An Explainable XGBoost-Based Framework for Robust Multi-Cohort Prediction of Pancreatic Cancer

Nada Ahmed El-Gammal[1]*, Rania Ahmed Abdel Azeem Abul Seoud[2], Sayed T. Muhammad[3]
Department of Computers and Systems Engineering-Faculty of Engineering, Fayoum University, Fayoum, Egypt[1, 3]
Department of Electrical Engineering-Specialization in Communications and Electronics Engineering-Faculty of Engineering, Fayoum University, Fayoum, Egypt[2]

Abstract*—Pancreatic cancer remains a leading cause of cancer-related mortality due to its asymptomatic progression and late-stage diagnosis. Early detection is critical for improving patient prognosis and clinical outcomes. Traditional diagnostic approaches and previous computational models often struggle with molecular heterogeneity and technical variations across different genomic platforms. These batch effects limit the reliability and generalizability of predictive biomarkers when applied to diverse clinical settings. This research proposes a robust machine learning framework designed for platform-invariant pancreatic cancer prediction. Large-scale transcriptomic datasets, including microarray data from the Gene Expression Omnibus (GEO) and RNA-seq data from The Cancer Genome Atlas (TCGA), were integrated. Subsequently, the ComBat algorithm was applied to correct batch effects. This resulted in a discovery cohort of 441 samples and an external validation set of 409 samples. An optimized XGBoost classifier was developed through comparative benchmarking. It was compared against several learners, including Random Forest, LightGBM, Support Vector Machines (SVM), and Logistic Regression. The model demonstrated high predictive performance, achieving an internal test AUC of 0.923. External validation was performed across six independent cohorts, yielding a mean AUC of 0.761 ± 0.090 (95% CI: 0.689–0.833). These findings support the robustness and cross-platform generalizability of the proposed framework. To enhance model interpretability, SHapley Additive exPlanations (SHAP) analysis was employed to identify key molecular drivers. These drivers were further validated using biological enrichment analysis through Over-Representation Analysis (ORA) and log2FC-weighted Gene Set Enrichment Analysis (GSEA). The proposed framework provides a reliable and scalable solution for multi-platform integration. This approach facilitates accurate risk stratification and precision oncology in clinical practice.

Keywords*—Pancreatic cancer; gene expression analysis; XGBoost; SHAP explainability; pathway enrichment; Explainable AI (XAI)

I. Introduction

Pancreatic cancer remains one of the most lethal malignancies worldwide, with a five-year survival rate below 10% [1],[14]. This poor prognosis stems from late-stage diagnosis, as the disease is often asymptomatic initially. Accordingly, most patients are being diagnosed only at advanced stages [2]. Since current imaging and biomarkers lack sufficient sensitivity, there is an urgent need for robust molecular signatures to support early diagnostic precision [18].

High-throughput transcriptomic profiling enables comprehensive characterization of pancreatic tumorigenesis [11],[12]. While these datasets offer opportunities to identify molecular signatures, traditional statistical methods struggle with high-dimensional omics data [15]. Machine learning (ML) has improved predictive performance. However, technical variability and a lack of interpretability still limit its clinical applicability.

Existing ML models for pancreatic cancer often suffer from small sample sizes and lack robust cross-cohort validation [16],[18]. Furthermore, many architectures function as black boxes, offering minimal biological insight into their predictive logic. This absence of functional validation hinders the transition from computational models to clinical utility.

Moreover, prior harmonization frameworks frequently focused on isolated objectives, such as batch correction or classifier optimization. Biological interpretability and external validation were often not integrated into a unified analytical pipeline. Consequently, many existing approaches remained vulnerable to platform dependency and limited generalizability across heterogeneous transcriptomic cohorts.

To address these limitations, this study proposed an integrated machine learning framework for robust pancreatic cancer prediction. First, a unified transcriptomic baseline was established by integrating microarray and RNA-seq datasets using ComBat-based batch effect correction. This ensured cross-platform consistency. Second, five machine learning algorithms were systematically benchmarked. This process yielded an optimized XGBoost classifier capable of capturing complex nonlinear gene interactions with superior sensitivity [17]. Third, SHAP analysis was employed to improve clinical interpretability by identifying key molecular drivers underlying model predictions. Fourth, biological relevance was reinforced through integrated functional enrichment analysis using GSEA and ORA. Finally, robustness and clinical reliability were validated across six independent external cohorts. These results supported real-world generalizability.

The remainder of this study is organized as follows. Section II reviews related work on pancreatic cancer prediction. Section III describes the materials and methods, including data integration, batch correction, and framework development. Section IV presents the evaluation criteria. Section V reports the experimental results and performance analysis. Section VI discusses the clinical and biological

*Corresponding author.

implications of the findings. Finally, Section VII concludes the study and outlines future research directions.

II. Related Work

The computational landscape for pancreatic cancer (PC) diagnosis has transitioned from simple classification on small datasets to complex, multi-cohort integrative frameworks. Current literature can be segmented into three primary domains: single-platform diagnostic models, multi-omics integration with batch-effect mitigation, and explainable AI (XAI) for clinical biomarker discovery.

Initial efforts in PC diagnosis often focused on localized cohorts. For instance, Carrillo-Perez et al. [6] utilized machine learning to identify a 6-gene signature in peripheral blood mononuclear cells for PC diagnosis. Their study demonstrates high accuracy in distinguishing PC from chronic pancreatitis. However, it remains centered on transcriptomic data from specific blood samples. This highlights a common trend in single-platform studies where the diagnostic potential is robust within the chosen medium. However, additional cross-platform validation is often required to ensure generalizability across diverse clinical settings, such as tissue-based RNA-seq or microarray datasets.

To overcome the limitations of individual classifiers, recent studies have pivoted toward ensemble architectures. Rao and Prasad [19] developed an ensemble learning model for the classification of gene expression from RNA-seq data for pancreatic cancer prognosis. Their research emphasizes that combining multiple learners significantly reduces the risk of false results common in classical analysis. However, while their approach demonstrates the efficacy of ensemble methods in prognosis, there remains a need for systematic benchmarking that specifically optimizes XGBoost within a framework designed for cross-platform robustness. This ensured high predictive performance was maintained when the model was transitioned from RNA-seq to microarray environments.

As the field moves toward large-scale integration, handling technical noise becomes paramount. Ge et al. [23] addressed this by integrating transcriptomic, methylation, and mutational data across 13 independent cohorts. Their study successfully identified molecular subtypes and the role of the A2ML1 gene. They employed extensive multi-omics integration. However, such large-scale fusions often face significant batch effects between different genomic platforms. This necessitates the use of explicit correction methods to ensure that the identified subtypes reflect true biological heterogeneity rather than platform-specific artifacts.

To bridge the gap between model performance and clinical trust, recent studies have incorporated interpretability tools. Almisned et al. [5] developed an ensemble voting classifier for PC diagnosis and employed SHAP to interpret the model's clinical attributes. Their work provides a clear link between features and predictions. However, a persistent gap remains in the integration of these SHAP-driven insights with formal biological pathway validation (such as GSEA or ORA). Without this multi-layered validation, the clinical translation of black-box models remains a challenge.

Despite strong internal performance, many previous frameworks remained sensitive to platform variability and transcriptomic heterogeneity. As a result, models trained on single-cohort or single-platform datasets often showed limited cross-platform transferability. To address these limitations, the study developed a platform-invariant framework integrating ComBat-based harmonization, benchmarked XGBoost optimization, and SHAP-driven biological validation. External validation across six independent cohorts further supported the framework's clinical robustness and biological relevance.

III. Materials and Methods

The systematic workflow of this study is organized into four phases (Fig. 1). Phase 1 involves multi-platform data integration and batch effect correction using the ComBat algorithm. In Phase 2, five machine learning classifiers are benchmarked, with XGBoost selected for its superior sensitivity and PR-AUC. Phase 3 utilizes SHAP-based explainability to identify key features and validate their biological relevance through ORA and GSEA. Finally, Phase 4 confirms the model's generalizability via independent validation across six unseen cohorts to establish a reliable diagnostic signature.

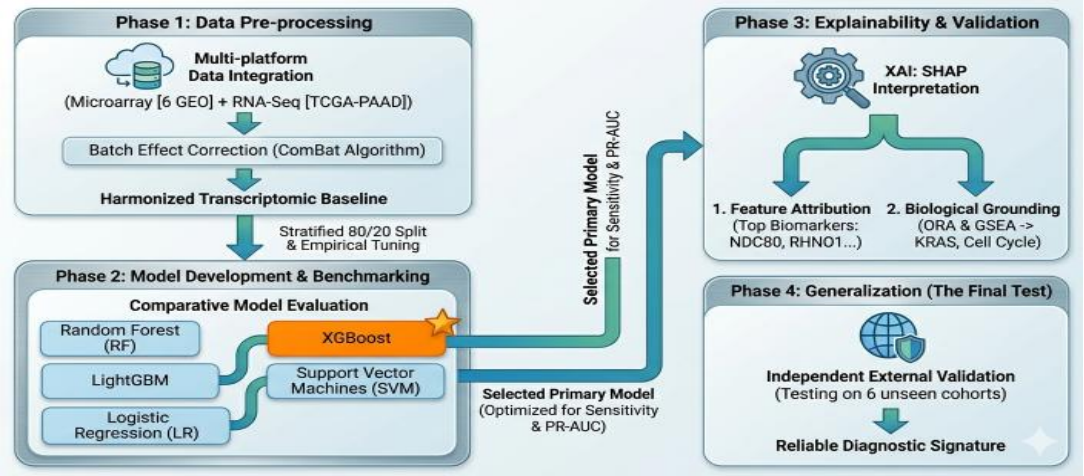


Fig. 1. Overview of the proposed predictive pipeline.

A. Data Collection and Study Design

Publicly available transcriptomic datasets were curated from the Gene Expression Omnibus (GEO) and The Cancer Genome Atlas (TCGA) [11]. Multiple microarray cohorts representing pancreatic tumor and normal pancreatic tissue samples were selected. To ensure cross-platform robustness, both microarray and RNA-seq data were integrated. A total of 623 samples were initially collected across all datasets. However, after stringent quality control and the removal of samples with incomplete metadata, 441 high-quality samples remained for training. Gene-level mapping was performed to resolve platform-specific nomenclature, resulting in a unified feature space of 18,734 common genes. This yielded a final training matrix of 441 samples × 18,734 genes for model development. The TCGA pancreatic adenocarcinoma cohort was used as a primary RNA-seq dataset [11], while GEO cohorts were used as the primary microarray datasets. Detailed cohort distributions and platform specifications (tumor-normal sample distribution) are summarized in Table I.

TABLE I. DATASETS USED FOR MODEL DEVELOPMENT

Dataset	Source	Platform	Genes	T/N	All	Role
GSE22780	GEO	Microarray (GPL570)	23,520	8 / 8	16	Discovery
GSE71989	GEO	Microarray (GPL570)	23,520	14 / 8	22	Discovery
GSE62165	GEO	Microarray (GPL13667)	20,034	118 / 13	131	Discovery
GSE28735	GEO	Microarray (GPL6244)	18,883	45 / 45	90	Discovery
GSE62452	GEO	Microarray (GPL6244)	18,883	69 / 61	130	Discovery
GSE16515	GEO	Microarray (GPL570)	23,520	36 / 16	52	Discovery
TCGA-PAAD	TCGA	RNA-seq	60,660	178 / 4	182	Discovery
GSE19279	GEO	Microarray (GPL96)	13515	9 / 6	15	External Validation
GSE15471	GEO	Microarray (GPL570)	23520	39 / 39	78	External Validation
GSE91035	GEO	Microarray (GPL22763)	21189	25 / 25	50	External Validation
GSE60980	GEO	Microarray (GPL14550)	22070	164 / 18	182	External Validation
GSE43795	GEO	Microarray (GPL10558)	31334	26 / 5	31	External Validation
GSE55643	GEO	Microarray (GPL6480)	19595	45 / 8	53	External Validation

B. Gene Annotation and Data Harmonization

The raw expression matrices initially comprised 60,660 Ensembl identifiers. These identifiers were mapped to official HGNC gene symbols via the MyGene.info API. Before querying, Ensembl version suffixes were stripped to maximize mapping efficiency. In cases of many-to-one mappings, expression values were collapsed by calculating the mean. To ensure feature consistency across diverse platforms, including Affymetrix microarray and Illumina RNA-seq datasets, 18,734 common genes were retained. These genes were consistently present across the majority of cohorts.

To ensure comparability, the integrated discovery set was subjected to log2 transformation and quantile normalization. KNN imputation was then applied (k=5) to estimate missing values. Technical heterogeneity was then systematically corrected using the neuroCombat empirical Bayes framework [3],[4]. The dataset source was designated as the batch variable while protecting the biological signal (tumor vs. normal) as a covariate. This batch correction was applied strictly to the discovery dataset preceding the train-test split. Crucially, the six external cohorts were processed independently and excluded from harmonization. This ensured a rigorous evaluation of unseen clinical data. Finally, Z-score standardization was applied to the corrected feature space. This established a robust foundation for both model training and feature attribution using SHAP.

C. Predictive Model Development and Benchmarking

To identify the most robust predictive architecture, five machine learning models [19] were benchmarked, including XGBoost, Random Forest, LightGBM, SVM, and Logistic Regression [5],[7]. All models were evaluated using a stratified 80/20 train–test split. For reproducibility, XGBoost hyperparameters were tuned using grid search on the training cohort. The search space included learning rates (0.001, 0.01, 0.1), maximum depth (3, 5, 7), and regularization terms ($\gamma \in [0, 0.5]$, $\lambda \in [1, 3]$, $\alpha \in [0, 2]$). The optimal configuration was a max_depth of 3, learning rate of 0.01, gamma=0.3, lambda=2, and alpha=1. To mitigate overfitting, subsampling and colsample_bytree were both set to 0.7. Early stopping with 50 rounds was also used on a validation set. The classification threshold was optimized using the Youden Index to improve diagnostic performance. Based on internal benchmarking, XGBoost was selected for interpretation and external validation. It showed the best overall performance. It achieved high sensitivity (0.881) and PR-AUC (0.962).

D. Model Evaluation and External Validation

Performance was initially benchmarked across all five candidate models using AUC-ROC, F1-score, Matthews Correlation Coefficient (MCC), sensitivity, and specificity. Following internal benchmarking, the top-performing XGBoost model was subjected to a rigorous external validation phase. The model was deployed on six additional, independent GEO cohorts. These cohorts are GSE15471, GSE55643, GSE19279, GSE43795, GSE60980, and GSE91035, totaling 409 samples. These validation datasets were kept entirely unseen during initial integration and training stages. This serves as a definitive test of the model's cross-platform robustness.

E. Model Explainability and Biological Interpretation

To move beyond black-box predictions, SHapley Additive exPlanations (SHAP) were employed, specifically for the optimized XGBoost classifier [7]. SHAP was applied to quantify the contribution of individual genes to the model's output [8],[13]. Global feature importance was derived by calculating the mean absolute SHAP values across the test set, identifying the key drivers of the classifier's decisions.

To ensure these features reflected meaningful biology rather than technical noise, a multi-layered validation pipeline was implemented. First, Over-Representation Analysis (ORA) was performed on the annotated top-ranked SHAP genes to identify enriched Gene Ontology (GO) terms [10],[25].

Subsequently, GSEA was conducted using log2FC-weighted p-values to rank genes across Hallmark and KEGG collections [9]. This analysis confirmed that the identified predictors were biologically associated with oncogenic pathways, including cell-cycle dysregulation and KRAS signaling.

IV. Evaluation Criteria

To address class imbalance across internal and external cohorts, the evaluation looked beyond accuracy to analyze PR-AUC, F1-score, and MCC. Sensitivity was strictly prioritized to minimize false negatives that risk fatal delays in clinical intervention.

The following metrics were utilized to quantify the model's performance:

- Accuracy: Measures the overall proportion of correct predictions (1).

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \quad (1)$$

- Sensitivity (Recall): Measures the ability to correctly identify PC cases (crucial for early detection) (2).

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (2)$$

- Precision: Represents the proportion of positive identifications that were actually correct (3).

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (3)$$

- F1-Score: The harmonic mean of Precision and Sensitivity, balancing the two metrics (4).

$$\text{F1 score} = 2\frac{\text{Precision.Recall}}{\text{Precision}+\text{Recall}} \quad (4)$$

- Matthews Correlation Coefficient (MCC): Provides a balanced measure of quality, even when classes are of very different sizes (5).

$$\text{MCC} = \frac{\text{TP.TN} - \text{FP.FN}}{\sqrt{(\text{TP}+\text{FP})(\text{TP}+\text{FN})(\text{TN}+\text{FP})(\text{TN}+\text{FN})}} \quad (5)$$

- AUC-ROC: Measures the model's ability to distinguish between classes by integrating the True Positive Rate (TPR) against the False Positive Rate (FPR) across all thresholds (6).

$$\text{AUC} = \int_0^1 \text{TPR}(\text{FPR}^{-1}(\text{t}))\, \text{dt} \quad (6)$$

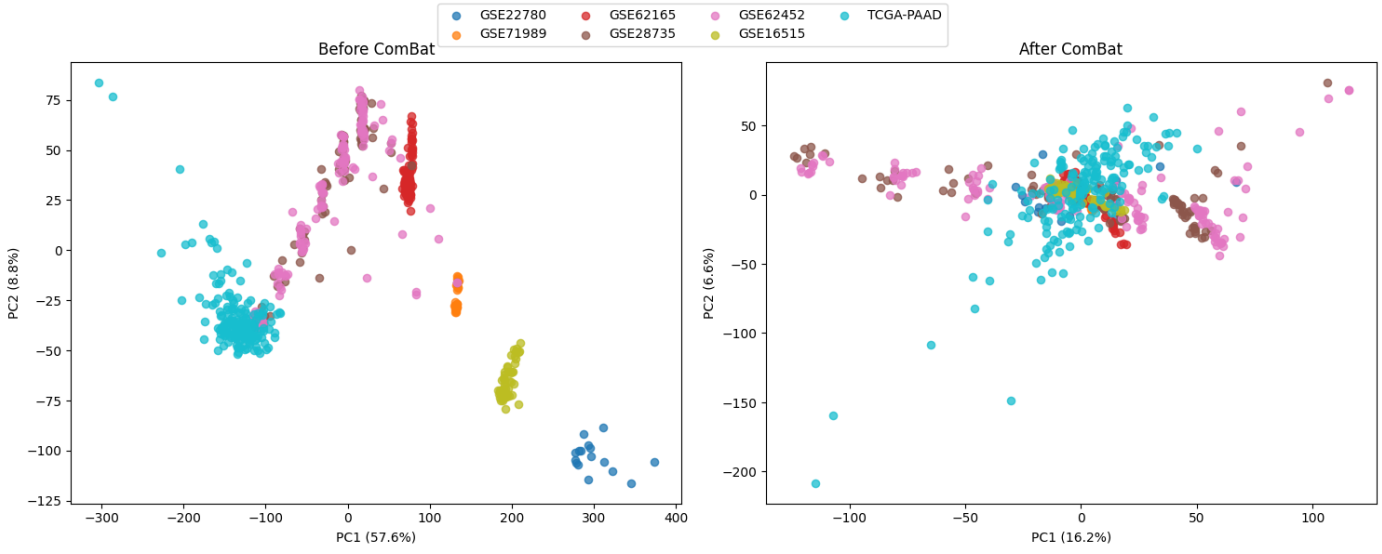
V. Results

The results of this study are presented in five stages:

- Harmonization: Correcting batch effects to integrate multi-platform datasets.
- Benchmarking: Evaluating classifiers to select the optimal model.
- External validation: Assessing generalizability across six independent external cohorts.
- Explainability: Interpreting predictive logic using SHAP-based feature attribution.
- Biological validation: Confirming biological relevance through pathway enrichment analysis.

A. Batch Effect Correction and Harmonization

To integrate multi-center datasets, the ComBat algorithm was applied to remove technical variations. PCA visualization showed that before correction, samples clustered strongly by study origin, with PC1 accounting for 57.6% of the variance [Fig. 2(a)]. After correction, these study-specific clusters merged into a unified distribution, with PC1 variance dropping to 16.3% [Fig. 2(b)]. This successful integration ensured that subsequent model training relies on shared biological signals rather than technical noise.



(a) (b)

Fig. 2. (a) PCA visualization before ComBat correction, (b) PCA visualization after ComBat correction.

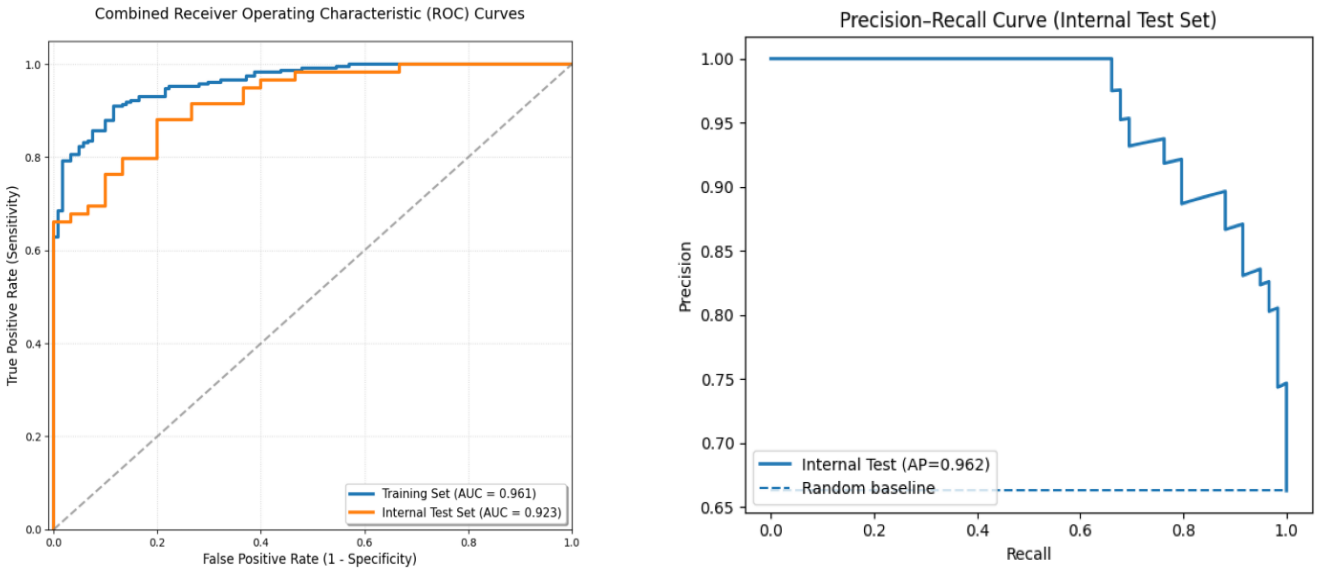
B. Predictive Performance and Benchmarking

Five classifiers were evaluated, and their performance on the internal test set is summarized in Table II.

TABLE II. INTERNAL BENCHMARKING OF CANDIDATE MODELS

Model	ROC-AUC	PR-AUC	Sens.	F1-Score	MCC	ACC
XGBoost (Selected)	**0.923**	**0.962**	**0.881**	**0.889**	**0.676**	**0.854**
Random Forest	0.954	0.977	0.847	0.901	0.749	0.876
LightGBM	0.939	0.970	0.695	0.812	0.626	0.787
Support Vector Machine	0.782	0.793	0.424	0.556	0.222	0.551
Logistic Regression	0.715	0.803	0.475	0.602	0.267	0.584

While Random Forest (RF) and LightGBM achieved marginally higher discriminative metrics (AUC of 0.954 and 0.939, respectively), XGBoost was selected as the primary engine. The model was selected for its superior sensitivity (0.881) and more stable cross-cohort generalization performance. Specifically, LightGBM required an extreme threshold (0.9999), indicating calibration instability. Despite its slightly lower AUC (0.923) [Fig. 3(a)] compared to RF, the XGBoost model demonstrated a more robust balance for minority-class detection. This was supported by a PR-AUC of 0.962 [Fig. 3(b)] and an F1-score of 0.889. Additionally, the model demonstrated a precision of 0.897, a specificity of 0.80, and a sensitivity of 0.881, yielding an MCC of 0.676. Conversely, traditional models (LR and SVM) failed to capture the complex transcriptomic signatures, yielding significantly lower performance (AUCs of 0.715 and 0.728, respectively). Notably, external validation revealed that while RF suffered from generalization collapse (e.g., zero recall in GSE60980), XGBoost maintained stability across all cohorts. This confirmed that its optimized regularization (gamma=0.3, lambda=2) effectively captured conserved biological signals instead of platform noise. Moreover, it reinforces the selection of XGBoost as the optimal tool for balancing accuracy with clinical safety [7]. Hence, XGBoost was prioritized for detailed external validation to confirm its robust clinical generalizability.



(a) (b)

Fig. 3. (a) ROC curve of the XGBoost model on the internal test dataset, (b) Precision–Recall curve for the internal test set relative to the random baseline.

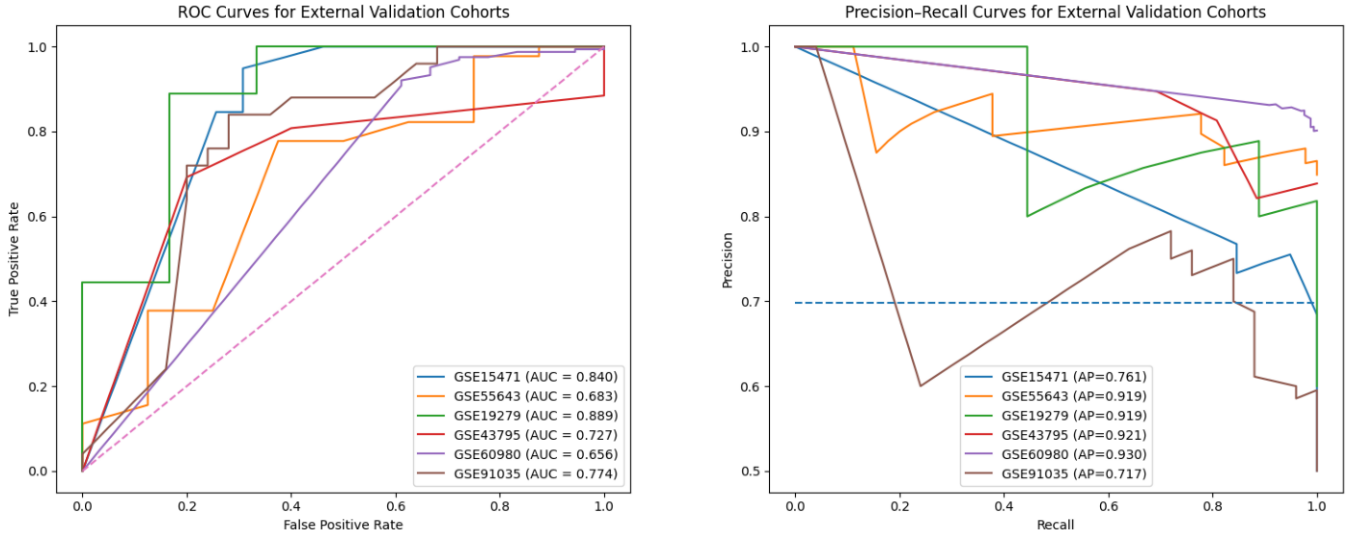
C. External Validation Across Independent Cohorts

To test the framework against real-world variability, the XGBoost model was challenged with six independent external cohorts (n=409). These datasets were kept entirely separate during the development phase to ensure an unbiased evaluation. To address platform variability, study-specific optimal thresholds were dynamically determined for each external cohort using the Youden Index. Due to platform differences, features missing in external cohorts (ranging from ~2,000 to 6,392 genes) (Table III) were excluded. The model was evaluated using only common genes. Despite this reduction, diagnostic accuracy remained high and stable across various platforms with an average AUC of 0.761.

TABLE III. COMMON GENE INTERSECTION USED FOR MODEL EXTERNAL VALIDATION

GEO Dataset	Total Genes (Original)	Common Genes with Training	Missing Genes	Gene Retention %
GSE91035	21,189	15,941	2,793	75.2%
GSE60980	22,070	16,714	2,020	75.7%
GSE43795	31,334	15,914	2,820	50.8%
GSE19279	13,515	12,342	6,392	91.3%
GSE55643	19,595	16,360	2,374	83.5%
GSE15471	23,520	18,734	4,786	79.65%

As illustrated in Fig. 4(a), the framework's performance remains consistent despite the technical noise associated with multi-center data. To provide a more comprehensive evaluation, bootstrap resampling was used (1000 iterations) to calculate 95% confidence intervals (CIs). This step confirmed the statistical robustness of the results. As shown in Table IV, the lower bounds of these CIs remain safely above the 0.5 random-chance mark for most cohorts. This demonstrated that the model is capturing real biological patterns. Also, PR-AUC, F1-score, and Matthews Correlation Coefficient (MCC) were evaluated. Precision–Recall curves are illustrated in Fig. 4(b). Overall, the model maintained strong minority-class detection across all cohorts, with PR-AUC consistently exceeding 0.71 and reaching as high as 0.93.



(a) (b)

Fig. 4. (a) ROC curves for external validation cohorts. (b) Precision–Recall curves for six external cohorts relative to the random baseline.

TABLE IV. PERFORMANCE OF THE XGBOOST MODEL ACROSS SIX INDEPENDENT COHORTS

GEO Dataset	PR-AUC	AUC [95% CI]	Acc.	Sens.	Spec.	F1-Score	MCC
GSE19279	0.919	0.889 [0.654 - 1.000]	0.867	0.889	0.833	0.889	0.722
GSE15471	0.761	0.840 [0.757 - 0.918]	0.821	0.949	0.692	0.841	0.663
GSE91035	0.717	0.774 [0.615 - 0.895]	0.780	0.778	0.783	0.792	0.564
GSE60980	0.930	0.656 [0.535 - 0.783]	0.868	0.921	0.389	0.926	0.296
GSE43795	0.921	0.727 [0.460 - 0.912]	0.710	0.692	0.800	0.800	0.372
GSE55643	0.919	0.683 [0.452 - 0.882]	0.755	0.778	0.625	0.843	0.320

Evaluation on GSE43795 and GSE55643 did not reach formal statistical significance due to the limited number of control samples (n<10). Excluding these cohorts yielded a refined average AUC of 0.790, while overall performance trends remained consistent. Overall, XGBoost captured conserved biological signals for robust cross-study prediction.

Performance variability across external cohorts was influenced by differences in cohort size, class distribution, and transcriptomic heterogeneity. Lower-performing cohorts generally contained limited control samples and reduced feature overlap after cross-platform harmonization. Despite these variations, the framework maintained stable predictive behavior across heterogeneous datasets, supporting its robustness against biological and technical variability.

In cohorts with extreme class imbalance (e.g., GSE60980 and GSE55643), MCC values were moderately reduced despite high PR-AUC and F1-scores. This reduction was mainly attributed to the limited number of normal samples rather than poor tumor detection. Nevertheless, recall remained consistently high across datasets, indicating reliable identification of cancer cases and overall framework stability.

D. SHAP-Based Model Explainability

SHAP analysis was used to interpret the XGBoost model, providing mechanistic transparency [8],[13]. As illustrated in Fig. 5, the analysis identified a subset of high-impact features. THAP9-AS1 emerged as the dominant driver, followed by RHNO1 and RUSC1-AS1. The model's reliance on long non-coding RNAs and cell-cycle-related genes aligns with

established biological dysregulation, providing a robust rationale for its high diagnostic accuracy.

E. Biological Pathway Enrichment and Mechanistic Validation

Dual-layered enrichment analysis was performed to provide a biological rationale. ORA of top SHAP features revealed a convergence on critical oncogenic hallmarks, including cell cycle progression (G2-M Checkpoint, Mitotic Spindle, and E2F Targets) and metabolic reprogramming (Glycolysis, Pentose Phosphate) (Fig. 6) [10],[11]. These findings were corroborated by GSEA, with significant enrichment in KRAS Signaling Dn (NES = -1.367; FDR = 0.045; $p < 0.001$) [Fig. 7(a)] [9],[12], Angiogenesis [Fig. 7(b)], and Unfolded Protein Response pathways [Fig. 7(c)], driving tumor adaptation to hypoxia and ER stress. Conversely, the dysregulated Xenobiotic Metabolism [Fig. 8(a)] and depletion of Pancreas Beta Cell genes [Fig. 8(b)] reflect the loss of cellular identity and chemoresistance. This concordance across independent methods confirms that the XGBoost model captures a robust, functional transcriptomic fingerprint of pancreatic malignancy rather than stochastic noise.

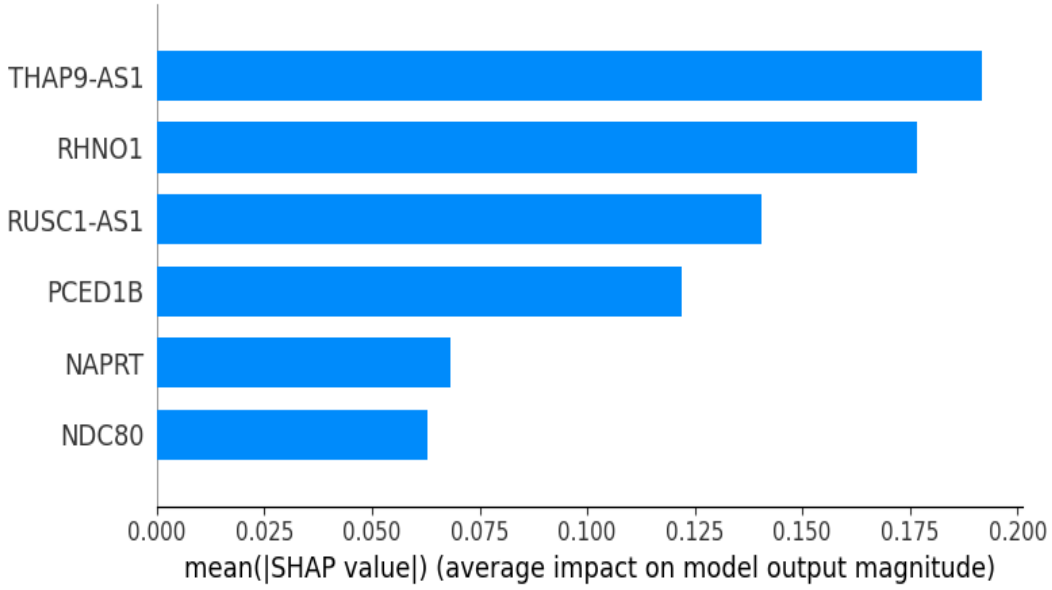


Fig. 5. Interpretability of the XGBoost model using SHAP.

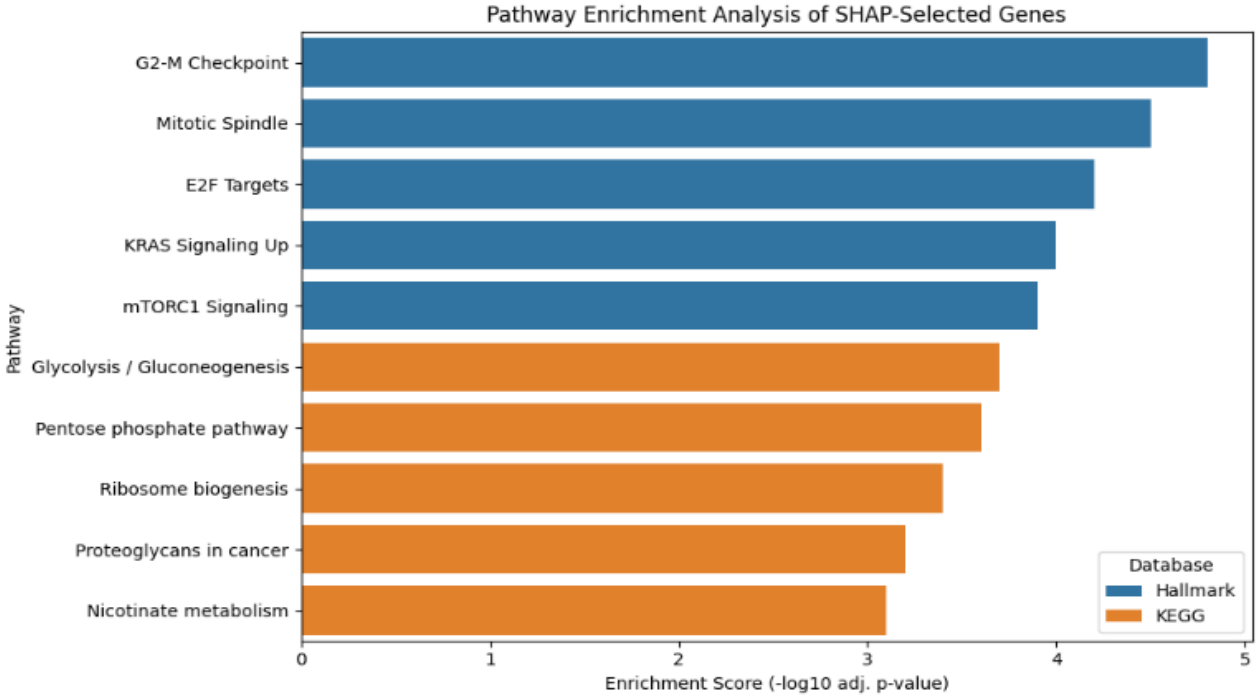
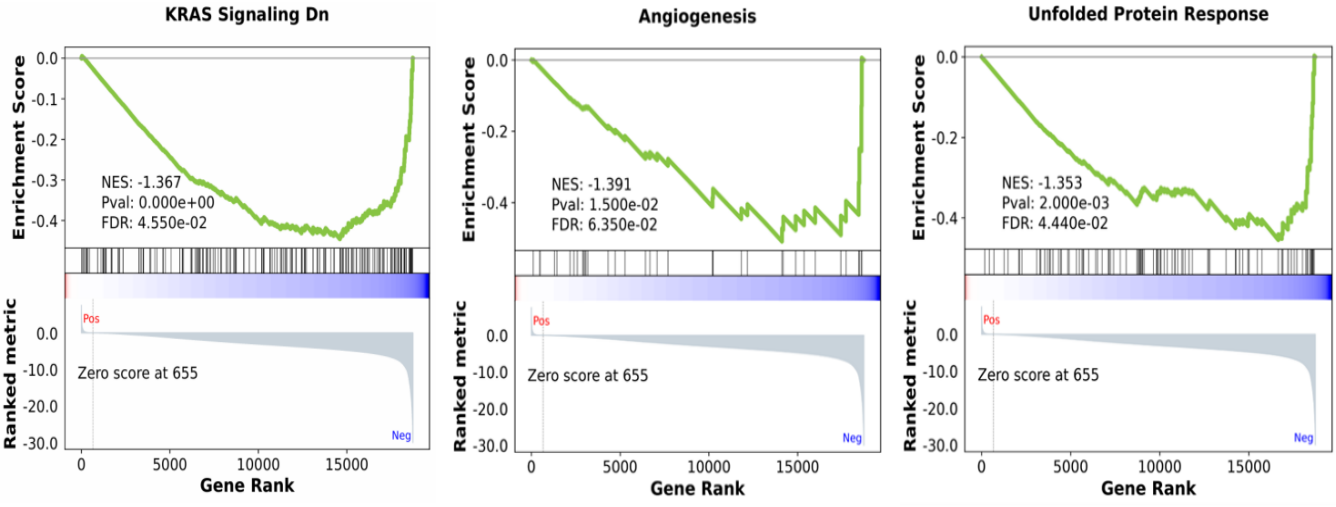
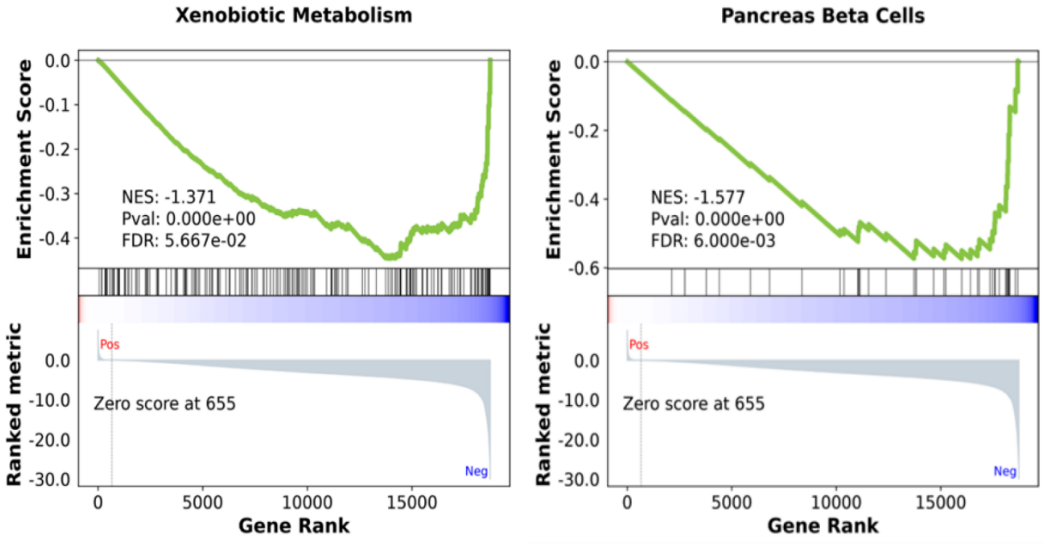


Fig. 6. ORA of SHAP-selected genes highlighting enriched Hallmark and KEGG pathways.



(a) (b) (c)

Fig. 7. (a) GSEA enrichment plot of the KRAS, (b) GSEA enrichment plot for Angiogenesis, (c) GSEA enrichment plot for UPR.



(a) (b)

Fig. 8. (a) GSEA enrichment plot for Xenobiotic metabolism, (b) GSEA enrichment plot for pancreatic beta cell signature.

VI. Discussion

XGBoost outperformed linear baselines and ensemble methods [15]. While RF and LightGBM showed higher internal AUCs, they suffered from overfitting on unseen cohorts. Given the study's emphasis on external generalizability and clinical sensitivity, final model selection was not based solely on internal discrimination performance. Internal comparison metrics alone may fail to reflect out-of-distribution overfitting across heterogeneous cohorts. Therefore, XGBoost was selected based on its superior sensitivity (0.881) and more stable predictive behavior during external validation. While class imbalance caused some variability in sensitivity, the consistently high AUC across cohorts underscores its potential for early diagnostic tools.

A key strength of this study lies in the alignment between model-driven feature importance and biological mechanisms. SHAP analysis identified influential genes [20], notably NDC80, RHNO1, and NAPRT [21]. The model also assigned importance to several long non-coding RNAs, such as THAP9-AS1 and RUSC1-AS1. However, protein-coding drivers were prioritized to facilitate a clear functional interpretation [11],[12]. NDC80, in particular, is a known regulator of cell cycle and chromosomal stability, playing a critical role in pancreatic tumor progression [21].

GSEA further validated the findings, revealing dysregulation in the cell cycle, metabolism, and KRAS signaling [12],[15],[16]. The convergence of ML feature attribution with established oncogenic mechanisms confirmed that the model captures disease-related patterns rather than technical artifacts [9],[10],[25].

Unlike in single-cohort studies, this signature remained robust across 409 unseen samples from six GEO datasets.

Prioritizing sensitivity proved effective. XGBoost maintained a stable 0.761 average AUC across all platforms. However, two cohorts (GSE43795 and GSE55643) contain fewer than ten control samples, limiting their statistical power. To prevent inflating the generalizability claims, the model was re-evaluated by excluding these two small datasets. The remaining four high-powered cohorts (325 samples) yielded a refined average AUC of 0.790. This stable performance confirmed that the model captured a fundamental biological pattern rather than sample-size artifacts or platform-specific noise.

Despite these results, limitations exist. While ComBat reduced batch effects [3],[4], residual variability may persist due to platform-specific differences in probe sensitivity. Nevertheless, ComBat remains the gold standard for multi-platform data harmonization. Recent benchmarks demonstrate its superior ability to preserve biological signals compared to complex deep-learning alternatives [24]. Additionally, reliance on public datasets limits performance assessment in prospective clinical settings.

The study lacks direct experimental validation; findings for NDC80 and THAP9-AS1 remain entirely in silico. To guide future wet-lab translation, a concrete three-phase workflow is proposed.

First, baseline expression levels will be quantified via RT-qPCR across pancreatic cancer cell lines, such as PANC-1 and MIA PaCa-2, versus normal HPDE controls. Western blotting will further confirm NDC80 protein translation. Second, functional loss-of-function assays will evaluate phenotypic changes using siRNA or shRNA-mediated knockdown. Specifically, MTT/CCK-8 proliferation and Transwell invasion assays will determine if depleting these targets suppresses oncogenic capabilities. Finally, ex vivo validation using immunohistochemistry (IHC) on patient tissue microarrays will map NDC80 expression across histological tumor grades. Mechanistically, future assays should clarify whether the lncRNA THAP9-AS1 acts as a competitive endogenous RNA (ceRNA) to regulate NDC80.

Furthermore, integrating clinical covariates like CA 19-9 could mitigate specificity fluctuations in imbalanced cohorts [14],[17]. Future work incorporating longitudinal data and multimodal biomarkers is essential to enhance predictive robustness and realize the framework's full translational potential [22].

VII. Conclusion

In this study, an interpretable machine learning framework was presented for pancreatic cancer prediction. Systematic batch-effect correction and comparative benchmarking of five classifiers identified XGBoost as the most clinically robust model. The proposed framework achieved strong predictive performance and maintained generalizability across multiple external cohorts. SHAP-based explainability enabled the identification of key molecular features associated with pancreatic tumorigenesis. These findings were further supported through pathway enrichment analyses, highlighting established oncogenic mechanisms including cell cycle dysregulation and KRAS signaling.

Overall, the findings demonstrated the potential of explainable machine learning for extracting biologically grounded signatures from high-dimensional transcriptomic data. However, the proposed framework should currently be considered an exploratory computational foundation. Prospective clinical validation and further clinical integration studies remain necessary before translational application in real-world diagnostic settings.

Data and Code Availability

The datasets analyzed in this study are publicly available from the Gene Expression Omnibus (GEO) and The Cancer Genome Atlas (TCGA). The code used in this study is available from the corresponding author upon reasonable request.

References

[1] H. Sung et al., “Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries,” CA: A Cancer Journal for Clinicians, vol. 71, no. 3, pp. 209–249, 2021.

[2] Siegel, R. L., Miller, K. D., Wagle, N. S., & Jemal, A., “Cancer statistics, 2023,” CA: A Cancer Journal for Clinicians, vol. 73, no. 1, pp. 17-48, 2023.

[3] J. T. Leek et al., “The sva package for removing batch effects and other unwanted variation in high-throughput experiments,” Bioinformatics, vol. 28, no. 6, pp. 882–883, 2012.

[4] W. E. Johnson, C. Li, and A. Rabinovic, “Adjusting batch effects in microarray expression data using empirical Bayes methods,” Biostatistics, vol. 8, no. 1, pp. 118–127, 2007.

[5] F. A. Almisned, N. Usanase, D. U. Ozsahin, and I. Ozsahin, "Incorporation of explainable artificial intelligence in ensemble machine learning-driven pancreatic cancer diagnosis," *Scientific Reports*, vol. 15, no. 1, p. 14038, 2025, doi: 10.1038/s41598-025-98298-0.

[6] F. Carrillo-Perez et al., "Machine learning identifies 6-gene signature in peripheral blood for pancreatic cancer diagnosis," Heliyon, vol. 11, no. 1, p. e43138, Jan. 2025, doi: 10.1016/j.heliyon.2025.e43138.

[7] T. Chen and C. Guestrin, “XGBoost: A scalable tree boosting system,” in Proc. 22nd ACM SIGKDD Int. Conf., pp. 785–794, 2016.

[8] S. M. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” in Advances in Neural Information Processing Systems, pp. 4765–4774, 2017.

[9] A. Subramanian et al., “Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles,” Proc. Natl. Acad. Sci. USA, vol. 102, no. 43, pp. 15545–15550, 2005.

[10] M. Ashburner et al., “Gene ontology: Tool for the unification of biology,” Nature Genetics, vol. 25, no. 1, pp. 25–29, 2000.

[11] Cancer Genome Atlas Research Network, “Integrated genomic characterization of pancreatic ductal adenocarcinoma,” Cancer Cell, vol. 32, no. 2, pp. 185–203, 2017.

[12] P. Bailey et al., “Genomic analyses identify molecular subtypes of pancreatic cancer,” Nature, vol. 531, no. 7592, pp. 47–52, 2016.

[13] S. M. Lundberg et al., “From local explanations to global understanding with explainable AI for trees,” Nature Machine Intelligence, vol. 2, no. 1, pp. 56–67, 2020.

[14] L. Rahib et al., “Projected incidence and deaths of pancreatic cancer in the United States, 2010–2030,” Cancer Research, vol. 74, no. 11, pp. 2913–2921, 2014.

[15] A. M. Waters and C. J. Der, “KRAS: The critical driver and therapeutic target in pancreatic cancer,” Cold Spring Harbor Perspectives in Medicine, vol. 8, no. 9, a031435, 2018.

[16] J. Gu et al., "Selection of key biomarkers and functions of pancreatic adenocarcinoma based on multiple internal and external data sets," Frontiers in Genetics, vol. 12, p. 745370, 2022.

[17] L. Chen et al., "A systematic evaluation of methods for correcting batch effects in transcriptomic data," Briefings in Bioinformatics, vol. 22, no. 4, bbaa348, 2021.

[18] X. Chen et al., "Identification of a cell cycle-related gene signature for predicting survival in pancreatic cancer," Cancer Medicine, vol. 11, no. 8, pp. 1914-1926, 2022.

[19] G. J. Rao and A. S. Prasad, "Classification of gene expression from RNA-seq data for pancreatic cancer prognosis using ensemble learning," Journal of Applied Biology & Biotechnology, vol. 12, no. 3, pp. 45–53, May-Jun. 2024, doi: 10.7324/JABB.2024.171755.

[20] K. S. Babu et al., "SHAP-based explanation of a machine learning model for predicting pancreatic cancer," Informatics in Medicine Unlocked, vol. 32, p. 101026, 2022.

[21] T. Liu et al., "NDC80 promotes proliferation and metastasis of pancreatic cancer through the Wnt/β-catenin signaling pathway," Journal of Cancer, vol. 12, no. 16, pp. 4843–4853, 2021.

[22] A. K. Shukla et al., "Integrative analysis of multi-omics data for biomarker discovery in cancer," Briefings in Bioinformatics, vol. 22, no. 4, bbaa212, 2021.

[23] J. Ge, J. Cai, G. Zhang, D. Li, and L. Tao, "Multi-omics integration and machine learning uncover molecular basal-like subtype of pancreatic cancer and implicate A2ML1 in promoting tumor epithelial-mesenchymal transition," Journal of Translational Medicine, vol. 23, no. 1, p. 741, 2025.

[24] Yu, Y., Zhang, N., Mai, Y. et al. Correcting batch effects in large-scale multiomics studies using a reference-material-based ratio method. Genome Biol 24, 201 (2023).

[25] The Gene Ontology Consortium, "The Gene Ontology knowledgebase in 2023," *Genetics*, vol. 224, no. 1, p. iyad031, 2023, doi: 10.1093/genetics/iyad031.