Briefings in Bioinformatics, 2025, **26(6)**, bbaf583
https://doi.org/10.1093/bib/bbaf583
Problem Solving Protocol

A machine learning framework using urinary biomarkers for pancreatic ductal adenocarcinoma prediction with *post hoc* validation via single-cell transcriptomics

Dahlak D. Solomon[1,2,‡], Ching-Chung Ko[3,4,5,‡], Hsin-Yi Chen[1], Sachin Kumar[1,6], Fitria Sari Wulandari[1], Do Thi Minh Xuan[7], Hung-Yun Lin[1,8,9,10,11], Hui-Ru Lin[12,13], Yung-Kuo Lee[12,14,15], Wen-Hsin Hsu[16,17], Yang Pei-Ming[1,8,18,19,20,*], Chih-Yang Wang[1,8,21,*], Ngoc Uyen Nhi Nguyen[22,23]

[1]Graduate Institute of Cancer Biology and Drug Discovery, College of Medical Science and Technology, Taipei Medical University, No. 301, Yuantong Road, Zhonghe District, New Taipei City 23561, Taiwan
[2]Yogananda School of AI Computers and Data Sciences, Shoolini University of Biotechnology and Management Sciences, Bajhol, Solan District, Himachal Pradesh 173229, India
[3]Department of Medical Imaging, Chi-Mei Medical Center, No. 901, Chung Hwa Road, Yongkang District, Tainan City 71004, Taiwan
[4]Department of Health and Nutrition, Chia Nan University of Pharmacy and Science, No. 60, Section 1, Erren Road, Rende District, Tainan City 71710, Taiwan
[5]School of Medicine, College of Medicine, National Sun Yat-Sen University, No. 70, Lienhai Road, Gushan District, Kaohsiung City 80424, Taiwan
[6]Faculty of Applied Sciences and Biotechnology, Shoolini University of Biotechnology and Management Sciences, Bajhol, Solan District, Himachal Pradesh 173229, India
[7]Faculty of Pharmacy, Van Lang University, 69/68 Dang Thuy Tram Street, Binh Loi Trung Ward, Ho Chi Minh City 70000, Vietnam
[8]TMU Research Center of Cancer Translational Medicine, Taipei Medical University, No. 301, Yuantong Road, Zhonghe District, New Taipei City 23561, Taiwan
[9]Traditional Herbal Medicine Research Center of Taipei Medical University Hospital, Taipei Medical University, No. 301, Yuantong Road, Zhonghe District, New Taipei City 23561, Taiwan
[10]Pharmaceutical Research Institute, Albany College of Pharmacy and Health Sciences, 1 Discovery Drive, Rensselaer, NY 12144, United States
[11]Cancer Center, Wan Fang Hospital, Taipei Medical University, No. 301, Yuantong Road, Zhonghe District, New Taipei City 23561, Taiwan
[12]Institute of Medical Science and Technology, National Sun Yat-Sen University, 70 Lienhai Road, Gushan District, Kaohsiung City 80424, Taiwan
[13]Nursing Department, Kaohsiung Armed Forces General Hospital, 2 Zhongzheng 1st Road, Lingya District, Kaohsiung City 80284, Taiwan
[14]Medical Laboratory, Medical Education and Research Center, Kaohsiung Armed Forces General Hospital, 2 Zhongzheng 1st Road, Lingya District, Kaohsiung City 80284, Taiwan
[15]Division of Experimental Surgery Center, Department of Surgery, Tri-Service General Hospital, National Defense Medical University, 325 Section 2, Chenggong Road, Neihu District, Taipei City 11490, Taiwan
[16]Department of Emergency Medicine, Kaohsiung Armed Forces General Hospital, 2 Zhongzheng 1st Road, Lingya District, Kaohsiung City 80284, Taiwan
[17]Department of Emergency Medicine, Tri-Service General Hospital, National Defense Medical University, 325 Section 2, Chenggong Road, Neihu District, Taipei City 11490, Taiwan
[18]Liver Medical Center, MacKay Memorial Hospital, 92 Section 2, Zhongshan North Road, Zhongshan District, Taipei City 10449, Taiwan
[19]Cancer Center, Wan Fang Hospital, Taipei Medical University, 111 Section 3, Xinglong Road, Wenshan District, Taipei City 11696, Taiwan
[20]TMU and Affiliated Hospitals Pancreatic Cancer Groups, Taipei Medical University, No. 301, Yuantong Road, Zhonghe District, New Taipei City 23561, Taiwan
[21]Ph.D. Program for Cancer Molecular Biology and Drug Discovery, College of Medical Science and Technology, Taipei Medical University, No. 301, Yuantong Road, Zhonghe District, New Taipei City 23561, Taiwan
[22]Center for Regenerative Medicine, University of South Florida Health Heart Institute, 560 Channelside Drive, Tampa, Florida 33602, United States
[23]Division of Cardiology, Department of Internal Medicine, Morsani School of Medicine, University of South Florida, 12901 Bruce B. Downs Boulevard, Tampa, Florida 33612, United States

[*]Corresponding authors. Yang Pei-Ming, Department of Pharmacology, College of Medicine, Taipei Medical University; TMU Research Center of Cancer Translational Medicine; Liver Medical Center, MacKay Memorial Hospital; Cancer Center, Wan Fang Hospital, Taipei Medical University, No. 301, Yuantong Road, Zhonghe District, New Taipei City 23561, Taiwan. E-mail: yangpm@tmu.edu.tw; Chih-Yang Wang, Graduate Institute of Cancer Biology and Drug Discovery, College of Medical Science and Technology, Taipei Medical University; TMU Research Center of Cancer Translational Medicine; Ph.D. Program for Cancer Molecular Biology and Drug Discovery, College of Medical Science and Technology, Taipei Medical University, No. 301, Yuantong Road, Zhonghe District, New Taipei City 23561, Taiwan. E-mail: chihyang@tmu.edu.tw
[‡]Dahlak D. Solomon and Ching-Chung Ko contributed equally to this work.

Abstract

Pancreatic ductal adenocarcinoma (PDAC) is a highly lethal cancer with a poor prognosis, thus emphasizing the need for early and accurate diagnostic tools. In this study, we propose a comparative study approach to understand how machine learning (ML) modeling using urinary biomarkers combined with demographic data can predict PDAC. The study also utilized a single-cell RNA sequencing

Received: July 16, 2025. **Revised:** September 9, 2025. **Accepted:** October 3, 2025

(scRNA-seq) analysis to assess and understand gene expressions of included biomarkers. With inclusion of available biomarkers and incorporation of demographic information, we employed different approaches for preprocessing techniques, normalization approaches, ML techniques, and deep learning (DL) approaches to provide a comprehensive prediction model. The scRNA-seq approach also highlighted the significance of the urinary biomarkers from the pancreatic single-cell sample. Based on this analysis, the marker was identified as one of the top three most highly expressed genes in PDAC tissues. The predictive modeling approach was conducted for both binary and multiclass classification using both ML and DL approaches. The comparative analysis using all included parameter combinations produced modeling settings, and among these parameters, the DL modeling approach using binary classification outperformed the other approaches by achieving 91% accuracy. This framework provided insights that highlighted the critical role of demographic data and potential approaches to include such features in the model without impacting the predictive accuracy. Future work will focus on examining the framework using different datasets, integrating additional omics data, and exploring advanced DL architectures to further improve predictive performances.

Keywords: pancreatic ductal adenocarcinoma (PDAC); single-cell RNA sequencing (scRNA-seq); machine learning; deep learning; biomarker discovery; urinary biomarkers; predictive modeling

Introduction

The poor prognosis and high mortality rate make pancreatic ductal adenocarcinoma (PDAC) one of the deadliest cancers with a low 5-year survival rate, and the complexity of the tissue components adds an additional layer of difficulty to achieving therapeutic effects with a single therapeutic method [1, 2]. A late diagnosis of PDAC identifies the tumor when it is in an advanced stage with an aggressive nature and limited options. Furthermore, the impact of PDAC in not only limited to the pancreas but is also a highly metastatic disease to the liver [3]. This highlights the urgent need for robust diagnostic tools that have the potential to enable detection of the disease at earlier stages to improve patient outcomes and prevent metastasis to other tissues. The traditional approach to diagnosing PDAC is based on imaging and/or invasive biopsy procedures; on top of the cost, this approach also has limited accuracy when it comes to detecting the disease in its nascent stages. Additionally, most existing predictive models using the same approach focused on a limited number of biomarkers and ignored demographic features that can help capture the full spectrum of disease-specific signals [4, 5]. Even if another proposal can produce better accuracy because several features included in the modeling have a high impact on the model accuracy, the proposed approaches still include a limited number of verified biomarkers [6, 7].

The involvement of data-oriented machine learning (ML) and deep learning (DL) approaches has been transformative in the healthcare sector [8–10]. Applying such technologies can encompass different perspectives and be geared towards specific or generalized goals. Depending on the available data and the complexity of the problem, multiple approaches can be used to provide accurate and modern solutions [11–15]. Using historical datasets and applying different approaches, models can enable pattern recognition, feature extraction, and predictive modeling. In the case of PDAC, ML and DL can be useful as they offer the potential to analyze complex biological data and construct the data into informative insights that can be used to diagnose and propose personalized treatment plans [16]. Recent advancements in single-cell analyses have opened several new avenues for exploring the molecular and cellular underpinnings of complex diseases for better understanding and insights [17, 18]. Using a single-cell approach within heterogeneous tissues, researchers can identify rare cell types, trace lineage trajectories, and uncover gene expression patterns that are masked in bulk analyses [19, 20]. Single-cell RNA sequencing (scRNA-seq) is one of the most popular and widely utilized approaches to single-cell analysis. scRNA-seq provides a mechanism for profiling gene expressions at the single-cell level [21–23]. This technology is used to unveil the genetic markup and capture the transcriptome of individual cells, whereby researchers can identify cell types and states, and family trees of heterogeneous tissues [24, 25].

This study was motivated to provide predictions for PDAC from new perspectives which unlock insights from demographic features on top of urine biomarkers and insights from scRNA-seq analyses to evaluate expression values of the provided biomarkers to understand their effectiveness. Using several different hyperparameter options for both binary and multiclass classification for ML and DL approaches, the study provides a deeper understanding of biomarker significance and predictive capabilities of each combination [26–28]. Our key contributions to the current domain knowledge are in: (i) developing a comprehensive prediction model that incorporates all available urinary biomarkers and demographic features, (ii) demonstrating the significance of the included biomarkers using scRNA-seq analyses to support the findings of this study, and (iii) conducting a comparative study of classical ML and DL approaches under various preprocessing strategies, including missing value imputation and data normalization. By providing these contributions in this study, we aimed to contribute to the advancement of PDAC research, offering a more inclusive and interpretable approach to disease predictions and biomarker evaluation. Rather than proposing a novel classifier, this work's contribution is methodological and translational to provide a comprehensive comparative analysis of preprocessing and modeling choices on urinary biomarker datasets by integrating demographic features in modeling, and biological *post hoc* validation with scRNA-seq to support biomarker selection and interpretation. Figure 1 illustrates PDAC and its subtypes.

Materials and methods

In this section, we introduce and explore the materials and methods that were used in the study. Starting with a high-level discussion of the overall study design, the section describes the research pipeline used in both predictive modeling experiments and the scRNA-seq analytical approaches. The section also gives a detailed explanation of the datasets utilized for both pipelines and how different preprocessing steps were performed on those datasets to carry out the experiment. Furthermore, the included technologies and approaches used in both pipelines are described, along with the evaluation metrics employed.

Overview of our study design

This study integrated ML approaches for predicting PDAC from urine biomarkers and demographic information. The framework also includes an assessment of the selected biomarkers using an scRNA-seq analysis to understand how the biomarkers are expressed in pancreatic cancer tissues, and this evaluation

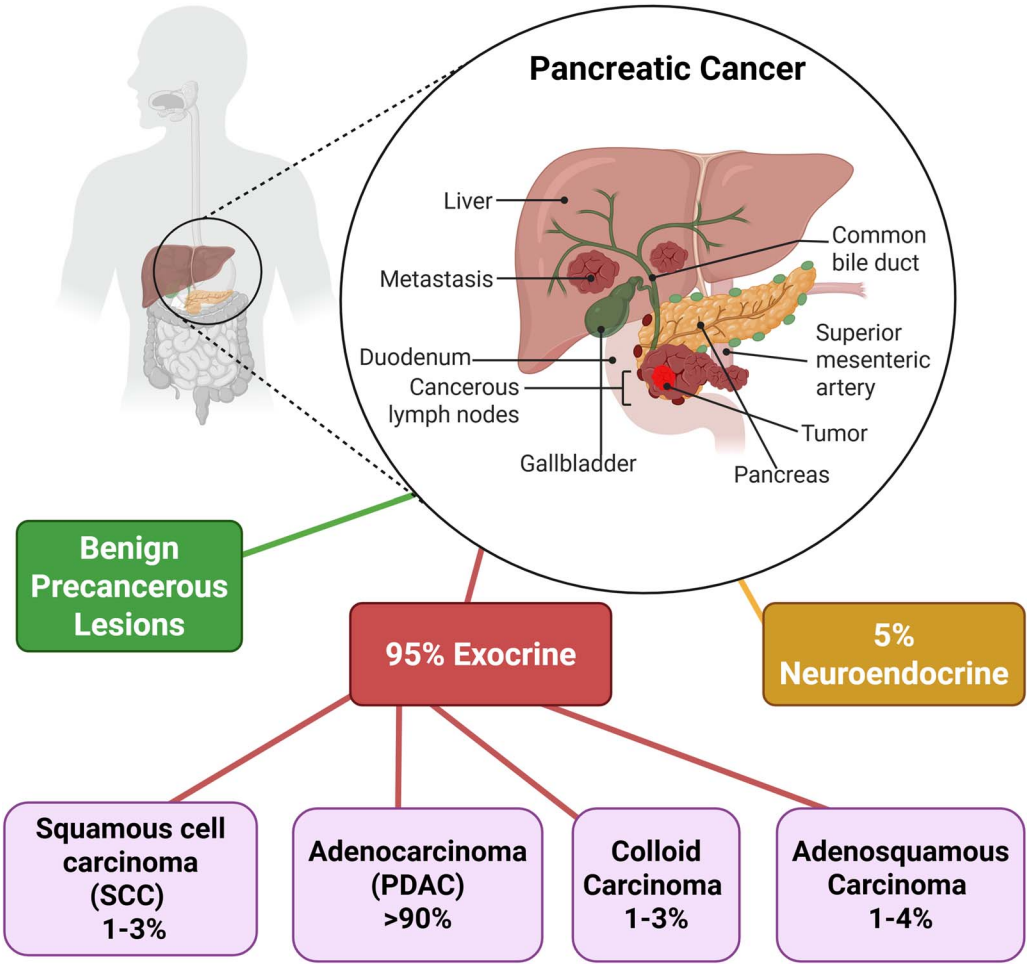

Figure 1. Illustration of pancreatic cancer and its subtypes, providing a visual overview of the disease context relevant to this study (Created in BioRender. Kumar, S. (2025) https://BioRender.com/9bolzl4).

approach on the biomarker's significance and gene expression patterns is used to support the included biomarkers in the ML prediction model. The study framework involves two distinct pipelines: one for an scRNA-seq analysis of biomarker significance assessments and the other for ML workflows aimed at predictive modeling using urinary biomarkers and demographic variables. The combination of these pipelines provides valuable insight into PDAC predictions and biomarker evaluation. An overview of the study is illustrated in Fig. 2. The scRNA-seq analysis was conducted in hierarchical-based steps in which each of the outputs from one step was used as the input for the next step of the pipeline. Initial preprocessing was performed by loading only reads with more than three cells, and >200 features were loaded to the Seurat object. Further quality control was performed to filter out low-quality cells and genes with high mitochondrial contents by taking only those that had fewer than 10,000 features and fewer than 60,000 counts, and range between 1.3 and 0.1 were selected for further study as we previous described [29–31]. Further processing such as normalization, dimensionality reduction using a principal component analysis (PCA), and data scaling, was conducted to ensure that the data used downstream met the quality standard. Dimensionality reduction was conducted using a linear transformation technique that identifies the orthogonal directions of maximum variance in the feature space [32–34]. The ML pipeline involves consecutive steps to process training and evaluate all of the prediction models to compare which approach is suitable. Determining the impact of carefully prepared data for such a comparative analysis is critical to understand the performance of each approach with no bias. Based on this aspect, an exploratory data analysis (EDA) was performed on urinary biomarkers and demographic information to assess the quality of the dataset and check how informative they were from a statistical perspective. Different preprocessing approaches, including imputation of missing values, feature engineering, and normalization, were included in the pipeline. Model performance was evaluated using a comprehensive set of metrics to understand how the model behaved with different parameters and prediction approaches.

Biological background of biomarkers used in pancreatic ductal adenocarcinoma prediction

The predictive modeling framework in this study focused on a curated set of biomarkers: *LYVE1* (lymphatic vessel endothelial hyaluronan receptor 1), *REG1A* (regenerating islet-derived protein 1 alpha), *REG1B*, *TFF1* (trefoil factor 1), and CA 19-9, each of which possesses established or emerging significance in PDAC biology. *LYVE1* encodes a surface receptor involved in hyaluronan transport and lymphangiogenesis. Although primarily expressed by lymphatic endothelial cells, its aberrant expression was implicated in tumor lymphatic invasion and metastasis, suggesting a potential role in the tumor microenvironment (TME) of PDAC. *REG1A* and its paralog *REG1B* are members of the regenerating (*REG*) gene family, typically associated with pancreatic tissue regeneration. Both proteins are secretory in nature and are

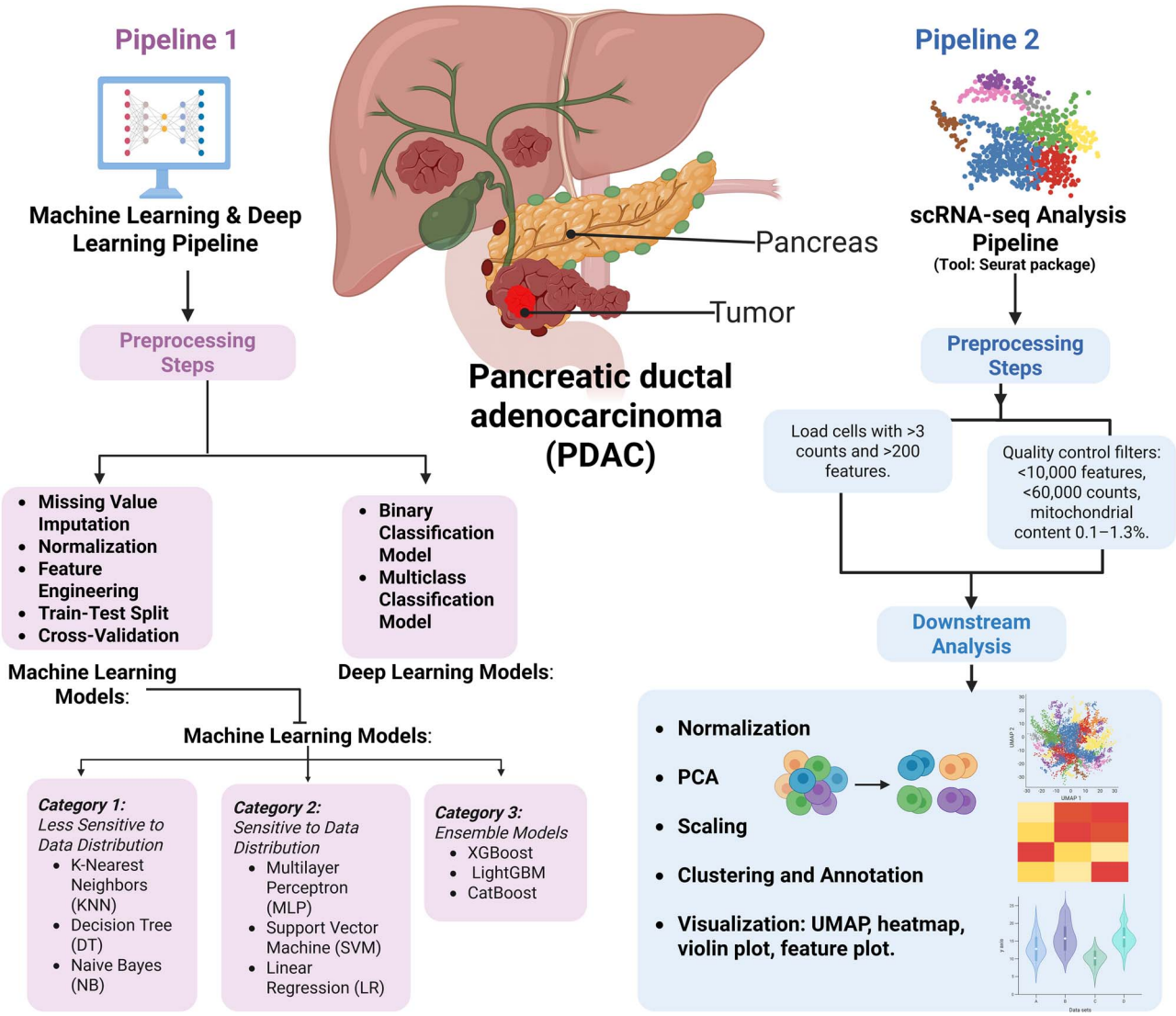

Figure 2. Overview of the study design, integrating ML workflows using urinary and demographic data with scRNA-seq-based biomarker validation for PDAC predictions (Created in BioRender. Kumar, S. (2025) https://BioRender.com/4okvmcy).

upregulated in inflammatory and neoplastic pancreatic conditions, with REG1A often highlighted for its diagnostic sensitivity in distinguishing PDAC from benign lesions. *TFF1* encodes a secreted peptide that contributes to epithelial restitution and mucosal protection. While it is more commonly associated with gastric and breast cancers, its aberrant upregulation in pancreatic neoplasms may contribute to epithelial remodeling and tumor progression. In contrast to these gene-based biomarkers, CA 19-9 is a tumor-associated carbohydrate antigen (sialyl-Lewis^A), not encoded by a single gene but synthesized via a series of glycosyltransferases, including fucosyltransferases (FUT3 and FUT6) and sialyltransferases (ST3GAL6). It is the most widely used clinical serum marker for PDAC and was included in this study as a quantitative plasma-derived variable. Together, these markers represent a hybrid panel of gene-encoded and glycan-based features, capturing both molecular and clinical dimensions of PDAC detection.

Description of datasets

Two different datasets were utilized in this study: a tabular-based dataset was utilized for the ML-based prediction model, and gene expression data were used to analyze biomarker significance using the scRNA-seq analysis. The dataset which was used for the prediction model was originally published in a study by Radon *et al.* [35]. The dataset includes biomarkers data such as *LYVE1*, *REG1B*, *TFF1*, *REG1A*, and *plasma CA19-9*, and demographic variables like age, gender, and patient cohort. Summaries of clinical characteristics and biomarker measurements for all patient samples are presented in Supplementary Table S1. The scRNA-seq dataset was sourced from the GEO website. It can be found under accession no. GSE274665 and was originally published in previous study [36]. Gene expression data consisted of data from four different donors.

Machine learning approaches

This study utilized a variety of classical ML models and DL approaches to predict PDAC from urine biomarkers and demographic data. The ML modeling techniques were classified into three categories based on their sensitivity to how data were distributed and based on missing values in the datasets. There were three categories in which all categories contained three ML techniques based on the categorization logic. Summaries of the included ML approach and their sensitivities to dataset distribution and missing values are given in Supplementary

Table S2. Binary classification was used to distinguish PDAC versus non-PDAC cases, providing a straightforward assessment of disease presence for potential screening or diagnostic applications. Multiclass classification differentiated among three clinically relevant groups, allowing the model to capture additional heterogeneity in patient presentations and evaluate biomarker performances across multiple disease states. Classes were defined based on clinical diagnoses and sample annotations from the original datasets, ensuring consistency with prior studies and enabling meaningful comparisons.

Models less affected by data distribution and missing values

The first category of ML techniques is widely known for their robustness to variations in data distributions, and they can tolerate missing values more easily than other ML approaches without significantly impacting the model performance. The included ML approaches under three categories were K-nearest neighbor (KNN) [37], decision tree (DT) [38], and naive Bayes (NB) [39]. While KNN predictions are based on most of the classes among the k-nearest neighbors at each data point, DTs follow a rule-based system that uses a tree structure to partition the data into expected classes. NB assigns a probability to each expected class work based on Bayes' theorem.

Models highly affected by data distribution and missing values

The second category of ML techniques requires careful preprocessing of the data as they are sensitive and highly exposed to biases based on how the data are distributed, and they perform poorly when the data contain missing values. The included ML techniques under this category were multilayer perceptron (MLP) [40], support vector machine (SVM) [41], and Logistic Regression (LR) [42]. The MLP contains a fully connected DL architecture with different ways to improve generalization and prevent overfitting. SVM is a margin-based classifier that highly depends on data scaling and requires an accurate imputation approach for missing values to perform well. LR is a regression-based classifier that assumes linear relationships between features, which means that it is much more sensitive to datasets that have missing values and when data are not normalized [43–46].

Ensemble methods for both data distribution and missing values

The third category contains ML techniques that are based on ensemble learning methods [47]. Ensemble methods follow an approach to take full advantage of multiple learning methods and combine them to utilize the strengths of each individual method. Ensemble methods are moderately affected by the data distribution and missing values in the dataset. The three ML techniques that were included in this third category were extreme gradient boosting (XGBoost) [48], light gradient boosting machine (LightGBM), and CatBoost. While all three methods follow an optimized gradient-boosting approach, there are a few things that differentiate each from the others [49, 50]. XGBoost uses tree-based learning with regularization to reduce overfitting, LightGBM directly supports categorical features and is suitable for larger datasets, and CatBoost improves the model performance through efficient encoding techniques. Using all nine ML approaches that were categorized under three categories based on how they performed on different data distributions and missing values, the study trained models on each nine models and used a dataset that was processed using different strategies. The classification intention was for both binary and multiclass classification. Binary classification focuses on predicting PDAC and non-PDAC cases, while multiclass classification aims to distinguish the three classes.

Each classifier $f : \mathbb{R}^d \rightarrow \{0, 1\}$ was trained on a feature matrix $X : \epsilon\ \mathbb{R}^{n \times d}$ with labels $y \epsilon \{0, 1^n$. The models were trained using cross-validation and multiple preprocessing conditions. The feature matrix included both continuous biomarker levels and categorical demographic encodings. To formalize model comparisons, each classifier f_θ was optimized over the empirical risk:

$$\hat{\theta} = \min_{\theta} \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}(y_i, f_\theta(x_i)) \tag{1}$$

where $\mathcal{L}$ denotes the selected loss function.

Deep learning model architectures

Besides the classical ML modeling techniques, this study also included basic DL modeling for each type of classification. Both architectures have the same input layer that accepts eight features, but while the output layer used for the multiclassification has three neurons with a SoftMax activation function, the binary classification architecture output layer consists of a single neuron with a Sigmoid activation function. Both architectures as illustrated and used for DL approaches are presented in Fig. 3.

For both the binary and multiclass classification tasks, we implemented two-layer MLPs to predict PDAC from urinary biomarkers and demographic data. The input layer accepted eight features, representing biomarker measurements and demographic variables. The first hidden layer contained 128 neurons with ReLU activation, followed by batch normalization and dropout (0.5) to improve model generalization. The second hidden layer contained 64 neurons with ReLU activation, followed by dropout (0.5). The output layer differed by task: the binary classification model used a single neuron with Sigmoid activation to predict PDAC versus non-PDAC, while the multiclass model used three neurons with SoftMax activation to classify patients into three clinically relevant groups. Models were trained on 1000 epochs with a batch size of 32. The RMSprop optimizer (with a learning rate of 0.001) was used for the binary model, and the Adaptive Moment Estimation (Adam) optimizer (with a learning rate of 0.001) for the multiclass model. Loss functions were binary cross-entropy and sparse categorical cross-entropy, respectively. Input features were standardized prior to training to ensure numerical stability. Total trainable parameters were 9729 for the binary model and 9859 for the multiclass model, with 256 nontrainable parameters in both. This architecture balanced predictive performance and interpretability while enabling reproducible training for both classification tasks.

DL models were structured as MLPs. For input $X\ \epsilon\ \mathbb{R}^8$, the forward pass through a two-layer MLP was defined as:

$$h = \sigma\left(W_1 x + b_1\right) \tag{2}$$

$$\hat{y} = \varnothing\left(W_1 h^{(1)} + b_2\right) \tag{3}$$

where σ is ReLU activation, $\varnothing$ is Sigmoid (for binary) or Softmax (for multiclass) output activation, and x is the input vector from urine biomarkers and demographics. Loss function cross-entropy

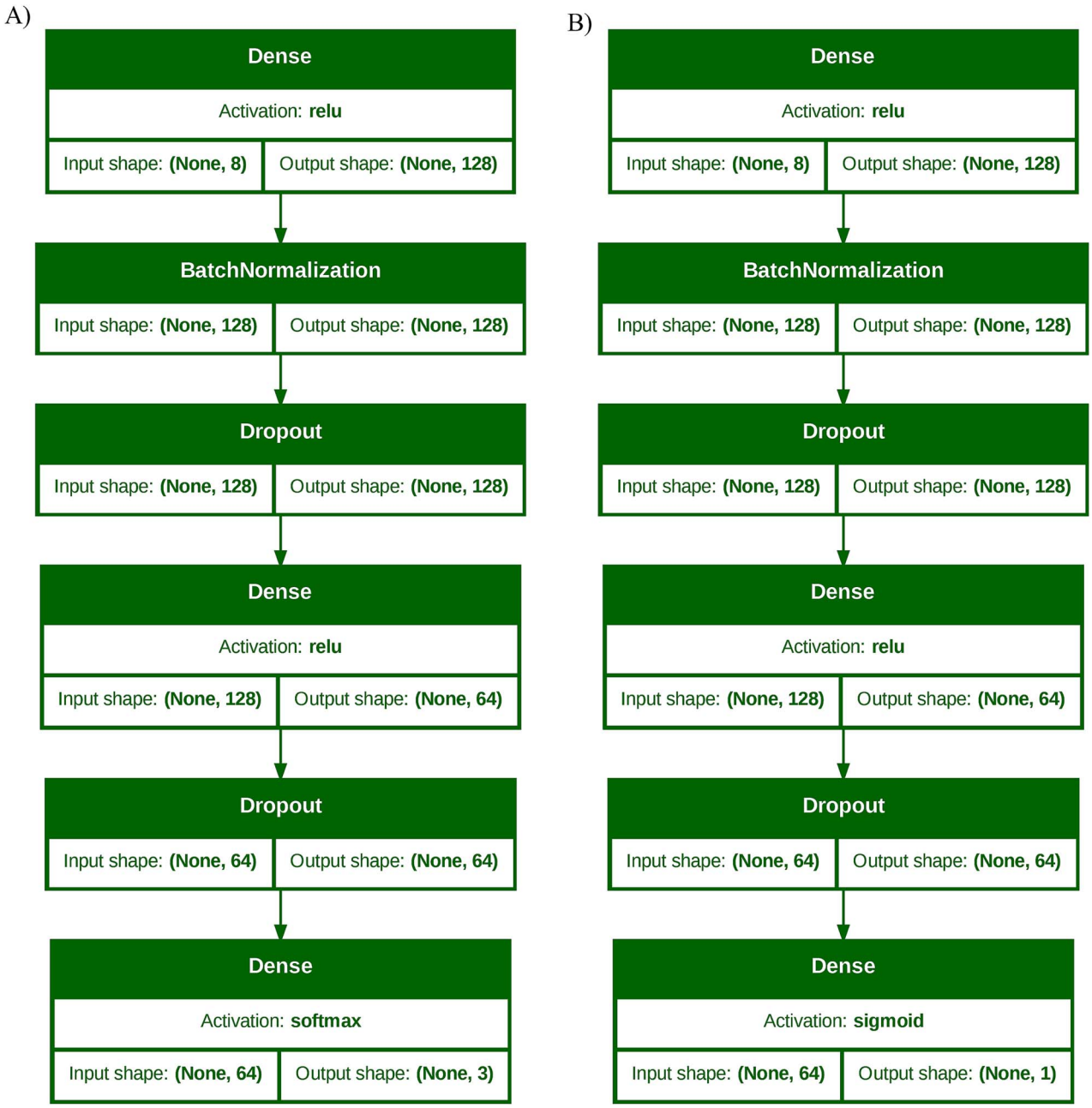

Figure 3. Deep learning model architectures used for multiclass and binary classification tasks, employing SoftMax and sigmoid activations, respectively.

for binary classification and categorical cross-entropy for multiclass classification were used.

$$\mathcal{L}_{\text{binary}} = -\frac{1}{n}\sum_{i=1}^{n}\left[y_i \log \hat{y}_i + (1 - y_i)\log\left(1 - \hat{y}_i\right)\right] \tag{4}$$

$$\mathcal{L}_{\text{multi}} = -\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K}\hat{y}_{ik}\log \hat{y}_{ik} \tag{5}$$

Preprocessing strategies

The effectiveness of preprocessing ensures that the data are clean and well structured and confirms the suitability of the data for the ML model. As one of the key aspects of this study was to conduct comparative approaches to find better approaches on each classification type using different parameters and address challenges associated with the dataset, the following key strategies were followed.

Data cleaning

Missing values in key biomarkers such as *REG1A* and *plasma CA19-9* were handled using four different imputation techniques, and then a comparative study was used to identify the ideal approach. Let $\mathbf{X} = [x_{ij}]$ represent the dataset, where x_{ij} denotes the jth feature of the ith sample. Missing values ($x_{ij} = NaN$) were handled using the following approaches:

The first approach is using KNN imputation. This method replaces missing values based on the mean value of the k-nearest neighbors, identified using a distance metric:

$$x_{ij} = \frac{1}{k}\sum_{l \,\epsilon\, N_k(i)} x_{lj} \tag{6}$$

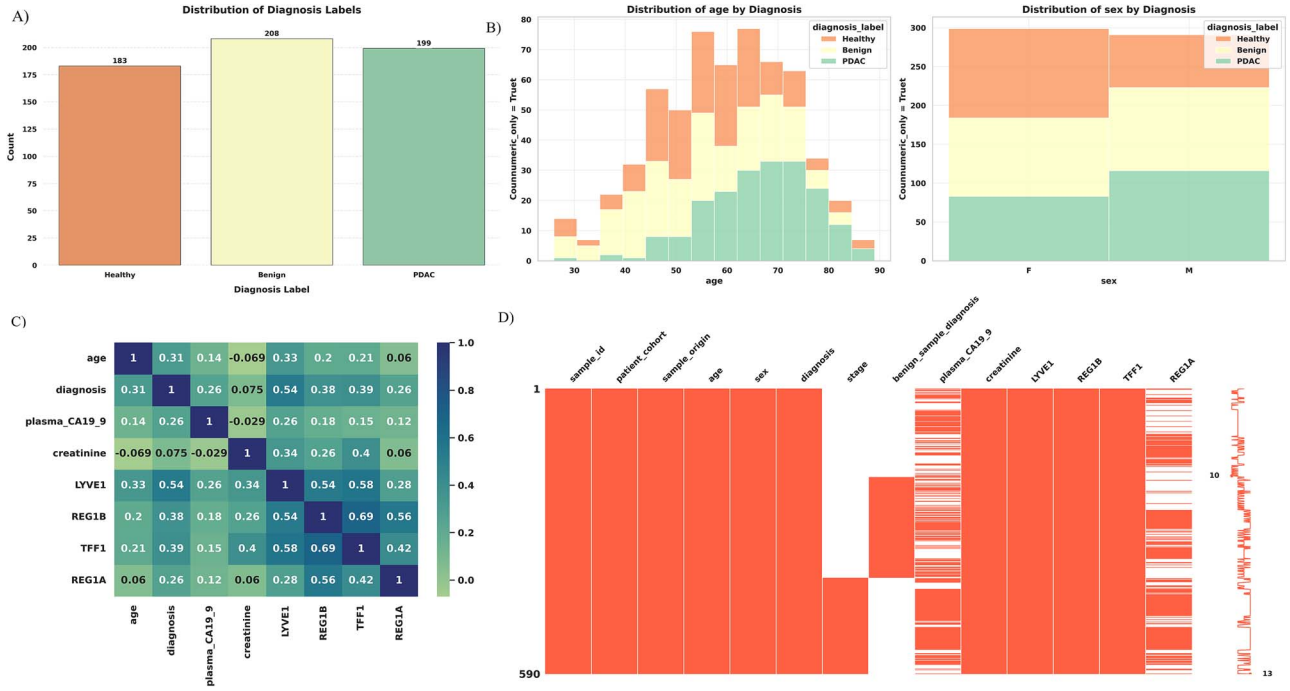

Figure 4. Exploratory data analysis results showing class distribution, demographic patterns, feature correlations, and missing value visualization.

The second approach used multivariate imputation by chained equation (MICE). This approach works by inputting the missing value iteratively by modeling each feature with missing values as a function of the other features. This method accounts for multivariate relationships in the data and reduces bias introduced by simpler imputation methods.

The third approach that was used for handling missing values replaced the missing value with the mean of the features across all samples:

$$\hat{x}_{ij} = \frac{1}{n}\sum_{i=1}^{n} x_{ij} \tag{7}$$

The last and fourth way that was used for benchmarking purposes was by just filling in missing value with 0 to see what happens to the model if no imputation is used.

Feature scaling and normalization

To assess the impact of normalization on the prediction models, the models were trained with and without data normalization. To normalize the data, this study used a log-transformation approach to stabilize variance and normalize skewed features. Log-transformation is represented as:

$$\hat{x}_{ij} = \log\left(x_{ij} + 1\right) \tag{8}$$

Feature engineering

Categorical variables are mapped to numerical variables as ML models were performed based on numerical values and the target diagnosis feature was also encoded to perform binary classification since the disease feature in the datasets has three classes. In the feature-engineered dataset, male values in sex features were represented by 1, and females were represented by 2. In the disease features, while the health class and benign class were encoded as 0 to show the absence of cancer, the rows that had PDAC were encoded as 1 to show the presence of cancer.

Train-test splits and cross-validation

To understand how the ML and DL techniques performed with different proportions, we split the dataset into training and test sets. For this comparative analysis, the dataset was partitioned into training and testing sets using three strategies 70/30, 75/25, and 80/20 to evaluate the stability of the models. To ensure the robust evaluation of each modeling approach and prevent overfitting, cross-validation was employed.

Single-cell analysis

To embed the single-cell data into a lower-dimensional space for visualization, Uniform Manifold Approximation and Projection (UMAP) was applied. UMAP minimizes cross-entropy between the high- and low-dimensional fuzzy simplicial sets:

$$\mathcal{L}_{\text{UMAP}} = \sum_{i \neq j} p_{ij} \log\left(\frac{p_{ij}}{q_{ij}}\right) + (1 - p_{ij}) \log\left(\frac{1 - p_{ij}}{1 - q_{ij}}\right) \tag{9}$$

where p_{ij}) and (q_{ij}) represent pairwise relationship in high and low dimensions, respectively.

Evaluation metrics

To evaluate the performances of the ML and DL prediction models, multiple metrics were employed to ensure the assessment of the models from different perspectives. Let $\boldsymbol{y}$ represents the true labels and $\hat{\boldsymbol{y}}$ the predicted labels for a dataset of size $\boldsymbol{n}$. The evaluation metrics we used are defined as follows:

A) Accuracy: The proportion of how many labels are correctly predicted from the samples:

$$Accuracy = \frac{\sum_{i=1}^{n} 1\left(y_i = \hat{y}_i\right)}{n} \tag{10}$$

B) Precision: The fraction of true positives (TPs) among all predicted positives:

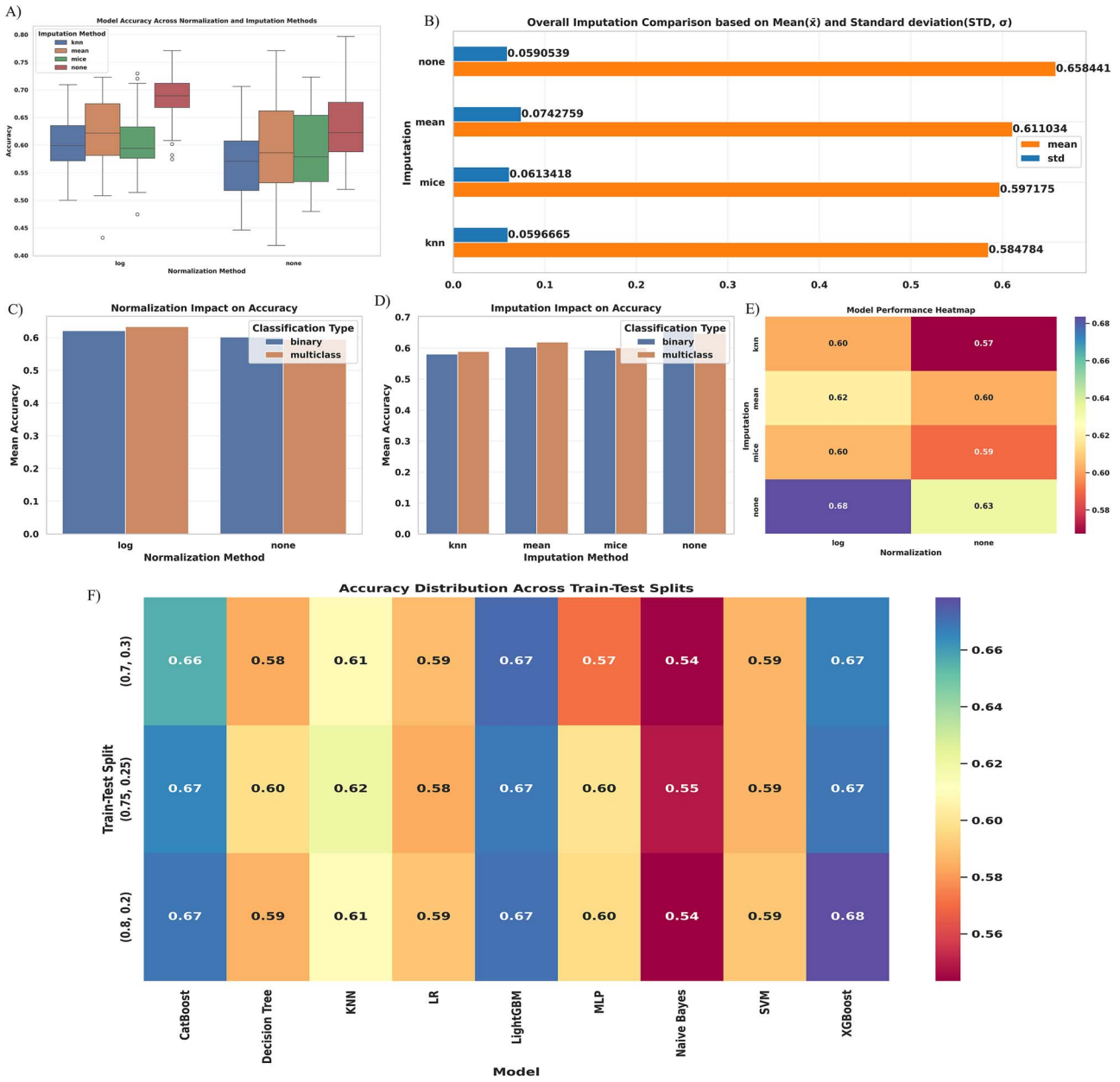

Figure 5. Comparative analysis of preprocessing strategies illustrating the effects of normalization, imputation methods, and train–test split strategies on model performance.

$$Precision = \frac{TP}{TP + FP} \tag{11}$$

C) Recall (Sensitivity/True Positive Rate): The division of TPs among all actual positives:

$$Recall = \frac{TP}{TP + FN} \tag{12}$$

D) F1-Score: The harmonic means of precision and recall:

$$F1 - Score = 2 \bullet \frac{Precision \bullet Recall}{Precision + Recall} \tag{13}$$

A) Confusion Matrix: A matrix representation of the TPs, false positives (FPs), true negatives (TNs), and false negatives (FNs):

$$C = \begin{bmatrix} TP & FP \\ FN & TN \end{bmatrix} \tag{14}$$

In addition to accuracy, we evaluated model performances using the area under the receiver operating characteristic (ROC) curve (AUROC), sensitivity, and specificity for both the binary and multiclass tasks. ROC curves with mean AUROC values were computed using five-fold cross-validation, and sensitivity/specificity were reported to align with clinical diagnostic evaluation standards.

Results

This section presents the findings of this study by dividing them into different aspects based on the results. The first subsection presents results related to the EDA. The distribution of the data, missing values, and other key insights about the data are included.

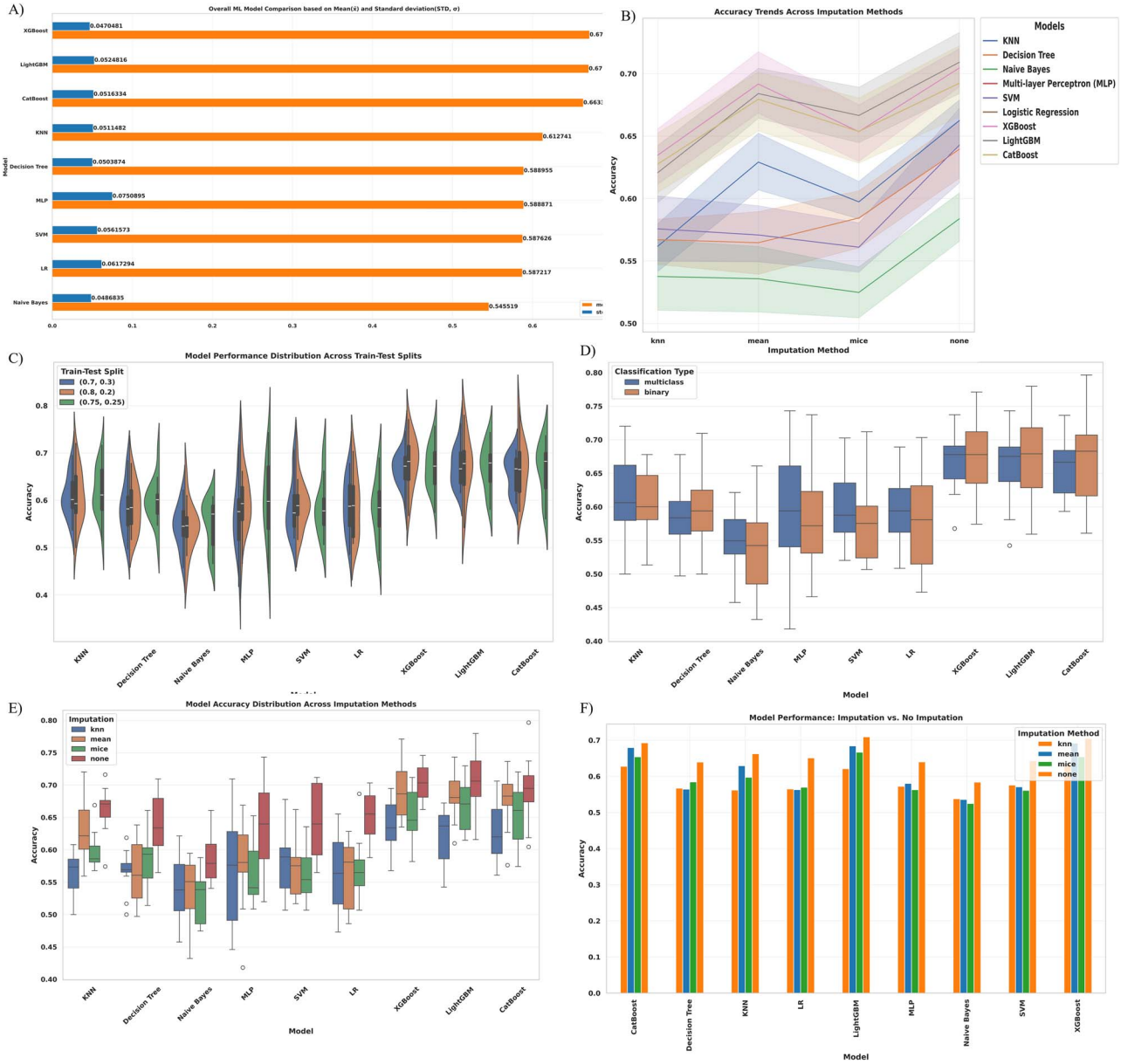

Figure 6. Comparative analysis of machine learning techniques illustrating accuracy trends, variability, data-splitting effects, and impacts of imputation methods (A–F).

Then results of the comparative analysis from the preprocessing strategies are presented. The following two sections present results related to binary classification and multiclass classification, respectively, using ML and DL. The last subsection provides results related to the significance of each biomarker using the scRNA-seq analysis.

Exploratory data analysis

A detailed analysis of the dataset structure was used to understand the distribution of biomarkers and demographic features in order to highlight the skewness of some features and which features have missing values. Correlation matrices were generated to examine the relationship of each numerical feature included in the dataset. Figure 4 summarizes results of the EDA. Further results of the EDA are provided in Supplementary Fig. S1.

Comparative study on preprocessing strategies

The preprocessing strategies described for the comparative analysis were implemented to comprehensively evaluate the trained models, aiming to determine which strategy performed most effectively and which classification type was most suitable for accurate cancer prediction. Notably, log normalization performed well rather than data normalization in almost all approaches for imputing missing values. The normalization impact was almost the same for both types of classification. For the included approaches for imputing missing value-based comparisons, from the average of all of the trained models, the well-known kNN and MICE performances were poor, while the simple imputation approach using 0 and mean values provided better performances. The three types of training-test split strategies revealed no great impacts, and a maximum of 2% accuracy variance was observed

Table 1. Top 10 best and worst performing ML modes the preprocessing strategies

Rank	Type	Model	Accuracy	Classification type	Imputation	Normalize
1	Best	CatBoost	0.79661	binary	none	none
2	Best	LightGBM	0.779661	binary	none	none
3	Best	LightGBM	0.771186	binary	none	log
4	Best	XGBoost	0.771186	binary	mean	none
5	Best	LightGBM	0.762712	binary	none	none
6	Best	XGBoost	0.756757	binary	mean	none
7	Best	XGBoost	0.745763	binary	none	none
8	Best	Neural Network (MLP)	0.743243	multiclass	none	log
9	Best	LightGBM	0.743243	multiclass	mean	none
10	Best	Neural Network (MLP)	0.737288	binary	none	log
1	Worst	Neural Network (MLP)	0.418079	multiclass	mean	none
2	Worst	Naive Bayes	0.432203	binary	mean	log
3	Worst	Neural Network (MLP)	0.445946	multiclass	knn	none
4	Worst	Naive Bayes	0.457627	multiclass	knn	none
5	Worst	Neural Network (MLP)	0.466102	binary	knn	none
6	Worst	Naive Bayes	0.466216	binary	mean	none
7	Worst	Linear Regression	0.472973	binary	knn	none
8	Worst	Naive Bayes	0.472973	binary	knn	none
9	Worst	Neural Network (MLP)	0.472973	binary	knn	none
10	Worst	Naive Bayes	0.474576	binary	mice	log

on the DT, while other ML approaches showed only 1% variance. Figure 5 summarizes results of the comparative analysis of the preprocessing strategies. Further model comparison results are provided in Supplementary Fig. S2.

Classical machine learning results

Binary classification models demonstrated various levels of performance depending on the imputation strategies and normalization techniques. Among other ML techniques, CatBoost achieved the highest accuracy with 79.66% accuracy, while NB demonstrated the lowest accuracy of 43.22% when mean imputation was used and log normalization was applied. On the other hand, for multiclass classification, LightGBM emerged as the top-performing model for multiclass classification, achieving an accuracy of 74.32% when imputed data were combined with no normalization, and MLP achieved the lowest accuracy of 41.81% when mean imputation and no normalization were applied. A comparative analysis of each ML technique using different aspects is presented in Fig. 6. Further results on the best and worst models are provided in Supplementary Fig. S3. The 10 best-performing ML approaches and preprocessing strategies are presented in Table 1. The top 10 best and worst performing ML approaches on each binary and multiclass classification were determined. Further numerical values for the best and worst model performances for binary class and multiclass classifications are respectively provided in Supplementary Tables S3 and S4.

Deep learning results

Compared to classical ML techniques, DL approaches yielded substantially higher performances in binary classification but showed limited improvements in multiclass classification. The DL model achieved 91% accuracy in binary classification, with dropout layers effectively mitigating overfitting. For multiclass classification, the model achieved 71% accuracy, where batch normalization facilitated better convergence. Furthermore, the ROC curve for binary classification with five-fold cross-validation showed a mean AUROC of 0.93, while in the multiclass setting, mean AUROC values of 0.86, 0.76, and 0.93 were, respectively, achieved for classes 0, 1, and 2. Results of the DL models for both classification tasks are presented in Fig. 7 and Supplementary Fig. S4.

As shown in Supplementary Fig. S4, the biomarkers demonstrated strong predictive potential in the binary classification model, with minimal variability (±1 SD) across k-folds. This reflects both high predictive power and robustness, independent of specific data partitions. In the multiclass model, urinary biomarkers maintained consistently high importance across PDAC subtypes with low variability, highlighting their value as generalizable biomarkers. The stability of their contributions across multiple subtypes further suggests diagnostic utility in diverse clinical contexts and supports the hypothesis that *REG1A* plays a central role in PDAC pathophysiology rather than being limited to subtype-specific mechanisms.

Single-cell RNA sequencing insights

Once scRNA data were processed to ensure quality control, different approaches were followed to find insights to support the biomarkers we used for the ML-based PDAC prediction model from urine and demographical data. A differential expression analysis highlighted *REG1A* as one of the top three most expressed genes in PDAC tissues. We used different visualization approaches, such as heatmap, UMAP, violin, and feature plots, to examine which biomarkers had higher expression. A summary of the scRNA-seq results is presented in two categories: Fig. 8 presents the PCA, clustering, cluster annotation, annotation distribution, and top 10 highly expressed genes, while Fig. 9 presents violin and feature plots showing expressions of urine biomarkers in PDAC tissues.

Bulk-RNA insights from TCGA

To further evaluate the predictive significance of the selected biomarkers, bulk RNA-seq data from TCGA were analyzed for both gene expression patterns and survival outcomes. As shown in Fig. 10, TFF1 exhibited significantly higher expression in tumor tissues compared with normal samples. Although REG1A and

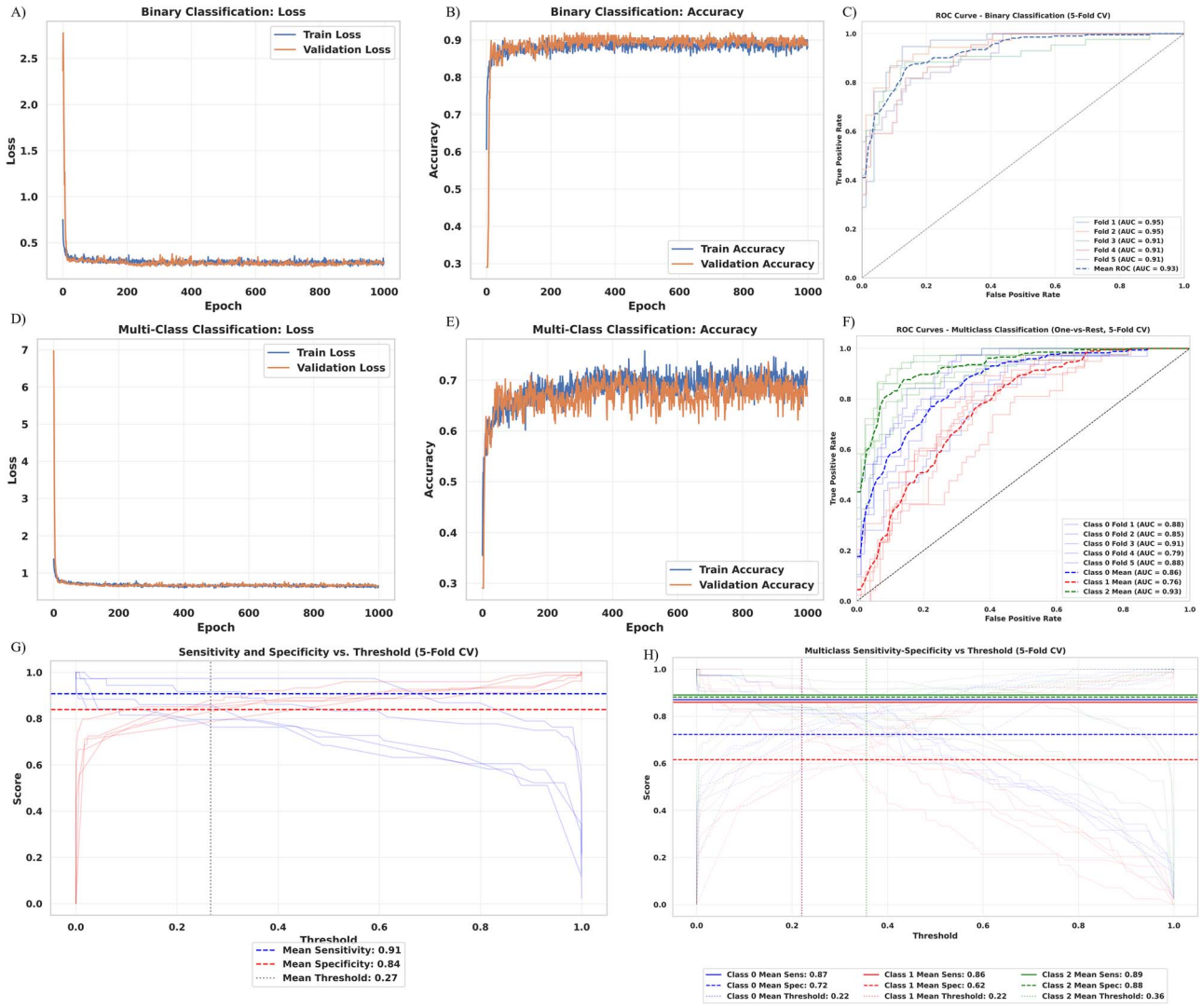

Figure 7. Deep learning–based prediction modeling results for PDAC showing training and validation performance, accuracy trends, ROC–AUROC metrics, and sensitivity–specificity evaluations for both binary and multiclass classifications.

REG1B did not show significant overexpression in tumor tissues, they were found to be highly expressed in normal tissues. This differential expression pattern suggests that their absence in tumors could serve as a potential predictive indicator. Furthermore, the survival analysis revealed that both REG1A and REG1B were strongly associated with patient prognoses, underscoring their relevance in survival prediction and risk stratification.

Discussion

The findings of this study offer valuable perspectives into the integration of scRNA-seq data insights to assess significant biomarkers utilized for PDAC predictions using ML. This study bridges the gap between existing research and highlights the significance of overlooked biomarkers and demographic features. The study demonstrated how all urinary biomarkers and demographic features can be included in ML predictive modeling to add an extra layer of confidence. Beyond the unique contributions of these features, the scRNA-seq analysis also reveals how individual urinary biomarkers are differentially expressed across distinct cell types [51–53].

This finding underscores the potential clinical relevance of urinary biomarkers in PDAC detection and their utility as diagnostic indicators. Our interpretability analysis (Supplementary Fig. S4) highlighted the included urinary biomarkers as key drivers of model predictions. Additionally, integrating demographic features enabled the model to extract insights from multiple perspectives. Previous studies primarily focused on a limited set of biomarkers, achieving high accuracies but neglecting the broader biological context. While this study's ML models achieved slightly lower predictive accuracies compared with prior models, the inclusion of all biomarkers and scRNA-seq data provided deeper insights into biomarker significance and model interpretability.

The proposed framework offers several advantages over traditional methods. First, it integrates scRNA-seq data to uncover the biological relevance of biomarkers that were previously overlooked. Second, it combines demographic features with molecular data, enabling a more- personalized approach to PDAC prediction. Third, the use of advanced preprocessing strategies, such as imputation and normalization, ensured robust model performance even with incomplete datasets. These strengths make the approach adaptable to other complex diseases where multidimensional data integration is required.

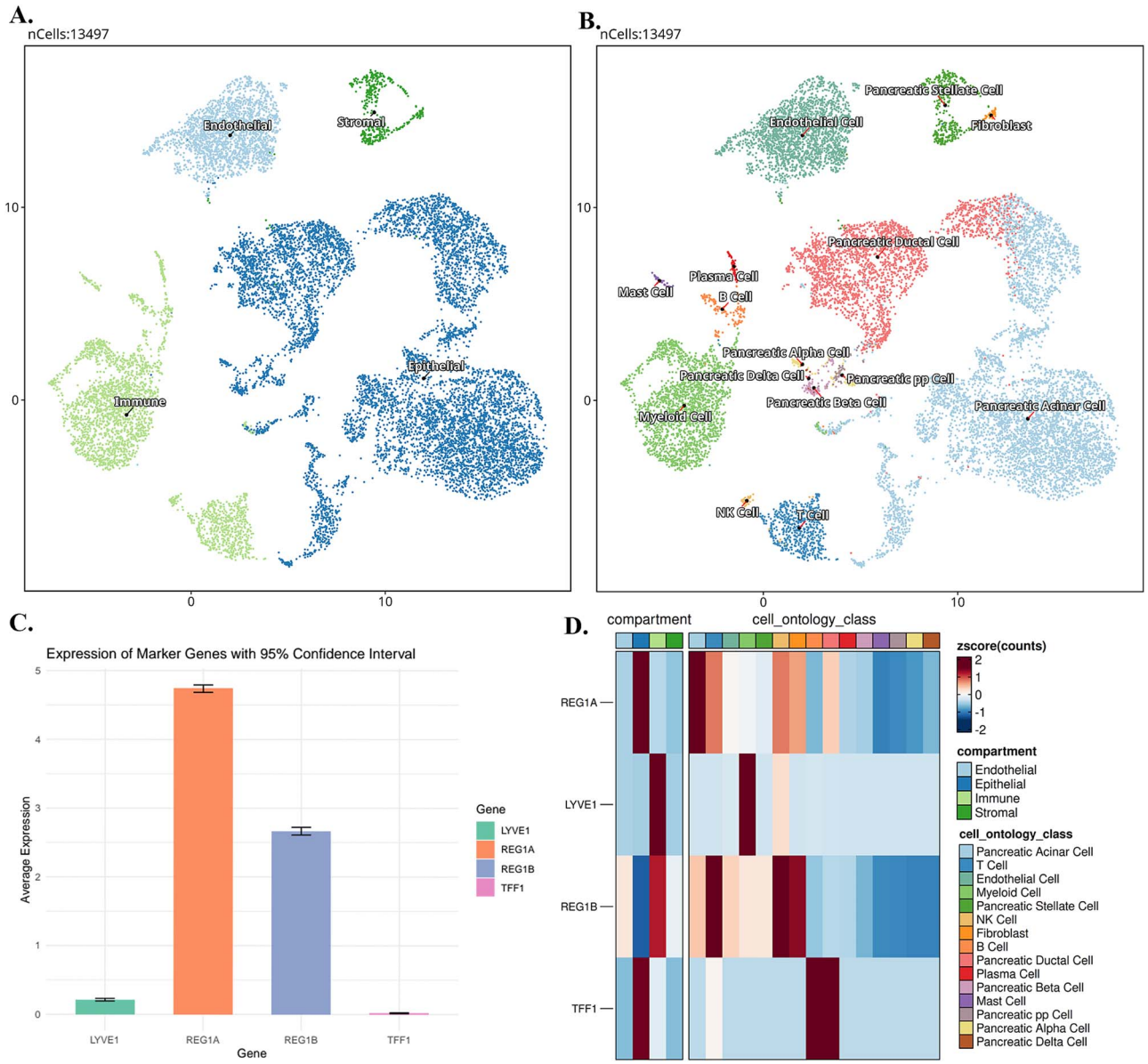

Figure 8. scRNA-seq analysis results illustrating cellular clustering, annotated cell-type distribution, and differential expression of urinary biomarkers across cell groups and types.

Before clinical implementation, prospective validation in multicenter cohorts will be necessary to confirm the robustness, generalizability, and reproducibility. Practical considerations include standardized sample collection protocols and handling, and biomarker assay consistency to ensure reliable input data for the models. Target performance metrics, such as sensitivity and specificity, should align with clinically actionable thresholds, ideally exceeding 85% to support screening and early detection. Regulatory and ethical compliance, including informed consent, data privacy, and reporting transparency, will be critical for translation, particularly when integrating molecular and demographic data for personalized predictions. By providing interpretable feature importance scores, this approach also facilitates clinical trust, enabling clinicians to understand which biomarkers drive predictions and how these align with the underlying biology, thereby advancing both model interpretability and clinical relevance beyond prior studies. Additionally, regulatory approval and ethical considerations, including patient consent, data privacy, and secure handling of molecular and demographic data, must be addressed before clinical deployment. While promising, real-world application may be limited by variabilities in biomarker expression across populations, assay availability, and the need for trained personnel to implement the predictive pipeline.

Similar to how cerebrospinal fluid metabolomic profiling identified vasoactive metabolites predictive of poor outcomes after brain injury [54], integrating urinary biomarkers in PDAC can reveal subtle but clinically relevant molecular signals that enhance predictive modeling [55–57]. Moreover, ML-driven integration of multi-omics datasets, including genomics, transcriptomics, proteomics, and metabolomics, was shown to improve personalized predictions and biomarker discovery in complex diseases, supporting the inclusion of diverse molecular and demographic features in PDAC modeling [58–60]. Finally, DL frameworks applied to single-cell multi-omics data can reconstruct cell-type-specific regulatory networks, highlighting the potential of scRNA-seq to uncover biologically meaningful

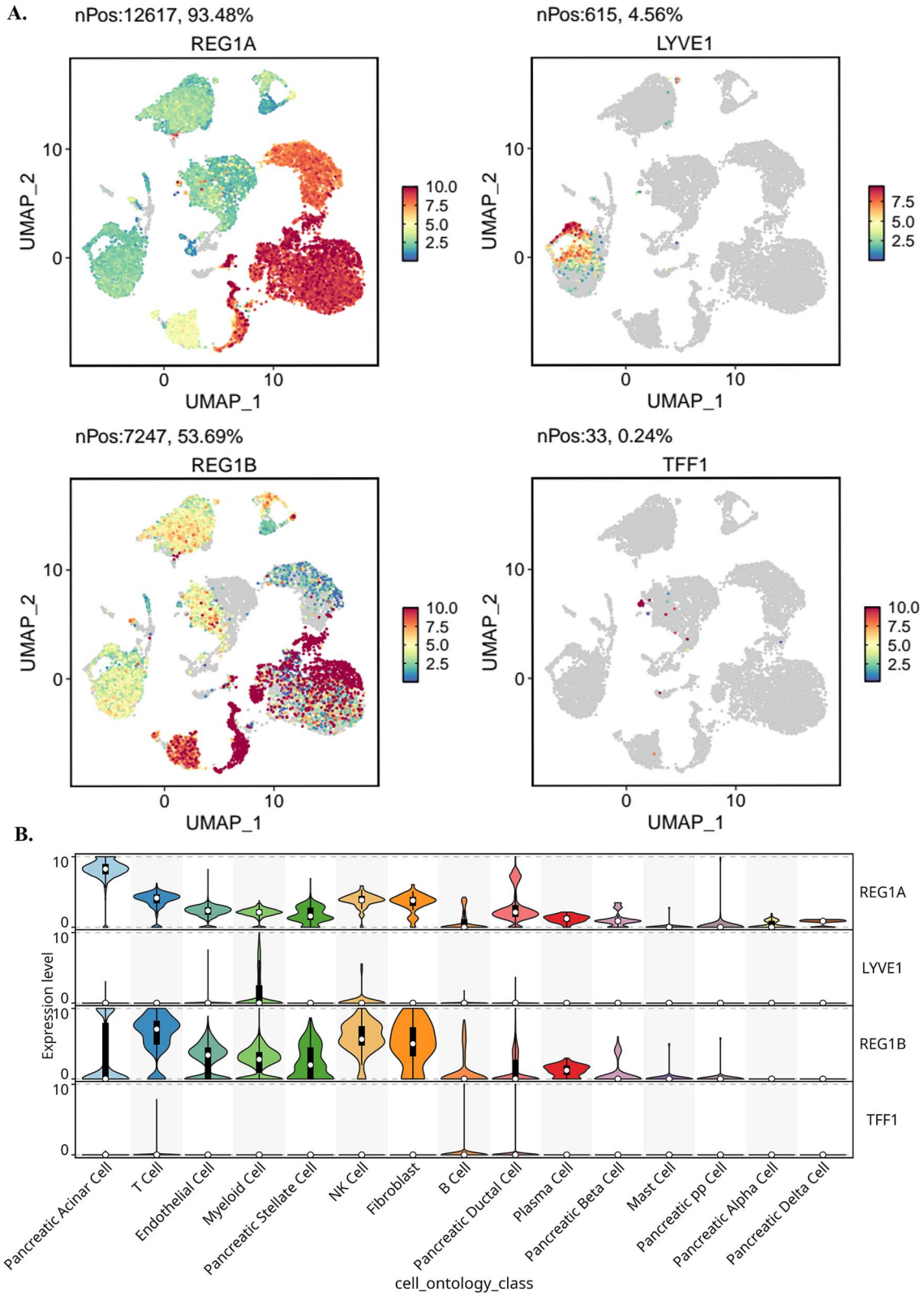

Figure 9. Comparative analysis of urinary biomarker expression from scRNA-seq data illustrating spatial distribution across cell clusters and expression variability among biomarkers.

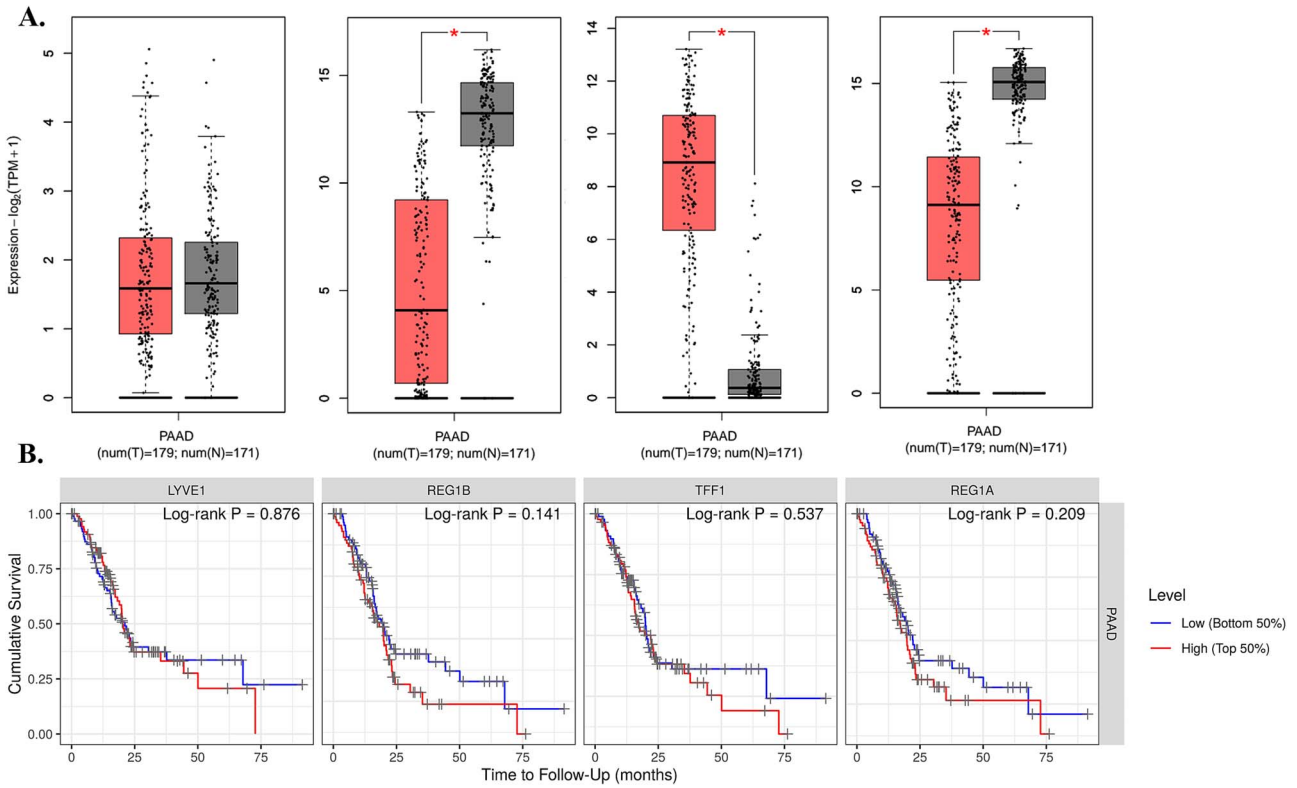

Figure 10. Expression and prognostic analysis of candidate genes in PDAC using TCGA data, showing differential expression between tumor and normal tissues and survival associations based on gene expression levels.

biomarkers that might be overlooked by standard analyses [61]. Together, these studies provide a strong rationale for integrating molecular, single-cell, and demographic data to achieve more interpretable and biologically informed predictive models for PDAC.

Despite its strengths, the study has certain limitations. The predictive accuracy of the models was slightly lower than that of some existing methods, likely due to the complexity introduced by including many biomarkers and demographic variables. Additionally, the computational demands of processing scRNA-seq data and training DL models were significant, requiring substantial resources. Another limitation was the reliance on publicly available datasets, which might not fully represent the heterogeneity of real-world PDAC cases, which may have introduced sampling biases and might not fully capture the heterogeneity of real-world PDAC cases, thereby limiting the generalizability of the models. Future studies incorporating larger, prospectively collected, and clinically diverse cohorts will be critical to improve the external validity. Tree-based gradient boosting methods often outperform shallow MLPs on small, low-dimensional tabular datasets because they capture complex feature interactions without large sample requirements. In our multiclass experiments, a class imbalance and limited feature dimensionality likely hampered the MLP. Thus, future work should consider attention/transformer architectures only when larger or multimodal datasets are available. Attention mechanisms can better model feature interactions and variable-length inputs and may improve performances when higher-dimensional inputs (full transcriptomes and radiomics) and larger sample sizes are available [62, 63].

The findings of this study have important implications for clinical practice. By establishing the significance of urinary biomarkers, this work provides a foundation for developing more inclusive and effective diagnostic panels for PDAC. The integration of demographic features also highlights the need for personalized prediction models that account for patient-specific variables. These advancements could lead to earlier and more-accurate PDAC detection, ultimately improving patient outcomes. In summary, this study demonstrated the potential of integrating scRNA-seq data and ML for PDAC predictions. By addressing key gaps in existing research, it provides a foundation for more comprehensive and interpretable prediction frameworks that can be applied to other complex diseases.

Conclusion

This study presents a comparative-based perspective towards predicting PDAC from urinary biomarkers and demographic features along with insights on the significance of the biomarkers we used through an scRNA-seq data analysis. The findings highlighted that feature importance analyses for ML approaches integrating such cutting-edge technologies such as scRNA-seq have great promise in improving model explainability and investigating the impacts of those biomarker features. Furthermore, the findings underscore the importance of combining ML with bioinformatics approaches which can help address key gaps in current research pipelines, and it will open new avenues for systematic ways for biomarker discovery and disease detection.

The integration of the scRNA-seq data analysis provided the significance of each biomarker in the dataset. By utilizing all available urinary biomarkers and demographic features such as gender and age of the patient, we conducted several approaches to determine better preprocessing strategies and which classification type was more suitable. The comparative study revealed that

for binary classifications, a DL approach is suitable and outperformed other techniques by achieving 91% accuracy. On the other hand, for multiclass classification, the higher accuracy achieved was 87% by XGBoost, and it was observed that ML was more suitable than the DL approach for multiclass classification. While slightly lower accuracy was achieved compared with prior studies that only focused on limited biomarkers and did not include demographic data, the added interpretability, demographic factors, and biological insights provided by this approach highlight its clinical relevance.

Future work based on this study will focus on conducting more-comprehensive analyses using multiple datasets, expanding the study preprocessing strategies to confirm the findings, exploring advanced DL architectures, such as transformers and attention-based models, to further improve predictive accuracy, and finally identifying a way to translate these findings into real-world practical diagnostic tools [64–66]. In conclusion, this study represents a step forward in PDAC predictions by integrating molecular, demographic, and computational insights into a unified framework. By prioritizing interpretability and inclusivity, it lays the foundation for more robust and clinically relevant prediction models, advancing the field of precision oncology.

Key Points

- We achieved 91% accuracy in pancreatic ductal adenocarcinoma (PDAC) predictions using a deep learning model that integrated urinary biomarkers and demographic data, surpassing all classical machine learning approaches.
- We identified *REG1A* as one of the top three most highly expressed genes in PDAC tissues through a single-cell RNA sequencing analysis, underscoring its diagnostic relevance despite its prior neglect.
- We established a biologically interpretable framework by combining molecular biomarkers, demographic features, and advanced computational modeling, paving the way for early and personalized PDAC detection.

Acknowledgements

We gratefully acknowledge the financial support by the "Taipe Medical Univeristy (TMU) Research Center of Cancer Translational Medicine" from The Featured Areas Research Center Program within the framework of the Higher Education Sprout Project by the Ministry of Education (MOE) in Taiwan.

Supplementary data

Supplementary data is available at *Briefings in Bioinformatics* online.

Conflict of interest: None declared.

Funding

This research was funded by the Ministry of Science and Technology (MOST), grant from MOST111-2314-B-038-105-MY3, and National Science and Technology Council (NSTC) 112-2320-B-038-056, 112-2320-B-038-059, 113-2320-B-038-011, 113-2320-B-038-014; 113-2320-B-393-001, 114-2320-B-393-003, 114-2320-B-393-004, 114-2320-B-038-004, 114-2811-B-038-046, and 114-2314-B-038-133-MY3. The Chi Mei Medical Center, grant no 113CM-TMU-08; Taipei Medical University, grant no TMU112-AE2-I16-4; and the Health and Welfare Surcharge of Tobacco Products (WanFang Hospital, Chi-Mei Medical Center, and Hualien Tzu-Chi Hospital Joint Cancer Center Grant - Focus on Colon Cancer Research), grant numbers MOHW110-TDU-B-212-144020, MOHW111-TDU-B-221-014013, and MOHW112-TDU-B-221-124013.

Data availability

The tabular dataset used for the ML prediction model is publicly available from the original study. The single-cell RNA-seq dataset used for biomarker significance analysis is publicly accessible from the Gene Expression Omnibus (GEO) under accession no. GSE274665.

References

1. Wei H, Ren H. Precision treatment of pancreatic ductal adenocarcinoma. *Cancer Lett* 2024;**585**:216636–6. https://doi.org/10.1016/j.canlet.2024.216636
2. George B, Kudryashova O, Kravets A. *et al.* Transcriptomic-based microenvironment classification reveals precision medicine strategies for pancreatic ductal adenocarcinoma. *Gastroenterology* 2024;**166**:859–871.e853. https://doi.org/10.1053/j.gastro.2024.01.028
3. Zhang S, Fang W, Zhou S. *et al.* Single cell transcriptomic analyses implicate an immunosuppressive tumor microenvironment in pancreatic cancer liver metastasis. *Nat Commun* 2023;**14**:1–19.
4. Ko CC, Yang PM. Hypoxia-induced MIR31HG expression promotes partial EMT and basal-like phenotype in pancreatic ductal adenocarcinoma based on data mining and experimental analyses. *J Transl Med* 2025;**23**:305.
5. Ko CC, Hsieh YY, Yang PM. Long non-coding RNA MIR31HG promotes the transforming growth factor β-induced epithelial-mesenchymal transition in pancreatic ductal adenocarcinoma cells. *Int J Mol Sci* 2022;**23**:6559. https://doi.org/10.3390/ijms23126559
6. Yokoyama S, Hamada T, Higashi M. *et al.* Predicted prognosis of patients with pancreatic cancer by machine learning. *Clin Cancer Res* 2020;**26**:2411–21.
7. Karar ME, El-Fishawy N, Radad M. Automated classification of urine biomarkers to diagnose pancreatic cancer using 1-D convolutional neural networks. *J Biol Eng* 2023;**17**:1–12.
8. Javaid M, Haleem A, Pratap Singh R. *et al.* Significance of machine learning in healthcare: features, pillars and applications. *Int J Intell Netw* 2022;**3**:58–73. https://doi.org/10.1016/j.ijin.2022.05.002
9. Shamshirband S, Fathi M, Dehzangi A. *et al.* A review on deep learning approaches in healthcare systems: Taxonomies, challenges, and open issues. *J Biomed Inform* 2021;**113**:103627.
10. Habehh H, Gohel S. Machine learning in healthcare. *Curr Genomics* 2021;**22**:291–300. https://doi.org/10.2174/1389202922666210705124359
11. Dhiman G, Juneja S, Viriyasitavat W. *et al.* A novel machine-learning-based hybrid CNN model for tumor identification in medical image processing. *Sustainability (Switzerland)* 2022;**14**:1447–7.

12. Ahsan MM, Siddique Z. Machine learning-based heart disease diagnosis: a systematic literature review. *Artif Intell Med* 2022;**128**:102289–9. https://doi.org/10.1016/j.artmed.2022.102289
13. Zheng J, Yu Z. A novel machine learning-based systolic blood pressure predicting model. *J Nanomater* 2021;**2021**: 9934998, 8 pages. https://doi.org/10.1155/2021/9934998
14. Kim WP, Kim HJ, Pack SP. *et al.* Machine learning–based prediction of attention-deficit/hyperactivity disorder and sleep problems with wearable data in children. *JAMA Netw Open* 2023;**6**:e233502–2. https://doi.org/10.1001/jamanetworkopen.2023.3502
15. Vatansever S, Schlessinger A, Wacker D. *et al.* Artificial intelligence and machine learning-aided drug discovery in central nervous system diseases: state-of-the-arts and future directions. *Med Res Rev* 2021;**41**:1427–73. https://doi.org/10.1002/med.21764
16. Elbadawi M, Gaisford S, Basit AW. Advanced machine-learning techniques in drug discovery. *Drug Discov Today* 2021;**26**:769–77. https://doi.org/10.1016/j.drudis.2020.12.003
17. Paraiso HC, Wang X, Kuo PC. *et al.* Isolation of mouse cerebral microvasculature for molecular and single-cell analysis. *Front Cell Neurosci* 2020;**14**:522647–7.
18. Kim N, Kim HK, Lee K. *et al.* Single-cell RNA sequencing demonstrates the molecular and cellular reprogramming of metastatic lung adenocarcinoma. *Nat Commun* 2020;**11**:1–15.
19. Slovin S, Carissimo A, Panariello F. *et al.* Single-cell RNA sequencing analysis: a step-by-step overview. *Methods Mol Biol* 2021;**2284**: 343–65.
20. Su BH, Kumar S, Cheng LH. *et al.* Multi-omics profiling reveals PLEKHA6 as a modulator of β-catenin signaling and therapeutic vulnerability in lung adenocarcinoma. *Am J Cancer Res* 2025;**15**: 3106–27. https://doi.org/10.62347/NVVF8441
21. Jovic D, Liang X, Zeng H. *et al.* Single-cell RNA sequencing technologies and applications: a brief overview. *Clin Transl Med* 2022;**12**:e694–4.
22. Kumar S, Wu CC, Wulandari FS. *et al.* Integration of multi-omics and single-cell transcriptome reveals mitochondrial outer membrane protein-2 (MTX-2) as a prognostic biomarker and characterizes ubiquinone metabolism in lung adenocarcinoma. *J Cancer* 2025;**16**:2401–20. https://doi.org/10.7150/jca.106902
23. Wu YJ, Chiao CC, Chuang PK. *et al.* Comprehensive analysis of bulk and single-cell RNA sequencing data reveals Schlafen-5 (SLFN5) as a novel prognosis and immunity. *Int J Med Sci* 2024;**21**: 2348–64. https://doi.org/10.7150/ijms.97975
24. Luo ZG, Peng J, Li T. Single-cell RNA sequencing reveals cell-type-specific mechanisms of neurological diseases. *Neurosci Bull* 2020;**36**:821–4. https://doi.org/10.1007/s12264-020-00496-5
25. Zhang MJ, Hou K, Dey KK. *et al.* Polygenic enrichment distinguishes disease associations of individual cells in single-cell RNA-seq data. *Nat Genet* 2022;**54**:1572–80. https://doi.org/10.1038/s41588-022-01167-z
26. Yang Y, Cheng F. Artificial intelligence streamlines scientific discovery of drug-target interactions. *Br J Pharmacol* 2025;**22**:1–18. https://doi.org/10.1111/bph.17427
27. Yang Y, Qiu Y, Hu J. *et al.* A deep learning framework combining molecular image and protein structural representations identifies candidate drugs for pain. *Cell Rep Methods* 2024;**4**:100865. https://doi.org/10.1016/j.crmeth.2024.100865
28. Matsuoka T, Yashiro M. Bioinformatics analysis and validation of potential markers associated with prediction and prognosis of gastric cancer. *Int J Mol Sci* 2024;**25**:5880.
29. Li CY, Anuraga G, Chang CP. *et al.* Repurposing nitric oxide donating drugs in cancer therapy through immune modulation. *J Exp Clin Cancer Res* 2023;**42**:22.
30. Hagerling C, Gonzalez H, Salari K. *et al.* Immune effector monocyte-neutrophil cooperation induced by the primary tumor prevents metastatic progression of breast cancer. *Proc Natl Acad Sci USA* 2019;**116**:21704–14. https://doi.org/10.1073/pnas.1907660116
31. Lawson DA, Bhakta NR, Kessenbrock K. *et al.* Single-cell analysis reveals a stem-cell program in human metastatic breast cancer cells. *Nature* 2015;**526**:131–5. https://doi.org/10.1038/nature15260
32. Xuan DTM, Yeh IJ, Liu HL. *et al.* A comparative analysis of Marburg virus-infected bat and human models from public high-throughput sequencing data. *Int J Med Sci* 2025;**22**:1–16. https://doi.org/10.7150/ijms.100696
33. Anuraga G, Lang J, Xuan DTM. *et al.* Integrated bioinformatics approaches to investigate alterations in transcriptomic profiles of monkeypox infected human cell line model. *J Infect Public Health* 2024;**17**:60–9. https://doi.org/10.1016/j.jiph.2023.10.035
34. Anuraga G, Wang WJ, Phan NN. *et al.* Potential prognostic biomarkers of NIMA (never in mitosis, gene a)-related kinase (NEK) family members in breast cancer. *J Pers Med* 2021;**11**:1089. https://doi.org/10.3390/jpm11111089
35. Radon TP, Massat NJ, Jones R. *et al.* Identification of a three-biomarker panel in urine for early detection of pancreatic adenocarcinoma. *Clin Cancer Res* 2015;**21**:3512–21. https://doi.org/10.1158/1078-0432.CCR-14-2467
36. Veghini L, Pasini D, Fang R. *et al.* Differential activity of MAPK signalling defines fibroblast subtypes in pancreatic cancer. *Nat Commun* 2024;**15**:1–20.
37. Kramer O. K-Nearest Neighbors. In: Kramer O. (ed.), *Dimensionality Reduction with Unsupervised Nearest Neighbors*. Berlin, Heidelberg: Springer Berlin Heidelberg; 2013, 13–23.
38. Quinlan JR. Induction of decision trees. *Mach Learn* 1986;**1**: 81–106. https://doi.org/10.1023/A:1022643204877
39. Webb GI. Naïve Bayes. In: Sammut C, Webb GI (eds.), *Encyclopedia of Machine Learning and Data Mining*. Boston, MA: Springer US; 2016, 1–2.
40. Haykin S. *Neural Networks: A Comprehensive Foundation*. 2nd ed. Upper Saddle River (NJ): Prentice Hall; 1999.
41. Shihong Y, Ping L, Peiyi H. SVM classification: its contents and challenges. *Appl Math* 2003;**18**:332–42.
42. Liaw A, Wiener M. Classification and regression by randomForest. *R News* 2002;**2**:18–22.
43. Chuang PK, Chang KF, Chang CH. *et al.* Comprehensive bioinformatics analysis of glycosylation-related genes and potential therapeutic targets in colorectal cancer. *Int J Mol Sci* 2025;**26**: 1648.
44. Xuan DTM, Yeh IJ, Wu CC. *et al.* Comparison of transcriptomic signatures between Monkeypox-infected monkey and human cell lines. *J Immunol Res* 2022;**2022**:3883822. https://doi.org/10.1155/2022/3883822
45. Chiao CC, Liu YH, Phan NN. *et al.* Prognostic and genomic analysis of proteasome 20S subunit alpha (PSMA) family members in breast cancer. *Diagnostics (Basel)* 2021;**11**:2220. https://doi.org/10.3390/diagnostics11122220
46. Xuan DTM, Yeh IJ, Su CY. *et al.* Prognostic and immune infiltration value of proteasome assembly chaperone (PSMG) family genes in lung adenocarcinoma. *Int J Med Sci* 2023;**20**:87–101. https://doi.org/10.7150/ijms.78590
47. Mienye ID, Sun Y. A survey of ensemble learning: concepts, algorithms, applications, and prospects, IEEE. *Access* 2022;**10**: 99129–49. https://doi.org/10.1109/ACCESS.2022.3207287
48. Chen T, Guestrin C. XGBoost: A scalable tree boosting system. *Proc 22nd ACM SIGKDD Int Conf Knowl Discov Data Min* 2016; 785–94. https://doi.org/10.1145/2939672.2939785

49. Friedman JH. Stochastic gradient boosting. *Comput Stat Data Anal* 2002;**38**:367–78. https://doi.org/10.1016/S0167-9473(01)00065-2
50. Ye J, Chow JH, Chen J. *et al.* Stochastic gradient boosted distributed decision trees. In: *Proceedings of the 18th ACM International Conference on Information and Knowledge Management (CIKM)*. Hong Kong, China: ACM; 2009 Nov 2–6.
51. Cui J, Wang M, Lin C. *et al.* Exploring machine learning strategies for single-cell transcriptomic analysis in wound healing. *Burns & Trauma* 2025;**13**:tkaf032.
52. Ran R, Brubaker D. Enhanced annotation of CD45RA to distinguish T cell subsets in single cell RNA-seq via machine learning. *J Immunol* 2024;**212**:0259_5381. https://doi.org/10.4049/jimmunol.212.supp.0259.5381
53. Wagle MM, Long S, Chen C. *et al.* Interpretable deep learning in single-cell omics. *Bioinformatics* 2024;**40**:btae374.
54. Koch M, Acharjee A, Ament Z. *et al.* Machine learning-driven Metabolomic evaluation of cerebrospinal fluid: insights into poor outcomes after aneurysmal subarachnoid Hemorrhage. *Neurosurgery* 2021;**88**:1003–11. https://doi.org/10.1093/neuros/nyaa557
55. Lin JC, Liu TP, Yang PM. CDKN2A-inactivated pancreatic ductal adenocarcinoma exhibits therapeutic sensitivity to paclitaxel: a bioinformatics study. *J Clin Med* 2020;**9**:4019.
56. Liu LW, Hsieh YY, Yang PM. Bioinformatics data mining repurposes the JAK2 (Janus kinase 2) inhibitor Fedratinib for treating pancreatic ductal adenocarcinoma by reversing the KRAS (Kirsten rat sarcoma 2 viral oncogene homolog)-driven gene signature. *J Pers Med* 2020;**10**:130. https://doi.org/10.3390/jpm10030130
57. Hsieh YY, Liu TP, Chou CJ. *et al.* Integration of bioinformatics resources reveals the therapeutic benefits of gemcitabine and cell cycle intervention in SMAD4-deleted pancreatic ductal adenocarcinoma. *Genes (Basel)* 2019;**10**:766.
58. Ma D, Fan C, Sano T. *et al.* Beyond biomarkers: machine learning-driven multiomics for personalized medicine in gastric cancer. *J Pers Med* 2025;**15**:166. https://doi.org/10.3390/jpm15050166
59. Chou CW, Hsieh YH, Ku SC. *et al.* Potential prognostic biomarkers of OSBPL family genes in patients with pancreatic ductal adenocarcinoma. *Biomedicines* 2021;**9**:1601.
60. Wang CY, Chao YJ, Chen YL. *et al.* Upregulation of peroxisome proliferator-activated receptor-α and the lipid metabolism pathway promotes carcinogenesis of ampullary cancer. *Int J Med Sci* 2021;**18**:256–69. https://doi.org/10.7150/ijms.48123
61. Xu J, Lu C, Jin S. *et al.* Deep learning-based cell-specific gene regulatory networks inferred from single-cell multiome data. *Nucleic Acids Res* 2025;**53**:gkaf138.
62. Abedi, Khoozani P, Bharmauria V, Schütz A. *et al.* Integration of allocentric and egocentric visual information in a convolutional/multilayer perceptron network model of goal-directed gaze shifts, cerebral cortex. *Communications* 2022;**3**:tgac026.
63. Wang S, Curry RN, McDonald MF. *et al.* Inferred developmental origins of brain tumors from single-cell RNA-sequencing data. *Neurooncol Adv* 2025;**7**:vdaf016. https://doi.org/10.1093/noajnl/vdaf016
64. Zeng M, Wu Y, Li Y. *et al.* LncLocFormer: a transformer-based deep learning model for multi-label lncRNA subcellular localization prediction by using localization-specific attention mechanism. *Bioinformatics* 2023;**39**:btad752. https://doi.org/10.1093/bioinformatics/btad752
65. Saleh H, El-Sappagh S, McCann M. *et al.* Multivariate multi-horizon time-series forecasting for real-time patient monitoring based on cascaded fine tuning of attention-based models. *Comput Biol Med* 2025;**194**:110406. https://doi.org/10.1016/j.compbiomed.2025.110406
66. Fan Z, Zhao H, Zhou J. *et al.* A versatile attention-based neural network for chemical perturbation analysis and its potential to aid surgical treatment: an experimental study. *Int J Surg* 2024;**110**:7671–86. https://doi.org/10.1097/JS9.0000000000001781

Briefings in Bioinformatics, 2025, **26(6)**, bbaf583
https://doi.org/10.1093/bib/bbaf583
Problem Solving Protocol