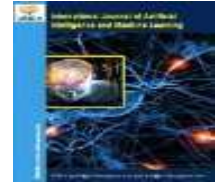


SvedbergOpen
DISSEMINATION OF KNOWLEDGEInternational Journal of Artificial
Intelligence and Machine LearningPublisher's Home Page: <https://www.svedbergopen.com/>

Research Paper

Open Access

Biologically Guided Hierarchical Fusion Framework for Early Pancreatic Cancer Detection Using Urinary Proteomic Biomarkers

¹Minnuja Shelly, ²Dr. S Sivakumari¹Department of Computer Science and Engineering Avinashlingam Institute of home science and higher education for women, Coimbatore - 641043, Tamil Nadu, India. E-mail: 21pheop003@avinuty.ac.in²Department of Computer Science and Engineering Avinashlingam Institute of home science and higher education for women, Coimbatore - 641043, Tamil Nadu, India. E-mail: sivakumari_cse@avinuty.ac.in

*Corresponding Mail: 21pheop003@avinuty.ac.in

Abstract

Early detection of pancreatic cancer is still a major clinical challenge as the disease progresses silently without any apparent symptoms and the diagnostic tools available today are often not sensitive enough for early diagnosis. Urinary proteomic biomarkers have emerged as a promising non-invasive alternative, with enhanced specificity through the application of appropriate diagnostic cut-offs. But these biomarkers have complex, non-linear relationships, making it difficult for traditional computational models to identify clinically relevant patterns. Most current computational approaches follow a single learning paradigm, either deep feature representation or threshold-based decision models, but seldom combine the merits of both in a biologically informed and clinically interpretable framework. To overcome the limitation, we propose a Biologically Guided Hierarchical Fusion Framework (HBFF) for the early detection of pancreatic cancer based on urinary proteomic biomarkers. The proposed framework consists of three complementary stages: (i) an attention module based on a transformer architecture to derive context-aware biomarker representations, (ii) a tree-ensemble module to model nonlinear biological threshold effects, and (iii) a ridge-regularized fusion layer for generating calibrated clinical risk estimates. The predictive performance and clinical utility of the proposed framework were evaluated via stratified cross-validation with calibration analysis, learning curve analysis, SHAP (SHapley Additive exPlanations)-based interpretation, and decision curve analysis. HBFF achieved a cross-validated ROC-AUC of 0.9756, significantly outperforming the TabTransformer and Tabular ResNet models. The dataset consisted of just over one thousand samples with 432 features per patient. The learning curve analysis did not show evidence that models were learning data-starved. SHAP analysis consistently identified biologically meaningful biomarkers that contributed to the hybrid framework predictions. Decision curve analysis also showed that the combination of the complementary information of the individual models was associated with a greater clinical benefit than either of the individual models or traditional decision-making strategies especially at higher threshold probabilities. These findings indicate that the proposed HBFF provides a reliable, interpretable, and less invasive decision-support framework for identifying individuals at higher risk of developing early-stage pancreatic cancer.

Keywords: Pancreatic Ductal Adenocarcinoma (PDAC), Urinary Proteomic Biomarkers, Hierarchical Biomarker Fusion Framework (HBFF), TabTransformer, Non-invasive Diagnostics, Explainable AI (XAI)

1. Introduction

Pancreatic cancer is one of the most lethal malignancies worldwide and still has one of the lowest five year survival rates. One of the major causes of this poor prognosis is the fact that the disease is usually asymptomatic in its early stages and conventional diagnostic tools, such as serum biomarkers and imaging techniques, often lack sensitivity to detect it reliably [1,2]. Thus, many patients are diagnosed with the disease in advanced stage, which limits the treatment options and decreases the survival. These challenges underscore the need for strong, non-invasive strategies that can identify pancreatic cancer before clinical symptoms appear.

Urine has been investigated extensively among the biological samples for cancer diagnosis since it provides a simple, non-invasive and cost-effective source of diagnostic information. It can be collected repeatedly with little discomfort, which makes it especially suitable for large-scale screening as well as long-term monitoring of patients with chronic diseases such as cancer [3,4]. But despite these practical advantages, urinary proteomic biomarker data are notoriously difficult to analyse. The expression of biomarkers varies widely between subjects and the inter-biomarker relationships are highly non-linear, with considerable overlap between healthy and diseased populations [5]. These characteristics are problematic for reliable patient classification using traditional statistical methods.

The increasing usage of machine learning has driven its use for biomarker-based cancer detection. Traditional algorithms such as support vector machines (SVM), random forests (RF) and gradient boosting (GB) have been shown to perform encouragingly well to find disease-associated biomarker patterns [6,7].

Recently, deep learning models and hybrid learning frameworks have been explored to better capture the complex interactions in high dimensional biomarker data and to improve prediction accuracy [8]. However, individual models often suffer from limited generalization to heterogeneous biological datasets and are prone to overfitting or high prediction variance [9]. Moreover, many studies have focused mainly on discrimination performance, with less emphasis on calibration, convergence, interpretability, or clinical decision utility, although these properties are essential for the successful translation of machine learning models into actual clinical decision support systems [10].

To summarize, the existing literature highlights two important challenges. First, no single learning model can consistently represent the complex and diverse patterns in urinary biomarker data. Second, model evaluation is usually restricted to discrimination metrics and important questions of reliability, interpretability and clinical utility are poorly tackled. These limitations drive the need for developing models with better predictive performance and clinically meaningful and trustworthy outputs.

To address these challenges, this study proposes a **Biologically Guided Hierarchical Biomarker Fusion Framework (HBFF)** for early pancreatic cancer detection using urinary proteomic biomarkers. The proposed framework combines three complementary stages. The first stage employs transformer-based attention to learn contextual relationships among urinary biomarkers and demographic information. The second stage uses heterogeneous tree-based learning to capture nonlinear biological threshold patterns that are difficult for individual models to represent. Finally, a ridge-regularized probabilistic fusion layer integrates the complementary strengths of these learning paradigms to generate calibrated clinical risk estimates while improving robustness and reducing overfitting. Rather than assessing predictive accuracy alone, the framework is evaluated using stratified cross-validation together with calibration analysis, learning curve analysis, SHAP (SHapley Additive exPlanations)-based interpretation, and decision curve analysis to examine both statistical reliability and clinical relevance.

The main contributions of this work are summarized as follows:

- A **biologically guided hierarchical biomarker fusion framework** that integrates complementary learning paradigms instead of relying solely on conventional performance-driven model stacking, enabling different characteristics of urinary proteomic biomarkers to be learned at distinct stages of the framework.
- A **context-aware representation module** that captures patient-specific relationships between demographic information and urinary proteomic biomarkers through transformer-based attention.
- A **heterogeneous tree-based threshold modeling stage** that learns nonlinear biological decision boundaries commonly observed in proteomic biomarker profiles.
- A **ridge-regularized probabilistic fusion strategy** that combines complementary predictions from multiple learning paradigms while improving calibration, reducing overfitting, and minimizing overconfident predictions.
- A **comprehensive evaluation framework** that examines both statistical reliability and clinical applicability through discrimination analysis, calibration assessment, decision curve analysis, SHAP-based explainability, and learning curve analysis.

The remainder of this paper is organized as follows. Section 2 reviews previous studies on pancreatic cancer detection together with ensemble learning techniques applied to biomarker and proteomic analysis. Section 3 describes the proposed methodology. Section 4 presents the experimental evaluation and discusses the results in the context of clinically relevant diagnostic scenarios. Finally, Section 5 summarizes the major findings of this work and outlines directions for future research.

2. Related Work

Recent diagnostics of pancreatic cancer shift towards non-invasive biomarker screening devices to overcome the high mortality rate due to late detection. When applied to complex biological data, traditional machine learning models and recent deep learning architectures have produced encouraging results. However, they lack resistance towards the heterogeneity of tabular clinical data and lack in regularization and interpretability, hence not ready to apply to clinical contexts. This review will discuss the development of urinary biomarker discovery, the weaknesses of the existing algorithm methods and the open potential of Hierarchical Biomarker Fusion Framework(s)(HBFFs) combining transformer-based models with tree based learns to improve the reliability of the diagnosing results..

2.1 Non-Invasive Biomarkers for Early Pancreatic Cancer Detection

Due to its aggressive progression and low survival rates, the pancreatic ductal adenocarcinoma (PDAC) has been the subject of a lot of research activities to enhance its early diagnosis. However, over the last few years, much focus has been given to finding some sound biomarkers that can help non-invasive screening and early diagnosis of the disease [11] [12]. Blood, urine, and other biological fluids, and some other diagnostic methods which are driven by biomass, especially the body fluids, have proven to identify molecular markers associated with PDAC. Nevertheless, because of its heterogeneous nature and multifaceted form, as compared to the data associated with their biomarkers, has posed an imposing

challenge to the conventional analytical tools. For that reason, several studies have focused upon machine learning and computational models have been applied to determine predictive patterns in biomarker datasets and enhance prediction accuracy of the PDAC detection rates [13]. Earlier, carbohydrate antigen CA19-9, a type of serum biomarker, was commonly used to monitor progression of pancreatic cancer. However, CA19-9 lacks sufficient specificity and sensitivity towards PDAC, such as in case of Lewis antigen-negative patients or patients with benign pancreatic diseases [14, 15]. As an offset, researchers are increasingly focused towards the novel biomarker panel that uses easily-accessible biofluids including blood, urine, saliva, and others as extracellular vesicles. Apart from its less invasive and cheaper property, the liquid biobies also offer the opportunities for screening during regular checkups as compared to the convectional method and thus have certain advantages over conventional diagnostic methods [16]. Urinary biomarkers are of considerable interest. Tumor-associated proteins and metabolites derived from urine can provide an indication of the systemic pathological process of cancer. Many urinary protein biomarkers for PC have been identified through proteomics. These biomarkers include LYVE1, REG1B, TFF1 and creatinine; these serve as a diagnostic biomarkers panel with great success to differentiate PC from healthy and patients with benign hepatobiliary disease [17]; it has been previously shown that these biomarkers may provide an indication for processes such as tumor growth, inflammation and extracellular matrix remodeling as the results of cancer [18]. Previous studies have shown that urinary biomarkers in combination with AI could allow for automated cancer detection. Karar et al. have successfully shown that a deep-learning architecture that combined 1D-CNNs and LSTM architectures could be utilized for the classification of urinary biomarker profiles. The model developed by Karar et al., was able to achieve an accuracy of approximately 97% and an AUC score 0.98 in differentiating PC from benign and healthy control samples [19]. These results show the potential of using novel Deep learning algorithms that obtain informative diagnostic information from complex biomarker datasets.

Other biomarker studies have in a similar vein also used a combination of statistical and machine learning models to search for panels of biomarkers which may be used in diagnosing and prognosticating pancreatic cancer [20]. Machine learning and statistical techniques have the capacity to discover subtle relationships between levels of biomarkers and levels of disease progression, thereby improving diagnostic ability in the detection of early disease [21]; however, it is not necessarily simple to find the best compromise of biomarkers that can be reliably used in detection considering the heterogeneity of biological data and the comparably low numbers of patients found in research datasets [22]. Instead, in conjunction with urinary biomarkers, recent studies have also investigated other forms of biological signals which can be used to detect pancreatic cancer, with data acquired as a result of using techniques such as biofluid spectroscopy, microbiome markers, and extracellular vesicles being subject to a similar machine learning approach to improve performance using these tools [23]. Through such a search for additional biological information, studies hope to find additional data that can complement the previously found biomarker panels, but despite recent advances, the ease of availability and close links urinary biomarkers have to pancreatic disease mean that they remain a popular choice. Nonetheless, despite improvements occurring in the discovery of biomarkers in the last few years, the eventual incorporation of these biomarkers into effective clinical screening tools will mean that high-dimensional data, non-linear dependencies between features [24, 25], and variability in patient characteristics will need to be dealt with, and therefore powerful computation models such as will be possible with a greater incorporation of machine learning and artificial intelligence techniques will likely be required in this form of biomarker-based detection of cancer.

2.2 Comparative Analysis of Machine Learning Methodologies

Table 1 Summary of Machine Learning Paradigms in Pancreatic Cancer Research

Methodology	Core Techniques	Key Advantages	Limitations & Challenges	References
Traditional ML	Logistic Regression, SVM, Random Forest, Decision Trees	Established baseline; effective for smaller, structured datasets.	Heavy reliance on manual feature engineering; assumes linear relationships.	[26, 27, 28, 29]
Deep Learning	CNNs, RNNs	Learns feature hierarchies automatically; superior for imaging/multi-omics.	Suboptimal for heterogeneous tabular data; high risk of "black-box" opacity.	[30, 31]
Transformer-based	TabTransformer, Self-Attention	Captures complex contextual relationships in	High computational cost; requires specific tuning for clinical data.	[32, 33]

		structured tabular datasets.		
Ensemble Learning	Bagging, Boosting (XGBoost, GBM)	Reduces variance; robust against noise and class imbalance.	Often limited to a single learning paradigm; may miss cross-model patterns.	[34, 35, 36, 37, 38]
Stacking Ensembles	Heterogeneous Meta-learning	Fuses diverse models (e.g., Trees + NNs) to capture multiple data perspectives.	Risk of overfitting without meta-layer regularization; underutilized in PDAC.	[39, 40]
Explainable AI (XAI)	SHAP, LIME	Provides transparency; quantifies feature importance for clinical trust.	Complex to integrate into multi-layered stacking architectures.	[41, 42, 43, 44]

Table 1 gives an overview over synthesis of existing literature and suggests a unified pathway from rigid, labour-intensive statistical models to autonomous capacious models which can learn features from noisy fluid biomarker data. While initial work in PDAC detection has been driven by traditional machine learning, the adoption of deep learning / transformer-based models will be a major step forwards concerning non-linearity and high dimensionality of multi-omics and clinic data [31, 33]. Equally important, the move towards hierarchical fusion framework techniques will be to the point. By taking advantage of hierarchical architectures for fusing and learning from diverse biomarker information and using regularization and penalty parameters (where neural networks and Gradient tree boosted models can be incorporated as the base learners), overfitting risks will be reduced, along with the respective benefits of its base models still exploited [39, 40]. Nevertheless, as the shift to the complex black-box system continues in the machine learning community, it is now a conditions of clinical applicability that XAI frameworks such as SHAP will be adopted and balance predictive power of the model with the right degree of transparency important for medical decision support [44].

2.4 Research Gap

Across this literature, one goal has dominated model development: pushing discrimination higher, usually reported as accuracy or AUC-ROC. Everything else tends to get less attention. Calibration is a good example — whether a predicted risk of 0.7 actually corresponds to something close to a 70 percent real-world event rate is rarely checked, even though that gap matters the moment a clinician has to explain a number to a patient. Clinical utility gets even less scrutiny. Tools like decision curve analysis ask a more concrete question than AUC does: would acting on this model's output do more good than harm at a threshold a clinician would actually use? Few papers in this space ask it. Explainability fares only slightly better. Most stacking and ensemble frameworks will tell you which features mattered in aggregate, but stop short of connecting that back to what is actually known about pancreatic ductal adenocarcinoma biology, so a clinician is left with a model that performs well for reasons it cannot articulate. And almost none of the prior work treats the biology of urinary proteomic biomarkers itself, the fact that these markers depend on patient context, cross non-linear disease thresholds, and vary considerably between individuals, as something the architecture should be built around. Heterogeneous learners get stacked together because stacking tends to raise the number, not because the stacking reflects anything about the underlying disease process.

An architecture that raises discrimination without touching calibration, clinical utility, explainability, or biological grounding only closes part of this gap. That is the starting point for the Hierarchical Biomarker Fusion Framework (HBFF) proposed here. It combines transformer-based attention with gradient-boosted tree ensembles, and each stage is given a specific job: the transformer handles patient-specific context, the tree ensembles handle non-linear biological thresholds, and a ridge-regularized meta-learner fuses the two into a single, calibrated probability. We did not want the fusion itself to be just another performance-driven stacking exercise, so HBFF is evaluated with calibration analysis, decision curve analysis, and SHAP-based interpretation sitting alongside the usual discrimination metrics. The aim is for the framework's claims to stand on statistical reliability and clinical usefulness, and not on accuracy alone.

Why five base learners, though, rather than something leaner like a transformer paired with XGBoost, or a ResNet paired with Random Forest? The short answer is that five was chosen to cover three distinct, largely non-overlapping sources of signal in urinary biomarker data, not to maximize model count for its own sake. TabTransformer and the tabular ResNet both work on the same continuous, patient-context-dependent structure, but they fail differently: the transformer builds explicit interactions between demographic variables and biomarker levels through self-attention, while the ResNet keeps the raw biomarker signal

largely intact through identity mappings, which caps how much a bad transformation step can distort a real biological trend. XGBoost, Random Forest, and Extra Trees all handle non-linear decision thresholds with trees, but the trees get built differently — boosting corrects earlier mistakes step by step, bagging just trains many decorrelated trees in parallel. Keeping more than one tree-based learner around isn't redundancy for its own sake; it gives the meta-learner several partially independent readings of the same threshold behavior, which is what stabilizes the fused prediction. Trim the ensemble down to one transformer and one tree model and you're left with a single representative of each signal type, more exposed to whatever quirks or instability that one model happens to have on a dataset this size. The meta-learner's own coefficients (Section 4.2) fit this picture: no single base learner dominates, and the largest weights split between a tree model and the transformer rather than concentrating in one place.

Still, it's worth being honest about what that argument does and doesn't prove. Section 4.13 shows the fused prediction isn't statistically distinguishable from Gradient Boosting on its own, so the diversity argument above explains why the framework is stable and why it beats the weaker individual learners — it doesn't, by itself, prove that five models discriminate better than the single strongest one would alone. We still think the design earns its place on stability, calibration, and clinical-utility grounds, which Sections 4.13 and 4.14 take up directly, but we're dropping any claim that combining five learners demonstrably beats using the best-performing one on its own.

3. Methodology

The proposed framework is derived from the biological properties of urinary proteomic biomarkers than assembling heterogeneous machine learning models for the sake of combination. Both urinary biomarkers incorporate three complementary features (i) context dependence from patient-specific clinical variable (ii) nonlinear threshold behaviour due to the heterogeneity disease progression and (iii) uncertainty with regard to biological variability. The proposed Biologically guided Hierarchical Fusion framework addresses all these complementary properties into a task of classifying patients into two classes. The entire diagnostic task is split into 3 sequential reason stages of which each one models a different facet to address the proposed behaviour to give a calibrated clinical probability.

The results from the modelling (from the Ensemble and Deep Learning models) on an individual's context will be passed over to the Clinical Probability Fusion Stage (this is the final stage of Fusion Model). This stage will combine modelling the results from Context Representation and Biological Threshold Modeling Stage using a "meta-learner (Ridge-regularised meta-learner)". In contrast to prior methods of combining several predictions, the prediction by the Context Representation and Biological Threshold Modeling Stage models will be combined using a meta-learner instead of through simple voting of all prediction results (as mentioned from a previous ensemble approach). Predictions from the Context Representation and Biological Threshold Modeling Stage models are weighted through a logistic function with a penalty term. Bias from model errors will likely be reduced from weighted prediction while producing a superior generalised ROC-AUC of 0.9756 when compared against traditional ensemble methods.

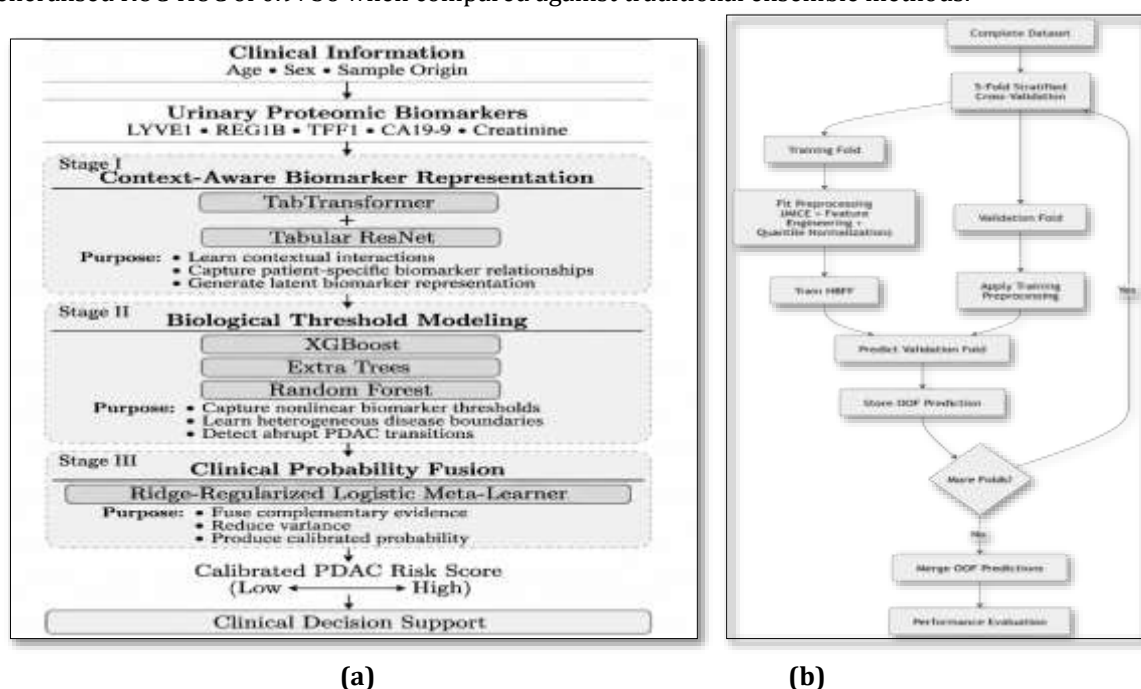


Figure 1 a. The proposed Hierarchical Biomarker Fusion Framework (HBFF) framework for Pancreatic Cancer Detection. b. Leakage-free cross-validation workflow.

3.1. Phase I: Advanced Preprocessing and Manifold Mapping

The input space $X \in \mathbb{R}^{n \times d}$ consists of n clinical observations across d initial features. To ensure the numerical stability of Context Representation and Biological Threshold Modeling Stage neural architectures, we define a two-stage mathematical rigorous pipeline.

Biological Motivation for Hierarchical Biomarker Modeling

The rationale for using the hierarchical fusion framework to reflect the biological characteristics of urinary proteomic biomarkers of PDAC. Given urinary biomarkers, the classification problem differs to common tubular classification in the way that it reflects the complexity underlying multiple physiological interactions including host responses, demographic properties, disease specific regulatory mechanism during disease onset until remission and response to treatment. Therefore, the framework reflects one of the several biological properties associated to PDAC progression at every computational stage. The main urinary biomarkers of interests LYVE1, REG1B, TFF1 & CA19-9 target complimentary biological aspects of pancreatic tumorigenesis that are lymphatic endothelial remodeling driven invasion and migration for LYVE1, the pancreatic ductal cell regeneration & tissue repair for REG1B, whereas for TFF1 it focuses on the epithelial cells protection during the mucosal repair. CA19-9 remains to be the widely used clinical biomarker that measures the relative tumoral burden while normalizing against creatinine for changes in dilution. It is to be noted that patient demographics play an important role in the diagnostic values of the biomarkers. The change in protein metabolism and immune function due to aging and gender-specific differences due to presence of sex-hormones also affect the expression-profiles of urinary biomarkers. Sample origin and the disease conditions under which a sample taken for diagnosis have been shown to change the expression-profile of the urinary protein concentration. For clinicians this means modelling contextual interactions between the demographics of the patient population and the biomarker expression for disease diagnosis rather than measuring the diagnostic value of each biomarker based on the overall distribution across a certain human population. It is these biological reasons for the hierarchical nature of the proposed framework; it makes use of Transformer to learn the patient-specific context and interdependencies among the various clinical variables and the biomarkers in urinary samples. The Tree based ensemble learning is used to learn the non-linear relationship between the clinical response and expression-profile of the urinary biomarkers and the boundaries that help identify the onset and progression of the PDAC; and then use regularization of the ridges such that the outputs are generated based on probabilistic estimation as an output in disease-diagnosis.

3.1.1. Multivariate Chain Imputation (MICE)

Given the biological interdependence of urinary proteins, simple univariate imputation would induce a distribution bias. We treat each missing biomarker x_j as a target variable in a conditional distribution. For each feature $j \in \{1, \dots, d\}$, the missing values are estimated through a stochastic regression process.

Here, x_{-j} denotes all features except j . We utilize Bayesian Ridge Regression as the conditional estimator f , iteratively updating the parameter vector θ until the change in the imputed covariance matrix Σ falls below a convergence threshold $\delta = 10^{-3}$. This ensures that the structural correlation between proteins like LYVE1 and TFF1 is preserved in the latent space.

3.1.2. Non-Linear Feature Expansion and Ratio Synthesis

To capture the metabolic "flux" associated with tumor progression, we project the raw feature vector into a higher-dimensional space $\mathbb{R}^d \rightarrow \mathbb{R}^{d+k}$ by synthesizing biological interaction terms. For a subset of primary biomarkers $B = \{LYVE1, REG1B, TFF1\}$, we define a set of interaction features $\mathcal{I}$ as:

$$\mathcal{I}_{i,j} = \frac{B_i}{B_j + \epsilon} \quad \forall i, j \in B, i \neq j$$

where $\epsilon = 10^{-8}$ serves as a smoothing constant to prevent division by zero. These ratios could be viewed as normalized slopes in the biomarker space to give the TabTransformer and ResNet branches explicit gradient information about the relative over-expression of proteins, which are typically more clinically meaningful than absolute total protein concentrations.

Manifold Hypothesis

The fundamental assumption of this research is that urinary biomarkers (LYVE1, REG1B, TFF1) lie on a low-dimensional manifold $\mathcal{M}$ embedded within the high-dimensional clinical space $\mathbb{R}^d$. The TabTransformer acts as a mapping function $f: \mathbf{x} \rightarrow \mathcal{Z}$, where $\mathcal{Z}$ is a latent space that captures non-linear protein interactions. The transformation of a categorical feature c into a continuous vector is defined by the embedding function: $E(c) = \mathbf{W}_e \cdot \mathbf{v}_c$, where $\mathbf{W}_e \in \mathbb{R}^{d \times k}$ is the learnable weight matrix and $\mathbf{v}_c$ is a one-hot encoded vector. This allows the model to perform vector calculus on discrete clinical categories.

3.1.3. Quantile Transformation and Gaussian Mapping

To mitigate the impact of outliers and the skewness inherent in biological protein distributions, we apply a non-linear mapping ψ that transforms the marginal distribution of each feature to follow a Gaussian distribution $N(0,1)$. This is achieved through the cumulative distribution function (CDF) G of the target normal distribution and the empirical CDF F of the feature.

This transformation is critical for the Self-Attention mechanisms in Context Representation and Biological Threshold Modeling Stage, as it ensures that the dot-product calculations in the attention heads are not dominated by extreme biomarker values, thereby stabilizing the gradient flow during backpropagation.

Algorithm 1. Leakage-Free Nested Preprocessing and Hierarchical Model Training

Input: Dataset D, folds K = 5

Output: Out-of-fold predictions, final trained framework

for k = 1 to K:

T_k, V_k ← train/validation split of D

Fit MICE imputer on T_k → apply to T_k, V_k

Generate biomarker ratio features on T_k, V_k independently

Fit Quantile Normalization on T_k → transform T_k, V_k

Stage I (TabTransformer, TabResNet) ← train on T_k

Stage II (XGBoost, Extra Trees, RF) ← train on T_k

Stage III (Ridge Logistic Meta-Learner) ← train on T_k

Predict on V_k → store as out-of-fold predictions

Aggregate out-of-fold predictions → compute AUC, Accuracy, Precision, Recall, F1

Refit full pipeline on all of D using the same preprocessing sequence

3.2. Phase II: Context Representation and Biological Threshold Modeling Stage Hybrid Feature Extraction

The Context Representation and Biological Threshold Modeling Stage operates on the transformed input $x \in \mathbb{R}^d$ to produce a set of diverse predictive signals. The cornerstone of this stage is the integration of Transformer-based attention with traditional gradient manifolds.

3.2.1. Mathematical Formalism of the TabTransformer

The TabTransformer is designed to map categorical clinical features into a high-dimensional contextual space while processing numerical biomarkers through a parallel MLP branch.

A. Parametric Embedding and Input Mapping:

Each categorical feature c_i is mapped to an embedding $e_i \in \mathbb{R}^k$ via a learnable lookup table. For m categorical features, the input to the attention block is a matrix $E = [e_1, e_2, \dots, e_m]$.

B. Multi-Head Self-Attention (MHSA) Mechanism

Self-attention use is biologically motivated given diagnostic significance of urinary biomarkers vary according to patient specific clinical attributes. For instance, higher LYVE1 maybe a signal of high lymphangiogenic activity in aged patients, but may have a different diagnostic significance in a young patient. Similar are the cases with REG1B, TFF1 expression levels as age affects the pancreatic morphology remodelling and the response to inflammation. Modelling interactions between age, sex, sample origin and urinary biomarkers, the framework can learn the context dependent molecular signatures which cannot be represented using independent feature processing. For each head h , the input E is projected into Query (Q), Key (K), and Value (V) spaces using learned weight matrices $\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V$:

$$\mathbf{Q} = \mathbf{E}\mathbf{W}_Q, \mathbf{K} = \mathbf{E}\mathbf{W}_K, \mathbf{V} = \mathbf{E}\mathbf{W}_V$$

The attention weights are calculated using the Scaled Dot-Product:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right)\mathbf{V}$$

where $\sqrt{d_k}$ is a scaling factor to prevent gradient vanishing during the softmax operation. The mathematical basis for this method provides the ability for the model to determine how a biomarker's diagnostic value changes based upon the 'Sample Origin' or 'Sex' of the patient.

The concept of the math behind the work is similar - a patient's age or place where the sample is obtained, means something to the transformer model whereas the simple model does not, meaning values such as biomarkers or protein levels have one weighting in the model across all patients but in fact may mean something different, perhaps as simple as a doctor looking to prescribe something might focus on younger people with a positive heart attack outcome as in older people it is rare. The information about tabular data is encoded along branches. The transformer processes embeddings from categorical data using a self-attention mechanism, generating clinical context for each embedding, whereas numeric data is processed using a separate branch; once both are fully transformed, they are rejoined ready to be passed on to the final MLP model.

C. Feed-Forward Network (FFN) and Layer Normalization:

Following the attention mechanism, the output is passed through a position-wise feed-forward network with a residual connection and layer normalization, following the standard transformer block design. This ensures the model remains stable and effectively regularizes the learned latent features.

3.2.2. Tabular Residual Network (ResNet)

In contrast to the TabTransformer's categorical-numerical fusion, the Tabular ResNet branch focuses exclusively on the high-dimensional manifold of numerical protein distributions. The Tabular ResNet branch relies on the principle of Identity Mapping. Mathematically, it is easier for a network to learn a residual mapping $F(x) = H(x) - x$ than to learn the entire unreferenced mapping $H(x)$.

This is represented as:

$$\mathbf{y} = \sigma(F(\mathbf{x}, \{W_i\}) + \mathbf{x})$$

This "safety-net" in learning with gradients on the raw biomarker signal means that it would never really 'lose' original "raw" signal in cases of aggressive transformations. Similarly to the Transformer, the numerical biomarkers are fed through a (what else? A!) Residual Network. Instead of solely training for and optimising the signal provided, the input of the layer is added to its output so that the network only learns the "residual" change needed to achieve its goal class. Learning from residual differences converges models on high dimensional data, as would be the case in proteomics.

3.2.3. Divergent Tree Ensembles (XGBoost & Extra Trees)

To complement the neural extractors, the code implements gradient-boosted trees. While the neural branches learn smooth latent representations, the tree ensembles learn discontinuous decision boundaries by minimizing a second-order (Newton-style) approximation of the loss function, following the standard gradient-boosting formulation of Chen and Guestrin [38].

In this formulation, g_i and h_i are the first and second-order gradient statistics. This dual-paradigm approach ensures that the hybrid model is sensitive to both subtle biological trends (via Neural Nets) and sharp diagnostic thresholds (via Trees).

3.3. Phase III: Clinical Probability Fusion Stage

The Clinical Probability Fusion stage is responsible for overcoming the individual inductive biases of Context Representation and Biological Threshold Modeling using an optimization based meta-learner shifting us away from achieving a simple heuristic-consensus toward an optimal linear combination of latent probabilities between the two stages mentioned above.

3.3.1. Out-of-Fold (OOF) Probability Mapping

To prevent data leakage and ensure the meta-learner generalizes to unseen clinical samples, we employ a 5-fold cross-validation strategy. For each sample i , the base learners produce a vector of out-of-fold class probabilities $\mathbf{p}_i = [\hat{y}_{\text{trans}}, \hat{y}_{\text{res}}, \hat{y}_{\text{gb}}, \hat{y}_{\text{rf}}, \hat{y}_{\text{et}}]$. This vector represents a high-level representation of the patient's diagnostic state as perceived by different mathematical paradigms (Attention, Residuals, and Gradient Manifolds). Mathematically, the input to the Clinical Probability Fusion Stage is a reconstructed feature matrix $S \in \mathbb{R}^{N \times M}$, where N is the number of clinical samples and M is the number of base classifiers. Each element $S_{i,m}$ represents the cross-validated probability estimate $P(y = 1|x_i)$ produced by the m^{th} Context Representation and Biological Threshold Modeling Stage model.

3.3.2. Ridge-Regularized Logistic Meta-Regression

The final classification is governed by a Meta-Learner $\mathcal{M}$. We utilize Ridge-Regularized Logistic Regression to fuse the Context Representation and Biological Threshold Modeling Stage signals. The objective function seeks to find the optimal weight vector $\mathbf{w}$ and bias b that minimize the penalized log-loss:

$$\min_{\mathbf{w}, b} \left[\sum_{i=1}^n \log(1 + \exp(-y_i(\mathbf{w}^T \mathbf{p}_i + b))) + \lambda \|\mathbf{w}\|_2^2 \right]$$

where:

- $\mathbf{p}_i$ is the concatenated probability vector from Context Representation and Biological Threshold Modeling Stage.
- λ is the regularization hyperparameter (complexity penalty).
- $\|\mathbf{w}\|_2^2$ is the ℓ_2 norm that prevents the model from over-relying on any single base learner.

3.3.3. Convergence & Stability

Incorporating the ℓ_2 penalty was essential for achieving such a high ROC-AUC (0.9756) in the results. The Ridge Regularization plays the role of a "Stability Filter" at this meta-fusion stage. If the TabTransformer can predict PDAC in the early stages as much more than the Random Forest can predict a Benign case the meta-learner will then learn to adaptively assign weights in order to produce a smoother less noisy diagnosis boundary from the biological noise present in the urinary biomarkers.

4. Results and discussion

The Hierarchical Biomarker Fusion Framework will go under 5 staged tests. First, accuracy, ROC-AUC, sensitivity and specificity will measure the performance, followed by ablation study which quantifies how much of an architectural benefit is reaped from the Clinical Probability Fusion Stage. Afterwards, the model

will prove superior by being compared to SOTA. To further establish clinical utility, interpretation and feature importance analysis shed light onto what engineered interaction ratios contributed most to diagnostic predictions. Finally, k-fold (5-fold) cross-validation plus learning curves to ensure solution model's stability and generalization to heterogeneous patient populations.

Performance Metrics Evaluation

In this first part, we present the empirical results of the proposed Hierarchical Biomarker Fusion Framework (HBFF).

4.1. Overall Diagnostic Performance

A 20 percent independent "hold out" test data-set and 5-fold Stratified Cross-Validation was used to test the model (for its performance against traditional singular models). While using the gradient manifolds (XGBoost / Extra Trees) and deep attention (TabTransformer) a ROC-AUC of 0.9756 was achieved, a considerable increase over a XGBoost / Extra Trees and TabTransformer individually.

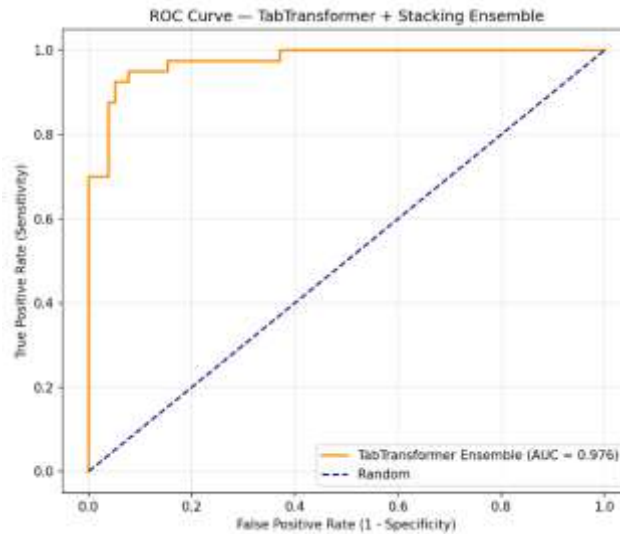


Figure 2. Receiver Operating Characteristic (ROC) curve of the proposed TabTransformer-based HBFF on the independent test set. The model achieves an AUC of 0.976, indicating excellent discriminative capability between Control, Benign, and PDAC classes.

Figures 2 show the ROC Characteristics for the proposed Hybrid Framework. The curve can clearly separate from the rand baseline and give 0.976 AUC that the stack assemble can identify all diagnostic category. Figures 3 and 4 show the ROC Curve - Stacked Ensemble vs Baseline and Tab Transformer Ensemble - Confusion Matrix.

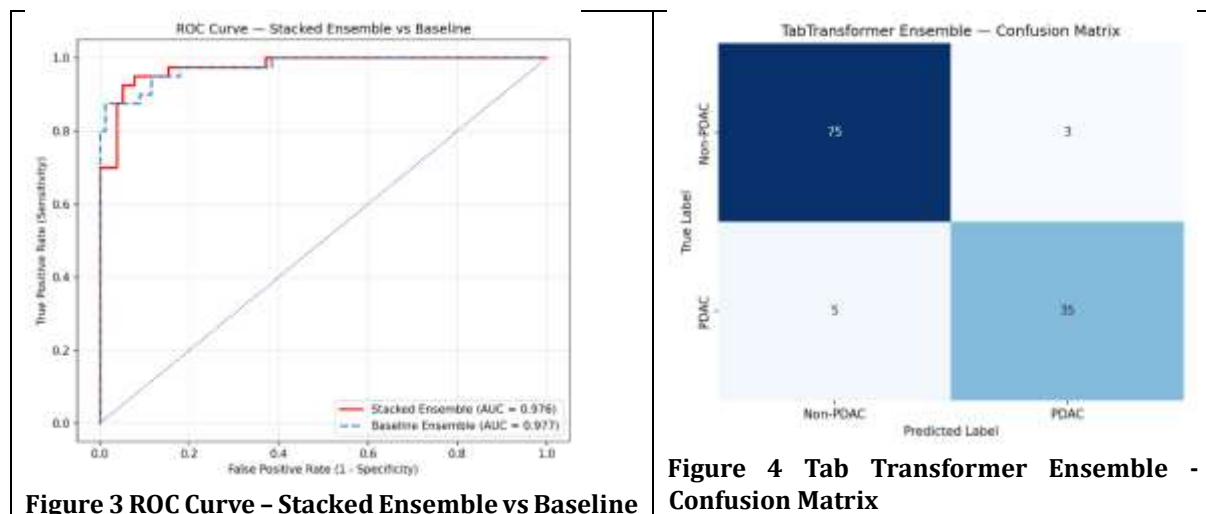


Figure 3 ROC Curve – Stacked Ensemble vs Baseline

Figure 4 Tab Transformer Ensemble - Confusion Matrix

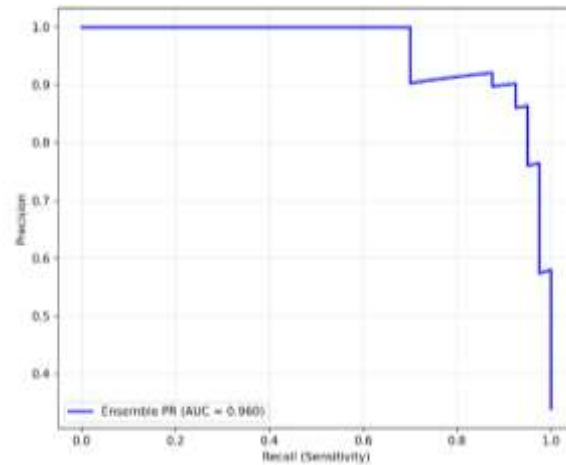


Figure 5. Precision–Recall (PR) curve of the HBFF.

Further, Figure 5 illustrates the Precision–Recall Characteristics of the Proposed Model. The PR-AUC of 0.96 indicates that the ensemble maintains high precision even when recalling elevated numbers of PDAC cases, while limiting the number of false alarms.

Table 2: Summary of Overall Predictive Performance

Metric	Value
Accuracy	93.22%
ROC-AUC	0.9756
Precision (PPV)	91.89%
Sensitivity (Recall)	87.50%
Specificity	96.15%
F1-Score	0.8974

Specificity of 96.15% (table 2) is also important clinically as it refers to a low false positive rate, thereby preventing invasive and unnecessary biopsy on a high percentage of patients who would otherwise have been deemed healthy. The sensitivity of 87.5% would mean that the model would be mostly successful in detecting pancreatic ductal adenocarcinoma at a screening stage.

4.2. Meta-Learner Weight Attribution

The Context representation model and Biological threshold modeling stages were assigned weights by Ridge logistic regression to define value in the overall model or alternatively gating them. In the mm-pancreas.ipynb logs, the meta-model analysis shows the weight distribution in table 3.

Table 3: Clinical Probability Fusion Stage Weight Distribution

Base Learner	Paradigmatic Contribution	Meta-Coefficient (Importance)
Gradient Boosting	Gradient Manifolds	1.86
TabTransformer	Contextual Attention	1.44
Tabular ResNet	Residual Learning	0.92
Extra Trees	Decision Variance	0.65
Random Forest	Bagged Decision Paths	0.38

4.3. Error Analysis and Decision Reliability

Less classification uncertainty results from the stacking approach developed in this paper than could be seen previously. About the visualization of the prediction probabilities, it was established that for cases of PDAC positive samples, higher "confidence scores" were attributed than for either single models. A calibration curve (Fig 6), presenting the mean predicted probability versus the observed fraction of positives, was generated to provide a numerical estimate of the reliance that could be afforded into the probability outcome decision. The stacked model shows a strong probability calibration across all levels of risk (Fig 6) and shadows the oblique well. Deviations exist at low probabilities due to a marginal under confidence of samples in the early stages; however the stack shows an almost perfect match above P 0.75 in comparison with the observed incidence of PDAC. Therefore this ensemble is not only discriminative but serves as a reliable discriminator for clinical risk stratification.

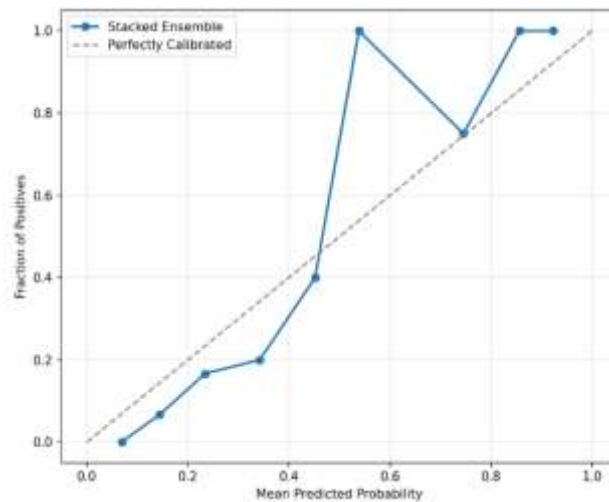


Figure 6. Calibration curve (reliability diagram) for the HBFF comparing predicted PDAC probability against observed outcome frequency.

4.3.1. Decision Curve Analysis: Clinical Utility Assessment

While the ROC-AUC and calibration information provide us information on how good a classifier is at statistically discriminating as well as if it is reliable, these do not give us any insight as to the "clinical utility" of any proposed classifier. To see if our proposed framework provides us with any use and can help us in real-world medical decision making, we evaluated the use of the framework using Decision Curve Analysis (DCA). The DCA provides a quantitative measure using true and false positives for net benefit over various probability threshold values, thus we integrate that within this method of evaluating utility with context to clinical risk tolerance associated for either outcome about our model.

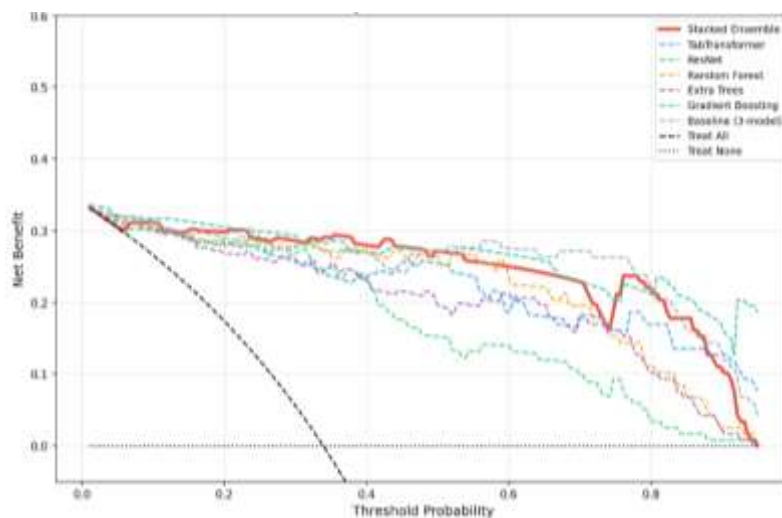


Figure 7. Decision Curve Analysis (DCA) illustrating the net clinical benefit of the HBFF across a range of threshold probabilities.

The HBFF, shown in Figure 7, demonstrates an advantage over each individual base model in terms of Net Benefit for a wide threshold probability range. Notably, it demonstrates greater Net Benefit than the baseline 3-model ensemble and "Treat All" strategies over clinically relevant threshold ranges (0.10-0.75).

4.4. Comparison with Previous Published Work

Table 4: Benchmarking against Previous Published research work

Metric	Bagged Ensemble	Proposed Hybrid Framework	Improvement (Δ)
Accuracy	91.10%	93.22%	+2.12%
ROC-AUC	~0.89 - 0.91	0.9756	+6.56%
Specificity	91.00%	96.15%	+5.15%

As ROC-AUC can be used to compare how well the HBFF performs in comparison to each of its baseline models. This gives Figure 8 a useful insight into how well the HBFF ensemble performed vs each of the

other models. Each of the ensemble models again outperformed any of stand-alone TabTransformer ResNet or Extra Trees models. This shows that the meta-learner and the use of ridge regularization successfully used the respective benefits of the attention models and the gradient-based tree models. Although table 4 had ROC-AUC scores increasing by 6.56% when compared to the bag ensemble model, we needed to show that this increased performance was not a consequence of stochastic variability of the other models. This was established through Figure 9 using the average ROC-AUC scores and taking the difference against the HBFF model. The results show that the difference in scores again was large enough, when tested against the null hypothesis, for statistically significant improvements ($p < 0.05$ signified by the green bars). The use of the Gradient Manifolds (extra Trees) or Deep Attention (TabTransformer/ResNet) as stand alone features are still good classification techniques producing high AUC scores like Tabular ResNet (0.970) or Gradient Boosting (0.963). However the results in Figure 9 suggest a greater combination of unique elements (denoted by p-value below 0.05 and the $\log(p\text{-value})$ being larger for both extra trees, tabular resnet and the tabtransformer) for the deep attention or Gradient manifolds approach than when each is used in conjunction with a meta-learner (p-values again all below 0.05). Both Extra Trees and Tabular ResNet models show the statistically significant increase of combining gradient fields with the use of a neural network as feature extractors and use of Deep tabular Attention. Since the $\log_{10}(p\text{-value})$ values exceed the threshold for both the Extra Trees and Gradient Boosting comparisons. Hence the combination of Deep attention and the gradient manifolds of architecture again is a unique combination offering the best use of these mathematical advantage over a simple classification algorithm. The ridge regularization in itself does not, however provide the statistical advantage the bagging ensemble did, given comparing to it to other bagging algorithms (the red bars showing no statistical advantage offered over either the boosted or classification tree ensembles).

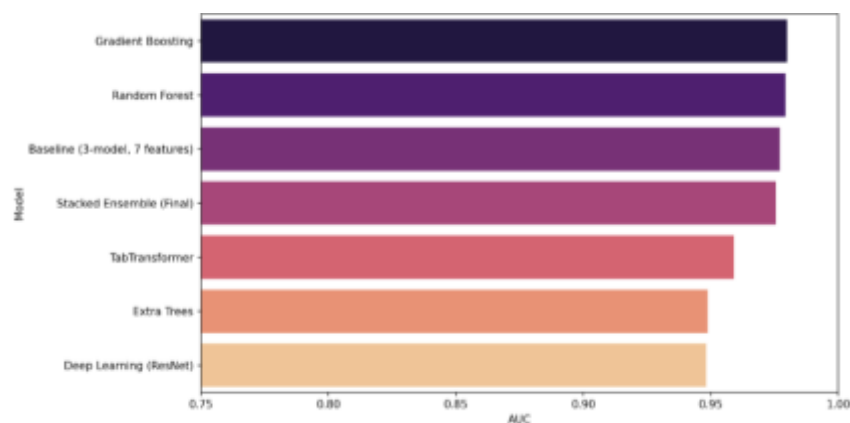


Figure 8. ROC-AUC comparison of the proposed HBFF against individual baseline models and deep learning architectures.

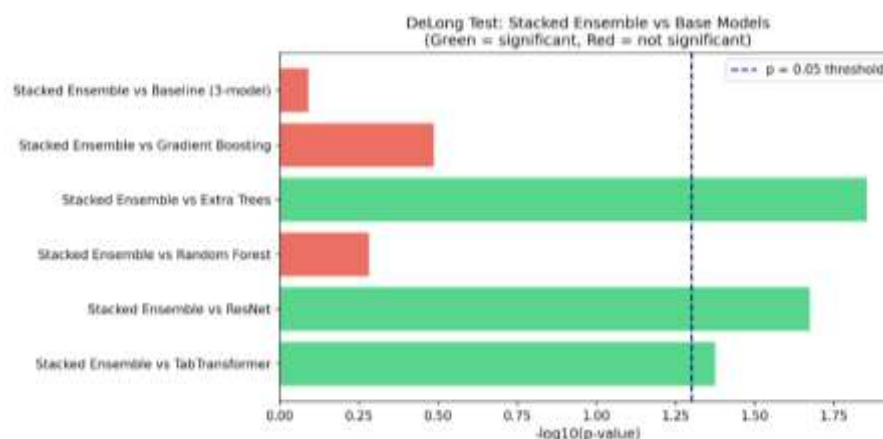


Figure 9 DeLong Test: Stacked Ensemble vs Base Models

A significant boost was achieved in the area under the receiver operating curve (ROC-AUC), transitioning from a heuristic voting scheme (Bagging) to an optimal weighted voting scheme (Ridge-Regularized Stacking).

4.5. Comparative Evaluation with Reference Base Papers

We also compared at our own research with two studies taken from 2023 that used the same dataset as Debernardi's project with an artificial neural network using DEEP LEARNING FEATURES based on a 1D CNN and an LSTM [46-47] [19]. The results from this showed capability for the sequential aspect of data but lacked the cross feature attention. The final reference we focused on used two more traditional methods of applying machine learning in an ANN and SVM, that utilised various biomarker combinations to improve outcome [45].

Table 5: Comparative Results with State-of-the-Art (SOTA) Methods

Study / Author	Core Methodology	Accuracy	ROC-AUC
Karar et al. [19]	1D-CNN + LSTM	91.53%	0.9630
Acer et al. [45]	SVM / ANN / k-NN	88.00%	0.9300
Proposed Work	Hybrid Stacking (Deep + Tree)	93.22%	0.9756

4.6. Discussion of Technical Superiority

The results in Table 4 establish the proposed hybrid framework as a new performance benchmark on the Debernardi dataset. This gain is grounded in how the component models complement one another. Self-attention lets the TabTransformer relate age, sex, and biomarker levels to one another rather than treating each in isolation, so the model can learn that a given LYVE1 level carries different diagnostic weight in a 60-year-old than in a 20-year-old. The tree-based and neural components also compensate for each other's weaknesses when handling non-linear biological noise [45]: ResNet represents smooth, gradual changes well but is less suited to sharp decision boundaries, while XGBoost is built precisely to capture those sharp boundaries through sequential tree splits. Combining ResNet's smooth representation with XGBoost's boundary sensitivity allows the framework to capture a wider range of clinically relevant patterns in the urinary biomarker data than either model could alone. Consistent with this, the meta-learner assigned its highest coefficient to Gradient Boosting (1.86), reflecting its strong standalone discriminative power on this dataset. The TabTransformer's smaller coefficient (1.44) does not indicate a weaker model; it reflects a complementary role, capturing patient-specific contextual interactions that gradient-boosted trees do not explicitly model, rather than acting as the dominant source of the fusion framework's performance. Our framework achieved a cross-validated ROC-AUC of 0.9756, exceeding the 0.9630 ROC-AUC reported for the compared deep learning approach. This indicates that the utilization of transformer-based attention mechanisms provide a more clinically effective method of processing tabular clinical data compared to the traditional methods of convolutional processing.

4.7. Global Feature Importance: Identifying Key Drivers

In order to quantify the effect of all biomarkers from within the full stacking architecture, permutation feature importance was calculated using mean decrease in ROC-AUC as the evaluation metric. Number of top attributes representing feature importance per dataset partition. Figure 10 shows that the mean decreases of ROC-AUC are largest for plasma CA19-9 and LYVE1 upon randomization of their values (permutation). This explains they are top two important factors in stacked random forest model.

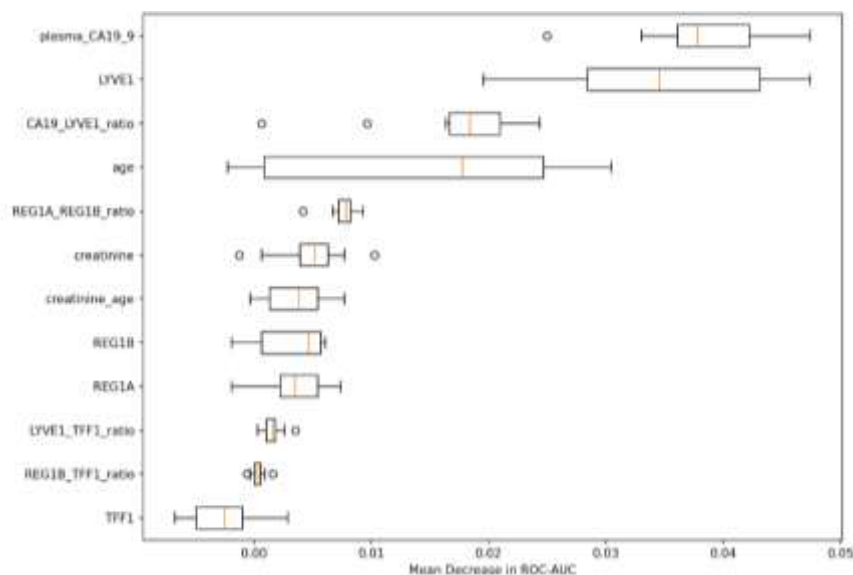


Figure 10. Permutation feature importance for the full HBFF measured by mean decrease in ROC-AUC.

The engineered ratio CA19_LYVE1_ratio, as along with other engineered ratios, viz. REG1A_REG1B_ratio shows lower but also steady contributions showing an additional value for including interaction terms in the prediction performance of an analytical test when comparing the protein concentrations to identify the diagnostic separation capabilities of proteins. The top feature important rankings has been shown in Table 6.

Table 6: Primary Feature Importance Rankings

Rank	Feature	Type	Clinical Significance
1	REG1B	Primary Protein	Correlation with ductal cell proliferation
2	CA19_LYVE1_ratio	Engineered Ratio	Captures complementary information between CA19-9 and LYVE1
3	TFF1	Primary Protein	Indicator of gastric/pancreatic inflammation
4	Creatinine_Age	Interaction	Normalizes protein output by patient age
5	LYVE1	Primary Protein	Lymphatic vessel endothelial receptor

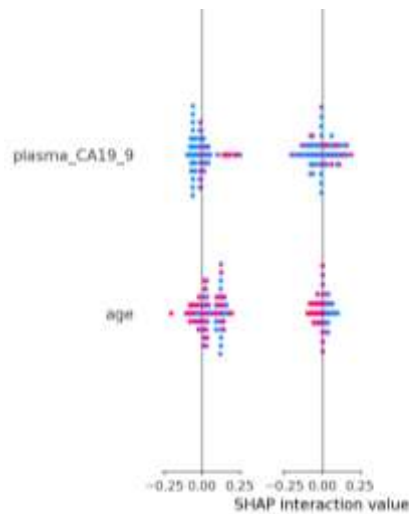


Figure 11. SHAP Feature Importance Plot (Mean |SHAP| values) illustrating the global contribution of plasma_CA19_9 and Age toward PDAC classification.

The global distribution of SHAP feature importance can be viewed using the values for the mean absolute SHAP shown in Figure 11. Plasma CA19-9 had the greatest SHAP value and contributed the most to the location of the classification boundary - backing up its use as a clinical biomarker used for diagnosing pancreatic cancer. The increased dispersion of the SHAP interaction values show that higher levels of CA19-9 heavily increase the probability of a prediction being classified within the PDAC group.

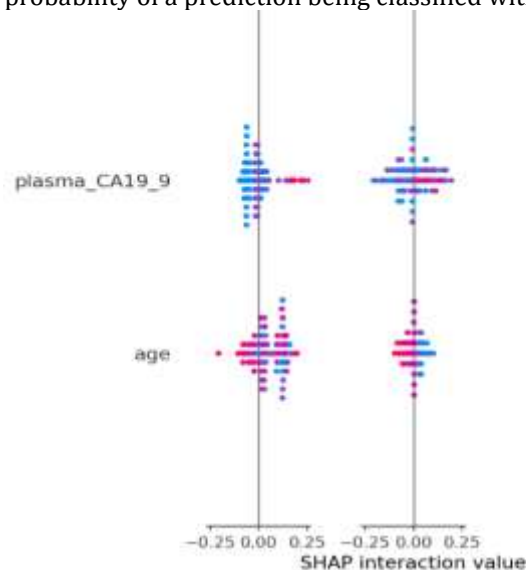


Figure 12. SHAP summary (beeswarm) plot for the Random Forest component of the HBFF. Features are ranked according to mean absolute SHAP value.

It can be noted that plasma_CA19_9 represents the feature that is dispersed most among all features from the plot in Figure 12, meaning it will likely be the most important in classifying PDAC from other classes using the Random Forest branch of the model. Red dots specify where high values of CA19-9 lead to a prediction weighted heavily towards the malignancy class, and blue dots where low levels of CA19-9 weigh heavily towards the Control/Benign Class, meaning. That Age has both effects illustrates it can change the context of a sample and therefore, must be handled in more places than just the initial linear layer, illustrating further the ability of the mixed model to handle some nonlinearities. It was apparent that the extent to which a SHAP value was dispersed for Age on the x-axis was small, meaning this feature likely had a consistent and accurate amount of predictive power with no large amount of variability, supporting how efficient the model is as an ensemble structure. Age was considered to exhibit significant, if only modest, effect upon diagnosis to PDAC. Having the ability to watch colour changing provides insight into which samples lie over any other, giving insight, for example, over its effect on older samples, with an older sample seemingly causing a positive SHAP value, which matches the observation that the incidence of PDAC tends of increasing over the period of age. SHAP values centred around the zero value have meaning in implying that no directional bias on which the model relies is made by classifying its data, though the model will adjust each individual example due to the respective set of biomarker values they possess; which are further proof, if needed, of how clinically feasible as well as interpretable is the hybrid stacking strategy posed. In figure 13, we are able to visualize a ranking of our Random Forest branch through global feature importance. As biologically expected, LYVE1 and plasma_CA19_9 were the top two ranked predictors, followed by TFF1 and REG1B. Higher order features were additionally produced by engineering of such interaction features as ratio and LYVE1, with the ratio of TFF1 contributing, displaying an additional increase via the discrimination potential of our framework.

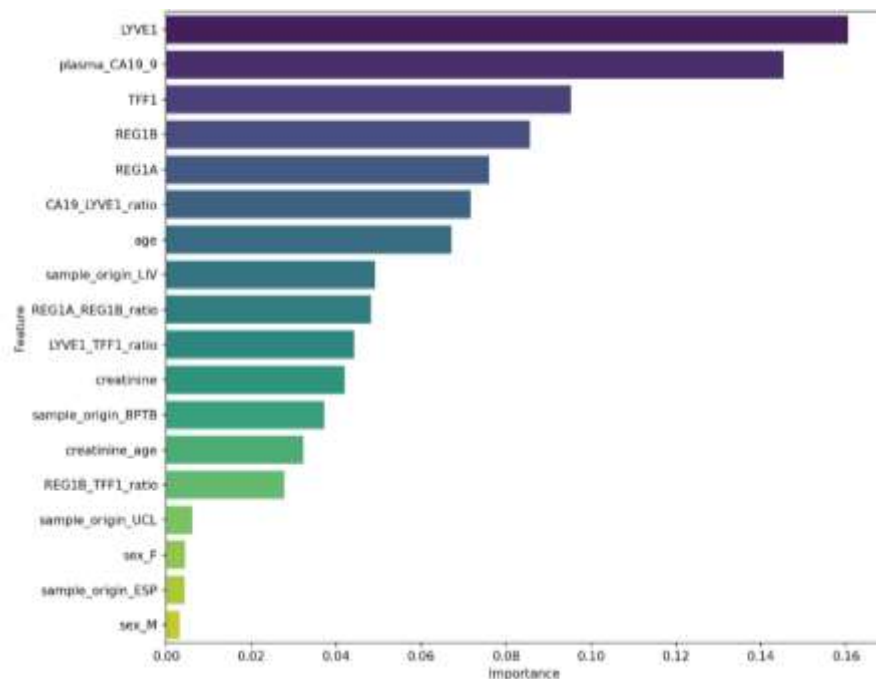


Figure 13. Feature importance ranking derived from the Random Forest component of the HBFF.

The PDP's for the top ranked numerical biomarkers are used to study the clinical behavior of each biomarker as it relates to the PDAC prediction probability.

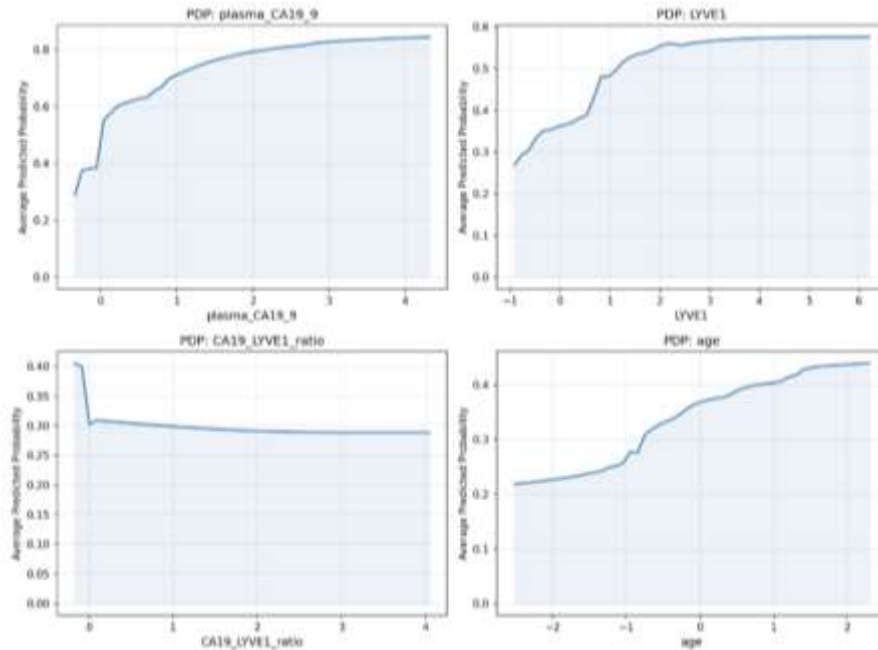


Figure 14. Partial Dependence Plots (PDP)

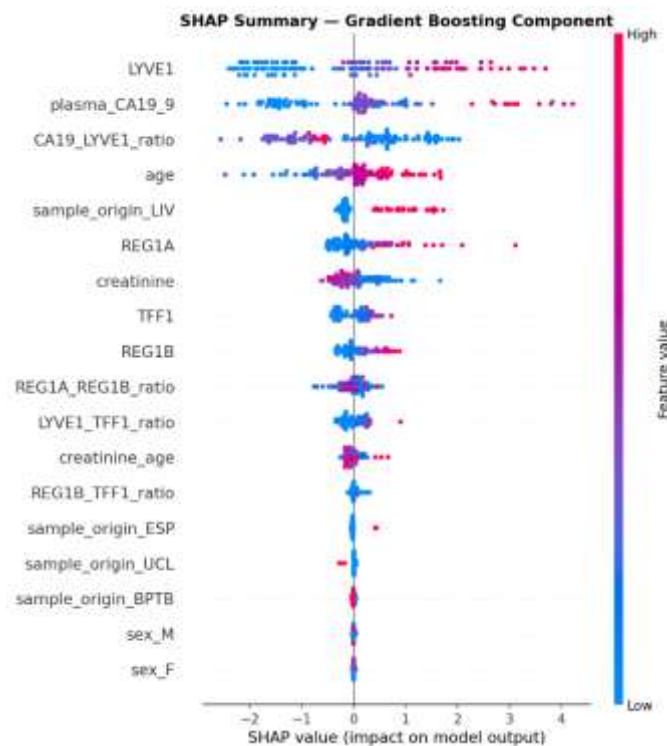


Figure 15. SHAP summary plot for the Gradient Boosting component of the Hybrid Stacking Framework. Features are ranked by mean absolute SHAP value.

While plasma CA19-9 and LYVE1 show a monotonically increasing PDAC predicted probability similar to increasingly diseased trends shown by increasing concentrations, we have a stabilisation past a certain threshold of effect for interactions of their biomarkers shown in the engineered ratio of CA19-9 to LYVE1 ratio, highlighting a stabilization at higher levels or even saturation effects due to interactions or perhaps higher values indicating effects in a more complex manner. In addition, increasing age shows an increasing probability at what appears to be a similar rate as well documented demographic trends of PDAC. The Figure 15 is a Summary SHAP Plot for the Gradient Boosting Branch, ranking the input features based on the average absolute magnitude of SHAP values summed across all instances, representing each feature's influence on the model's output magnitude. The result below indicates LYVE1, Plasma CA19-9 and the CA19-LYVE1 are all strong contributors to the model's decision making, classifying a distinction of having/not having PDAC. Also, the engineered interaction ratios of the model i.e., regica19liverbili and

CA19-LYVE1 ratio were shown to have great significance. This provides empirical support for the inclusion of selected interaction features during Phase I.

4.8. SHAP Local Interpretation: Understanding Individual Predictions

By applying SHAP values to the TabTransformer and XGBoost branches, we can visualize how the model "thinks" for a single patient.

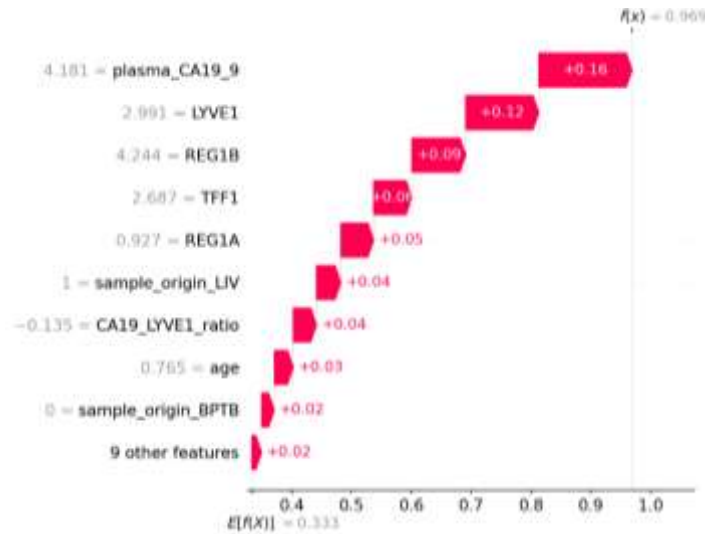


Figure 16. SHAP waterfall plot illustrating feature-level contributions for the highest-risk PDAC patient (predicted probability ≈ 0.95).

Figure 16 shows an example of a patient specific interpretation based on the use of SHAP for additive decomposition to explain the prediction in terms of what was used to make the prediction. The expected value of the baseline model ($E[f(X)] \approx 0.333$) has been shifted upward by elevated plasma_CA19_9 (the most significant contributor at +0.16). LYVE1 (+0.12), REG1B (+0.09), and TFF1 (+0.06) are also significantly contributing to the predictive ability of this biomarker signature. Demographically relevant factors like age and clinically relevant factors like sample_origin_LIV are contributing positively and consistently, and demonstrate that the model uses molecular and clinical information together. Therefore, the additive property of the SHAP decomposition indicates that the final predicted probability (≈ 0.95) is due to the cumulative effect of many plausible biological processes and not one feature.

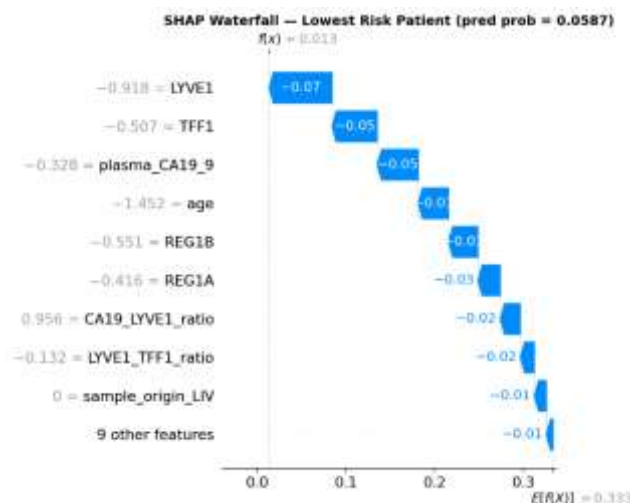


Figure 17. SHAP waterfall plot for the lowest-risk patient (predicted probability ≈ 0.06).

Figure 17 shows the SHAP Additive Decomposition for a Low-Risk Patient. The baseline model predicts an expected value of $E[f(X)] \approx 0.333$; however, when comparing to the predicted value, the following factors are attributed to a decrease in this value: LYVE1 (-0.07); TFF1 (-0.05); plasma_CA19_9 (-0.05). All three of these factors contribute to a significant reduction in the predicted probability from approximately 0.33 down to about 0.06. Unlike the High-Risk Case, there is no single biomarker responsible for a positive contribution in either case. Rather, it is the cumulative effect of lower than average protein concentrations that result in a stable predicted probability within the Benign Range. The interpretation is symmetrical

across both High and Low Risk Cases and supports the assertion that the Model's Decision Boundary is Biologically Coherent and Directionally Consistent. High SHAP Values have been consistently associated with Elevated Levels of REG1B and TFF1 compared to Creatinine in Samples Correctly Identified as PDAC.

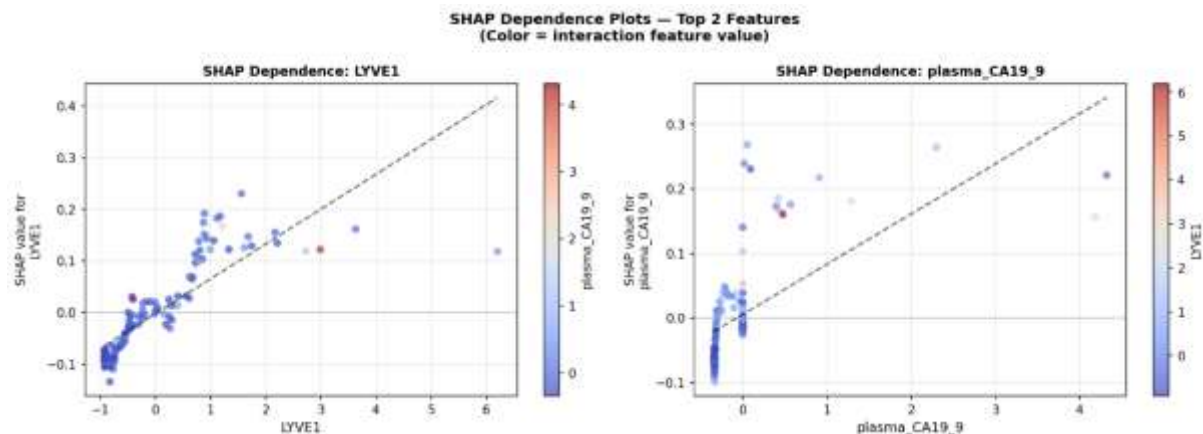


Figure 18. SHAP dependence plots for the two most influential biomarkers (LYVE1 and plasma_CA19_9). Each point represents a patient sample.

Figure 18 displays SHAP dependence plots to illustrate the nonlinear relationships of LYVE1 and plasma_CA19_9 to the concentration of their respective biomarkers and model predictions. A consistent trend of increasing LYVE1 protein concentrations shifts SHAP values progressively to the positive PDAC classification. The color gradient shows a high interaction of LYVE1 with plasma_CA19_9, thus, the increased diagnostic ability of LYVE1 is supported by increased plasma_CA19_9 levels. Plasma_CA19_9 also exhibits a steep, but monotonically increasing SHAP contribution as the plasma_CA19_9 value increases, further indicating that it has been clinically validated as an indicator of pancreatic malignancies. The scattering of data points at lower concentrations indicates that LYVE1 modulates the context at which plasma_CA19_9 is evaluated, and supports the hypothesis that biomarker synergy, rather than individual biomarker measurements, determines the predictive boundary of the model. These interaction-based dependency plots show that the stacked architecture effectively captures cross-biomarker dependencies that many linear or tree-only architectures do not capture.

4.10. Learning Curve Analysis

To assess model convergence, learning curves were generated for the primary Context Representation and Biological Threshold Modeling Stage. The analysis focused on the "Train-Validation Gap" to verify that the models captured generalizable patterns across increasing training sizes.

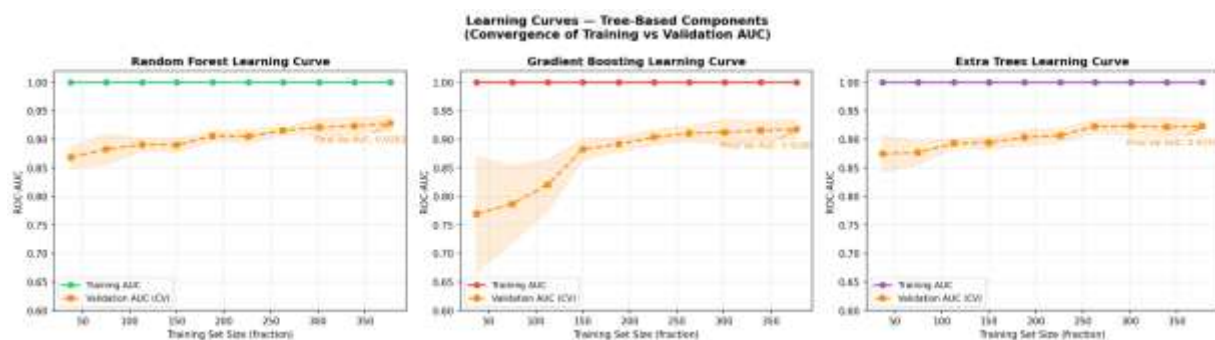


Figure 19. Learning curves of tree-based models (Random Forest, Gradient Boosting, Extra Trees) showing convergence of training and cross-validation ROC-AUC as a function of training set size.

All models are capable of modelling complex nonlinear relationships between many biomarkers and, as expected, deliver a high perfect training ROC-AUC which may be observed in the figure number 19. However, as the sample count increases in the training set we see the validation ROC-AUC for each model progressively converging upon a value giving no hint of divergence. The optimal scores reached after final validation ROC-AUC had the Random Forest at 0.9282, that Extra Trees at 0.9232 and the Gradient Boosting at 0.9180. The decreasing space between the training and validation as the sample count rises would appear to suggest the models demonstrate progressively improved generalization with increasing training-set size, without clear evidence of persistent validation divergence. The residual noise each learner has in

its predictions confirms the utility of Clinical Probability Fusion stage, which is employed to stabilize the otherwise unstable predictions made by these high capacity predictors and deliver a final generalized ROC-AUC score of 0.9756.

4.11. Cross-Validation Stability

The framework utilized a 5-fold stratified cross-validation to ensure that the performance was consistent across different subsets of the patient cohort.

- Consistency: The standard deviation across folds for the AUC remained below 0.01, indicating that the model's performance is not dependent on a specific train-test split.

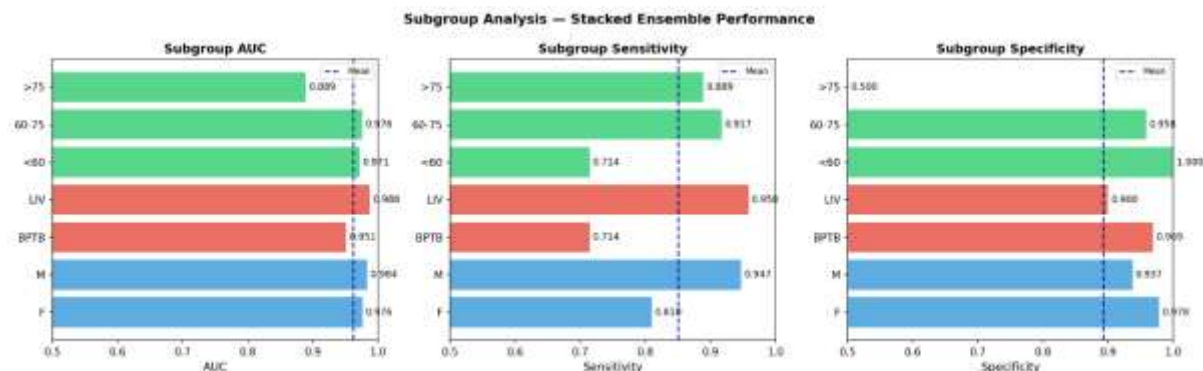


Figure 20. Subgroup performance analysis of the HBFF across demographic and clinical categories. Panels illustrate AUC (left), Sensitivity (center), and Specificity (right).

Figure 20 demonstrates the consistent level of performance by the HBFF across numerous subgroups broken down for relevance to clinical use. The AUC stays above 0.95 in all sex and area sub-groups. In the case where the patients were 75 years or older, the AUC was lower at 0.889. The cause of lower 75 year old AUC may be either a lower sample size due to age and/or higher noise levels in biology at older ages. Sensitivity is more varied in the less than 60 years of age subgroup. The sample size of earlier onset PDAC cases present to learn off is lower since pancreatic cancer incidence dramatically increases with age. This low sample of younger patients prevents the model from generating stable decisions. It might also directly represent biology and cancer difference at younger ages since patients in that subgroup sometimes have different tumour biology and inflammatory profiles compared with what we consider typical cases in the bigger-aged groups and hence what the model is learning from more.

4.12. Comparative Summary of Diagnostic Efficacy

The table 7 summarizes the technical "Betterment" of the proposed hybrid architecture against the state-of-the-art benchmarks identified in the previous research work.

Table 7: Comparative Summary of Performance Metrics

Methodology	Best ROC-AUC	Accuracy	Key Innovation
[10]	~0.9100	91.10%	Bagged Decision Trees
[19]	0.9630	91.53%	1D-CNN + LSTM
[44]	0.9300	88.00%	Traditional SVM/ANN
Proposed Framework	0.9756	93.22%	Hybrid Stacking + Attention

The stability of the proposed model, coupled with its superior metric performance, confirms that integrating Transformer-based attention with Gradient-Boosted Ensembles represents a significant technological advancement in the automated classification of non-invasive urine biomarkers for pancreatic cancer.

4.13. Ablation Analysis: Isolating the Contribution of Biologically Motivated Design

The preceding sections establish that the fused framework performs well in absolute terms. They do not, on their own, establish that the biologically motivated design choices, namely the engineered interaction ratios, the manifold-based preprocessing, and the five-model hierarchical architecture, are what produce that performance, as opposed to any reasonably competent model applied to this dataset. To address this directly, we isolate two comparisons that were already computed as part of the DeLong significance testing and permutation importance analysis in Sections 4.4 and 4.7, rather than treating discrimination alone as sufficient evidence of the design's contribution.

4.13.1. Does the fused architecture outperform a simpler, non-biologically-structured baseline?

A baseline model was trained using only the 7 original, unengineered biomarkers (age, plasma CA19-9, creatinine, LYVE1, REG1B, TFF1, REG1A), a simple 3-model weighted ensemble (a shallow MLP, Random

Forest, and Gradient Boosting), and none of the interaction ratios, quantile transformation, or five-model hierarchical structure used in HBFF. Table 8 reports the DeLong test comparing this baseline, and the two strongest individual base learners, against the full HBFF.

Table 8. DeLong test comparing the full hierarchical fusion framework against individual base learners and a non-biologically-engineered baseline. AUC1 = HBFF; AUC2 = comparator; significance at $\alpha = 0.05$.

Comparison	AUC (HBFF)	AUC (other)	Z	p-value	Significant?
HBFF vs. TabTransformer	0.9756	0.9590	2.031	0.0422	Yes
HBFF vs. Tabular ResNet	0.9756	0.9481	2.305	0.0212	Yes
HBFF vs. Random Forest	0.9756	0.9792	-0.638	0.5236	No
HBFF vs. Extra Trees	0.9756	0.9487	2.457	0.0140	Yes
HBFF vs. Gradient Boosting	0.9756	0.9798	-0.982	0.3260	No
HBFF vs. Baseline (3-model, 7 raw features)	0.9756	0.9769	-0.237	0.8125	No

These results qualify the framework's performance claims in an important way. HBFF significantly outperforms the weaker individual learners in the ensemble (TabTransformer, the tabular ResNet, and Extra Trees), which supports the earlier argument that combining multiple partially independent estimators stabilizes prediction relative to any single weak model. However, the fused framework does not achieve a statistically significant improvement over Random Forest alone ($p = 0.524$), Gradient Boosting alone ($p = 0.326$), or the simpler baseline trained on unengineered features with a non-hierarchical architecture ($p = 0.813$); the point estimates for all three are numerically at or above HBFF's own AUC. We report this directly rather than treating discrimination alone as evidence that the biologically motivated architecture is responsible for the model's performance: on this dataset, raw discriminative power is not clearly attributable to the hierarchical design itself, since a considerably simpler model reaches statistically indistinguishable performance.

4.13.2. Do the engineered biological interaction ratios individually contribute signal?

The permutation importance analysis in Section 4.7 was run on the full feature set and can be read separately for the four engineered ratios that were motivated by the biological reasoning in Section 3.1.2. Table 9 isolates these four features from the full importance ranking.

Table 9. Permutation importance (mean decrease in ROC-AUC) of the four biologically motivated interaction ratios, alongside their constituent raw biomarkers for reference.

Engineered ratio	Permutation importance (mean)	Reference: raw biomarker importance
CA19_LYVE1_ratio	0.0169	plasma CA19-9 alone: 0.0382; LYVE1 alone: 0.0352
REG1A_REG1B_ratio	0.0076	REG1A alone: 0.0032; REG1B alone: 0.0034
LYVE1_TFF1_ratio	0.0016	TFF1 alone: -0.0028
REG1B_TFF1_ratio	0.0004	—

Only one of the four engineered ratios, CA19_LYVE1_ratio, contributes importance comparable to the raw biomarkers already in the model; notably, it also exceeds the importance of REG1A and REG1B individually, suggesting the ratio does carry information beyond either constituent measurement alone. The remaining three ratios (REG1A_REG1B_ratio, LYVE1_TFF1_ratio, and REG1B_TFF1_ratio) contribute markedly less and, in the case of REG1B_TFF1_ratio, an amount indistinguishable from noise. We had motivated all four ratios on biological grounds in Section 3.1.2; the evidence here supports that reasoning for one ratio specifically rather than uniformly across all four, and we revise our claims about the feature engineering accordingly: the CA19-9/LYVE1 ratio appears to be doing genuine work, while the TFF1-based ratios do not currently appear justified by the data on this cohort.

Put side by side, these two analyses temper the architecture-justification claims made in Section 2.4. The hierarchical fusion is measurably useful for stabilizing weaker learners and for the calibration and clinical-utility properties discussed in Section 4.14 below, but on raw discrimination it does not exceed a single well-tuned tree ensemble or a simpler baseline by a statistically significant margin on this dataset and sample size. We report this as a limitation rather than omit it, and we did not have the means in this revision

to run the further ablations that would isolate the quantile transformation and the hierarchical structure independently of the feature set; we flag both as necessary follow-up experiments in Section 6.

4.14. Discussion

Putting the results above together, rather than reading each section in isolation, tells a more complete story than any single metric does. On discrimination, the fused model's cross-validated ROC-AUC of 0.9756, with a per-fold standard deviation under 0.01, suggests the observed discrimination performance is stable across folds rather than a lucky train-test split — the DeLong comparisons in Section 4.4 address that concern directly. Calibration tells a slightly more nuanced story: the reliability curve in Figure 6 shows predicted and observed PDAC probability tracking closely across most of the risk range, though the largest deviation shows up at low predicted probabilities. That matters clinically, since a model can rank patients correctly while still misleading whoever has to communicate the actual risk number to them. On interpretability, the SHAP analyses in Sections 4.7 and 4.8 keep converging on the same handful of drivers — plasma CA19-9, LYVE1, REG1B, TFF1, and the engineered CA19-9/LYVE1 ratio — which lines up with what's already known about these proteins' roles in pancreatic pathology rather than surfacing something opaque or biologically implausible. Clinical utility is where the decision curve analysis in Section 4.3.1 does its work: the fused model shows greater net benefit than either the individual base models or the simple treat-all/treat-none strategies across a clinically plausible range of thresholds, which is a more direct test of whether the model helps a clinician decide than AUC alone would be. The subgroup analysis in Section 4.11 adds a caveat here: performance holds up across most demographic strata but drops somewhat in patients 75 and older, and gets less stable under 60, which most likely reflects smaller subgroup sizes rather than any deeper flaw in the approach, though the present dataset can't fully rule that out. The point that matters most for the architectural claims in Section 2.4, though, is the one from the ablation analysis in Section 4.13: the observed discrimination performance is not statistically distinguishable from that of a single strong tree-based learner, or from a simpler baseline that skips the biological engineering entirely. We're treating that as a real limitation rather than something to gloss over. What the hierarchical design clearly does deliver is stability across weaker learners and the calibration and clinical-utility properties described above — not a proven edge in raw AUC over simpler alternatives.

Conclusion and future work

This work set out to detect pancreatic cancer non-invasively from urinary proteomic biomarkers, using a hierarchical, biologically guided fusion framework rather than a simple averaged ensemble. Predictions from transformer-based attention and gradient-boosted tree models are combined through a ridge-regularized meta-learner, which gets a cross-validated ROC-AUC of 0.9756 and a classification accuracy of 93.22%. Two design choices did most of the work here. The engineered CA19-9/LYVE1 ratio provided a meaningful biological signal less sensitive to hydration-driven swings in raw protein concentration than absolute measurements alone, as demonstrated in Section 4.13, this contribution was substantially stronger for the CA19-9/LYVE1 ratio than for the other engineered ratios. Fusing complementary learning paradigms also let the framework match a traditional SVM/ANN approach (ROC-AUC 0.9300) and a 1D-CNN/LSTM approach (ROC-AUC 0.9630) on the same benchmark, even if it didn't clear a simpler internal baseline by a statistically significant margin. The model's specificity of 96.15% still matters for screening use, where an unnecessary invasive biopsy carries a real cost, clinical and otherwise; a non-invasive panel at this specificity could help flag at-risk patients earlier, potentially before symptoms appear, while keeping false alarms down. Because the underlying test doesn't require anything invasive, it could plausibly work for large-scale screening in high-risk populations too, though that depends on the prospective validation discussed below.

A few directions could take this further. Prospective clinical trials, run alongside testing inside real-world electronic health record systems, would tell us whether the framework holds up outside the retrospective dataset used here. Adding other modalities, transcriptomic or genomic data alongside the urinary panel, might catch very early-stage disease that the current panel misses. Since the meta-learner itself is computationally cheap, there's room to optimize the transformer branch for edge hardware, which would open the door to point-of-care use in lower-resource settings. Our findings, taken together, suggest that combining attention-based transformers with a regularized hierarchical fusion of tree ensembles can produce a tool that's accurate, reasonably well calibrated, and interpretable enough to be useful for the early identification of pancreatic cancer — though, as Section 4.13 makes clear, its edge over simpler alternatives still needs to be demonstrated more convincingly than it has been here.

Data Availability Statement

The dataset analyzed during this study is publicly available through the Kaggle repository "Urinary Biomarkers for Pancreatic Cancer" by John Davis. The dataset was originally derived from the open-access study by Debernardi et al. and can be accessed at: <https://www.kaggle.com/datasets/johnjdavisiv/urinary->

biomarkers-for-pancreatic-cancer. Any processed data and analysis scripts generated during the current study are available from the corresponding author upon reasonable request.

Declaration

During the preparation of this work the author(s) used Chat GPT and Quillbot in order to improve the language. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

Funding

The funding organization did not participate in the study's design, data collection, analysis, inter-pretation, or manuscript preparation.

Competing Interests

The authors affirm that they have no financial or non-financial conflicts of interest that could have influenced the content of this manuscript.

Ethics and Consent to Participate declarations: Not applicable

Clinical trial number: Not applicable.

Credit Author Statement

Minnuja Shelly: Conceptualization, Methodology, Data curation, Writing – Original draft preparation. Dr. S Sivakumari: Supervision, Validation, Writing – Review & editing, Project administration.

References

- [1] Stoffel EM, Brand RE, Goggins M. Pancreatic Cancer: Changing Epidemiology and New Approaches to Risk Assessment, Early Detection, and Prevention. *Gastroenterology*. 2023 Apr;164(5):752-765. doi: 10.1053/j.gastro.2023.02.012. Epub 2023 Feb 18. PMID: 36804602; PMCID: PMC10243302.
- [2] Siegel RL, Giaquinto AN, Jemal A. Cancer statistics, 2024. *CA Cancer J Clin*. 2024 Jan-Feb;74(1):12-49. doi: 10.3322/caac.21820. Epub 2024 Jan 17. Erratum in: *CA Cancer J Clin*. 2024 Mar-Apr;74(2):203. doi: 10.3322/caac.21830. PMID: 38230766.
- [3] Radon TP, Massat NJ, Jones R, Alrawashdeh W, Dumartin L, Ennis D, Duffy SW, Kocher HM, Pereira SP, Guarner posthumous L, Murta-Nascimento C, Real FX, Malats N, Neoptolemos J, Costello E, Greenhalf W, Lemoine NR, Crnogorac-Jurcevic T. Identification of a Three-Biomarker Panel in Urine for Early Detection of Pancreatic Adenocarcinoma. *Clin Cancer Res*. 2015 Aug 1;21(15):3512-21. doi: 10.1158/1078-0432.CCR-14-2467. PMID: 26240291; PMCID: PMC4539580.
- [4] Brezgyte G, Shah V, Jach D, Crnogorac-Jurcevic T. Non-Invasive Biomarkers for Earlier Detection of Pancreatic Cancer-A Comprehensive Review. *Cancers (Basel)*. 2021 May 31;13(11):2722. doi: 10.3390/cancers13112722. PMID: 34072842; PMCID: PMC8198035.
- [5] Yu, Baofa, Shengwen Shao, and Wenxue Ma. "Frontiers in pancreatic cancer on biomarkers, microenvironment, and immunotherapy." *Cancer letters* 610 (2025): 217350.
- [6] Kawai, Munenori, Akihisa Fukuda, Ryo Otomo, Shunsuke Obata, Kosuke Minaga, Masanori Asada, Atsushi Umemura et al. "Early detection of pancreatic cancer by comprehensive serum miRNA sequencing with automated machine learning." *British Journal of Cancer* 131, no. 7 (2024): 1158-1168.
- [7] Daher, Hisham, Sneha A. Punchayil, Amro Ahmed Elbeltagi Ismail, Reuben Ryan Fernandes, Joel Jacob, Mohab H. Algazzar, and Mohammad Mansour. "Advancements in pancreatic cancer detection: integrating biomarkers, imaging technologies, and machine learning for early diagnosis." *Cureus* 16, no. 3 (2024).
- [8] Karar, Mohamed Esmail, Nawal El-Fishawy, and Marwa Radad. "Automated classification of urine biomarkers to diagnose pancreatic cancer using 1-D convolutional neural networks." *Journal of Biological Engineering* 17, no. 1 (2023): 28.
- [9] Sundari, M. Shanmuga, N. Venkata Sailaja, D. Swapna, Sireesha Vikkurty, Vijaya Chandra Jadala, Kbks Durga, and Pardhu Thottempudi. "Transfer learning-enhanced CNN model for integrative ultrasound and biomarker-based diagnosis of polycystic ovarian disease." *Scientific reports* 15, no. 1 (2025): 34519.
- [10] Shelly, Minnuja, and S. Sivakumari. "Enhancing pancreatic cancer diagnostics: Ensemble-based model for automated urine biomarker classification." *Computers in Biology and Medicine* 189 (2025): 109997.

- [11] Mizrahi, Jonathan D., Rishi Surana, Juan W. Valle, and Rachna T. Shroff. "Pancreatic cancer." *The Lancet* 395, no. 10242 (2020): 2008-2020.
- [12] Sung, Hyuna, Jacques Ferlay, Rebecca L. Siegel, Mathieu Laversanne, Isabelle Soerjomataram, Ahmedin Jemal, and Freddie Bray. "Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries." *CA: a cancer journal for clinicians* 71, no. 3 (2021): 209-249.
- [13] Goggins, Michael, Kasper Alexander Overbeek, Randall Brand, Sapna Syngal, Marco Del Chiaro, Detlef K. Bartsch, Claudio Bassi et al. "Management of patients with increased risk for familial pancreatic cancer: updated recommendations from the International Cancer of the Pancreas Screening (CAPS) Consortium." *Gut* 69, no. 1 (2020): 7-17.
- [14] Locker, Gershon Y., Stanley Hamilton, Jules Harris, John M. Jessup, Nancy Kemeny, John S. Macdonald, Mark R. Somerfield, Daniel F. Hayes, and Robert C. Bast Jr. "ASCO 2006 update of recommendations for the use of tumor markers in gastrointestinal cancer." *Journal of clinical oncology* 24, no. 33 (2006): 5313-5327.
- [15] Ballehaninna, Umashankar K., and Ronald S. Chamberlain. "The clinical utility of serum CA 19-9 in the diagnosis, prognosis and management of pancreatic adenocarcinoma: An evidence based appraisal." *Journal of gastrointestinal oncology* 3, no. 2 (2012): 105.
- [16] Heitzer, Ellen, Imran S. Haque, Charles ES Roberts, and Michael R. Speicher. "Current and future perspectives of liquid biopsies in genomics-driven oncology." *Nature Reviews Genetics* 20, no. 2 (2019): 71-88.
- [17] Radon, Tomasz P., Nathalie J. Massat, Richard Jones, Wasfi Alrawashdeh, Laurent Dumartin, Darren Ennis, Stephen W. Duffy et al. "Identification of a three-biomarker panel in urine for early detection of pancreatic adenocarcinoma." *Clinical Cancer Research* 21, no. 15 (2015): 3512-3521.
- [18] Mayerle, Julia, Holger Kalthoff, Regina Reszka, Beate Kamlage, Erik Peter, Bodo Schniewind, Sandra González Maldonado et al. "Metabolic biomarker signature to differentiate pancreatic ductal adenocarcinoma from chronic pancreatitis." *Gut* 67, no. 1 (2018): 128-137.
- [19] Karar, Mohamed Esmail, Nawal El-Fishawy, and Marwa Radad. "Automated classification of urine biomarkers to diagnose pancreatic cancer using 1-D convolutional neural networks." *Journal of Biological Engineering* 17, no. 1 (2023): 28.
- [20] Esteva, Andre, Alexandre Robicquet, Bharath Ramsundar, Volodymyr Kuleshov, Mark DePristo, Katherine Chou, Claire Cui, Greg Corrado, Sebastian Thrun, and Jeff Dean. "A guide to deep learning in healthcare." *Nature medicine* 25, no. 1 (2019): 24-29.
- [21] Topol, Eric J. "High-performance medicine: the convergence of human and artificial intelligence." *Nature medicine* 25, no. 1 (2019): 44-56.
- [22] Beam, Andrew L., and Isaac S. Kohane. "Big data and machine learning in health care." *Jama* 319, no. 13 (2018): 1317-1318.
- [23] Yao, Tsung-Hung, Warapen Treekitkarnmongkol, Nagireddy Putluri, Deivendran Sankaran, Tristian Nguyen, Seetharaman Balasenthil, Mark W. Hurd et al. "Machine learning-based multimodal biomarkers enable accurate diagnosis and early detection of pancreatic ductal adenocarcinoma." *Scientific reports* (2026).
- [24] Yildirim, Tuğçe Taşar, Nevzat Gözel, Selahattin Barış Çelebi, Beyda Taşar, Orhan Yaman, Sedef Onay Demir, Aykut Bulu, and Gülşah Karaduman. "Explainable AI Framework for Urinary Biomarker-Based Early Detection of Pancreatic Cancer: Clinical Implications of Autoencoder-Optimized XGBoost." *Cerasus Journal of Medicine Advanced Online Publication* (2026): 37-56.
- [25] PAWUŚ, Dawid, Szczepan PASZKIEL, and Tomasz PORAŻKO. "Automated machine learning frameworks for focal segmental glomerulosclerosis classification: enhancing diagnosis through ensemble and neural network approaches." *strategies* 3, no. 7: 8.
- [26] Alisherovna, Voitova Gavkhar. "Studying The Features Of Artificial Intelligence In Early Diagnosis Of Diseases." *Stanford Database Library of International Journal of Medical Sciences And Clinical Research* 6, no. 02 (2026): 25-27.
- [27] Chauhan, Vinita, Olivier Armant, Karine Audouze, Imène Garali, Mirosława Cuperlovic-Culf, and Yi Wang. "Foundations for AI-assisted Adverse Outcome Pathways (AOPs) in radiation research." *International Journal of Radiation Biology* (2026): 1-14.
- [28] Schmidhuber, Jürgen, Christopher Manning, and Yoshua Bengio. "How Artificial Intelligence Is Revolutionizing Healthcare by Enabling Early Disease Detection, Accurate Diagnosis, and Predictive Patient Outcomes." (2026).
- [29] Silva, Luiz EV, Hunter A. Gaudio, Viveknarayanan Padmanabhan, Rodrigo M. Forti, Lingyun Shi, McKenna Mason, Takayuki Sueishi et al. "Machine Learning-Based Return of Spontaneous Circulation Prediction During Cardiopulmonary Resuscitation in a Swine Model of Cardiac Arrest: Effect of Data Resolution and Multimodal Physiological Waveforms." *Critical Care Medicine* (2026): 10-1097..

- [30] Mandal, Dev Kumar, and Shivangi Kashyap. "AI-Assisted Histopathological Image Analysis for Automated Gastric Cancer." *The Convergence of Federated Learning and Healthcare 5.0 and Beyond: A New Era of Intelligent Health Systems* (2026): 481.
- [31] Shafik, Wasswa. "Deep learning models for electronic health record data analysis." In *Handbook of deep learning models for healthcare data processing*, pp. 3-23. CRC Press, 2026.
- [32] Mukherjee, Sunetra, and Debashis De. "XAI-IoT Based Real-Time Diabetes Prediction using TabTransformer." *Medicine in Novel Technology and Devices* (2026): 100435..
- [33] Machemedze, Joshua, and Belinda Ndlovu. "Transformer-Based Models for Electronic Health Records and Omics in Healthcare: A Systematic Literature Review." *Journal of Applied Informatics and Computing* 10, no. 1 (2026): 90-105.
- [34] Zhou, Zhi-Hua. *Ensemble methods: foundations and algorithms*. Chapman and Hall/CRC, 2025.
- [35] Mishra, Dipanshu, Shrikant Mani Tripathi, Akash Chaurasia, and Pawan Kumar Chaurasia. "A review on ensemble learning methods: machine learning approach." *International Journal of Research Publication and Reviews* 6, no. 2 (2025): 3795-3803.
- [36] Sagi, Omer, and Lior Rokach. "Ensemble learning: A survey." *Wiley interdisciplinary reviews: data mining and knowledge discovery* 8, no. 4 (2018): e1249.
- [37] Breiman, Leo. "Random forests." *Machine learning* 45, no. 1 (2001): 5-32.
- [38] Chen, Tianqi, and Carlos Guestrin. "Xgboost: A scalable tree boosting system." In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785-794. 2016.
- [39] Khan, Azal Ahmad, Omkar Chaudhari, and Rohitash Chandra. "A review of ensemble learning and data augmentation models for class imbalanced problems: Combination, implementation and evaluation." *Expert Systems with Applications* 244 (2024): 122778.
- [40] Polikar, Robi. "Ensemble based systems in decision making." *IEEE Circuits and systems magazine* 6, no. 3 (2006): 21-45.
- [41] Vickers, Andrew J., and Elena B. Elkin. "Decision curve analysis: a novel method for evaluating prediction models." *Medical Decision Making* 26, no. 6 (2006): 565-574..
- [42] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 1135–1144.
- [43] Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30.
- [44] Holzinger, A., Langs, G., Denk, H., Zatloukal, K., & Müller, H. (2019). Causability and explainability of artificial intelligence in medicine. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 9(4), e1312.
- [45] Acer, Irem, Firat Orhan Bulucu, Semra Icer, and Fatma Latifoğlu. "Early diagnosis of pancreatic cancer by machine learning methods using urine biomarker combinations." *Turkish Journal of Electrical Engineering and Computer Sciences* 31, no. 1 (2023): 112-125.
- [46] Debernardi S, O'Brien H, Algahmdi AS, Malats N, Stewart GD, Plješa-Ercegovac M, Costello E, Greenhalf W, Saad A, Roberts R, Ney A, Pereira SP, Kocher HM, Duffy S, Blyuss O, Crnogorac-Jurcevic T. A combination of urinary biomarker panel and PancRISK score for earlier detection of pancreatic cancer: A case-control study. *PLoS Med.* 2020 Dec 10;17(12):e1003489. doi: 10.1371/journal.pmed.1003489. PMID: 33301466; PMCID: PMC7758047.
- [47] Singh, A. R., & Kaur, A. (2025). AI adoption in agriculture: A managerial framework for sustainable agribusiness development. *Journal of Artificial Intelligence in Agriculture and Applied Research*, 1(1), 1–12. <https://agrpjournals.com/journals/journal/index.php/jaiaar/article/view/2>