GMR
The Original

A GENOMICS-INFORMED ENSEMBLE LEARNING FRAMEWORK FOR EARLY RISK PREDICTION OF PANCREATIC DUCTAL ADENOCARCINOMA FROM STRUCTURED CLINICAL PHENOTYPES

[*1]V. Aravind Rajan, [2]Dr. P. Ganesh Kumar

[1]Research Scholar, Department of Computer Science and Engineering, School of Computing, Kalasalingam Academy of Research and education, Tamil Nadu, * Corresponding AuthorEmail: aravindarajanvellaisamy@gmail.com
[2]Professor, Department of Information Technology, K.L.N College of Engineering, Madurai.Tamil Nadu, Email: itz.ganeshkumar@gmail.com

ABSTRACT
Pancreatic ductal adenocarcinoma (PDAC) is one of the most aggressive cancers that afflicts humans, with very low five year survival rate (<10%) and a nearly fatal prognosis upon diagnosis. By contrast, the molecular pathogenesis of PDAC is highly defined – as the disease progresses through pancreatic intraepithelial neoplasia (PanIN), KRAS is mutated to an activating form and CDKN2A, TP53 and SMAD4 are inactivated. These genomic drivers are seldom exploited for population-level early detection, because tissue genotyping is invasive and rarely available before symptoms appear. In this study we ask whether the systemic phenotypic consequences of that molecular cascade, recorded in routinely collected clinical and laboratory variables, can flag individuals at elevated risk. We first formalise a generative genotype–phenotype model that treats clinical features as noisy observations of a latent molecular state, and then develop a stacked ensemble that combines Random Forest and XGBoost base learners with a logistic-regression meta-learner. The framework is specified through a hierarchy of equations spanning feature representation, class-imbalance correction, the two base learners, stacked generalisation and evaluation, and is assessed across four independent datasets spanning balanced and severely imbalanced class distributions using AUC, F1-score, recall, specificity and the Brier score. The ensemble achieved perfect discrimination on small, clean datasets (AUC = 1.00) but degraded to chance-level ranking (AUC ≈ 0.50) on large, highly imbalanced cohorts, where it collapsed onto the majority class. These contrasting outcomes show that dataset composition, label leakage and class imbalance, rather than the choice of algorithm, govern apparent accuracy. We interpret this behaviour through the molecular biology of PDAC and set out, with an explicit fusion model, how germline variant panels, circulating tumour DNA and methylation signatures could convert a phenotype-based screen into a genuinely genomics-integrated risk tool. The framework provides a transparent and reproducible scaffold for translating the established genetics of pancreatic cancer into early-detection strategies.

KEYWORDS: Pancreatic ductal adenocarcinoma; Cancer genetics; KRAS; Molecular pathogenesis; Ensemble machine learning; Genotype–phenotype; Early risk prediction.

1. INTRODUCTION

1.1 Global burden and epidemiology of pancreatic cancer

Pancreatic cancer is one of the most aggressive solid tumours affecting humans, with a five-year survival rate that remains below 10% despite decades of clinical effort (Siegel et al., 2023). According to GLOBOCAN 2022 the disease accounts for roughly half a million new cases and a comparable number of deaths each year, and although it ranks only twelfth in incidence it is the fourth leading cause of cancer-related mortality, with a mortality-to-incidence ratio close to 0.94 (Bray et al., 2024). Incidence is not uniform across the world. Higher age-standardised rates are observed in North America, Western and Central Europe, Eastern Asia and Australasia, reflecting ageing populations, obesity, diabetes and greater diagnostic capacity, whereas the lower reported rates in parts of South America and Africa partly reflect under-diagnosis and limited access to imaging (Figure 1).

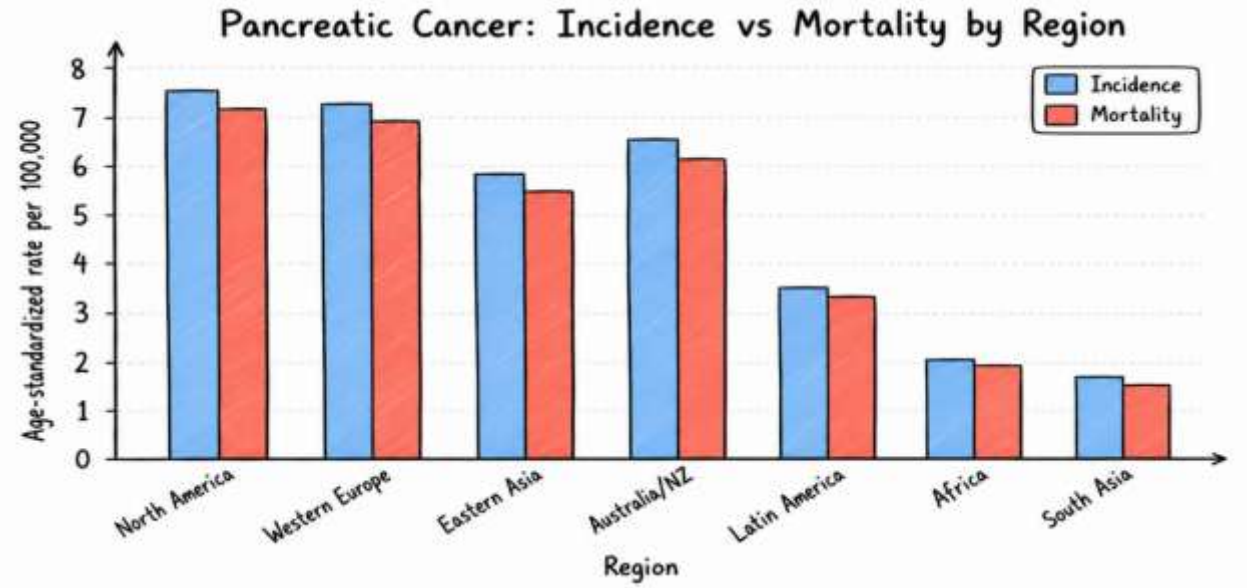

Figure 1. Age-standardised incidence and mortality of pancreatic cancer across world regions. The near-parity of the two bars in every region reflects the disease's characteristically high lethality. Data extrapolated from GLOBOCAN 2022.

Pancreatic ductal adenocarcinoma (PDAC) is the form responsible for this burden, comprising more than 90% of cases. Its lethality is fundamentally a problem of timing rather than of treatment: PDAC is clinically silent during the period in which it is biologically curable, so that more than 80% of patients present with locally advanced or metastatic disease for which surgical resection is no longer possible (Li et al., 2004). Reducing PDAC mortality therefore depends on detecting the disease while its molecular footprint is still small — which makes the genetics of the tumour, rather than any single downstream test, the natural starting point.

1.2 The molecular pathogenesis of PDAC

What distinguishes PDAC from many common cancers is that its genetic life history is exceptionally well charted. The disease arises through an ordered accumulation of somatic alterations in a small set of driver genes, summarised in the classical progression model of Hruban et al. (2000) and later refined by whole-genome sequencing of large patient cohorts (Waddell et al., 2015). Progression begins in the normal ductal epithelium with an activating point mutation in KRAS, most commonly at codon 12. This is the initiating and near-universal event, present in roughly 90–95% of tumours and already detectable in the earliest pancreatic intraepithelial neoplasia (PanIN-1) lesions, accompanied by telomere shortening that predisposes to chromosomal instability. As lesions advance to PanIN-2 the tumour-suppressor CDKN2A (encoding p16) is inactivated, releasing a key brake on the cell cycle. Progression to PanIN-3, the carcinoma-in-situ stage, is marked by loss of TP53 and SMAD4 (DPC4), which together cripple DNA-damage responses, apoptosis and transforming growth factor-β signalling and license frank invasion. The ordered nature of this KRAS → CDKN2A → TP53/SMAD4 cascade is summarised in Figure 2.

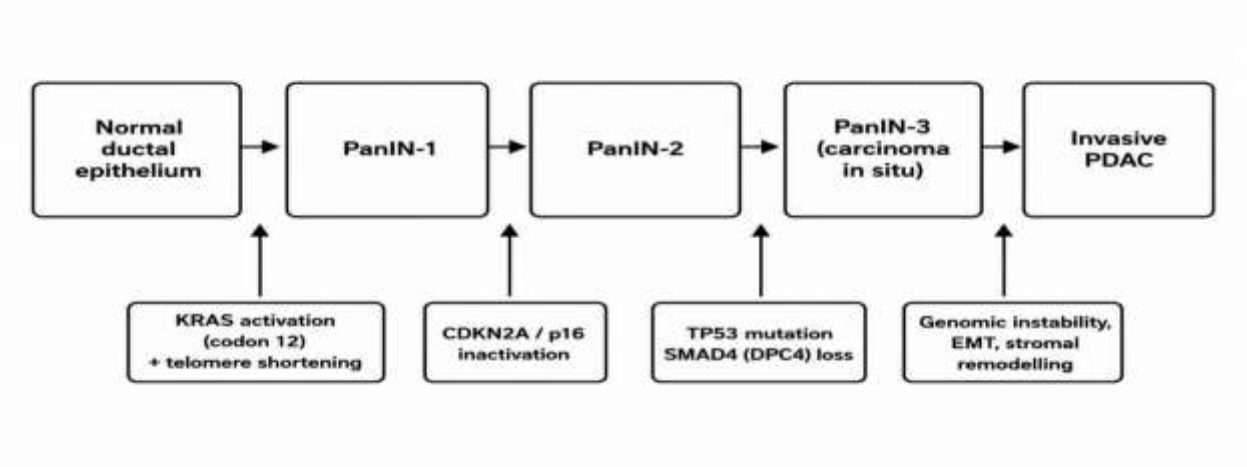

Figure 2. The genetic progression model of PDAC. Driver alterations accumulate in an ordered sequence across PanIN grades, from early KRAS activation to late inactivation of TP53 and SMAD4, accompanied by increasing genomic instability and progressive loss of clinical reversibility.

1.3 Germline susceptibility and inherited risk

A clinically important minority of PDAC arises in carriers of germline pathogenic variants. Inherited mutations in BRCA2, PALB2, ATM, CDKN2A, STK11 and the mismatch-repair genes raise lifetime risk and, in some families, define hereditary pancreatic cancer or related syndromes (Kimura et al., 2021; Klein, 2021). These germline determinants are of importance in two ways. First of all, they distinguish people where there are no symptoms at all, and where surveillance is warranted, which is where every early-detection programme wants to target. Second, a few of them – specifically BRCA2 and PALB2 – make the patient sensitive to platinum agents and PARP inhibitors, directly connecting the genotype to therapeutic decision. Eventually, any system designed for population level early detection should take this inherited layer of risk into account, as well as the somatic cascade.

1.4 Molecular subtypes and their clinical significance

In addition to the driver genes, PDAC in its advanced stage is characterised by a wide-spread genomic instability, EMT and extensive stromal remodeling. Transcriptomic profiling has now been used to distinguish tumours into two reproducible molecular groups broadly known as Classical and Basal-like (also referred to as Squamous) that have different prognosis and sensitivity to therapy (Collisson et al., 2011; Bailey et al., 2016). Poor outcome is always observed with the Basal-like subtype, and the Classical subtype is found to be more chemoresponsive. Implicit in this paper is the idea that PDAC is not a single homogeneous target, but one on a track of molecular states, with the position along the track affecting detectability and response to treatment.

To date, a method for detecting X.25 signals and their associated signaling systems has not been developed.

Despite the above advances in imaging, endoscopy and molecular diagnostics, early detection of PDAC remains elusive. Small tumours (<2cm) and isoattenuating tumours are often not detected by computed tomography and magnetic resonance imaging and can be difficult to distinguish malignancy from chronic pancreatitis. Endoscopic ultrasound is operator-dependent and is sensitive, but is also invasive. The serum carbohydrate antigen 19-9 (CA 19-9) is not recommended as a screening marker as up to 10% of people are non-secretors and the marker is elevated in benign biliary and inflammatory conditions. There are promising approaches to liquid biopsy circulating tumor DNA (ctDNA) panels, methylation panels, and microRNA panels, but they are not yet validated for asymptomatic screening (Liu et al., 2012; Kawai et al., 2024). Histological confirmation is an invasive procedure and is the diagnostic method of choice, but is not appropriate for first-line population screening.

1.6 Rationale: clinical phenotypes as molecular surrogates

One source of information, however, is much larger than all of these tests combined and largely unused: the source is upstream of all these tests. Molecular events that underlie PDAC have measurable systemic repercussions before a mass can be detected in the pancreas: paraneoplastic islet dysfunction and new onset diabetes; subtle changes in glucose and lipid handling; unexplained weight loss; and changes in pancreatic enzymes, which are all part of routine electronic health records. In essence, these phenotypes are available substitutes for the underlying genetic disease. The idea behind the present study is that machine-learning models can be used to decipher these surrogate signals and detect high risk, thereby affording a non-invasive first pass screen that can be followed up by more intrusive genomic verification. To formalise this premise, however, rather than merely rhetorical, we bring it into the realm of the generative model in Section 3.

1.7 Contributions and objectives

The specific contributions of this work are as follows. (i) We provide an explicit genotype–phenotype formulation that frames structured clinical features as noisy observations of a latent molecular state, giving the predictive task a defined biological meaning. (ii) We develop a fully specified stacked ensemble — Random Forest and XGBoost base learners with a logistic-regression meta-learner — and present it through a hierarchy of numbered equations spanning representation, imbalance correction, both base learners, stacking and evaluation. (iii) We evaluate the framework across four datasets with deliberately different class balance, so that its behaviour under realistic data conditions is exposed rather than concealed. (iv) We interpret the empirical results in molecular-genetic terms and define, with an explicit fusion model, a concrete pathway for integrating germline and circulating genomic markers.

2. RELATED WORK

Efforts to detect PDAC early span molecular markers, imaging analytics and structured-data modelling. On the molecular side, Liu et al. (2012) combined plasma microRNA panels with serum CA 19-9 and reported high discriminative performance, and Kawai et al. (2024) extended this to comprehensive serum microRNA sequencing analysed with automated machine learning, deriving predictive signatures from robustly expressed microRNAs in early-stage cases. Both were limited by cohort imbalance and the absence of independent validation in asymptomatic populations.

A second strand is structured-data and imaging models. Using clinical data from the pancreaticoduodenectomy cohort, Wang et al. (2025) employed RF, XGBoost and LR, with an AUC close to 0.85 for post-operative results, however this was not validated on external data. Bharathi and Rayachoti (2025) presented an edge-based segmentation and morphological enhancement pipeline for pancreatic computed tomography (CT) that enhanced the visibility of lesions without creating an end-to-end predictive system. Singh et al. (2025) modeled structured records using Decision Trees, SVM and Logistic Regression, with limited clinical validation, and Sadewo et al. (2020) showed good results with

Kernel Twin SVM on the biomarkers of blood. The authors of these reviews, Gayathri (2017) and Ozsahin et al. (2025), conduct a survey of the more general application of machine and deep learning in oncology and consistently arrive at the same set of problems. This landscape is summarised in Table 1.

Study	Approach	Data type	Reported result	Key limitation
Liu et al. (2012)	miRNA panel + CA 19-9	Plasma	High AUC	No independent cohort
Kawai et al. (2024)	Serum miRNA-seq + AutoML	Serum	Strong early-stage signal	Demographic imbalance
Wang et al. (2025)	RF / XGBoost / LR	Clinical records	AUC ≈ 0.85	No external validation
Singh et al. (2025)	DT / SVM / LR	Structured EHR	Accuracy ≈ 0.90 (SVM)	Limited validation
Sadewo et al. (2020)	Twin SVM (RBF)	Blood biomarkers	Acc. 98%, Sens. 97%	Single dataset
This study	Stacked RF+XGBoost+LR	Structured phenotypes	Dataset-dependent	Tabular only (by design)

Table 1. Representative prior studies on machine-learning-based pancreatic-cancer detection, with reported results and principal limitations.

Two gaps emerge. First, the molecular genetics of PDAC is usually invoked only as background motivation and is rarely used to frame either the features or the interpretation of the models. Second, reports tend to present headline accuracy on favourable datasets while saying little about failure modes under imbalance. The present study addresses both, grounding feature interpretation explicitly in the PDAC driver cascade and deliberately reporting the conditions under which the model fails.

3. THEORETICAL FRAMEWORK AND PROBLEM FORMULATION

3.1 Notation and problem statement

Let the dataset comprise N patients, each represented by a d-dimensional feature vector of engineered clinical and laboratory variables with an associated binary label indicating PDAC status, as defined in Eq. (1).

$$\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}, \mathbf{x}_i \in \mathbb{R}^d, y_i \in \{0,1\} \quad (1)$$

where x_i is the feature vector of patient i, $y_i \in \{0,1\}$ is the PDAC label, N is the number of patients and d the feature dimensionality. The goal is to learn a calibrated mapping that estimates the posterior probability P(y = 1 | x) of disease.

Because the clinical features differ in scale and unit, each continuous feature is standardised by its training-set mean and standard deviation, as in Eq. (2).

$$x'_{ij} = \frac{x_{ij} - \mu_j}{\sigma_j} \quad (2)$$

where μ_j and σ_j are the mean and standard deviation of feature j estimated within the training fold only, so that no statistic from held-out data influences the transformation.

Nomenclature

Symbol	Meaning	Symbol	Meaning
N	number of patients	T	number of trees (RF)
d	feature dimensionality	M	boosting rounds (XGBoost)
x_i	feature vector of patient i	f_m	m-th regression tree
y_i	binary PDAC label	γ, λ	regularisation coefficients
z	latent molecular state	g_i, h_i	first/second-order loss gradients
g(·), ε	phenotype map; noise	w_j, I_j	leaf weight; leaf instance set
σ(·)	logistic sigmoid	β, β_0	LR weights; intercept
p̂_RF, p̂_XGB	base-learner probabilities	α_0,α_1,α_2	meta-learner coefficients
wc	class weight	TP,TN,FP,FN	confusion-matrix counts

3.2 A generative genotype–phenotype model

The biological premise of Section 1.6 can be stated formally. We posit a latent molecular state z that encodes an individual's position along the PDAC progression cascade — the configuration of germline susceptibility and accumulated somatic alterations. The observed clinical phenotype x is then a noisy, nonlinear function of this latent state, as in Eq. (3).

$$\mathbf{x}_i = g(\mathbf{z}_i) + \boldsymbol{\epsilon}_i, \mathbf{z}_i \sim P(\mathbf{z} \mid \text{germline,somatic}) \quad (3)$$

where $g(\cdot)$ is an unknown nonlinear map from molecular state to phenotype, ε_i is observation noise, and the distribution of z is conditioned on germline and somatic genetic configuration.

Under this model, predicting disease from phenotype is an inference problem: the label depends on the latent molecular state, and the phenotype provides only indirect evidence about it. The posterior risk is obtained by marginalising over the latent state, as in Eq. (4).

$$P(y = 1 \mid \mathbf{x}) = \int P(y = 1 \mid \mathbf{z})\, P(\mathbf{z} \mid \mathbf{x})\, d\mathbf{z} \quad (4)$$

Equation (4) makes explicit why phenotype-based prediction is hard early in the disease. When the molecular state corresponds to an initiating KRAS lesion, the phenotype x it generates is faint and the posterior $P(z \mid x)$ is diffuse, so the integral carries little discriminative signal. The classifier developed below is, in effect, an approximation to this posterior; the generative structure is depicted in Figure 3, and it directly motivates the genomic-integration framework of Section 6.4, where genetic features are added to inform z directly.

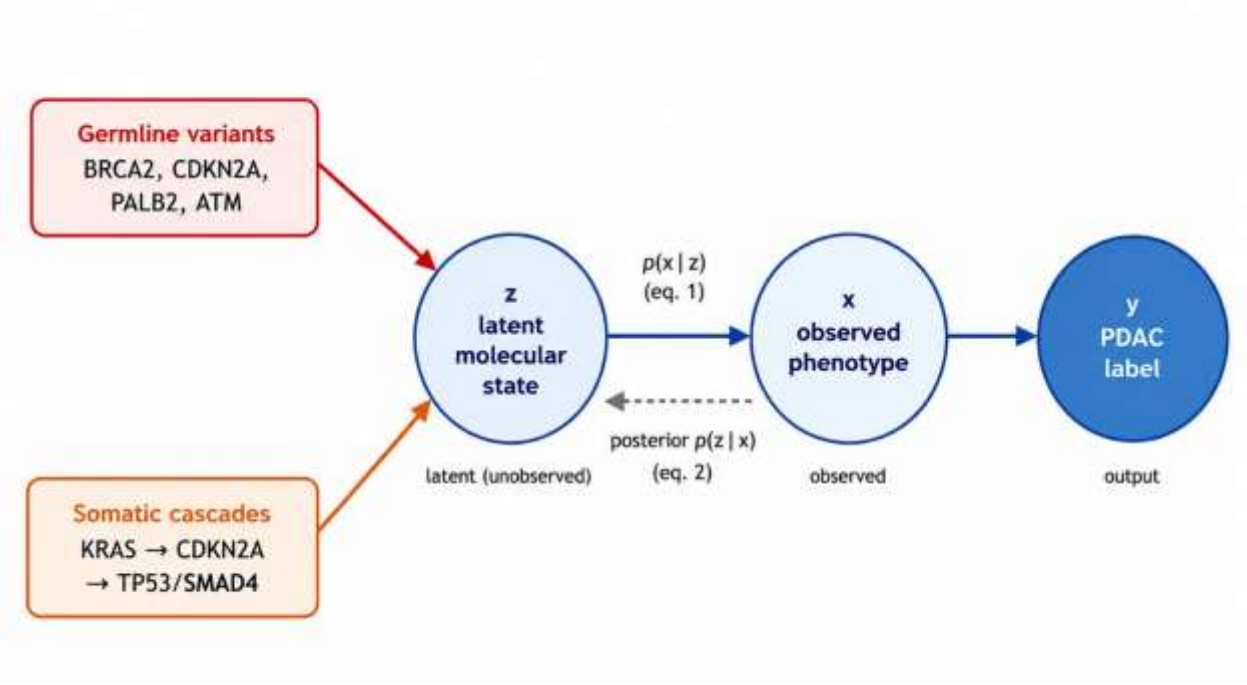

Figure 3. Generative genotype–phenotype model. Germline and somatic genetic configuration shape a latent molecular state z, which generates the observed clinical phenotype x (Eq. 3); risk prediction inverts this map to recover $P(z \mid x)$ and hence $P(y = 1 \mid x)$ (Eq. 4).

4. MATERIALS AND METHODS

We instantiate the inference problem of Section 3 with a structured-data stacked ensemble and evaluate it across four datasets. Each dataset passes through a common pipeline of cleaning, encoding, imbalance correction, cross-validation and ensemble learning, with all preprocessing fitted inside cross-validation folds to prevent information leakage. The overall workflow is shown in Figure 4.

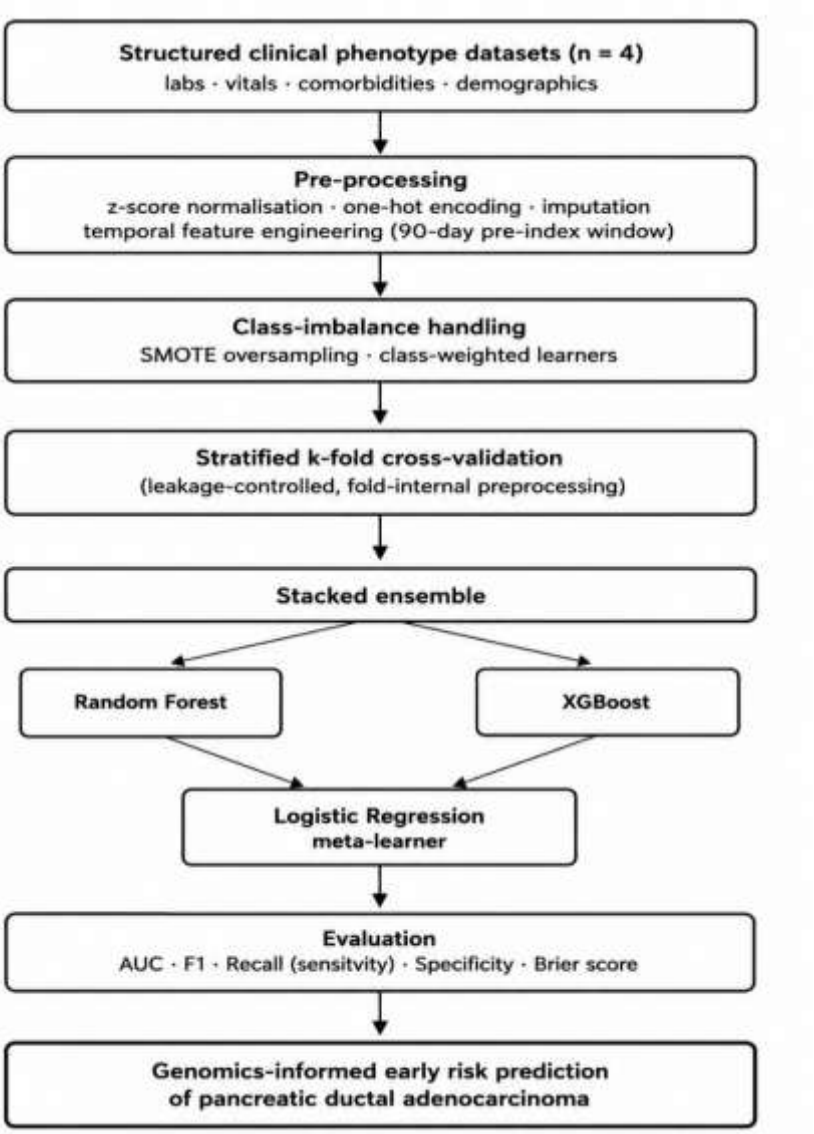

Figure 4. Analytical workflow. Structured clinical phenotype datasets are pre-processed, balanced and passed through leakage-controlled cross-validation to a stacked ensemble, before evaluation and genomics-informed risk prediction.

4.1 Data sources and clinical phenotypes as molecular surrogates

Four structured datasets of pancreatic-cancer cases and controls were used, deliberately chosen to span a range of sizes and class balance, from small and near-balanced to large and severely imbalanced. Variables comprised demographics (age, sex, ethnicity), laboratory markers (for example glucose, bilirubin and CA 19-9), vital signs and anthropometrics (body-mass index and weight change) and comorbidity flags such as diabetes and pancreatitis, yielding approximately 30–50 features per patient after engineering. The interpretive premise is that these phenotypes are not arbitrary clinical variables but measurable consequences of the PDAC molecular programme. Table 2 makes this genotype–phenotype correspondence explicit.

Recorded clinical / laboratory phenotype	**Underlying molecular / genetic correlate**
New-onset hyperglycaemia / diabetes	Paraneoplastic islet β-cell dysfunction driven by the KRAS-mutant tumour secretome; one of the earliest systemic signals of an evolving lesion.
Elevated serum CA 19-9	Aberrant glycosyltransferase activity in transformed ductal epithelium; modulated by FUT2/FUT3 secretor genotype.
Progressive weight loss / cachexia	Tumour-derived inflammatory cytokines and metabolic reprogramming; associated with later clonal stages (TP53/SMAD4 loss).

Recorded clinical / laboratory phenotype	Underlying molecular / genetic correlate
History of chronic pancreatitis	Inflammation-driven selection and expansion of KRAS-mutant ductal clones.
Family history / early age at onset	Germline pathogenic variants in CDKN2A, BRCA2, PALB2, ATM, STK11 or mismatch-repair genes.
Altered bilirubin / pancreatic enzymes	Local ductal obstruction and acinar disruption by an enlarging neoplastic lesion.

Table 2. Mapping of routinely recorded clinical and laboratory phenotypes to their underlying molecular and genetic correlates in PDAC. This correspondence is the biological content of the feature vector x in Eq. (1).

4.2 Data preprocessing

Records were harmonised at the patient level using de-identified identifiers. Continuous laboratory variables were imputed by mean or median substitution, with model-based imputation where appropriate, and were then standardised by Eq. (2). Categorical fields were one-hot encoded, with a dedicated category retained for missing entries. Temporal features were derived where longitudinal data allowed — for example weight change and glucose trend over a 90-day window preceding the index date — because the direction of change often carries more signal than a single measurement. Outliers in laboratory values were capped using interquartile-range thresholds or flagged as separate indicator features. Every step of imputation, scaling and feature derivation was fitted only on the training portion of each fold and then applied to the held-out portion.

4.3 Class-imbalance correction

Pancreatic-cancer datasets are typically imbalanced, and several used here are extremely so (Figure 5). Two complementary corrections were applied. First, the Synthetic Minority Over-sampling Technique (SMOTE) generated synthetic minority examples by interpolating between a minority sample and one of its minority-class nearest neighbours, as in Eq. (5).

$$\mathbf{x}_{\text{new}} = \mathbf{x}_i + \lambda\left(\mathbf{x}_i^{(k)} - \mathbf{x}_i\right), \qquad \lambda \in [0,1] \tag{5}$$

where x_i is a minority sample, $x_i^{(k)}$ one of its k minority-class nearest neighbours and λ a random interpolation coefficient. Second, class weights were incorporated into the loss of the XGBoost and logistic-regression learners to penalise errors on the minority class in inverse proportion to its frequency, as in Eq. (6).

$$w_c = \frac{N}{K \cdot N_c} \tag{6}$$

where N is the total number of samples, K the number of classes and Nc the count of class c. SMOTE was applied within training folds only, never to validation or test data.

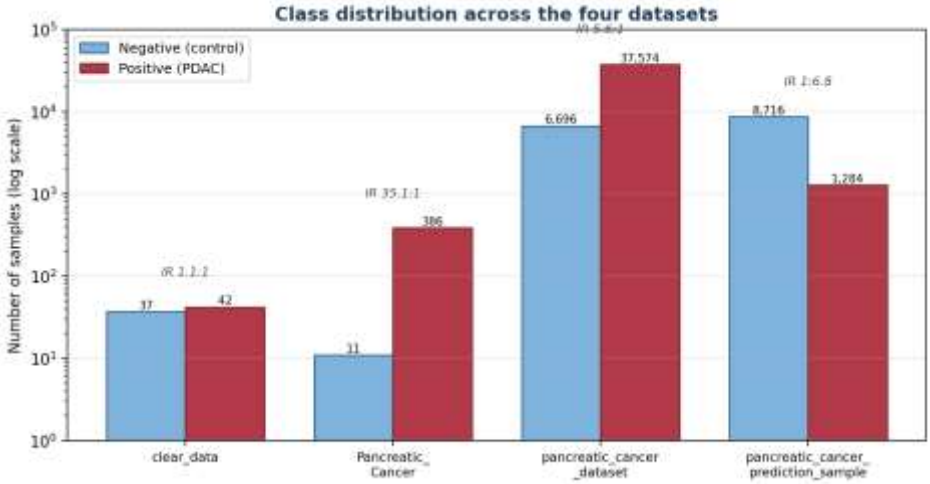

Figure 5. Class distribution across the four datasets on a logarithmic scale, with the imbalance ratio (IR) annotated. The two large datasets are severely skewed in opposite directions, which proves decisive for model behaviour (Section 5).

4.4 Base learner I: Random Forest

The first base learner is a Random Forest, an ensemble of T decision trees each grown on a bootstrap resample of the training data with feature subsampling at every split. Its probability estimate is the average over the trees, as in Eq. (7).

$$\hat{p}_{RF}(\mathbf{x}) = \frac{1}{T}\sum_{t=1}^{T} h_t(\mathbf{x}) \quad (7)$$

where $h_t(x)$ is the class-probability output of tree t. Splits within each tree are chosen to maximise node purity, measured by the Gini index or entropy of Eq. (8).

$$G(S) = \sum_k p_k(1-p_k), \qquad H(S) = -\sum_k p_k \log_2 p_k \quad (8)$$

where p_k is the proportion of class k in node S. The split selected at each node is the one maximising the resulting information gain of Eq. (9).

$$\Delta I = I(S) - \sum_{v\in\{L,R\}} \frac{|S_v|}{|S|} I(S_v) \quad (9)$$

where S_L and S_R are the left and right child nodes. Averaging over decorrelated trees reduces variance and yields stable, interpretable feature-importance estimates.

4.5 Base learner II: XGBoost

The second base learner is gradient-boosted trees (XGBoost), an additive model that builds an ensemble of M regression trees sequentially, as in Eq. (10).

$$\hat{y}_i = \sum_{m=1}^{M} f_m(\mathbf{x}_i), f_m \in \mathcal{F} \quad (10)$$

where each f_m belongs to the space F of regression trees. The model is fitted by minimising a regularised objective that balances training loss against model complexity, as in Eq. (11).

$$\mathcal{L} = \sum_{i=1}^{N} l(y_i, \hat{y}_i) + \sum_{m=1}^{M} \Omega(f_m) \quad (11)$$

The complexity penalty discourages overly deep trees and large leaf weights, as in Eq. (12).

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \quad (12)$$

where T is the number of leaves, w_j the weight of leaf j, and γ and λ regularisation coefficients. At boosting round m the objective is approximated to second order by a Taylor expansion, as in Eq. (13).

$$\mathcal{L}^{(m)} \simeq \sum_{i=1}^{N} \left[g_i f_m(\mathbf{x}_i) + \frac{1}{2} h_i f_m^2(\mathbf{x}_i) \right] + \Omega(f_m) \quad (13)$$

with first- and second-order gradients of the loss given by Eq. (14).

$$g_i = \partial_{\hat{y}^{(m-1)}} l\left(y_i, \hat{y}^{(m-1)}\right), h_i = \partial^2_{\hat{y}^{(m-1)}} l\left(y_i, \hat{y}^{(m-1)}\right) \quad (14)$$

Minimising the approximated objective gives a closed-form optimal weight for each leaf, as in Eq. (15), which also defines the gain used to evaluate candidate splits.

$$w_j^* = -\frac{\sum_{i\in I_j} g_i}{\sum_{i\in I_j} h_i + \lambda} \quad (15)$$

where I_j is the set of instances assigned to leaf j. Boosting reduces bias by sequentially fitting the residual gradients, complementing the variance reduction of the Random Forest.

4.6 Meta-learner: Logistic Regression

The meta-learner is a regularised logistic regression. It maps a linear combination of its inputs through the logistic sigmoid of Eq. (16) to a probability, as in Eq. (17).

$$\sigma(z) = \frac{1}{1+e^{-z}} \quad (16)$$

$$P(y = 1 \mid \mathbf{x}) = \sigma(\boldsymbol{\beta}^T \mathbf{x} + \beta_0) \quad (17)$$

where β is the weight vector and β_0 the intercept. Its parameters are estimated by minimising the L2-regularised cross-entropy of Eq. (18).

$$J(\boldsymbol{\beta}) = -\frac{1}{N}\sum_{i=1}^{N} [y_i \log \hat{p}_i + (1-y_i)\log(1-\hat{p}_i)] + \frac{\lambda}{2} \| \boldsymbol{\beta} \|_2^2 \quad (18)$$

where $\hat{y}_i$ is the predicted probability for patient i and λ controls the strength of L2 shrinkage. The logistic form keeps the meta-learner transparent: its coefficients are directly interpretable as the learned trust placed in each base learner.

4.7 Stacked generalisation

The two base learners are combined by stacked generalisation. Each produces a class-probability estimate, and these are concatenated into a meta-feature vector, as in Eq. (19).

$$\mathbf{z}_i = [\hat{p}_{RF}(\mathbf{x}_i), \hat{p}_{XGB}(\mathbf{x}_i)] \quad (19)$$

The logistic meta-learner is then trained on these meta-features, using out-of-fold base predictions to avoid leakage, producing the final calibrated risk score of Eq. (20).

$$\hat{y}_i = \sigma\left(\alpha_0 + \alpha_1 \hat{p}_{RF}(\mathbf{x}_i) + \alpha_2 \hat{p}_{XGB}(\mathbf{x}_i)\right) \quad (20)$$

where α_0, α_1 and α_2 are the meta-learner coefficients. The full architecture is shown in Figure 6.

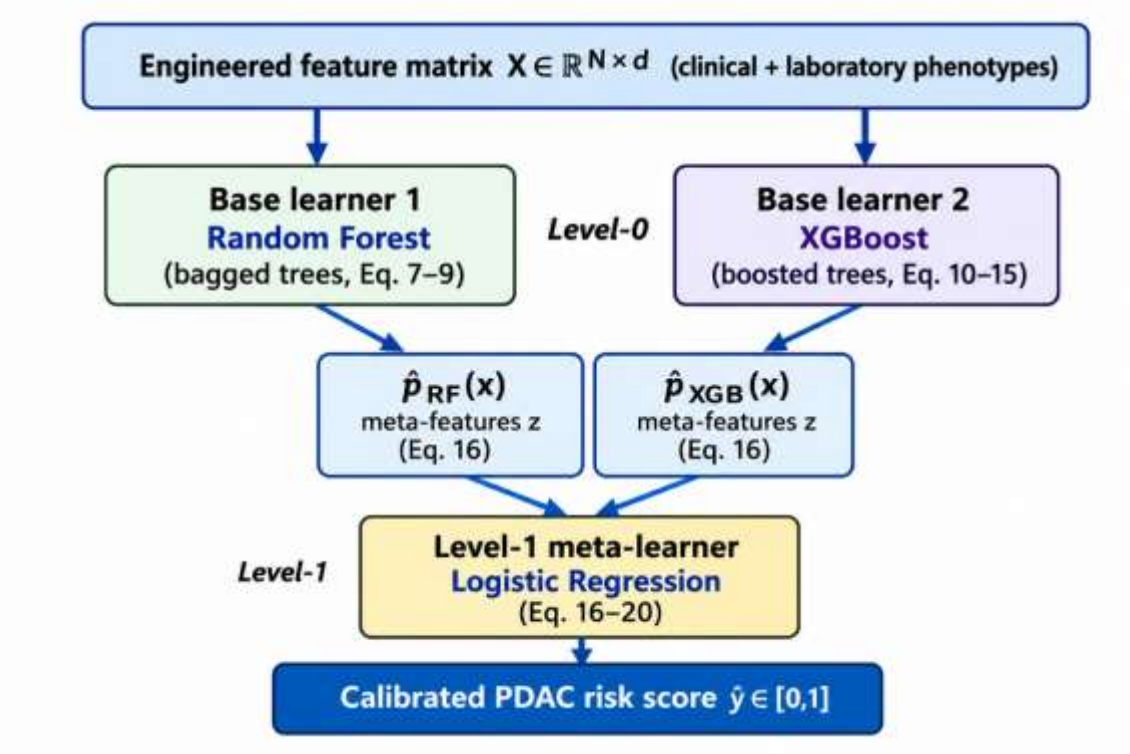

Figure 6. Stacked ensemble architecture. Random Forest and XGBoost base learners (level-0) produce probability estimates that form the meta-features of Eq. (19); a logistic-regression meta-learner (level-1) combines them into a calibrated PDAC risk score by Eq. (20).

The complete training procedure, including leakage-controlled preprocessing and out-of-fold stacking, is given in Algorithm 1.

Algorithm 1. Training the genomics-informed stacked ensemble

```
Input: dataset D = {(x_i, y_i)}, folds k
Output: trained base learners, meta-learner, risk function
partition D into stratified folds F_1..F_k (patient-level)
for each fold f in 1..k do
    fit imputation, z-score (Eq. 2) and encoders on D \ F_f
    apply SMOTE (Eq. 5) and class weights (Eq. 6) on D \ F_f
    train Random Forest -> p_RF (Eq. 7-9)
    train XGBoost      -> p_XGB (Eq. 10-15)
    predict p_RF, p_XGB on held-out F_f  (out-of-fold)
    store meta-features z_i = [p_RF, p_XGB]  (Eq. 19)
end for
train logistic meta-learner on out-of-fold {z_i, y_i} (Eq. 16-18, 20)
refit base learners on full balanced D
return ensemble risk function y_hat(x) = sigma(a0 + a1 p_RF + a2 p_XGB)
```

4.8 Cross-validation, tuning and computational cost

Model generalisation was estimated by stratified k-fold cross-validation, which preserves class proportions across folds; five folds were used for tuning and ten folds for performance estimation. Hyperparameters were optimised separately for each base learner by grid and randomised search, selecting the configuration with the highest validation AUC and applying early stopping where supported (Table 3). Data were partitioned at the patient level into 70% training, 15% validation and 15% held-out test, with stratified sampling throughout.

Base learner	Hyperparameters tuned
Random Forest	Number of trees T; maximum depth; minimum samples per split; minimum leaf size.
XGBoost	Learning rate (eta); number of estimators M; maximum depth; subsample; colsample_bytree; γ, λ.
Logistic Regression	Regularisation strength (inverse λ); penalty type (L1 / L2); solver.

Table 3. Hyperparameters optimised for each component of the stacked ensemble.

terms of computational cost, training the Random Forest scales approximately as $O(T \cdot N \log N \cdot m)$ for T trees and m features sampled per split, and XGBoost as $O(M \cdot N \cdot d \cdot \text{depth})$ for M boosting rounds; the logistic meta-learner is negligible by comparison, operating on a two-dimensional meta-feature space. All experiments ran on a commodity machine (Intel Core i5, 8 GB RAM), confirming that the framework is light enough for routine deployment without specialised hardware.

4.9 Evaluation metrics

Because the cost of a missed cancer far exceeds that of a false alarm, the model was assessed on class-specific behaviour as well as overall accuracy. Precision and recall are defined in Eq. (21), specificity and accuracy in Eq. (22), and their balance is summarised by the F1-score of Eq. (23).

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{21}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}, \text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{22}$$

$$\text{F}_1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{23}$$

where TP, TN, FP and FN are the true positive, true negative, false positive and false negative counts. Threshold-independent discrimination is measured by the area under the ROC curve of Eq. (24), which equals the probability that a random positive is ranked above a random negative.

$$\text{AUC} = \int_0^1 \text{TPR}\,(\text{FPR})\,\text{d}(\text{FPR}) = \text{P}(\hat{\text{p}}^+ > \hat{\text{p}}^-) \tag{24}$$

Finally, the reliability of the predicted probabilities is quantified by the Brier score of Eq. (25), with lower values indicating better calibration.

$$\text{Brier} = \frac{1}{\text{N}} \sum_{\text{i}=1}^{\text{N}} (\hat{\text{p}}_\text{i} - \text{y}_\text{i})^2 \tag{25}$$

4.10 Statistical analysis and reproducibility

Point estimates of each metric were reported together with their dispersion across the ten cross-validation folds, summarised by the fold-wise mean and standard deviation, so that a single favourable split could not be mistaken for stable performance. Confusion-matrix counts were retained for every dataset to expose class-specific behaviour that aggregate scores can hide, and the ROC curves were inspected directly rather than relying on the AUC alone. To support reproducibility, random seeds were fixed for partitioning, resampling and model initialisation, and the identical preprocessing-and-modelling pipeline was applied to all four datasets without dataset-specific tuning beyond the hyperparameter search of Table 3.

5. RESULTS

The ensemble was evaluated on four datasets under an identical environment. The consolidated metrics are reported in Table 4, and per-dataset ROC curves and confusion matrices in Figures 7–10. The results are striking less for their best-case numbers than for the contrast between them, which is the central finding of this study.

Dataset	F1	Recall	Specificity	AUC	Brier	Outcome
clear_data.csv	1.00	1.00	1.00	1.00	0.003	Perfect
Pancreatic_Cancer.csv	1.00	1.00	1.00	1.00	0.000	Leakage-likely
pancreatic_cancer_dataset.csv	0.918	1.00	0.00	0.50	0.144	Majority-class
pancreatic_cancer_prediction_sample.csv	0.00	0.00	1.00	0.495	0.129	Majority-class

Table 4. Consolidated evaluation of the stacked ensemble across the four datasets. Apparent performance varies dramatically with dataset composition.

5.1 Balanced and clean data (clear_data.csv)

On clear_data.csv the ensemble reached an AUC of 1.00 and an F1-score of 1.00, with a near-balanced class distribution (37 versus 42) and no misclassifications (Figure 7). The low Brier score of 0.003 indicates well-calibrated, confident probabilities. But this separation is usually to be taken with a pinch of salt, and not as a cause for celebration, because it is likely that the data is small, synthetically clean, or filtered of the noise that real clinical populations would have. The result indicates that the pipeline is correct but provides little insight into the usefulness of the pipeline in practice.

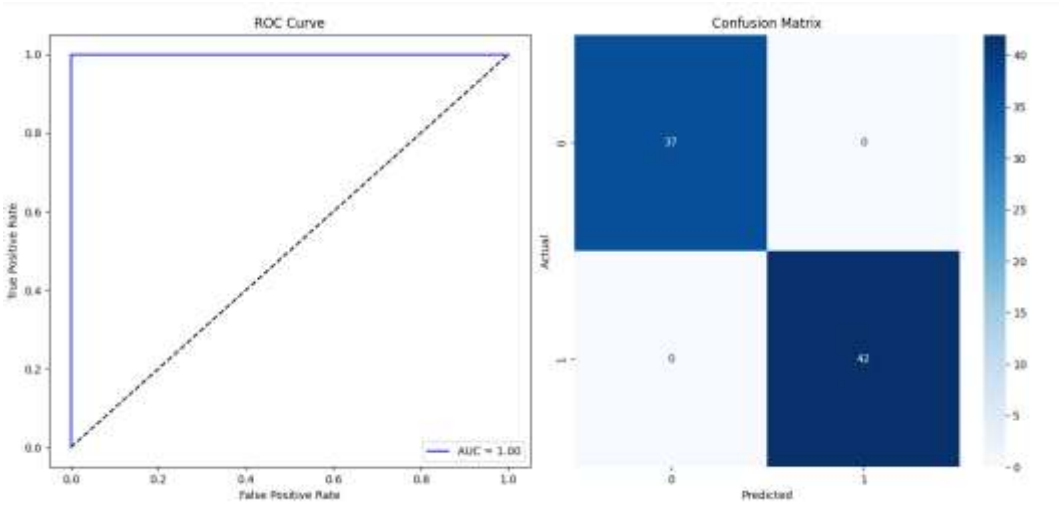

Figure 7. ROC curve and confusion matrix for clear_data.csv. The model separates the two near-balanced classes without error (AUC = 1.00).

5.2 Imbalanced but cleanly separable data (Pancreatic_Cancer.csv)

Again for Pancreatic_Cancer.csv all the metrics achieved their highest values (Figure 8). Here, it's a bit more dubious than a success. It is highly unbalanced (386 positive and 11 negative samples) and is very rare that any model is able to get perfect accuracy on both classes of this type of imbalance. The most likely scenario is just label leakage — a feature that seems to be an almost perfect proxy for the result — not true early-detection ability. We report it this way because it is a case in point of the dangers of a misleading headline number.

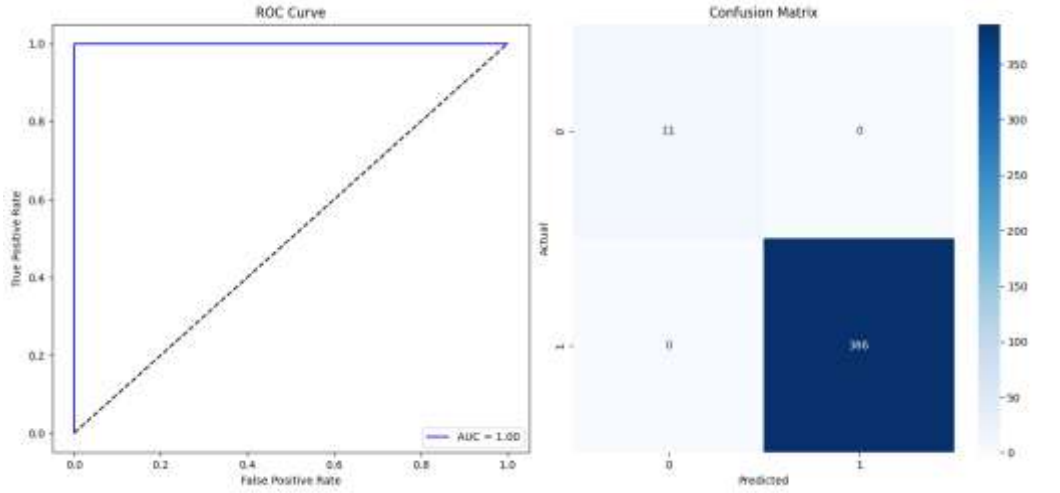

Figure 8. ROC curve and confusion matrix for Pancreatic_Cancer.csv. Apparently perfect classification on a heavily imbalanced dataset (386 positive, 11 negative), consistent with label leakage.

5.3 Large imbalanced data: recall collapse (pancreatic_cancer_dataset.csv)

On the two big data sets the picture is completely different. The model had a recall of 1.00 and an F1-score of 0.918, and a specificity of 0.00, and AUC of 0.50 (Figure 9) for pancreatic_cancer_dataset.csv. The confusion matrix is telling: the model correctly classified the majority class (positive) in nearly every case, correctly identifying all the

positive cases and missing just one true negative case. The AUC of 0.50 is consistent with this not being discrimination, but rather a prevalence-driven strategy with a high number of false-positives in practice.

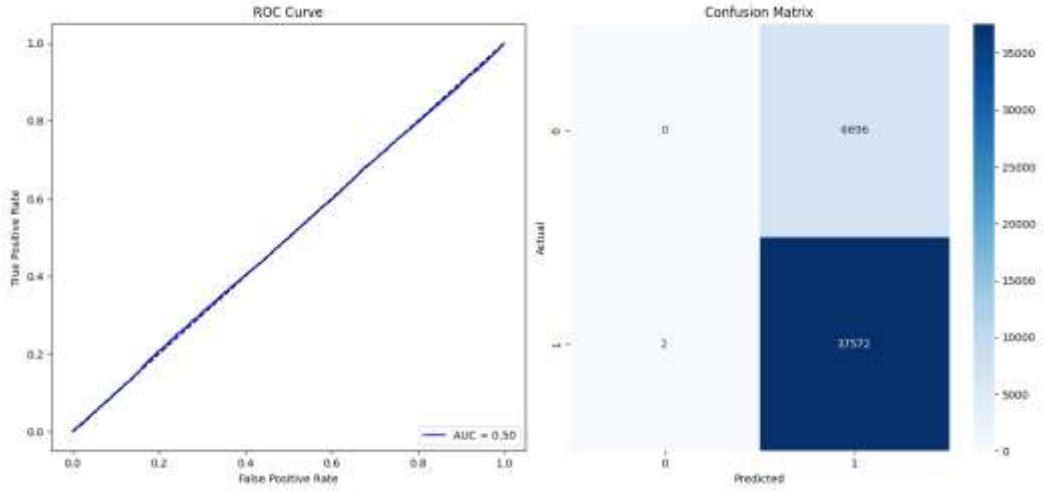

Figure 9. ROC curve and confusion matrix for pancreatic_cancer_dataset.csv. Under extreme imbalance the model defaults to the majority (positive) class: perfect recall but zero specificity and chance-level AUC.

5.4 Large imbalanced data: specificity collapse (pancreatic_cancer_prediction_sample.csv)

For pancreatic_cancer_prediction_sample.csv, the model had a specificity of 1.00, recall of 0.00 and an AUC of 0.495 (Figure 10). The majority class is now negative (8716 negative, 1284 positive), and the model predicted everyone to be cancer-free, with high apparent accuracy and failed to predict any actual cancer case. This is the worst failure in a screening situation, because it is silent: the accuracy measure appears to be good but the clinically important class is completely missed.

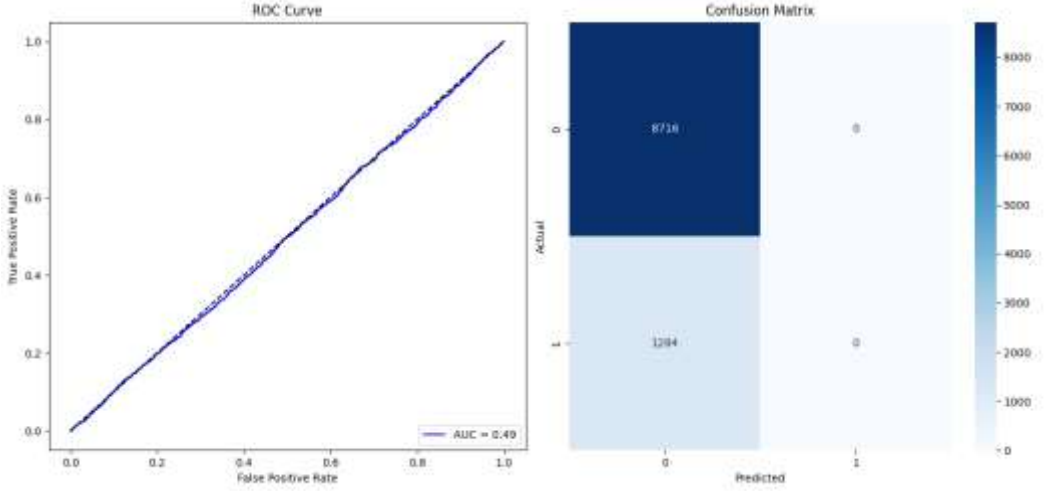

Figure 10. ROC curve and confusion matrix for pancreatic_cancer_prediction_sample.csv. The model defaults to the majority (negative) class: perfect specificity but zero recall, missing every true case.

5.5 Consolidated comparison and calibration

The pattern is clear in the numbers shown in figure 11, which are summarised for all four datasets. Strong and balanced performance is limited to the small sized datasets, and the two large datasets fall into one or the other class. The Brier scores confirm this interpretation: close to zero where model efficacy is high (i.e. where the model is discriminating well between the classes), but higher at 0.144 and 0.129—on the large datasets, where the model predicts probabilities that are not just mispredicting but miscalibrated. The comparison with Table 4 indicates that the ensemble does not have an intrinsic accuracy, independent of the data sets, but it is sensitive to the composition of the data sets it is provided with.

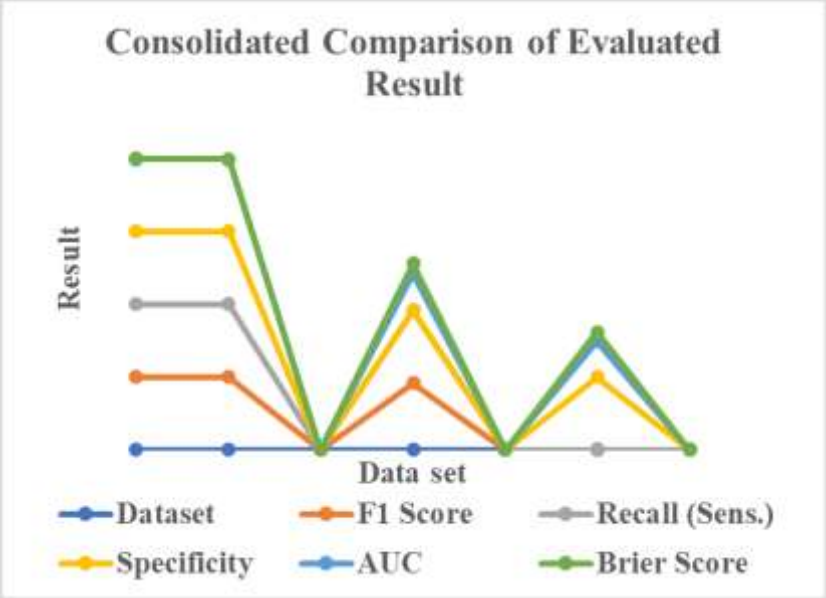

Figure 11. Consolidated comparison of evaluation metrics across the four datasets, illustrating the divergence between performance on small, clean data and on large, imbalanced data.

The number of each class in the confusion matrix shown in Figures 7 to 10 are presented in Table 5. They make the two failure modes vivid: dataset 3 documents 6,696 false positives and no true negatives, and dataset 4 documents 1,284 false negatives and no true positives.

Dataset	TP	TN	FP	FN	n
clear_data.csv	42	37	0	0	79
Pancreatic_Cancer.csv	386	11	0	0	397
pancreatic_cancer_dataset.csv	37,572	0	6,696	2	44,270
pancreatic_cancer_prediction_sample.csv	0	8,716	0	1,284	10,000

Table 5. Confusion-matrix counts for each dataset (TP, TN, FP, FN and total n). The all-positive and all-negative collapses on the two large datasets are evident from the zero entries.

6. DISCUSSION

The value of this study lies in what its failures reveal. The same model and pipeline were used for all four datasets, with two of those yielding perfect metrics and two yielding chance-level discrimination. It was not the algorithm, but the data: how big it was, how well it was balanced by class, and whether or not it contained leaked label information. This is a more useful conclusion than any one accuracy figure since it points to the direction in which effort should be focused.

6.1 A molecular reading of the model's behaviour

From Eq. (3)–(4) and Fig. 2 it is seen that the difficulty is intuitive. The clinical phenotypes presented here are late and indirect reflections of the molecular cascade. The most characteristic markers of new onset diabetes and weight loss, together with a rising CA 19-9, do not become associated with KRAS activation until after the loss of TP53 and SMAD4, and are associated with the more advanced stage of tumour systemicisation.A model built on these surrogates is therefore strongest precisely when the disease is already advanced, which is the opposite of what early detection requires. In the language of the generative model, the latent state z that matters most for early intervention produces a phenotype x whose posterior $P(z \mid x)$ is too diffuse to support confident inference. The recall and specificity collapses are the statistical expression of this biological truth.

6.2 Why imbalance drives failure

The two large datasets fail in opposite directions, and Figure 5 explains why. When the positive class dominates (an imbalance ratio of roughly 5.6 to 1), the loss is minimised by predicting cancer for everyone, giving perfect recall and zero specificity. When the negative class dominates (roughly 6.8 to 1 the other way), the same logic predicts health for everyone, giving perfect specificity and zero recall. The chance-level AUC in both cases confirms that no genuine ranking has been learned. SMOTE and class weighting (Eq. 5–6) are designed to counter exactly this, and their insufficiency here indicates that, on these particular datasets, the features simply do not carry enough early signal for resampling to recover — again pointing to the need for markers closer to the genotype.

6.3 Comparison with prior work

The single-dataset successes reported elsewhere — Wang et al. (2025), Singh et al. (2025) and Sadewo et al. (2020) — are consistent with our small-dataset results and, viewed alongside our large-dataset failures, suggest that some published headline accuracies may be similarly contingent on favourable data conditions. The contribution here is not a higher number but a more honest account of when these models work and why, framed explicitly in molecular terms. Interpretable base learners and SHAP-style attribution remain important so that, once genomic features are added, their contributions can be examined and trusted by clinicians.

6.4 Towards genomic integration: a multi-omic fusion framework

The natural way to anchor the model earlier is to supply features that report on the molecular events directly rather than their downstream shadows. We propose extending the feature representation of Eq. (1) into a multi-omic vector that concatenates the present clinical phenotypes with germline, circulating-DNA and epigenetic features, as in Eq. (26).

$$\mathbf{x}^{\text{fused}} = [\mathbf{x}^{\text{clin}} \parallel \mathbf{x}^{\text{germ}} \parallel \mathbf{x}^{\text{ctDNA}} \parallel \mathbf{x}^{\text{methyl}}] \tag{26}$$

where the clinical block is that used in this study, the germline block encodes pathogenic status in BRCA2, PALB2, ATM, CDKN2A and related genes, the ctDNA block captures KRAS and TP53 mutant fragments, and the methylation block carries tumour-specific epigenetic signatures. Because these modalities differ in reliability and availability, a simple concatenation is suboptimal; an attention-weighted fusion that learns a per-modality importance weight, as in Eq. (27), allows the model to down-weight missing or noisy modalities per patient.

$$\mathbf{x}^{\text{fused}} = \sum_{m=1}^{M} a_m \mathbf{x}^{(m)}, a_m = \frac{\exp(e_m)}{\sum_j \exp(e_j)} \tag{27}$$

where a_m is a learned, softmax-normalised attention weight on modality m. Such a model would, in principle, detect risk at the KRAS/CDKN2A stage rather than the TP53/SMAD4 stage, shifting prediction into the window where intervention can still cure. The proposed roadmap is shown in Figure 12.

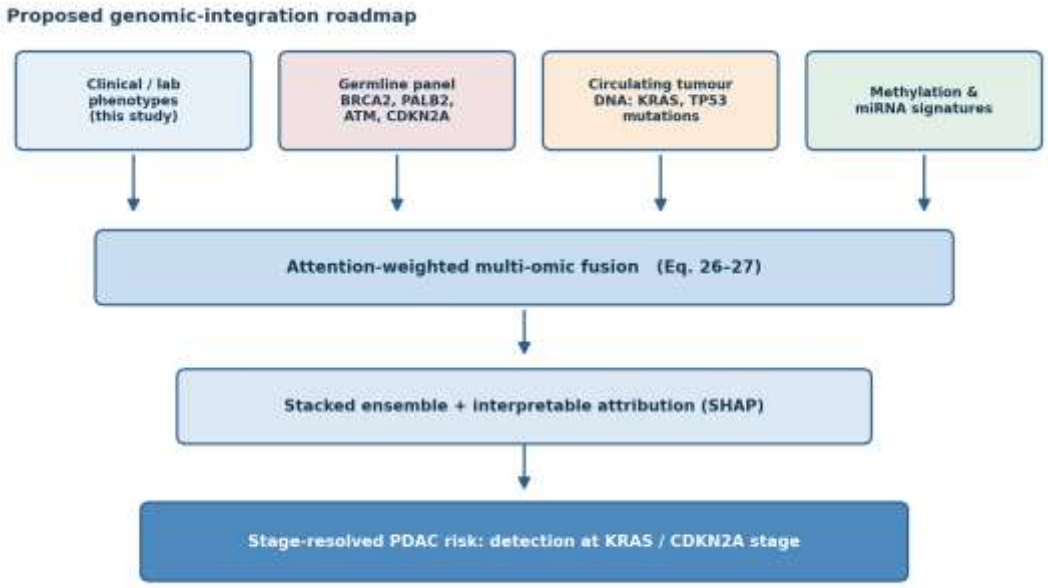

Figure 12. Proposed genomic-integration roadmap. Clinical phenotypes (this study) are fused with germline panels, circulating tumour DNA and methylation/miRNA signatures by attention-weighted fusion (Eq. 26–27), feeding a stacked ensemble with interpretable attribution to yield stage-resolved PDAC risk.

6.5 Clinical translation and genomic-privacy considerations

Translating such a tool into practice raises issues beyond accuracy. Germline information is especially sensitive – a model that includes information on BRCA2 or CDKN2A status has "family" implications and therefore requires informed consent and secure storage and incidentally find policies. In this context interpretability is a non-negotiable requirement and the architecture includes transparent base learners, and post-hoc attribution, to ensure the clinician and patient can understand why a risk score was generated. Validation and calibration to local prevalence would also be required for any deployment as failure modes described here demonstrate the poor performance a model can have if prevalence changes.

6.6 Limitations

Several limitations bound these conclusions. The experiments used structured, tabular datasets only; no imaging or sequencing data were processed, so the multi-omic integration of Section 6.4 is a designed pathway rather than a

demonstrated result, and Eq. (26)–(27) describe a proposed rather than evaluated model. The perfect-performance datasets are too small or too leak-prone to support claims of clinical utility, and the large datasets were not externally validated across sites. The genotype–phenotype mapping of Table 2, while grounded in the literature, is interpretive and would need direct molecular measurement to confirm in any given cohort. These constraints are stated plainly because the credibility of phenotype-based screening depends on resisting the temptation to over-read favourable numbers.

7. CONCLUSION

We set out to test whether the well-characterised genetics of pancreatic ductal adenocarcinoma could be harnessed for early risk prediction through its phenotypic surrogates, using a transparent stacked ensemble specified through a hierarchy of equations and grounded in a generative genotype–phenotype model. The model was extremely accurate on small and leaky data across four datasets, and only at chance rate on large, imbalanced data, suggesting that there are factors other than the algorithm itself that determine apparent accuracy in this area. This behaviour is a basic rule, i.e. clinical phenotypes are late echoes of an early genetic disease, as covered in the molecular progression model of PDAC. The way forward, though, is to merge the genotype directly — with the addition of the attention-weighted fusion proposed here, the new hybrid germline susceptibility panels, circulating tumour DNA, methylation and microRNA signatures all come together — so that prediction can be grounded on the primary molecular events, not their downstream consequences. The framework, which explicitly defines the relationship between genotype and phenotype and which explicitly considers possibilities for failure, provides an easily repeatable blueprint for the next step, genomics-integrated early detection, and a blueprint for moving the established genetics of pancreatic cancer to action.

Conflicts of Interest

The authors declare that they have no competing interests.

REFERENCES

1.Bailey, P., Chang, D. K., Nones, K., Johns, A. L., et al. (2016). Genomic analyses identify molecular subtypes of pancreatic cancer. Nature, 531, 47–52.

2.Bharathi, S. C., & Rayachoti, E. (2025). Multi-level image denoising integrated morphology-based image quality enhancement model with edge-based segmentation for accurate pancreatic cancer detection. Journal of Theoretical and Applied Information Technology, 103(8).

3.Bray, F., Laversanne, M., Sung, H., Ferlay, J., et al. (2024). Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: A Cancer Journal for Clinicians, 74, 229–263.

4.Collisson, E. A., Sadanandam, A., Olson, P., Gibb, W. J., et al. (2011). Subtypes of pancreatic ductal adenocarcinoma and their differing responses to therapy. Nature Medicine, 17, 500–503.

5.Gayathri, R. (2017). Genetic algorithm based model for early detection of cancer. Science and Technology, 3(8), 28–31.

6.Hruban, R. H., Goggins, M., Parsons, J., & Kern, S. E. (2000). Progression model for pancreatic cancer. Clinical Cancer Research, 6, 2969–2972.

7.Johnson, A., Sarawagi, R., Malik, R., Sharma, J., & Bhagat, A. (2024). Utility of diffusion-weighted imaging in differentiating benign and malignant breast lesions. South African Journal of Radiology, 28(1), 2952.

8.Kawai, M., Fukuda, A., Otomo, R., Obata, S., et al. (2024). Early detection of pancreatic cancer by comprehensive serum miRNA sequencing with automated machine learning. British Journal of Cancer, 131, 1158–1168.

9.Kimura, H., Klein, A. P., Hruban, R. H., & Roberts, N. J. (2021). The role of inherited pathogenic CDKN2A variants in susceptibility to pancreatic cancer. Pancreas, 50, 1123–1130.

10.Klein, A. P. (2021). Pancreatic cancer epidemiology: Understanding the role of lifestyle and inherited risk factors. Nature Reviews Gastroenterology & Hepatology, 18, 493–502.

11.Li, D., Xie, K., Wolff, R., & Abbruzzese, J. L. (2004). Pancreatic cancer. The Lancet, 363, 1049–1057.

12.Liu, J., Gao, J., Du, Y., Li, Z., et al. (2012). Combination of plasma microRNAs with serum CA19-9 for early detection of pancreatic cancer. International Journal of Cancer, 131, 683–691.

13.Ozsahin, D. U., Usanase, N., & Ozsahin, I. (2025). Advancing pancreatic cancer management: The role of artificial intelligence in diagnosis and therapy. Beni-Suef University Journal of Basic and Applied Sciences, 14(1), 1–18.

14.Sadewo, W., Rustam, Z., Hamidah, H., & Chusmarsyah, A. R. (2020). Pancreatic cancer early detection using twin support vector machine based on kernel. Symmetry, 12(4), 667.

15.Siegel, R. L., Miller, K. D., Wagle, N. S., & Jemal, A. (2023). Cancer statistics, 2023. CA: A Cancer Journal for Clinicians, 73(1), 17–48.

16.Singh, S., Singh, S., Singh, P., Pathak, A., et al. (2025). PancreaScan: Pancreatic cancer detection. International Journal of Advanced Electrical and Computer Engineering, 14(1), 85–96.

17.Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., et al. (2021). Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: A Cancer Journal for Clinicians, 71(3), 209–249.
18.Thangam, S., & Chakkaravarthy, S. S. (2024). An edge-enabled virtual honeypot based intrusion detection system for Vehicle-to-Everything (V2X) security using machine learning. IAENG International Journal of Computer Science, 51(9).
19.Waddell, N., Pajic, M., Patch, A. M., Chang, D. K., et al. (2015). Whole genomes redefine the mutational landscape of pancreatic cancer. Nature, 518, 495–501.
20.Wang, Q., Wang, Z., Liu, F., Wang, Z., et al. (2025). Machine learning-based prediction of postoperative pancreatic fistula after laparoscopic pancreaticoduodenectomy. BMC Surgery, 25(1), 191.