutic targets. Sixth, no functional validation of candidate genes was performed; biological associations are correlative. Seventh, the classification threshold (0.40) was optimised on training data; threshold-dependent metrics in validation cohorts may be mildly optimistic. Eighth, the survival endpoint differed between cohorts (overall survival in training, disease-free survival in validation), though these are considered functionally concordant in resected lung cancer. Ninth, the sex imbalance between cohorts (50% male in training vs more than 85% male in both validation subsets) may contribute to performance differences. Finally, prospective clinical validation has not been performed, and these findings do not support immediate clinical implementation without independent confirmation in prospective cohorts.

Conclusions

An interpretable machine learning analysis reveals that the molecular biology of lung cancer mortality is, in substantial part, conserved across histological subtypes. A classifier trained exclusively on adenocarcinoma identifies near-identical biological features as the dominant predictors of mortality across squamous cell carcinoma, neuroendocrine tumours, and additional subtypes, pointing to shared rather than histotype-specific mechanisms of lung cancer aggressiveness. The identity of that shared signal, centred on fibrinogen chain gene expression and kynurenine pathway immune evasion, has direct implications for prognostic stratification and provides mechanistic justification for further investigation of these pathways as therapeutic targets.

Beyond the biological discovery, the practical clinical finding is clear: gene expression substantially reduces the rate at which staging-based classification incorrectly labels patients with favourable outcomes as high-risk, with a 20 percentage-point improvement in specificity over clinical staging alone. The integration of molecular profiling with clinical staging has the

potential to meaningfully improve the precision of lung cancer prognostication and to reduce the burden of treatment given to patients whose tumour biology is more favourable than their anatomical stage implies.

Author Contributions

SGM: research conceptualisation; methodology; investigation; data curation; formal analysis; visualisation; manuscript writing, review and editing.

Acknowledgements

Acknowledgement and gratitude are extended to the Director's Challenge Consortium for the Molecular Classification of Lung Adenocarcinoma for generating and publicly depositing the GSE68465 dataset, and Rousseaux S and colleagues for generating and depositing the GSE30219 dataset. Both datasets were accessed through the NCBI Gene Expression Omnibus (GEO), a publicly available data repository maintained by the National Center for Biotechnology Information. The contributions of all original study participants and research teams whose data underpin these analyses are also gratefully acknowledged.

The preparation of this manuscript was supported by artificial intelligence (AI)-assisted services used for purposes including manuscript drafting, structural development and language editing. All scientific content, analytical decisions, data interpretation, and conclusions are the sole responsibility of the author, who has reviewed and approved the manuscript in its entirety.

Conflict of Interest

None.

Funding

None.

Data Availability Statement

All gene expression data are publicly available through the NCBI Gene Expression Omnibus under accession numbers GSE68465 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE68465) and GSE30219 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE30219).

Analysis code is available from the corresponding author upon reasonable request.

References

1. Sung H, Ferlay J, Siegel RL, et al. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA Cancer J Clin*. 2021;71(3):209-249. doi:10.3322/caac.21660

2. Goldstraw P, Chansky K, Crowley J, et al. The IASLC lung cancer staging project: Proposals for revision of the TNM stage groupings in the forthcoming (eighth) edition of the TNM

Classification for lung cancer. *Journal of Thoracic Oncology*. 2016;11(1):39-51. doi:10.1016/j.jtho.2015.09.009

3. Chen T, Guestrin C. XGBoost: A scalable tree boosting system. In: *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Vols 13-17-August-2016. Association for Computing Machinery; 2016:785-794. doi:10.1145/2939672.2939785

4. Lundberg SM, Allen PG, Lee SI. A Unified Approach to Interpreting Model Predictions. In: Guyon I, Luxburg UV, Bengio S, et al., eds. *Proceedings of the 31st International Conference on Neural Information Processing Systems*. Vol 30. Neural Information Processing Systems; 2017:4768-4777.

5. Lundberg SM, Erion G, Chen H, et al. From local explanations to global understanding with explainable AI for trees. *Nat Mach Intell*. 2020;2(1):56-67. doi:10.1038/s42256-019-0138-9

6. Edgar R. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res*. 2002;30(1):207-210. doi:10.1093/nar/30.1.207

7. Shedden K, Taylor JMG, Enkemann SA, et al. Gene expression-based survival prediction in lung adenocarcinoma: A multi-site, blinded validation study. *Nat Med*. 2008;14(8):822-827. doi:10.1038/nm.1790

8. National Center for Biotechnology Information. GSE68465. Gene Expression Omnibus (GEO). May 2, 2015. Accessed June 6, 2026. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE68465

9. Rousseaux S, Debernardi A, Jacquiau B, et al. Ectopic activation of germline and placental genes identifies aggressive metastasis-prone lung cancers. *Sci Transl Med*. 2013;5(186). doi:10.1126/scitranslmed.3005723

10. National Center for Biotechnology Information. GSE30219. Gene Expression Omnibus (GEO). May 24, 2013. Accessed June 6, 2026. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE30219

11. Sean D, Meltzer PS. GEOquery: A bridge between the Gene Expression Omnibus (GEO) and BioConductor. *Bioinformatics*. 2007;23(14):1846-1847. doi:10.1093/bioinformatics/btm254

12. Ritchie ME, Phipson B, Wu D, et al. Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res*. 2015;43(7):e47. doi:10.1093/nar/gkv007

13. Vickers AJ, Elkin EB. Decision curve analysis: A novel method for evaluating prediction models. *Medical Decision Making*. 2006;26(6):565-574. doi:10.1177/0272989X06295361

14. Kaplan EL, Meier P. Nonparametric Estimation from Incomplete Observations. *J Am Stat Assoc*. 1958;53(282):457-481.

15. Davidson-Pilon C. lifelines: survival analysis in Python. *J Open Source Softw*. 2019;4(40):1317. doi:10.21105/joss.01317

16. Mantel N. Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemother Rep*. 1966;50(3):163-170.

17. Pedregosa F, Michel V, Grisel O, et al. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*. 2011;12:2825-2830.

18. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD Statement. *BMC Med*. 2015;13(1). doi:10.1186/s12916-014-0241-z

19. Zhong H, Qian Y, Fang S, Wang Y, Tang Y, Gu W. Prognostic value of plasma fibrinogen in lung cancer patients: A meta-analysis. *J Cancer*. 2018;9(21):3904-3911. doi:10.7150/jca.26360

20. Zhang K, Xu Y, Tan S, Wang X, Du M, Liu L. The association between plasma fibrinogen levels and lung cancer: A meta-analysis. *J Thorac Dis*. 2019;11(11):4492-4500. doi:10.21037/jtd.2019.11.13

21. Palumbo JS, Talmage KE, Massari J V., et al. Platelets and fibrin(ogen) increase metastatic potential by impeding natural killer cell–mediated elimination of tumor cells. *Blood*. 2005;105(1):178-185. doi:10.1182/blood-2004-06-2272

22. Opitz CA, Litzenburger UM, Sahm F, et al. An endogenous tumour-promoting ligand of the human aryl hydrocarbon receptor. *Nature*. 2011;478(7368):197-203. doi:10.1038/nature10491

23. Opitz CA, Litzenburger UM, Sahm F, et al. *An Endogenous Ligand of the Human Aryl Hydrocarbon Receptor Promotes Tumor Formation*.

24. Pignon JP, Tribodet H, Scagliotti G V., et al. Lung adjuvant cisplatin evaluation: A pooled analysis by the LACE collaborative group. *Journal of Clinical Oncology*. 2008;26(21):3552-3559. doi:10.1200/JCO.2007.13.9030

Tables

Table 1. Patient Characteristics

Characteristic	Training GSE68465 LUAD (n=440)	LUAD Validation GSE30219 (n=85)	Mixed-Histology Validation GSE30219 (n=202)
Sex, n (%)			
Male	221 (50.2)	66 (77.6)	179 (88.6)
Female	219 (49.8)	19 (22.4)	23 (11.4)
Age (years)			
Mean ± SD	64.44 ± 10.11	61.49 ± 9.28	61.53 ± 12.35
Range	33.0-87.0	44.0-84.0	15.0-82.0

Characteristic	Training GSE68465 LUAD (n=440)	LUAD Validation GSE30219 (n=85)	Mixed-Histology Validation GSE30219 (n=202)
Histological subtype, n (%)			
Adenocarcinoma	440 (100)	85 (100)	—
Squamous cell carcinoma	—	—	61 (30.2)
Large cell neuroendocrine carcinoma	—	—	55 (27.2)
Basaloid carcinoma	—	—	39 (19.3)
Carcinoid	—	—	23 (11.4)
Small cell carcinoma	—	—	17 (8.4)
Large cell carcinoma	—	—	3 (1.5)
Other	—	—	4 (2.0)
pT stage, n (%)			
T1	149 (33.9)	71 (83.5)	95 (47.0)
T2	251 (57.0)	12 (14.1)	57 (28.2)
T3	28 (6.4)	2 (2.4)	29 (14.4)
T4	12 (2.7)	—	21 (10.4)
pN stage, n (%)			
N0	299 (68.0)	82 (96.5)	116 (57.4)
N1	88 (20.0)	3 (3.5)	50 (24.8)
N2	53 (12.0)	—	27 (13.4)
N3	—	—	9 (4.5)
Vital status, n (%)			
Deceased	235 (53.4)	45 (52.9)	149 (73.8)
Surviving	205 (46.6)	40 (47.1)	53 (26.2)
Microarray platform	Affymetrix U133A	Affymetrix U133 Plus 2.0	Affymetrix U133 Plus 2.0

SD, standard deviation. —, not applicable or not observed in this cohort. The substantially more favourable stage distribution of the LUAD validation subset relative to the training cohort (pT1: 83.5% vs 33.9%; pN0: 96.5% vs 68.0%) is acknowledged as a factor affecting cross-cohort comparability of SHAP feature rankings.

Table 2. Three-Model Ablation Study: Performance on Held-Out Training Test Set (GSE68465, n=88)

Metric	Clinical Only	Gene Expr. Only	Full Multimodal Model	Change: Clinical to Full
Accuracy	0.60	0.60	0.70	+0.10
Precision (PPV)	0.62	0.62	0.73	+0.11
Sensitivity (Recall)	0.68	0.66	0.70	+0.02
Specificity	0.51	0.54	0.71	+0.20
F1 Score	0.65	0.64	0.72	+0.07
AUC	0.65	0.67	0.73	+0.08
AUPRC (baseline: 0.53)	0.72	0.68	0.75	+0.03
CV Mean AUC (SD)	0.675 (0.041)	0.682 (0.046)	0.707 (0.051)	+0.032
CV fold AUC scores	0.744/0.668/0.657 /0.618/0.688	0.629/0.635/0.704 /0.690/0.752	0.632/0.664/0.722 /0.767/0.751	—

AUC, area under the ROC curve; AUPRC, area under the precision-recall curve; CV, stratified 5-fold cross-validation on training partition only (n=352); PPV, positive predictive value; SD, standard deviation. All threshold-dependent metrics at probability threshold=0.40. The specificity improvement (+0.20) is the clinically most consequential metric: the clinical-only model misclassified 49% of survivors as high-risk vs 29% for the full model, corresponding to approximately 93 fewer incorrect high-risk designations per 1,000 LUAD patients.

Table 3. External Validation Performance of the Full Multimodal XGBoost Classifier

Metric	Training Test Set GSE68465 (n=88)	LUAD Validation GSE30219 (n=85)	Mixed-Histology Validation GSE30219 (n=202)
Accuracy	0.70	0.66	0.66
Precision (PPV)	0.73	0.69	0.83
Sensitivity (Recall)	0.70	0.64	0.67
Specificity	0.71	0.68	0.62
F1 Score	0.72	0.67	0.74
AUC (95% bootstrap CI)	0.73 (0.61-0.83)	0.71 (0.58-0.81)	0.69 (0.60-0.78)
AUPRC	0.75	0.69	0.82
Prevalence (AUPRC baseline)	0.53	0.53	0.74
High-risk group, n	—	42	115
Low-risk group, n	—	43	73
High-risk cumulative event rate	—	42.9% (18/42)	52.2% (60/115)
Low-risk cumulative event rate	—	20.9% (9/43)	32.9% (24/73)
Log-rank p-value	—	0.044†	0.011

AUC, area under the ROC curve; AUPRC, area under the precision-recall curve; CI, 95% confidence interval (1,000 stratified bootstrap iterations, seed=42); PPV, positive predictive value. All threshold-dependent metrics at probability threshold=0.40. Survival endpoint: disease-free survival (time to first relapse). The AUC declines by only 0.04 across three independent evaluations (0.73→0.71→0.69). †LUAD subset result should be interpreted as exploratory given n=85.

Figures:

Figure 1:

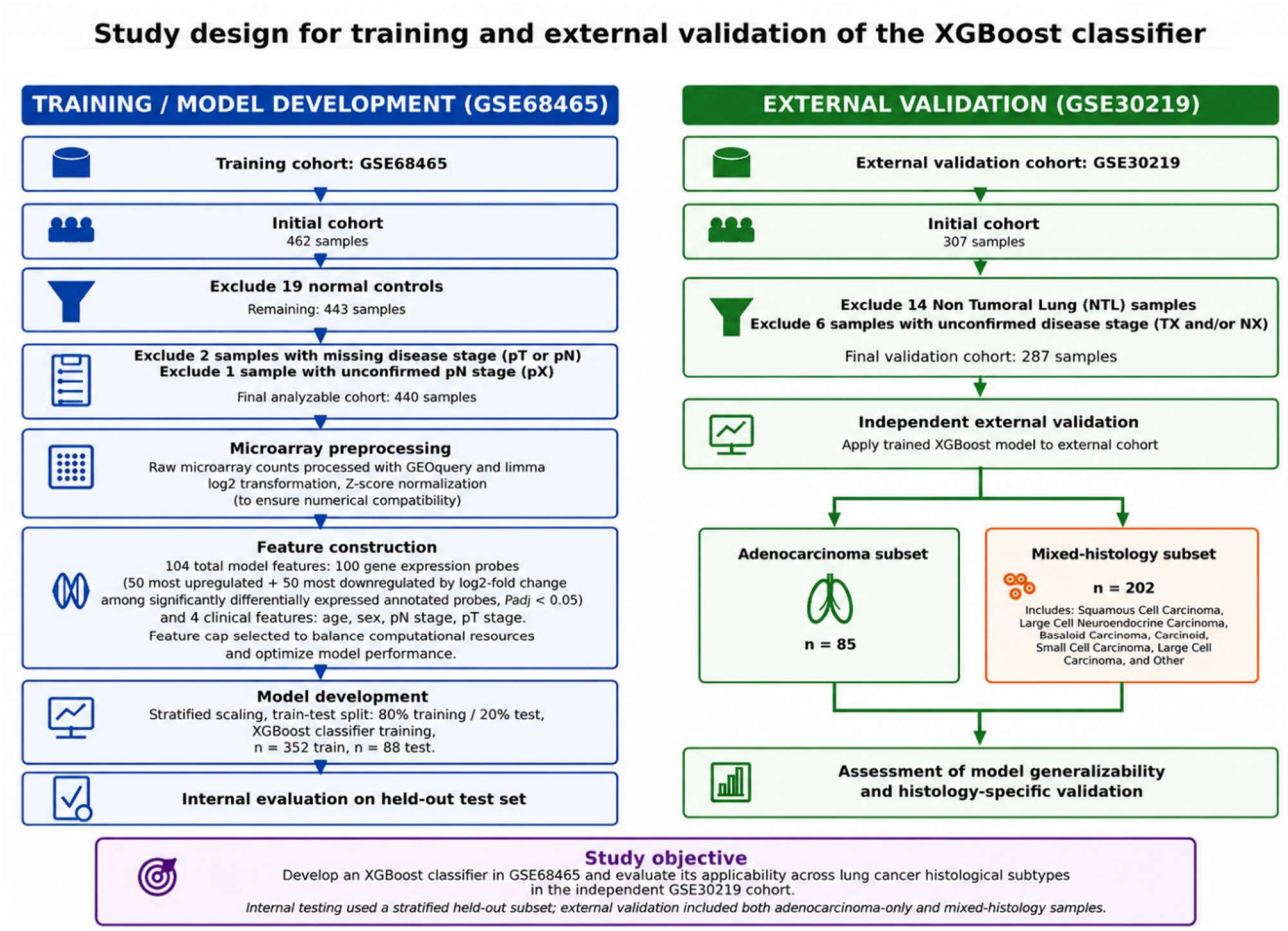


Figure 1. Study design schematic. Overview of the training and external validation pipeline. Left panel: training cohort (GSE68465) inclusion and exclusion criteria, preprocessing steps, feature construction, and model development workflow. Right panel: external validation cohort (GSE30219) exclusion criteria and stratification into an adenocarcinoma subset (n=85) and a mixed-histology subset (n=202) with histotype composition. The study objective is stated at the bottom.

Figure 2:

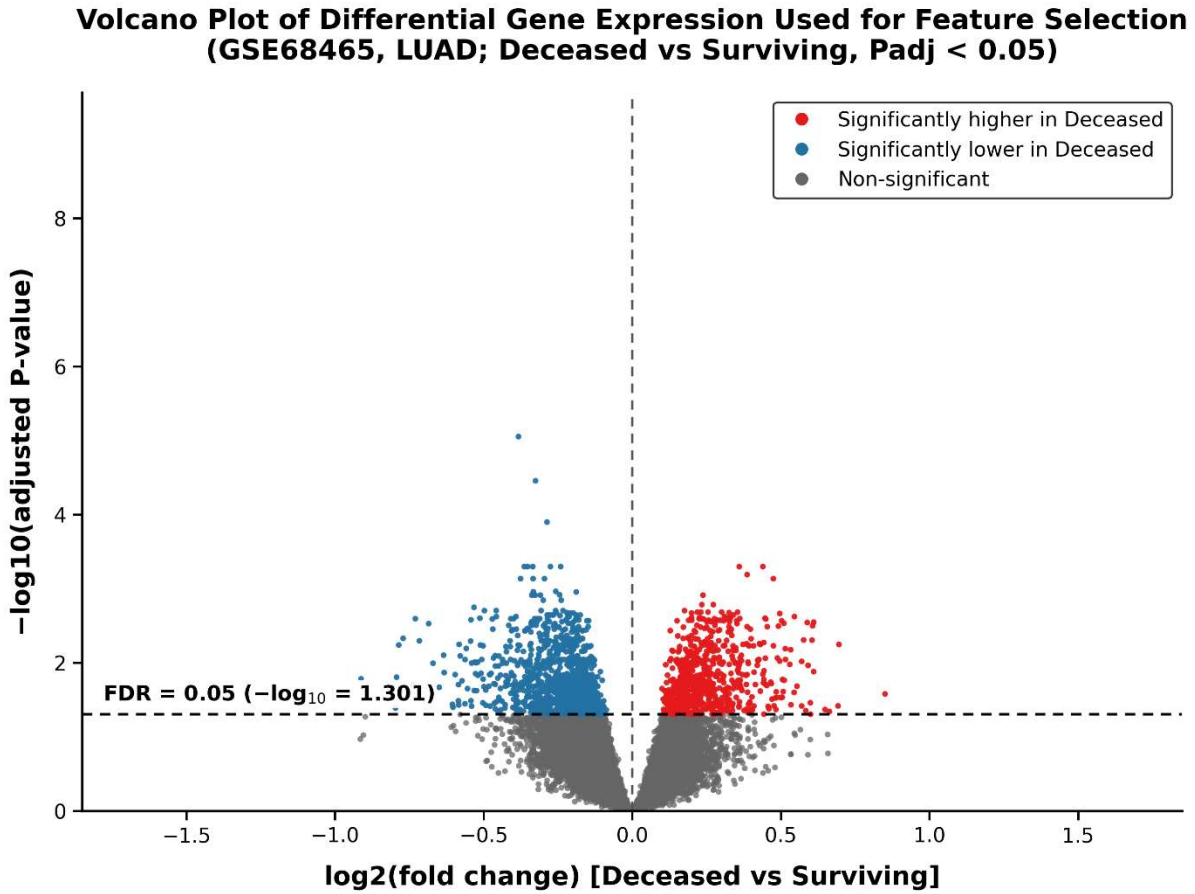


Figure 2. Differential expression and feature selection (GSE68465, n=440). Volcano plot of differential expression between deceased and surviving patients. Red: significantly upregulated in deceased patients ($log_2FC > 0$, Padj < 0.05, n=752 probes); blue: significantly downregulated ($log_2FC < 0$, Padj < 0.05, n=1,128 probes); grey: non-significant. The horizontal dashed line marks the FDR=0.05 threshold ($-log_{10}=1.301$). Affymetrix internal control probes and unannotated probes were excluded prior to feature selection and are not shown. The 50 probes with the largest positive and 50 with the largest negative log_2 fold change were selected as gene expression features.

Figure 3:

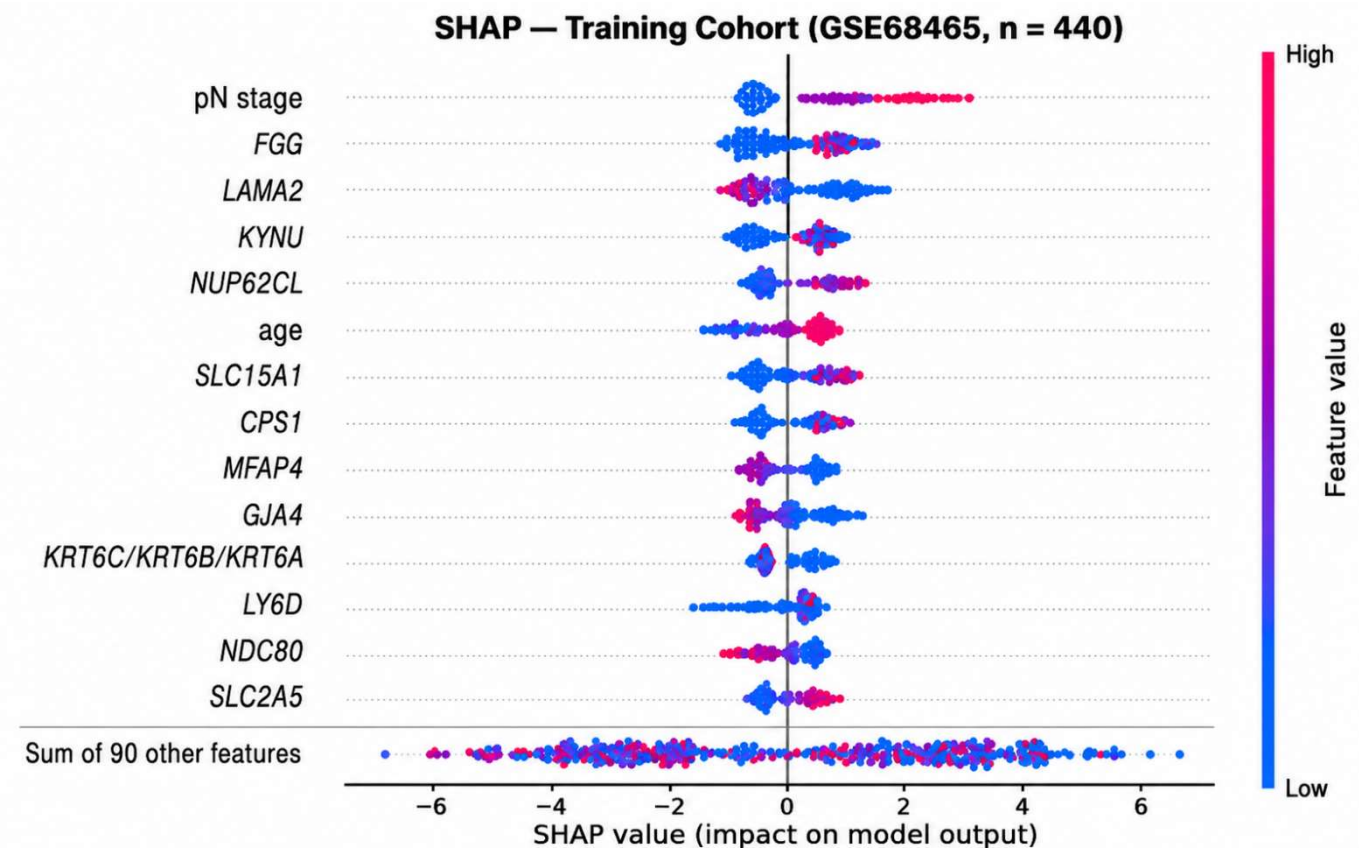
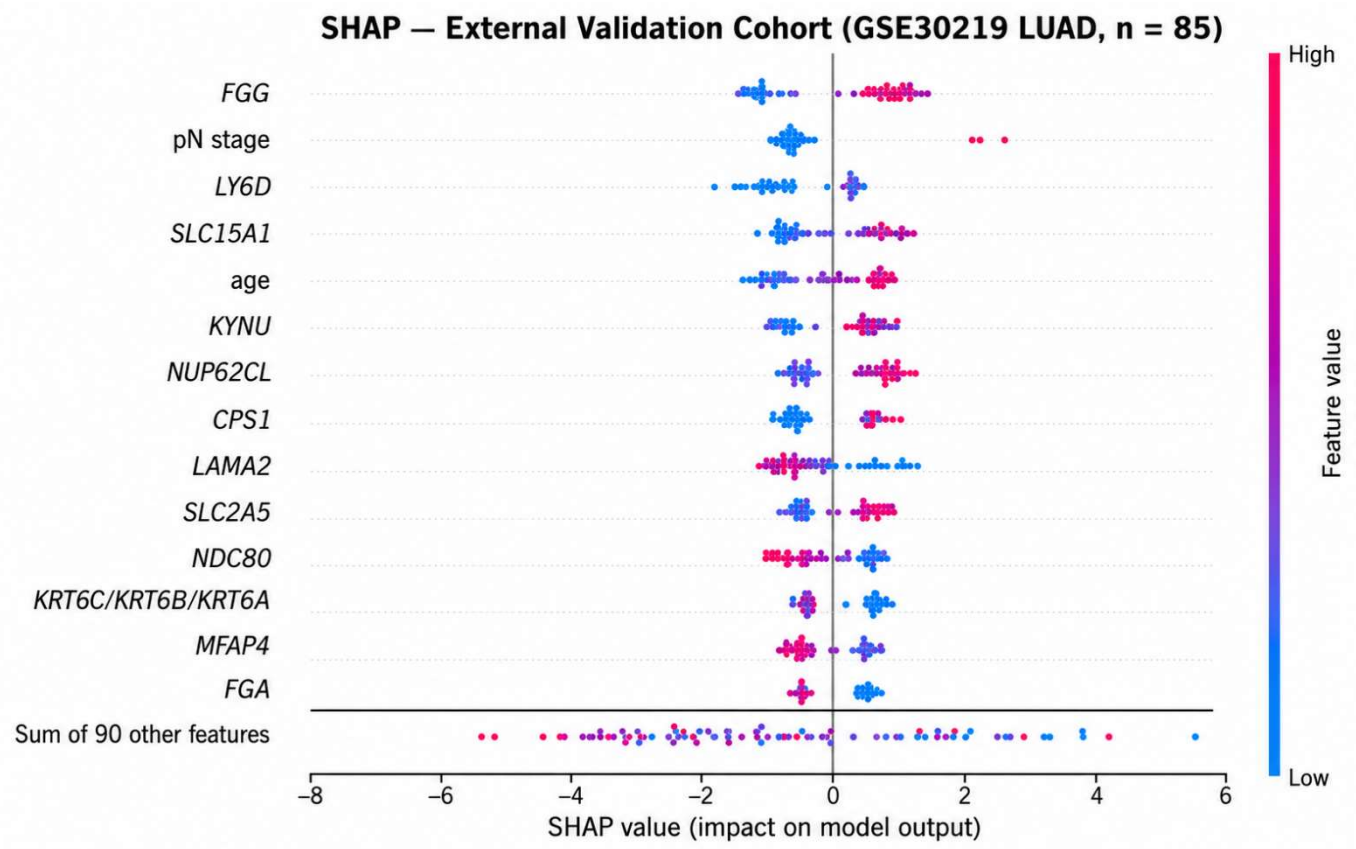


Figure 3. SHAP beeswarm plot, training cohort (GSE68465, n=440). Top 14 features ranked by mean absolute SHAP value. Each point represents one patient; position on the x-axis indicates direction and magnitude of contribution to predicted mortality; colour indicates feature value (red: high; blue: low). Gene symbols but not clinical variables are in italic. pN stage ranks first overall, validating the model's face-validity. FGG ranks second, consistent with its emergence as the dominant feature in both external validation cohorts.

Figure 4:

A)


SHAP — External Validation Cohort (GSE30219 LUAD, n = 85)
FGG
pN stage
LY6D
SLC15A1
age
KYNU
NUP62CL
CPS1
LAMA2
SLC2A5
NDC80
KRT6C/KRT6B/KRT6A
MFAP4
FGA
Sum of 90 other features
−8
−6
−4
−2
0
2
4
6
SHAP value (impact on model output)
High
Feature value
Low

B)

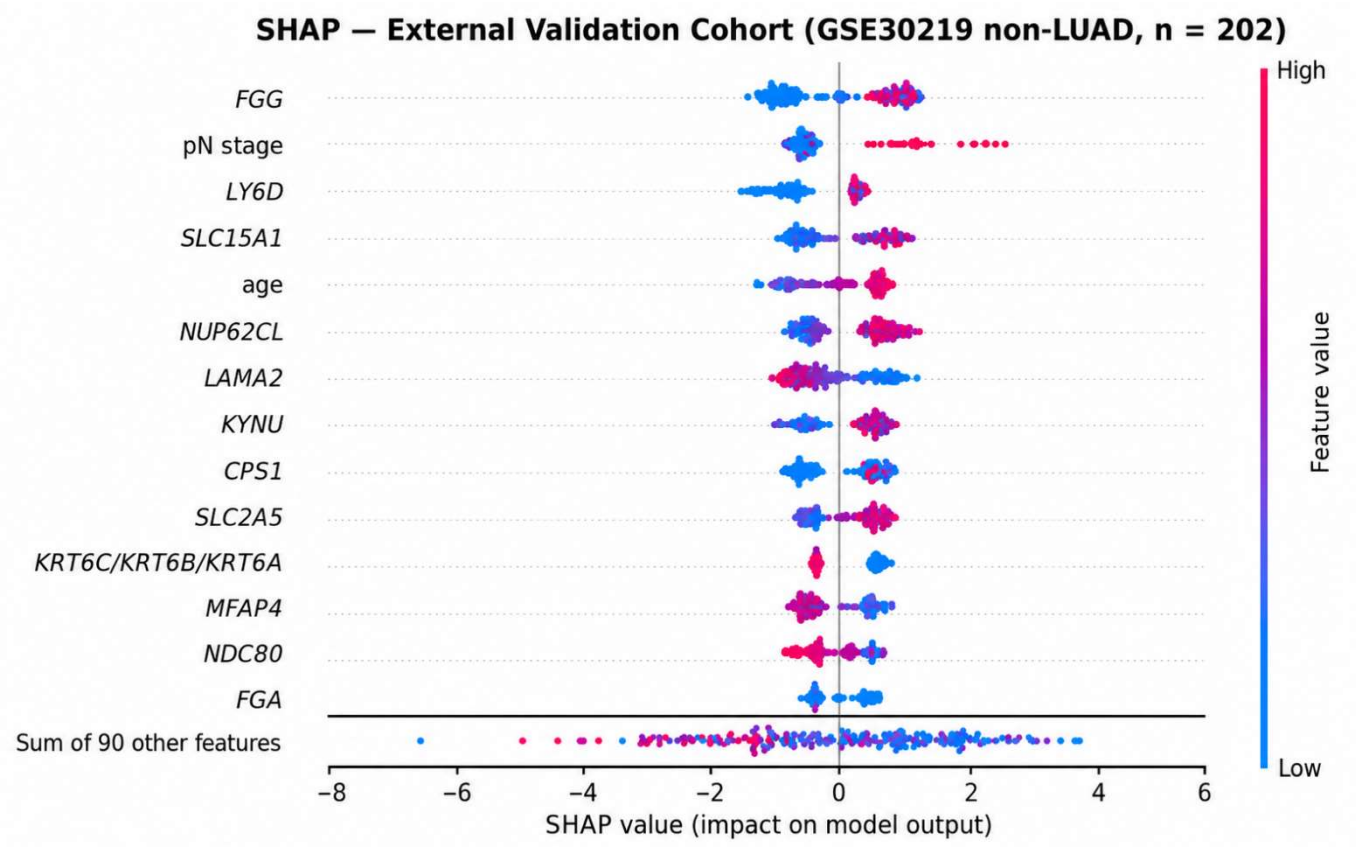
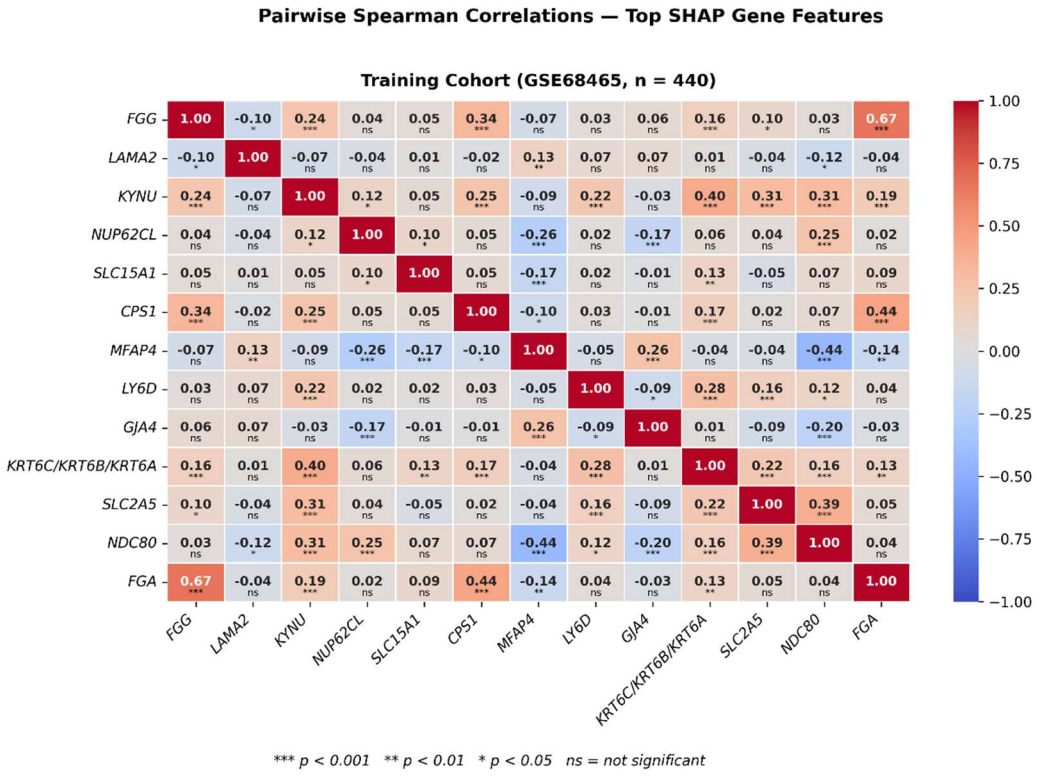


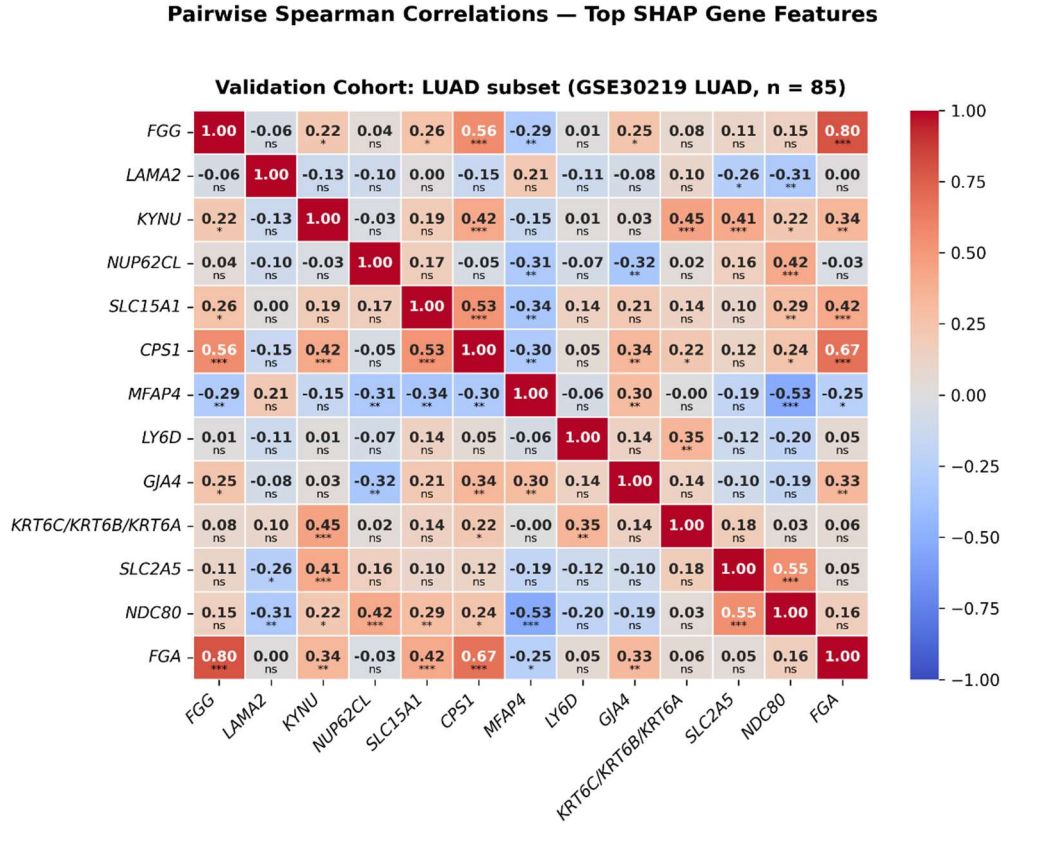
Figure 4. SHAP beeswarm plots, external validation cohorts. Panel A: LUAD validation subset (GSE30219, n=85). Panel B: Mixed-histology validation subset (GSE30219, n=202). Both panels share identical x-axis scales (-8 to +6) to facilitate cross-cohort comparison. FGG ranks first in both validation cohorts, outranking pN stage; this pattern is consistent with the near-uniform pN0 staging in the LUAD validation subset (96.5%) reducing the discriminative contribution of nodal stage relative to the training cohort. Twelve of 14 top features are shared across training, LUAD validation, and mixed-histology validation cohorts.

Figure 5:

A)


Pairwise Spearman Correlations — Top SHAP Gene Features
Training Cohort (GSE68465, n = 440)
FGG
LAMA2
KYNU
NUP62CL
SLC15A1
CPS1
MFAP4
LY6D
GJA4
KRT6C/KRT6B/KRT6A
SLC2A5
NDC80
FGA
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
*** p < 0.001 ** p < 0.01 * p < 0.05 ns = not significant

B)


Pairwise Spearman Correlations — Top SHAP Gene Features
Validation Cohort: LUAD subset (GSE30219 LUAD, n = 85)
FGG
LAMA2
KYNU
NUP62CL
SLC15A1
CPS1
MFAP4
LY6D
GJA4
KRT6C/KRT6B/KRT6A
SLC2A5
NDC80
FGA
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
*** p < 0.001 ** p < 0.01 * p < 0.05 ns = not significant

C)

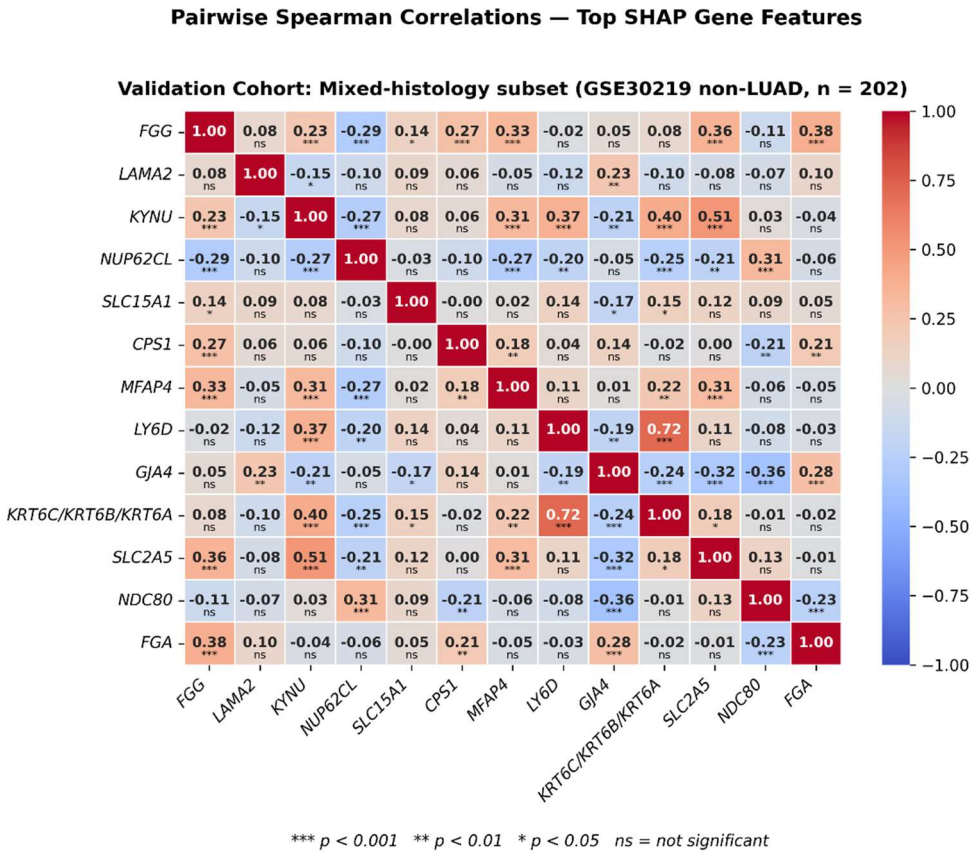
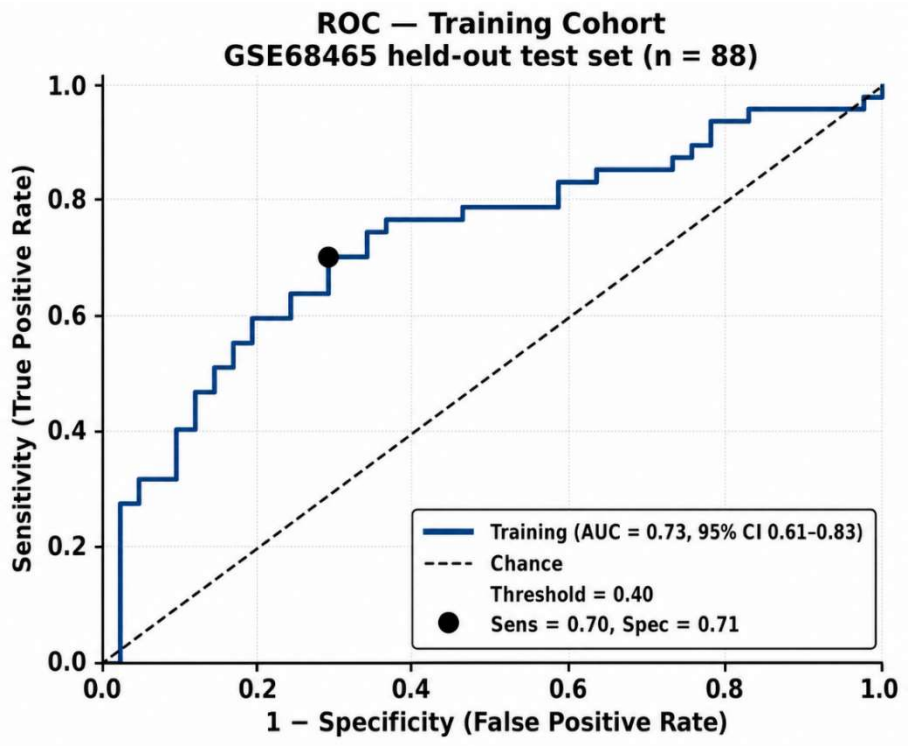


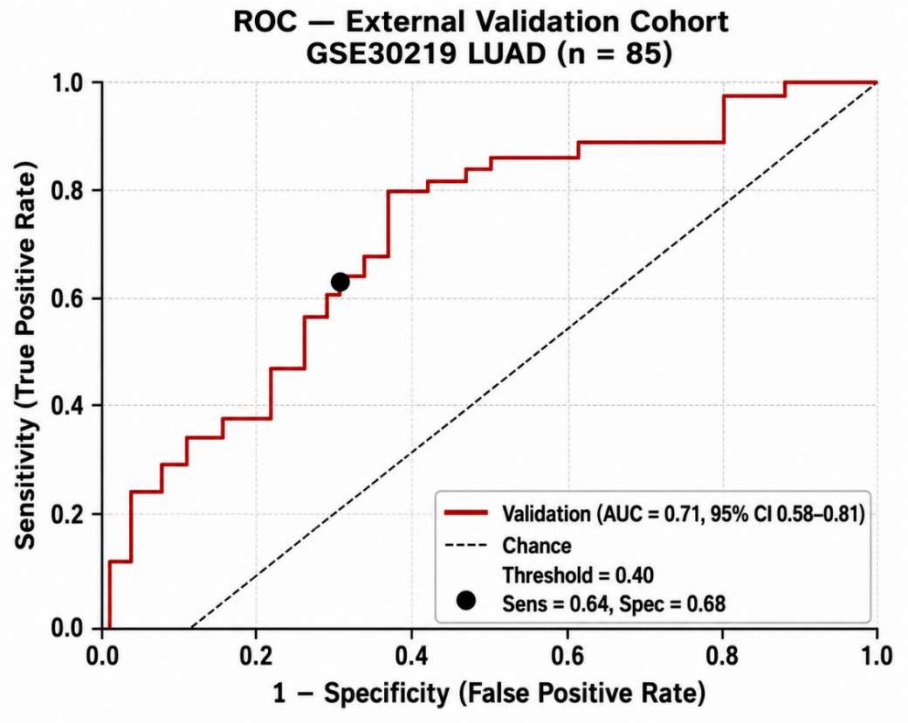
Figure 5. Pairwise Spearman correlation matrices, top SHAP gene features. Panel A: Training cohort (GSE68465, n=440). Panel B: LUAD validation subset (GSE30219, n=85). Panel C: Mixed-histology validation subset (GSE30219, n=202). Colour indicates Spearman correlation coefficient (red: positive; blue: negative). Significance annotations: *** p<0.001, ** p<0.01, * p<0.05, ns=not significant. Gene names are in italic. Thirteen gene features are included, representing the union of top SHAP-contributing genes across all three cohorts. The colour scale (-1 to +1) is consistent across all panels.

Figure 6:

A)


ROC — Training Cohort
GSE68465 held-out test set (n = 88)
Sensitivity (True Positive Rate)
1 − Specificity (False Positive Rate)
0.0
0.2
0.4
0.6
0.8
1.0
Training (AUC = 0.73, 95% CI 0.61–0.83)
Chance
Threshold = 0.40
Sens = 0.70, Spec = 0.71

B)


ROC — External Validation Cohort
GSE30219 LUAD (n = 85)
Sensitivity (True Positive Rate)
1 − Specificity (False Positive Rate)
0.0
0.2
0.4
0.6
0.8
1.0
Validation (AUC = 0.71, 95% CI 0.58–0.81)
Chance
Threshold = 0.40
Sens = 0.64, Spec = 0.68

C)

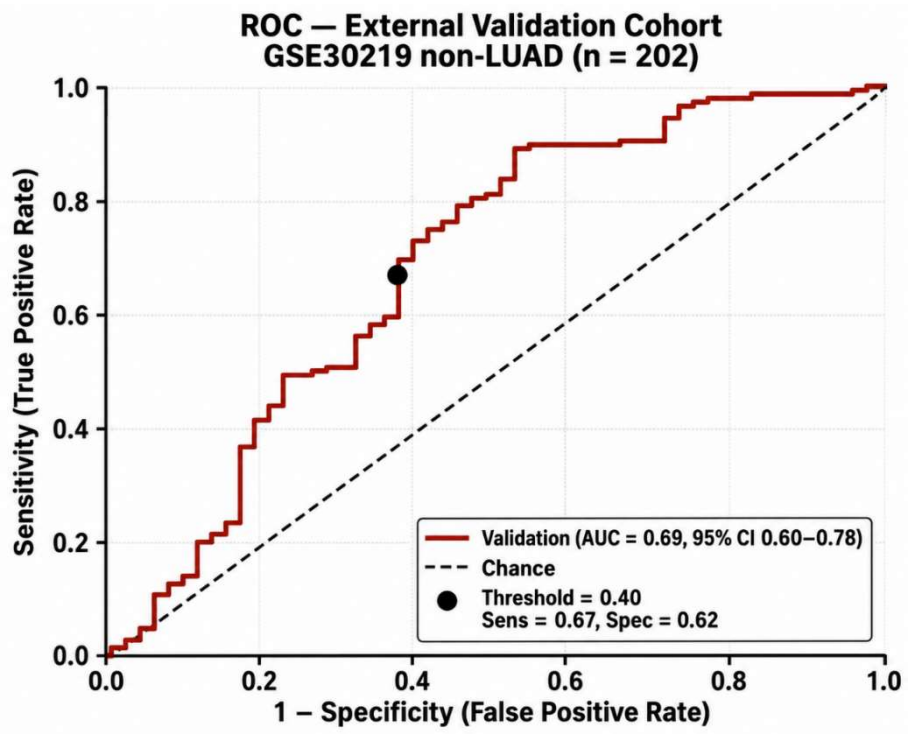
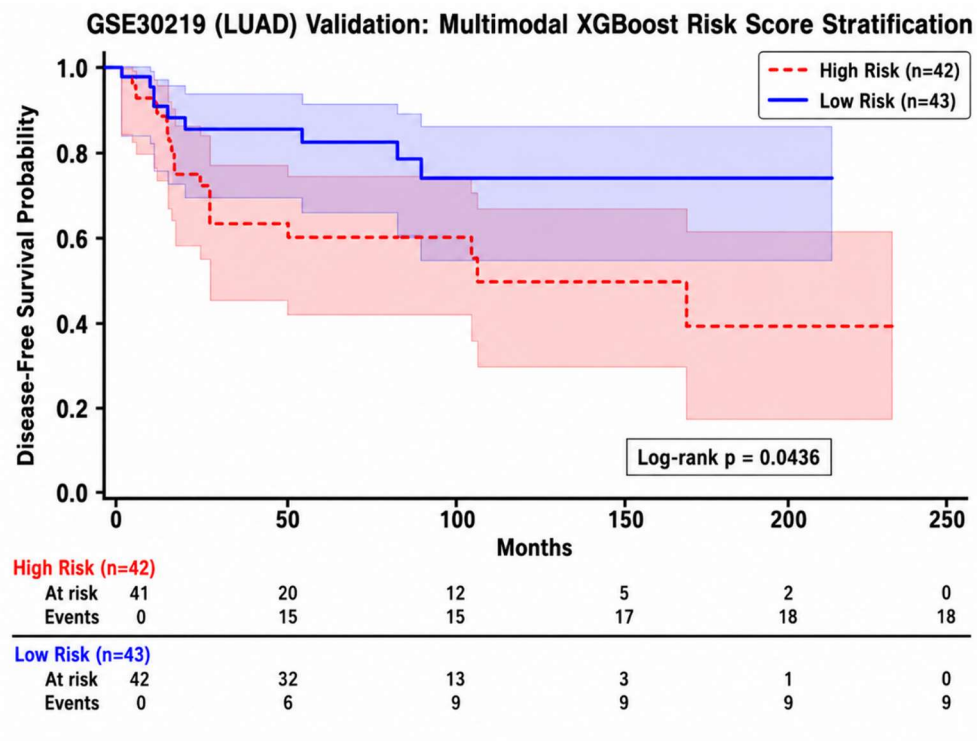


Figure 6. Receiver operating characteristic curves. Panel A: Training cohort held-out test set (GSE68465, n=88). Panel B: LUAD validation subset (GSE30219, n=85). Panel C: Mixed-histology validation subset (GSE30219, n=202). AUC and 95% bootstrap confidence intervals are shown. The operating point at probability threshold=0.40 is marked. The dashed diagonal line represents chance discrimination (AUC=0.50). The AUC decreases by only 0.04 across three independent evaluations (0.73 to 0.71 to 0.69).

Figure 7:

A)


GSE30219 (LUAD) Validation: Multimodal XGBoost Risk Score Stratification
High Risk (n=42)
Low Risk (n=43)
Disease-Free Survival Probability
1.0
0.8
0.6
0.4
0.2
0.0
Log-rank p = 0.0436
0
50
100
150
200
250
Months
High Risk (n=42)
At risk 41 20 12 5 2 0
Events 0 15 15 17 18 18
Low Risk (n=43)
At risk 42 32 13 3 1 0
Events 0 6 9 9 9 9

B)

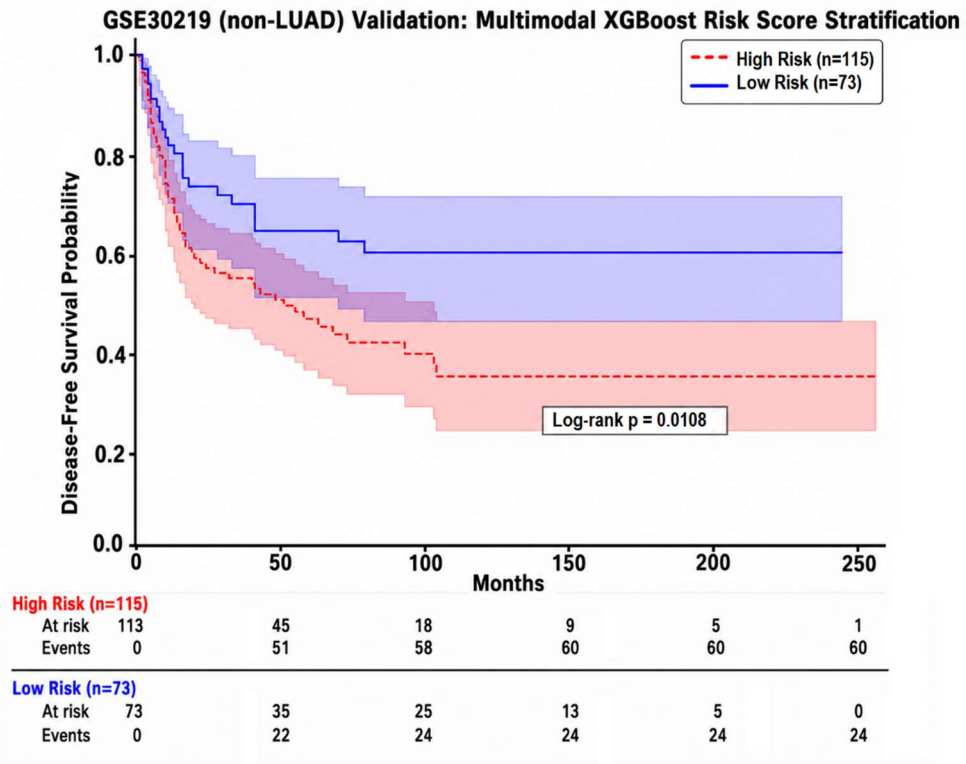
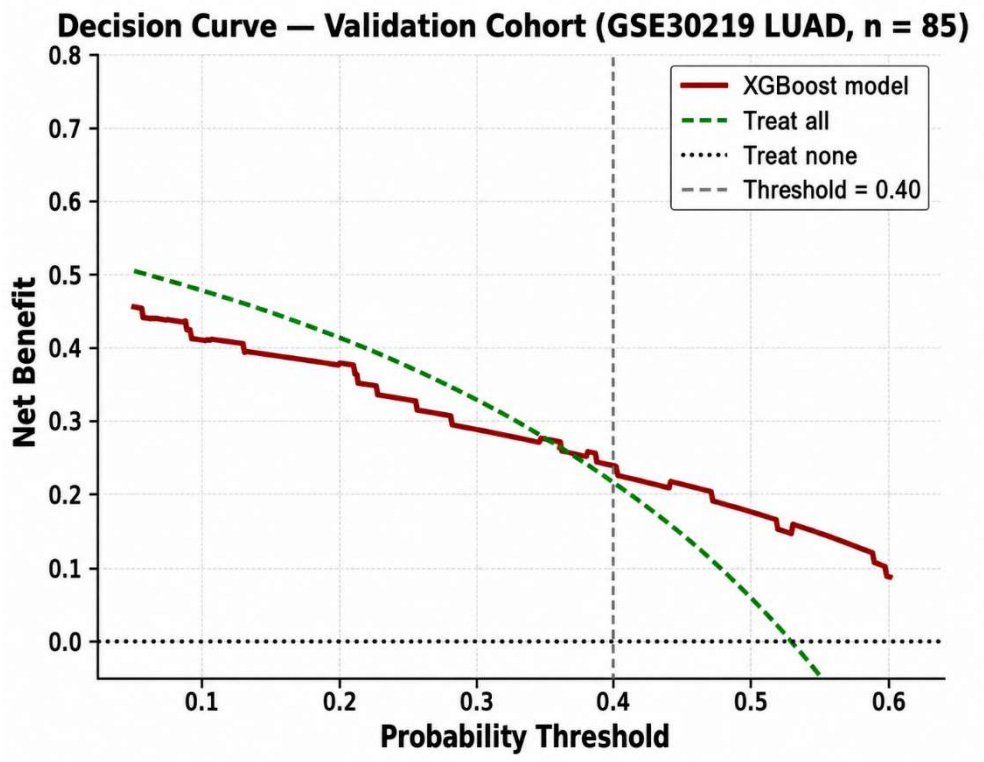

GSE30219 (non-LUAD) Validation: Multimodal XGBoost Risk Score Stratification
High Risk (n=115)
Low Risk (n=73)
Disease-Free Survival Probability
1.0
0.8
0.6
0.4
0.2
0.0
Log-rank p = 0.0108
0
50
100
150
200
250
Months
High Risk (n=115)
At risk 113 45 18 9 5 1
Events 0 51 58 60 60 60
Low Risk (n=73)
At risk 73 35 25 13 5 0
Events 0 22 24 24 24 24

Figure 7. Kaplan-Meier survival stratification, external validation cohorts. Panel A: LUAD validation subset (GSE30219, n=85). Panel B: Mixed-histology validation subset (GSE30219, n=202). Patients are stratified by XGBoost risk score into high-risk (predicted probability ≥0.40, dashed red line) and low-risk (< 0.40, solid blue line) groups. Shaded areas represent 95% confidence intervals. Log-rank p-values are annotated. At-risk and event tables are shown at 50-month intervals. Survival endpoint: disease-free survival. †The LUAD result (p=0.044) should be interpreted as exploratory given n=85.

Figure 8:

A)



B)

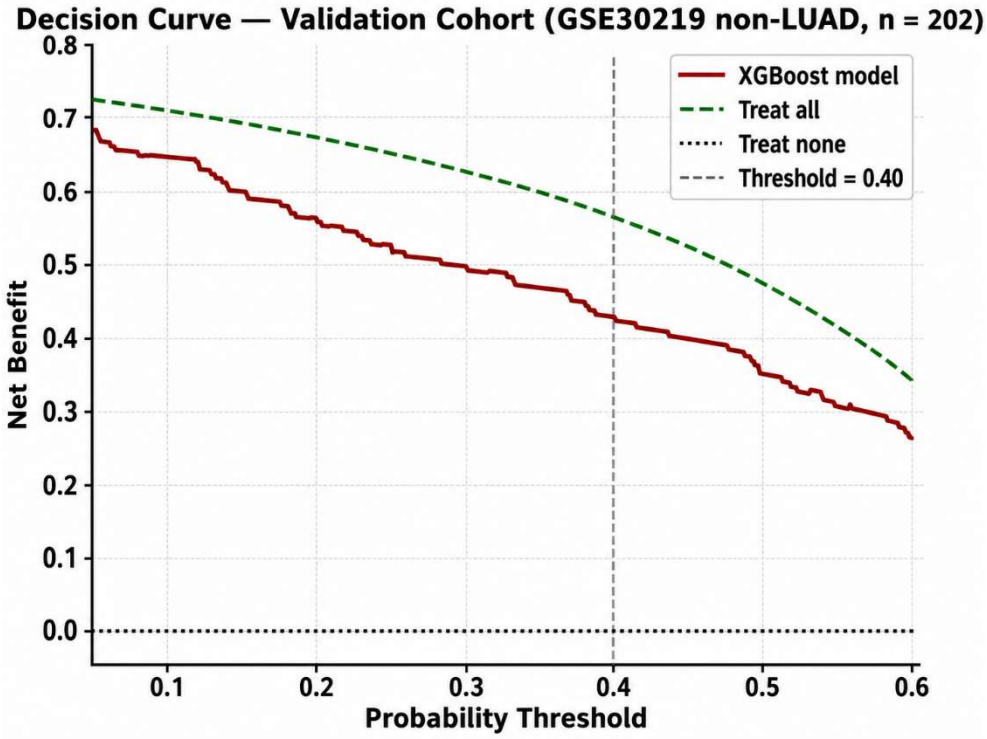


Figure 8. Decision curve analysis, external validation cohorts. Panel A: LUAD validation subset (GSE30219, n=85). Panel B: Mixed-histology validation subset (GSE30219, n=202). Net benefit is plotted across probability thresholds of 0.05 to 0.60 for the XGBoost model (solid red), treat-all (dashed green), and treat-none (dotted black). The vertical dashed grey line marks the 0.40 operating threshold. Both panels share a common y-axis scale (0 to 0.80). In Panel B, the model runs below the treat-all line; at a mortality prevalence of 73.8%, the treat-all strategy carries an inherently high net benefit that reflects a property of decision curve analysis in high-prevalence cohorts rather than a limitation of model discrimination.