Machine Learning-Based Identification of Blood Biomarkers that Distinguish Precachectic and Cachectic Patients with Pancreatic Ductal Adenocarcinoma

Kayode D. Olumoyin[1], Magaret Park[2,3], Evan W. Davis[2,4], Jennifer B. Permuth[2,4], Katarzyna A. Rejniak[1,5,*]

[1]Department of Integrated Mathematical Oncology, H. Lee Moffitt Cancer Center & Research Institute, Tampa, FL, USA
[2]Department of Gastrointestinal Oncology, H. Lee Moffitt Cancer Center & Research Institute, Tampa, FL, USA
[3]Department of Biostatistics and Bioinformatics, H. Lee Moffitt Cancer Center & Research Institute, Tampa, FL, USA
[4]Department of Cancer Epidemiology, H. Lee Moffitt Cancer Center & Research Institute, Tampa, FL, USA
[5]University of South Florida, Morsani College of Medicine, Department of Oncologic Sciences, Tampa, FL, USA

ABSTRACT

Background: Identification of minimally invasive biomarkers of different stages of cachexia (Ca), and precachexia (PCa) in particular, might help clinicians in treating patients with pancreatic ductal adenocarcinoma (PDAC) at high risk of progressing to a more severe cachectic stage. In this work, we developed a machine-learning (ML) model optimized to blood biomarkers data that identifies precachectic and cachectic patients.

Methods: Blood and clinical data was collected from treatment-naïve patients with PDAC through the Florida Pancreas Collaborative (FPC), a multi-institutional cohort study and biobanking initiative. Blood was processed into serum and assayed for a total of 35 candidate biomarkers. Participants were classified as having noncachexia (NCa), precachexia, or cachexia according to modified criteria by Vigano and colleagues which consider unintentional weight loss and biochemical data. Using these data, we designed ML algorithms to: (i) pre-select predictive blood biomarker candidates using a combination of mutual information method together with the leave-one-feature-out (LOFO) feature importance approach; (ii) identify the minimal combination of predictive biomarkers using the forward feature selection method; (iii) determine the optimal classification hyperparameters for the support vector machine using a cross-validation technique; and (iv) adjust the decision-boundary threshold for imbalanced data using the Matthews correlation coefficient. Three ML-based binary predictors were designed to determine patients' cachexia status: NCa vs. Ca; PCa vs. Ca; and PCa vs. NCa.

Results: The biomarker levels from 184 patients (28 NCa, 53 PCa, and 103 Ca) were used in this study. The NCa vs. Ca predictor identified a set of 6 biomarkers and yielded area under the curve (AUC) of 0.835. The PCa vs. Ca predictor identified a set of 6 biomarkers and yielded AUC of 0.810. The PCa vs. NCa predictor identified a set of 5 biomarkers and yielded AUC of 0.771.

Conclusions: The developed ML models that use blood biomarker data provided effective predictions of patient's cachexia stage that can help clinicians to diagnose PCa.

Keywords:
Cancer-associates cachexia; machine-learning predictor; cachexia patients' stratification; pancreatic cancer;

1. Background

Cancer-associated cachexia (CC) is a multifactorial syndrome observed in up to 80% of PDAC patients characterized by unintentional weight loss, muscle wasting in the presence or absence of fat loss, and fatigue [1], which can lead to a reduction in quality of life and poor clinical outcomes. [2, 3]. To distinguish between cachexia stages, we followed the Florida Pancreas Collaborative and used criteria described in [4], which are based on the Vigano classification [5]. This classification system is comprised of the following types of data: (a) biochemistry (level of C-reactive protein (CRP) or albumin, or hemoglobin, or white blood cell count), (b) changes in food intake, (c) minimal or significant weight loss (WL), and (d) changes in daily activities based on the Patient-Generated Subjective Global Assessment (PG-SGA) performance status [6]. The recognized 4 cancer cachexia stages are: noncachexia (NCa), precachexia (PCa)—an early stage of the syndrome characterized by abnormal food intake or blood chemistry but no significant weight loss, cachexia (Ca), and refractory cachexia (RCa)—a stage that is largely irreversible [7]. However, several criteria in the classification used by FPC are based on patient-reported outcomes, which may be quite subjective, and thus differentiation between CC stages is difficult. This suggests the need for more precise tools for distinguishing patients with different cachexia stages, such as machine-learning (ML) methods, that can simultaneously handle complex data and focus on patterns between multiple data features.

Moreover, there is a dire need to develop a minimally invasive approach to identify CC earlier. Since blood is routinely collected clinically as a part of standard of care, identifying novel blood-based biomarkers of different stages of cachexia could be worthwhile. In previous work, certain blood biomarkers were identified as prognostic for CC stages for PDAC patients [8-12]. These include CRP, interleukin-6 (IL-6), interleukin-8 (IL-8), tumor necrosis factor alpha (TNF-α), monocyte chemoattractant protein-1 (MCP-1), transforming growth factor beta (TGF-β), and growth/differentiation factor (GDF-15). Our previous work [12] also identified GDF-15 as a marker of CC that is predictive of survival, but only among Hispanic and Non-Hispanic White populations. However, these analyses use predominantly single feature correlation with the target outcome or pairwise data comparisons. Since CC is a complex multifactorial syndrome, there are potentially multidimensional and nonlinear interactions between different candidate biomarkers for CC. Single feature correlations may not fully capture these complex data relationships [13, 14]. In contrast, ML methods can successfully utilize multi-dimensional data and identify data interconnections that yield non-intuitive predictions of target outcomes.

Developing tools that can distinguish between NCa, PCa, and Ca patients may aid in earlier diagnosis of the disease and may allow for earlier therapeutic interventions. In particular, differentiating between PCa and NCa stages is important for early detection of cancer patients who may not show symptoms of Ca (i.e. weight loss) but may be on a trajectory towards Ca. This will benefit the patient, since interventions are more likely to be effective at the early stage. Such tools can also help clinicians in designing improved surveillance protocols for at-risk patients.

The goal of our study was to develop a ML-based framework that determines the nonlinear interconnectivity between patients' blood-based biomarker data and identifies a minimal set of biomarkers predictive of the different CC stages (NCa, PCa, or Ca) in PDAC patients. Our classifier has been applied to data collected by the Florida Pancreas Collaborative (FPC), a multi-institutional state-wide cohort study [15]. As a result, we provide three tools to distinguish between NCa and Ca stages, PCa vs. Ca stages, and PCa vs. NCa stages.

2. Methods

2.1. Study population and data collection

This study included patient data collected by the Florida Pancreas Collaborative (FPC), a multi-institutional prospective cohort study and biobanking initiative between 2018 and 2021, and approved by the Moffitt Cancer Center Scientific Review Committee (MCC19717, Pro00029598),

and Advarra IRB (IRB00000971). All patients provided informed consent for participation [15]. Pre-treatment serum biomarker levels that comprised of cytokines, chemokines, adipokines, lipoproteins, glycans, and other analytes (35 biomarkers in total) were available for 202 patients [12]. Patients' CC stage was determined using the modified Vigano criteria [4, 5] and categorized into 4 CC stages: NCa, PCa, Ca, and refractory cachexia (RCa). However, due to low number of cases (n=18) unsuitable for ML algorithms, patients with RCa status were excluded from this computational study.

2.2. Description of the ML classification problem

Our goal was to identify a predictive subset of features that differentiates between two targets in a binary classification task, and to provide metrics of success for such data stratification. Let the dataset $\boldsymbol{X}$ consists of M data points and P features: $\boldsymbol{X} = [\boldsymbol{X}_1, \boldsymbol{X}_2, \cdots, \boldsymbol{X}_M]^{\intercal}$, where each data point $\boldsymbol{X}_i = \left(\boldsymbol{x}_i^1, \boldsymbol{x}_i^2, \cdots, \boldsymbol{x}_i^P\right)$ for $i \in \{1,2,\cdots,M\}$, and for each i, the corresponding target is the binary class $y_i \in \{-1, 1\}$. We consider three different machine learning predictors: NCa vs. Ca, PCa vs. Ca, and PCa vs. NCa. In each case, the goal was to find the minimal subset of features of size Q, where $Q < P$, that divides the dataset $\boldsymbol{X}$ into distinct binary classes. We used a data-informed approach. First, we split $\boldsymbol{X}$ into training and testing cohorts. Using the training cohort, we implemented feature selection method to identify predictive features. Next, we determined an optimal machine learning classifier using a cross-validation technique that finds optimal hyperparameters for the classifier and also learns the optimal decision boundary threshold that corrects for imbalance in the binary classes in $\boldsymbol{X}$. Finally, this classifier was applied to the testing cohort to assess the prediction metrics.

2.3. Mutual information measure for nonlinear dependencies

Mutual information (MI) is a non-parametric measure of statistical dependency between the dataset $\boldsymbol{X} \in \mathbb{R}^{M\times P}$ and the predicted binary class $\boldsymbol{Y} \in \mathbb{R}^{M\times 1}$, where M is the number of patients and P is the number of features. MI captures nonlinear dependencies in high-dimensional data, which makes them robust for measuring feature relevance in discrete or categorical datasets [16, 17]. A discrete MI is computed as follows: $MI(\boldsymbol{X}, \boldsymbol{Y}) = \sum_{x \in \boldsymbol{X}} \sum_{y \in \boldsymbol{Y}} p(x,y) \log \frac{p(x,y)}{p(x)\, p(y)}$. Where $p(x,y)$ is the joint probability of $\boldsymbol{X}$ and $\boldsymbol{Y}$, while $p(x)$ and $p(y)$ are the marginal probabilities of $\boldsymbol{X}$ and $\boldsymbol{Y}$, respectively.

2.4. Leave-one-feature-out for determining feature importance

The leave-one-feature-out (LOFO) approach measures the importance of a given feature in a dataset by retraining a model without that specific feature and recording the model's performance to assess whether it declines, stays the same, or improves. Specifically, LOFO computes the marginal contribution of each feature by measuring the change in model performance when the feature is omitted. An increase in model performance indicates that the excluded feature had a favorable impact on predictive accuracy. If the change in model performance is zero or negative, the feature is either redundant or does not improve model performance [18, 19]. Unlike other feature importance methods, LOFO is model-agnostic and computationally easy to implement.

2.5. Multiple imputation framework for data retention

Multiple imputation is a statistical method to handle missing data. Let $\boldsymbol{X} \in \mathbb{R}^{M\times P}$ be a data matrix with observed entries $\boldsymbol{X}_{obs}$ and missing entries $\boldsymbol{X}_{miss}$. Under the assumption that data are missing at random (MAR) or completely at random (MCAR), the multiple imputation replaces all entries in $\boldsymbol{X}_{miss}$ with imputed values that preserve the interrelations in $\boldsymbol{X}_{obs}$. We use here the multiple imputation with denoising autoencoders (MIDAS) method [20], which is a scalable deep learning-based technique that employs a class of unsupervised neural networks known as denoising

autoencoders [21] and Monte Carlo dropout to generate multiple imputation of the missing data with realistic uncertainty quantification.

2.6. Data normalization

Data normalization is performed to ensure that all features can contribute equally. We first split data into the training and testing cohorts (70/30). In order to prevent data leakage into the testing cohort, we implement 'Pipeline', a functionality available in Scikit-Learn, a free software machine learning library [22]. We use the z-score to normalize each biomarker in the training and testing cohort. The use of 'Pipeline' allows us to apply identical scaling for both cohorts.

2.7. Forward feature selection method for identification of predictive blood biomarkers

The feature selection is performed to identify the predictive subset of blood biomarkers and to reduce the number of biomarkers used in the data classifier. Here we used the Forward Feature Selection (FFS) method [23]. We started by obtaining a pre-ranking of the collected blood biomarkers by applying the MI and LOFO algorithms. Next, we used a 10-fold cross-validation to partition the training data into two subsets. In each of the 10 folds, we used one subset to train a random forest (RF) classifier, and the second subset as a validation set. The biomarkers were added sequentially in the order of pre-ranking to train RF classifier at each step, and its predictive accuracy was evaluated using the validation subset. To mitigate fluctuations due to noisy biomarkers, accuracy curves were converted to a monotonic increasing curve, retaining maximum accuracy observed up to each step. The optimal number of features for each fold was determined using an elbow-point detection method [24, 25], which identifies the point beyond which additional biomarkers provide minimal incremental gain in accuracy. The mean accuracy curve and 95% confidence interval across all folds were used to visualize biomarker contribution to predictiveness of the model and to determine final biomarker selection. This approach eliminates multicollinearity of blood biomarkers for prediction purposes.

2.8. Support vector machine model for learning the optimal classification rules

The binary data classification was performed using the support vector machine (SVM) model [26, 27] with the nonlinear Radial Basis Function (RBF) kernel: $K(\boldsymbol{X}_i, \boldsymbol{X}_j) = exp\left(-\gamma \left\| \boldsymbol{X}_i - \boldsymbol{X}_j \right\|^2\right)$. First, all data was split 70/30 into the training and testing cohorts. Then, the optimal classification hyperparameters were identified by the RBF-SVM based on the training cohort.

The linear SVM generates a predicted binary class $\hat{y}_i \in \{-1, 1\}$ for each data point $\boldsymbol{X}_i = \left(\boldsymbol{x}_i^1, \boldsymbol{x}_i^2, \cdots, \boldsymbol{x}_i^P\right)$ in $\boldsymbol{X} = [\boldsymbol{X}_1, \boldsymbol{X}_2, \cdots, \boldsymbol{X}_M]^\intercal$ by computing a decision function $\boldsymbol{w}^\intercal \boldsymbol{X}_i + b = w^1 \boldsymbol{x}_i^1 + \cdots + w^P \boldsymbol{x}_i^P + b$ for each $\boldsymbol{X}_i$. The corresponding predicted class label $\hat{y}_i$ is assigned according to the following rule: $\hat{y}_i = -1$ for $\boldsymbol{w}^\intercal \boldsymbol{X}_i + b < 0$, and $\hat{y}_i = 1$ for $\boldsymbol{w}^\intercal \boldsymbol{X}_i + b > 0$. The objective function that is minimized during the training of the linear SVM is $\min_{\boldsymbol{w}, b} \left\{ \frac{1}{2} \|\boldsymbol{w}\|^2 + C \sum_{i=1}^{M} H(y_i, \boldsymbol{w}^\intercal \boldsymbol{X}_i + b) \right\}$, where $H(y_i, \boldsymbol{w}^\intercal \boldsymbol{X}_i + b) = max(0, 1 - y_i \times \boldsymbol{w}^\intercal \boldsymbol{X}_i + b)$ is the convex hinge loss [28]. The parameter $\boldsymbol{w}$ is the learnable weight vector, b is the bias value, C specifies the width of the margins for avoiding data misclassification. This linear SVM was extended to a nonlinear SVM by using the RBF kernel that transforms each $\boldsymbol{X}_i$ into a higher dimensional space. This allowed us to capture interactions between predictive feature values, that are potentially multidimensional and nonlinear.

The optimal RBF-SVM hyperparameters (C, γ) were identified by performing k-fold cross-validation (CV) on the training dataset, for k = 2, …,10. Here γ controls the nonlinearity of the decision boundary hyperplane. For each k, the best C and γ were determined by (i) minimizing the difference between the cross-validation training accuracy and testing accuracy; (ii) maximizing the cross-validation training accuracy; and (iii) minimizing the value of C. The optimal values of C and γ were obtained by computing the Matthews correlation coefficient (MCC) for each cross-validation (k = 2, … ,10) and then choosing C and γ for which MCC is maximal, and C is minimal.

2.9. Matthews correlation coefficient algorithm to account for data imbalance

The Matthews correlation coefficient (MCC) statistical test [29, 30] was used to determine the optimal decision boundary threshold to account for true and false positives and negatives in imbalanced training dataset. Usually, this threshold is set to 0.5, because it is assumed that SVM works with balanced data (i.e., similar numbers of data fall into each class). For the imbalanced dataset, this threshold has to be adjusted. The magnitude of the decision function $\boldsymbol{w}^{\mathrm{T}}\boldsymbol{X}_i + b$ for each $\boldsymbol{X}_i$ was extended to probability estimates using the Scikit-Learn library [31], with option 'probability=True'. Next, the thresholds between 0.1 and 0.9 were tested by separating the training cohort data into binary classes based on whether their RBF-SVM–generated prediction probabilities exceeded the given threshold. For each threshold, MCC was calculated on the labeled prediction probabilities. The threshold with the maximum MCC was called optimal and was used as the decision boundary threshold to generate predictions on the testing cohort.

2.10. Performance metrics

To assess performance of the classification protocol on the testing dataset, the following performance metrics were used for evaluation: (1) true positive rate (TPR) or sensitivity, is the percentage of correctly classified positive instances: TPR=TP/(TP+FN); (2) true negative rate (TNR) or specificity, is the percentage of correctly classified negative instances: TNR=TN/(FP+TN); (3) accuracy is the percentage of correctly classified positive and negative instances: accuracy=(TP+TN)/(TP+FN+FP+FN); (3) area under the receiver operating characteristics curve (AUC/ROC or AUC) measures the ability to discriminate between positive and negative cases and ranges from 0.5 (coin toss) to 1.0 (perfect classification), when ROC curve shows tradeoffs between TP and FP. Here, TP (true positive) is the correctly classified data, TN (true negative) is the correctly classified data, FP (false positive FP) is the misclassification of the positive class, and FN (false negative) is the misclassification of the negative class.

3. Results

3.1. Computational study design

A total of 202 PDAC patients from the Florida Pancreas Collaborative (FPC) [15] had available pre-treatment blood biomarker data [12] and CC status assessed using the modified Vigano classification [4, 5]. The reported CC stages were along the cachexia continuum from NCa to PCa, to Ca, and to RCa. However, patients classified as RCa were excluded from this study due to small sample size and high pairwise positive correlations among several blood biomarkers (**Supplemental Figure 1**), which makes the RCa data insufficient for ML classification. After removing RCa cases due to low cell sizes, samples from 184 PDAC cases were included in our study; 28 NCa, 53 PCa, and 103 Ca. For each patient, 35 blood biomarkers were considered, however, for some cases the blood biomarkers values were missing, likely because they were outside the assessable range (**Table 1.**)

Our approach started with a pre-selection step to retain features that are essential for model predictions. This comprised of an ablation study (leave-one-feature-out, LOFO) identifying the redundant features and the Mutual Information (MI) method which demonstrates that none of the features was singlehandedly predictive of CC status. This was followed by using the multiple imputation with denoising autoencoders (MIDAS) method for learning admissible values for the missing data. The final step of pre-processing was data normalization applied separately to the training and testing datasets using the *z-score* approach. The normalized training dataset was then used as an input for the forward feature selection (FFS) method for the 3 predictors: NCa vs. Ca, PCa vs. Ca, and PCa vs. NCa. For each predictor, the optimal hyperparameters for the support vector machine (SVM) classifier were identified and the optimal decision boundaries determined using the Matthews correlation coefficient (MCC) method. Finally, the performance of

each predictor was analyzed on the testing dataset with the confusion matrices and AUC/ROC curves. The detailed flowchart for this study is shown in **Figure 1**.

Table 1: Blood biomarkers for PDAC patients from the Florida Pancreas Collaborative[a]

	Missing[b]	Overall	Ca	NCa	PCa
Sample size		184	103	28	53
ENA.78 [c]	0	13.2 ± 1.2	13.2 ± 1.2	13.2 ± 1.1	13.2 ± 1.2
IFN.y	0	6.5 ± 1.3	6.7 ± 1.2	6.4 ± 1.1	6.2 ± 1.6
IL.10	1	2.7 ± 1.6	3.1 ± 1.4	2.0 ± 1.5	2.2 ± 1.7
IL.6	0	5.0 ± 1.4	5.2 ± 1.4	4.3 ± 1.0	4.9 ± 1.4
IL.8	0	7.9 ± 1.2	8.2 ± 1.3	7.3 ± 0.7	7.8 ± 0.9
MCP.1	0	11.5 ± 0.6	11.5 ± 0.6	11.3 ± 0.5	11.5 ± 0.5
MDC	0	13.6 ± 0.5	13.6 ± 0.6	13.5 ± 0.4	13.6 ± 0.5
MIP.1a	6	7.7 ± 1.3	7.9 ± 1.3	7.6 ± 1.6	7.3 ± 1.2
TNF.a	0	5.4 ± 0.8	5.7 ± 0.8	5.0 ± 0.7	5.1 ± 0.7
C.peptide [c]	1	14.2 ± 1.0	14.2 ± 1.0	14.3 ± 1.2	14.0 ± 1.0
G.CSF	0	6.9 ± 1.0	6.9 ± 1.1	6.8 ± 0.7	7.0 ± 1.0
IL.22	9	2.3 ± 1.7	2.7 ± 1.5	1.2 ± 1.2	2.1 ± 2.0
Insulin	3	5.8 ± 1.4	5.8 ± 1.3	6.3 ± 1.9	5.7 ± 1.2
Leptin	3	16.2 ± 2.6	15.5 ± 2.8	17.1 ± 1.8	16.9 ± 2.2
GRO.a [c]	0	12.0 ± 1.1	12.0 ± 1.1	11.5 ± 1.0	12.0 ± 1.1
HGF	0	9.5 ± 0.9	9.6 ± 0.8	9.1 ± 0.5	9.5 ± 1.1
MMP.2	2	15.0 ± 0.8	15.1 ± 0.8	14.9 ± 0.6	14.8 ± 0.9
Adiponectin	0	24.1 ± 0.8	24.1 ± 0.7	24.0 ± 0.8	24.1 ± 0.9
CRP	0	21.9 ± 2.5	22.1 ± 2.7	20.2 ± 1.4	22.3 ± 2.4
GDF.15	0	10.8 ± 1.0	11.1 ± 1.0	10.1 ± 0.7	10.6 ± 0.9
TIMP.1	0	18.6 ± 0.7	18.8 ± 0.8	18.2 ± 0.5	18.6 ± 0.7
TGF.B2	20	6.9 ± 1.3	7.1 ± 1.3	6.4 ± 1.0	7.0 ± 1.2
TGF.B1	0	15.9 ± 0.6	16.0 ± 0.6	15.8 ± 0.5	16.0 ± 0.6
PPAR.y [c]	1	1.7 ± 0.9	1.8 ± 0.9	1.7 ± 1.1	1.7 ± 0.9
HIF.1a [c]	12	9.4 ± 1.8	9.4 ± 1.9	9.9 ± 1.6	9.2 ± 1.8
Laminin [c]	0	10.8 ± 0.7	10.8 ± 0.7	10.7 ± 0.5	11.0 ± 0.6
HbA1c [c]	0	9.0 ± 0.7	8.9 ± 0.8	9.1 ± 0.6	9.1 ± 0.6
CA19.9	14	4.8 ± 3.4	4.7 ± 3.6	4.8 ± 3.7	4.9 ± 2.8
Glucose	5	6.6 ± 0.6	6.7 ± 0.5	6.6 ± 0.6	6.5 ± 0.6
HDL	0	9.2 ± 0.8	9.1 ± 0.7	9.3 ± 0.8	9.2 ± 0.8
CCK	1	8.3 ± 0.8	8.3 ± 0.8	8.3 ± 0.6	8.5 ± 0.8
LDL	0	14.6 ± 0.9	14.7 ± 0.9	14.4 ± 0.7	14.6 ± 0.9
Triglyceride	21	5.1 ± 0.8	5.2 ± 0.8	5.2 ± 0.8	5.1 ± 0.8
Lumican [c]	0	20.9 ± 0.7	20.9 ± 0.7	21.0 ± 0.7	20.9 ± 0.7
ZAG [c]	0	22.1 ± 0.6	22.1 ± 0.7	22.1 ± 0.7	22.2 ± 0.5

Table 1: Data are presented as mean ± SD. Abbreviations: Ca, Cachexia; PCa, Precachexia; NCa, Noncachexia. [a]All values are Log2 transformed. [b]The number of patients missing data for the indicated blood biomarker. [c]Features that were removed from analysis due to the pre-selection process.

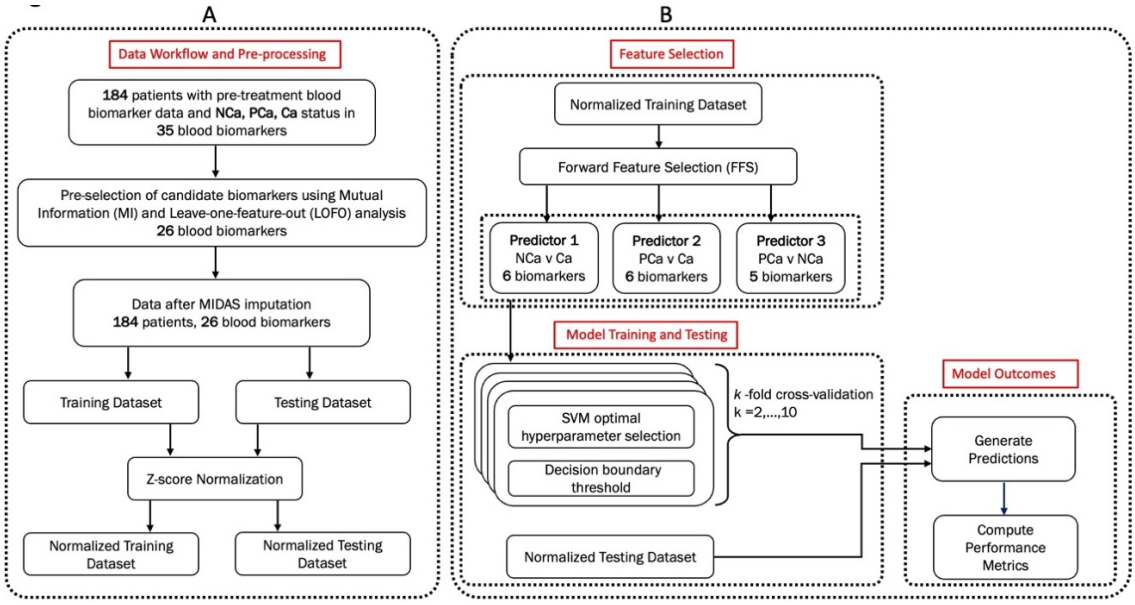

Figure 1. Machine learning flowchart for this study. A. The data workflow and pre-processing consists of a blood biomarker pre-selection step (mutual information and leave-one-feature-out analysis), data imputation (MIDAS), and data normalization (*z-score*). **B.** Predictor structure is comprised of a feature selection step (FFS) for each of the 3 predictors: NCa vs. Ca, PCa vs. Ca, and PCa vs. NCa, support vector machine (SVM) classifier for each predictor, decision boundary adjustment (MCC), and performance analyses (confusion matrices and AUC/ROC curves).

3.2. Blood Biomarker predictive analysis and ablation study

The first step in our data pre-processing stage, was to test all 35 candidate blood biomarkers to identify whether any of them was predictive of CC status by itself. We accomplished this by computing the MI score [32], which measures the amount of information each blood biomarker provides about CC status. We followed [33] in assuming that a blood biomarker could be predictive by itself if its MI value is high. Across all 35 candidate blood biomarkers, MI values were less than 0.10 (**Figure 2**), which demonstrates that no blood biomarker is singlehandedly predictive of cancer cachexia status.

Next, we conducted an ablation study to assess whether any of the 35 biomarkers were either redundant or essential for the performance of the CC classification. We used the LOFO test [18, 19] by training a multi-class (NCa, PCa, Ca) random forest (RF) classifier model [34, 35] with all 35 blood biomarkers and comparing its performance against the performance of RF with one biomarker removed in each instance. If the difference in the prediction accuracy between these two models is positive, this indicates that the removed biomarker is helpful in CC classification and its removal reduces predictiveness of the model (those biomarkers are shown as blue dots in **Figure 2**). If the difference is zero ('C-peptide' and 'PPAR-gamma'), it indicates that the biomarker is redundant. If the difference is negative, removing that biomarker improves predictiveness of the model, and it is recommended that this biomarker should be excluded from further analysis. Those biomarkers are indicated by red dots in **Figure 2**. After performing the mutual information and ablation studies on each of the 35 blood biomarkers, we identified 9 blood biomarkers that did not improve predictiveness of CC classification. These are: ENA-78, C-peptide, GRO-alpha, PPAR-gamma, HIF1alpha, Laminin, HbA1c, Lumican, and ZAG; they are indicated by [c] in **Table 1**. These blood biomarkers were excluded from the rest of this study.

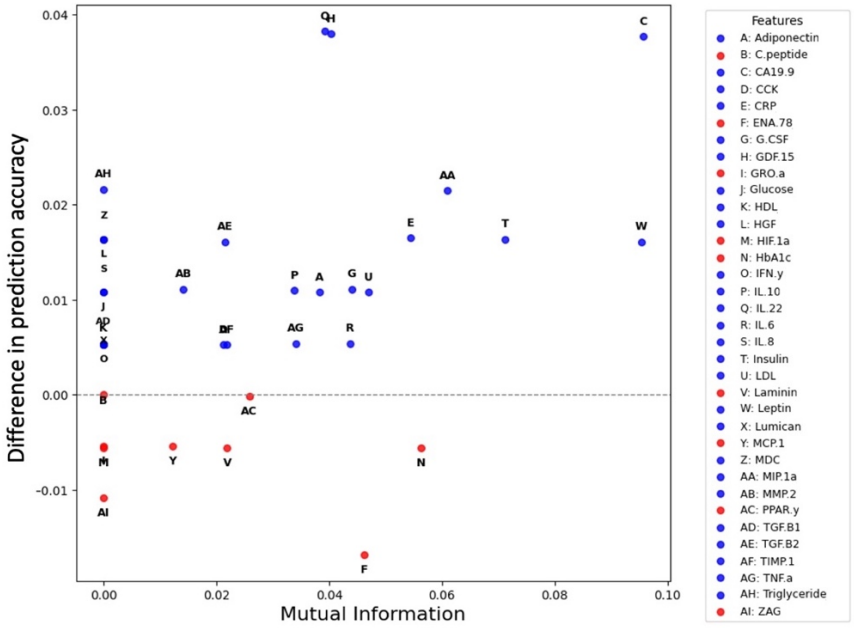

Figure 2. Criteria for determining predictiveness of each blood biomarker and its contribution to model performance. The leave-one-feature-out (LOFO) ablation study determined that the 9 biomarkers (red dots) below the horizontal dash line do not contribute to model performance and should be removed from the data workflow. The remaining biomarkers (blue dots) contribute to model predictions.

3.3. Multiple imputation of missing blood biomarker data

After the data pre-processing step, the list of blood biomarkers under consideration was reduced from 35 to 26. The total proportion of missingness in the new dataset was (1.7%), with 11 biomarkers that have between 1 to 20 missing entries (**Table 1**). To address the missing entries issue, we used the MIDAS method, a deep learning approach that generated imputed values preserving interrelations among all biomarkers. The imputed values were constrained to be consistent with the observed values in each biomarker containing missing entries. We observed that for the dataset with imputed values the AUC values were higher than in the case without imputation for all 3 predictors: 0.835 compared to 0.810 for NCa vs. Ca predictor; 0.81 compared to 0.779 for PCa vs. Ca predictor; and 0.771 compared to 0.658 for PCa vs. NCa predictor.

3.4. A computational predictor that distinguishes non-cachectic from cachectic patients

131 PDAC patients' data collected through the FPC biobank were classified either as Ca (103) or NCa (28). This dataset was split 70/30 into training (91 patients) and testing (40 patients) cohorts. Each patients' entry contained 26 blood biomarkers identified by the MI and LOFO methods in the pre-processing step with values imputed using the MIDAS technique, when needed.

Using the training cohort, we implemented the FFS method to identify a subset of the blood biomarkers that have high predictive accuracy in differentiating between NCa and Ca status. FFS calculated a pre-ranking of the 26 blood biomarkers using a binary class (NCa, Ca) RF classifier for 5 random subsamplings of 70% of data (**Figure 3A**). The ranking was calculated according to feature importance score (FIS [35, 36]). Next, a 10-fold cross-validation was used to train RF by sequentially adding each biomarker in the order of their pre-determined rank. With each added biomarker, we evaluated the predictive accuracy of RF on the validation set. This process of adding new features yields a non-strictly increasing curve of the accuracy, as there may be local regions of downward fluctuations due to noisy biomarkers. To mitigate this, the curve of accuracy was converted to a monotonically increasing curve by passing it through a cumulative

transform, retaining maximal accuracy observed up to each step. The optimal number of features for each fold was determined using an elbow-point detection method [24, 25], that identifies the point beyond which additional biomarkers provide incremental gain in accuracy (**Figure 3B**).

For the NCa vs. Ca case, FFS identified a set of 6 robust biomarkers (out of initial set of 26) that together have a high predictive power. This set includes, according to their individual ranking: Insulin, GDF-15, tissue inhibitor of metalloproteinases-1 (TIMP-1), CRP, TNF-α, and interleukin-22 (IL-22) (**Figure 3A,B**). For these 6 biomarkers, we implemented the RBF-SVM method using the training cohort to identify the optimal hyperparameters (C = 6.42, γ = 1.85). Next, the MCC approach was used to adjust for imbalance in the training dataset yielding the optimal decision boundary threshold (0.6640). Finally, the model predictability was evaluated using the testing cohort with 40 patients' data. The obtained confusion matrix of the model performance is shown in **Figure 3C**. The model generated the accuracy of 0.800, sensitivity (rate at which the model correctly predicts Ca) of 0.871, and specificity (the rate at which the model correctly predicts NCa) of 0.556. The area under the ROC curve (AUC) for the testing cohort was 0.835 (**Figure 3D**).

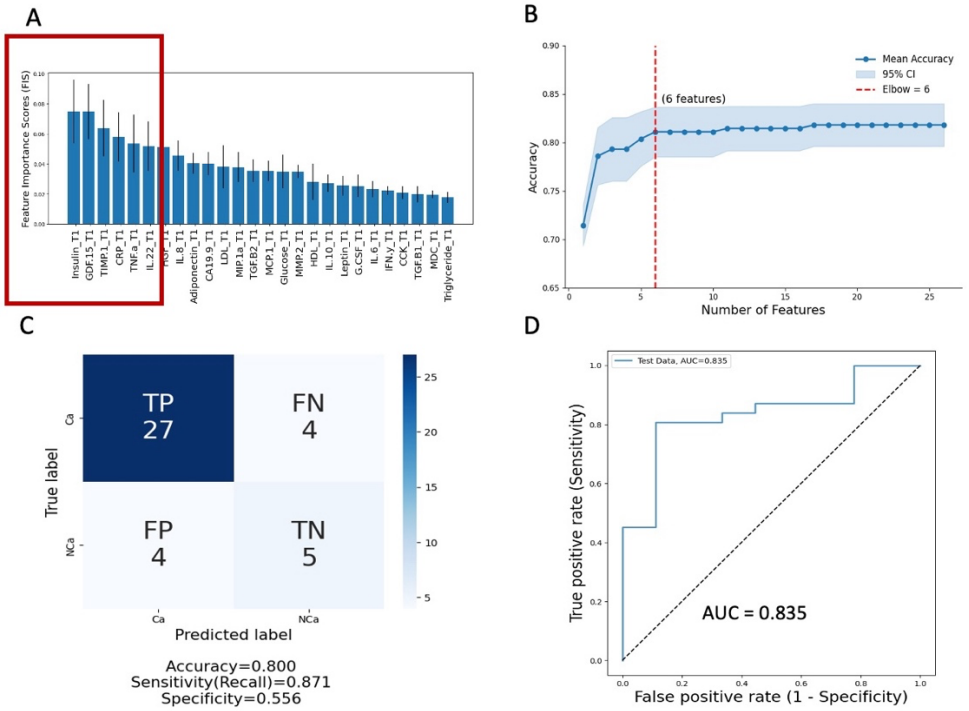

Figure 3. Feature selection and performance analysis for NCa vs. Ca predictor. A. The 26 candidate blood biomarkers were pre-ranked using the FIS values. **B.** The cumulative curve of accuracy identified a subset of 6 robust predictive biomarkers indicated by the red box in **A**. **C.** The confusion matrix and performance metrics of the SVM-RBF predictor. **D.** The AUC/ROC curve generated for predictions of NCa vs. Ca status for the testing set.

3.5. A computational predictor for stratification of pre-cachectic vs. cachectic patients

In this case, there were 156 PDAC patients in the FPC database that were classified either as Ca (103) or PCa (53). This dataset was split into training (109 patients) and testing (47 patients) cohorts. Using the training cohort, the FFS method identified 6 predictive blood biomarkers that together discriminate between PCa and Ca stages. This set includes: TNF-α, interferon gamma (IFN-γ), IL-22, interleukin-10 (IL-10), macrophage inflammatory protein (MIP-1α), and glucose, listed here according to their ranking (**Figure 4A,B**). Next, these 6 biomarkers were used in the

RBF-SVM implementation for the training cohort, and the optimal hyperparameters (C = 1.41, γ = 0.37) were identified. Subsequently, the MCC method yielded an optimal decision threshold of 0.3440 to adjust for imbalance in the training dataset. The confusion matrix for the testing cohort showed accuracy of 0.723, sensitivity of 0.562, and specificity of 0.806 (**Figure 4C**). The corresponding AUC yielded the value of 0.810 (**Figure 4D**).

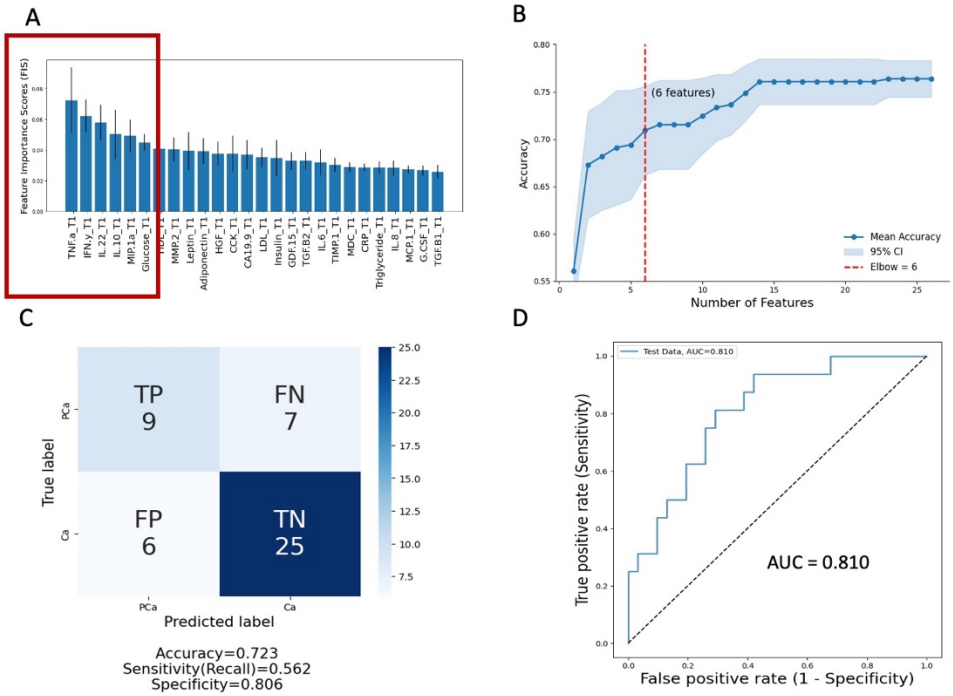

Figure 4. Feature selection and performance analysis for PCa vs. Ca predictor. A. The 26 blood biomarker candidates were pre-ranked using the FIS values obtained by the RF algorithm. **B.** The FFS approach with the cumulative curve of accuracy identified 6 predictive biomarkers indicated by the red box in **A. C.** The confusion matrix and performance metrics for the testing set. **D.** The AUC/ROC curve generated for predictions of PCa vs. Ca status for the testing set.

3.6. A predictor to differentiate between pre-cachectic and non-cachectic patients.

For the PCa vs. NCa cases, the FPC dataset contained 53 PCa patients and 28 NCa patients. This dataset was split into training (56 patients) and testing (25 patients) cohorts. Using the training cohort, the FFS method identified the following set of 5 predictive biomarkers that are listed according to their ranking: CRP, Insulin, TGF-β2, MCP-1, granulocyte colony-stimulating factor (G-CSF) (**Figure 5A,B**). Using these 5 biomarkers and the training cohort, the RBF-SVM method identified the optimal hyperparameters (C = 3.41, γ = 0.12) for model predictions, and the MCC algorithm was used to adjust for data imbalance and determine the optimal decision threshold (0.6194). The resulting confusion matrix for the testing cohort is shown in **Figure 5C** and yielded the accuracy of 0.720, the sensitivity of 0.875, and the specificity of 0.444. The AUC of model performance for the testing cohort was 0.771 (**Figure 5D**).

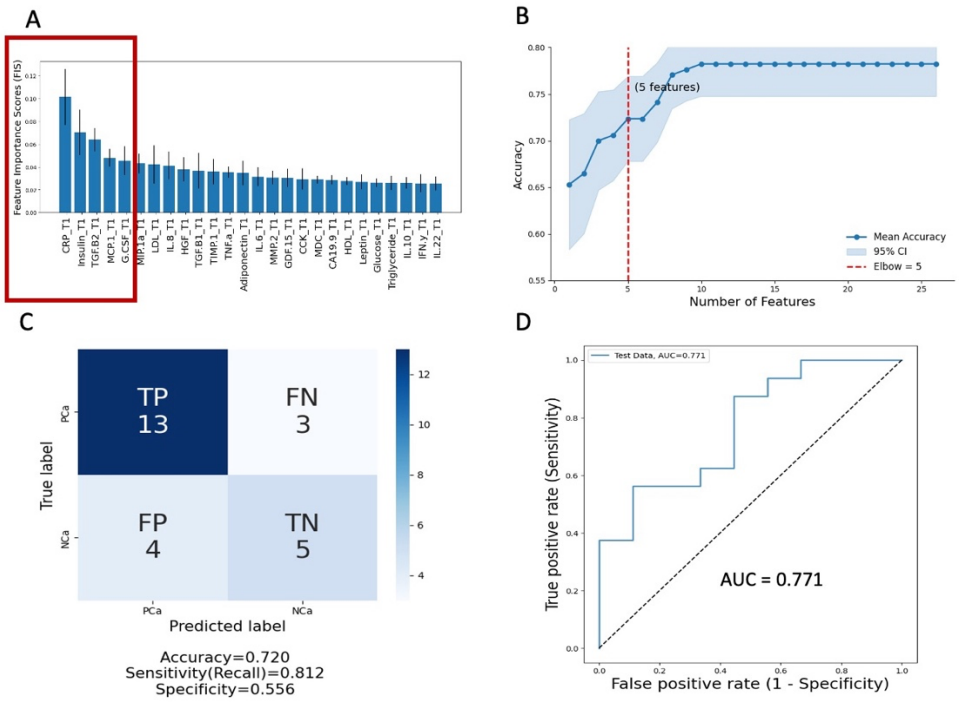

Figure 5. Feature selection and performance analysis for PCa vs. NCa predictor. A. The 26 blood biomarker candidates were pre-ranked according to the FIS values. **B.** FFS approach identified 5 predictive biomarkers indicated by the red box in **A**. **C.** The confusion matrix and performance metrics for the testing set. **D.** The AUC/ROC curve generated for predictions of PCa vs. NCa status for the testing set.

4. Discussion

In this study, we aimed to predict patients' cachexia stage using the blood biomarker data and the CC classification criteria from the Florida Pancreas Collaborative. To do so, we first developed a data-informed machine learning protocol that pre-selected a set of 26 blood biomarkers essential for model predictions of different cachexia stages. However, none of those biomarkers was singlehandedly predictive of CC status, as demonstrated by the mutual information method. Because there are potentially complex interactions between different candidate biomarkers for CC, we focused on identifying a minimal set of biomarkers that together had predictive power. For each of the three classification tasks, the forward feature selection method narrowed the number of predictive biomarkers to 5-6 that together were optimal in distinguishing between two different cachexia stages. These sets of biomarkers were then used to train three RBF-SVM (with MCC adjustment) binary classifiers to stratify PDAC patients' data by differentiating between (i) NCa and Ca stages with the accuracy of 80%; (ii) PCa and Ca stages with the accuracy of 72.3%; and (iii) PCa and NCa stages with 72% accuracy. In particular, we identified that a set of 6 biomarkers: Insulin, GDF-15, TIMP-1, CRP, TNF-α, and IL-22 was optimal in together distinguishing between Ca and NCa stages. In the case of PCa vs. Ca, we found the following set of 6 biomarkers to be predictive when used together: TNF-α, IFN-γ, IL-22, IL-10, MIP-1α, and glucose. Moreover, we demonstrated that a set of 5 biomarkers: CRP, Insulin, TGF-β2, MCP-1, and G-CSF can together effectively distinguish the PCa status from NCa. For each of the three predictors the AUC was near or greater than 0.80. Among the identified predictive blood biomarkers used in all three predictors, there is some overlap. Two blood biomarkers: IL-22 and TNF-α are used in PCa vs. Ca and NCa vs. Ca predictors. Similarly, two blood biomarkers: CRP

and Insulin are employed in NCa vs. Ca and PCa vs. NCa predictors. Interestingly, there is no overlap in blood biomarkers used in PCa vs. Ca and PCa vs. NCa predictors.

We recognize that CRP is also used as one of the criterium of the modified Vigano system for CC classification [4, 5, 12]. However, there is no one-to-one correlation between CRP levels and CC status, since this criterium is complex and contains thresholds for levels of either CRP or albumin, or hemoglobin, or white blood cell count. Similarly, in our classification, CRP is one of the biomarkers with multidimensional and nonlinear interactions that have predictive value only in combination with 5 other blood biomarkers. Moreover, several blood-based biomarkers that were previously reported to correlate with CC status for pancreatic cancer patients in our studies and by others [8-12], were also identified as predictive in our approach. These include CRP, TNF-α, MCP-1, TGF-β, and GDF-15, however another two—interleukin-6 (IL-6) and interleukin-8 (IL-8)—were not selected as necessary in any of our three predictors.

In our previous work [12], we analyzed the AUC for several analytes which were significantly different between NCa and Ca patient groups. These blood-based biomarkers included WBC count, albumin, and hemoglobin that are typically associated with cachexia status, asl well as GDF-15 and TNF-α that were identified as significantly higher in patients with Ca compared with those with NCa. For WBC count, albumin, and hemoglobin, the AUC values were between 57% and 63%, for GDF-15 or TNF-α alone, or both combined, the AUC values were between 71% and 76% [12]. However, our ML predictor that uses a combination of multiple blood-based biomarkers shows AUC of 83.5% which indicates that by considering more complex interconnectivity between patients' blood-based biomarkers may increase their predictability.

There are a few published studies that used machine learning-based approaches to identify possible biomarkers for Ca from clinical data. In [37], the authors used demographic, clinical, and patient reported outcomes (PRO) from a multi-center patient cohort study to identify biomarkers that predict Ca and PCa status. One of the factors identified to be predictive was the C-reactive protein (CRP*),* which was also selected by our approach. The model developed in [37] reported an AUC value similar to our study for differentiation between NCa vs. Ca (~0.83) and slightly lower for PCa vs. NCa (0.701 in [37] and 0.771 in our study), despite using multi-modal features and a dataset that was an order of magnitude larger than our data. Similarly, *CRP* was identified as one of the 15 top predictive biomarkers of Ca in the case when weight loss information is not available [38]. The data used by this ML-based model consisted of demographic, cancer-related clinical data, PRO related to GI symptoms, and blood-based biomarkers. The model showed good performance for predicting Ca in the validation set with the AUC of 0.763. However, this model did not address PCa status. Another ML-based approach by the same authors was used to predict potentially reversible cancer cachexia, which was defined as a cachexia diagnosis at baseline that turned negative one month later [39]. This model used clinical and demographic data for 16 different tumors, but no blood-based biomarkers. This model showed very good predictability with the AUC of 0.887 for the holdout test set and AUC of 0.863 for the external validation set. It was also suggested that the generated results can provide insights into symptoms that can be addressed to prevent or treat PCa.

The machine learning classifier presented here was used to stratify patients' blood biomarkers data into different stages of cachexia based on the modified Vigano system [4, 12]. However, similar computational frameworks can be developed for other cachexia classification criteria, such as Fearon et al. [40], Vigano et al. [5], or Martin et al. [41].

One of the limitations of this study is lack of independent dataset for external validation. Since several of blood-based biomarkers used in our predictors are not collected as a part of the standard of care procedures for the PDAC patients, further studies are needed to identify a suitable dataset to validate our findings externally.

In summary, this study showed that the integration of machine learning and deep learning techniques, such as multiple imputation method to handle missing data, feature selection techniques to identify a minimal predictive subset of blood biomarkers, and machine learning

classification that incorporates algorithms for handling data imbalance and cross-validation for optimal hyperparameter tuning, can identify predictive blood-borne biomarkers and stratify PDAC patients into different cachexia stages. Thus, the developed method has a high translational potential and could be used as a supportive tool in earlier diagnosis of the disease for cancer patients who may not show symptoms but may be on a trajectory towards CC. It may also aid in improving supportive care and clinical outcomes in the treatment of PDAC patients who are at risk of CC. Finally, it can be used as part of surveillance strategy for patients at risk of progressing to a more severe cachectic stage.

5. Conclusions

Our study established three ML models to identify cachexia stages based on blood biomarkers. The data features used by these models are minimally invasive and easily accessible. This may inform and assist clinicians in diagnosing early-stage cachexia and help guide treatment strategies for patients at risk of progressing to a more severe cachectic stage.

Ethics approval and consent to participate

This study was conducted according to guidelines of the Declaration of Helsinki, and approved by the Moffitt Cancer Center Scientific Review Committee (MCC19717, Pro00029598), and the Institutional Review board Advarra IRB (IRB00000971). All patients provided informed consent for participation.

Data and software availability

The normalized data and computational code used in this study are available from the following GitHub links: github.com/okayode/Cancer-Cachexia-PDAC-project and github.com/rejniaklab/Cancer-Cachexia-PDAC-project

Competing interests

The authors declare no competing interests.

Funding

This work was supported in part by the Department of Defense Health Program Congressionally Directed Medical Research Program grants: W81XWH-22-1-0340 (to KAR), and W81XWH-22-1-1021 LOG#PA210192 (to JBP), by the US National Institutes of Health, National Cancer Institute grant R01-CA259387 (to KAR), and by the James and Esther King Biomedical Research Program, Florida Department of Health grants #8JK02 and #24K03 (to JBP). This work was supported in part by the Shared Resources at the H. Lee Moffitt Cancer Center & Research Institute an NCI designated Comprehensive Cancer Center under the grant P30-CA076292 from the National Institutes of Health. The funders played no role in study design, data collection, analysis and interpretation of data, or the writing of this manuscript.

Authors' contributions

KDO and KAR conceptualized the project. KDO designed the presented work, developed the computational model, and created the software. JBP, EWD, and MP prepared and provisioned the data and provided recommendations on interpretation of the results. KDO and KAR drafted the manuscript. All authors revised the manuscript.

Acknowledgements

All data used in this computational study was provided by the Florida Pancreas Collaborative (FPC), a state-wide initiative and biobank. We would like to thank all participating FPC institutions, members, and patients.

References

1. Baracos, V.E., et al., *Cancer-associated cachexia.* Nat Rev Dis Primers, 2018. **4**: p. 17105.
2. Sun, L., X.Q. Quan, and S. Yu, *An Epidemiological Survey of Cachexia in Advanced Cancer Patients and Analysis on Its Diagnostic and Treatment Status.* Nutr Cancer, 2015. **67**(7): p. 1056-62.
3. Yu, Y.C., et al., *Review of the endocrine organ-like tumor hypothesis of cancer cachexia in pancreatic ductal adenocarcinoma.* Front Oncol, 2022. **12**: p. 1057930.
4. Permuth, J.B., et al., *Leveraging real-world data to predict cancer cachexia stage, quality of life, and survival in a racially and ethnically diverse multi-institutional cohort of treatment-naive patients with pancreatic ductal adenocarcinoma.* Front Oncol, 2024. **14**: p. 1362244.
5. Vigano, A.A.L., et al., *Use of routinely available clinical, nutritional, and functional criteria to classify cachexia in advanced cancer patients.* Clin Nutr, 2017. **36**(5): p. 1378-1390.
6. Gabrielson, D.K., et al., *Use of an abridged scored Patient-Generated Subjective Global Assessment (abPG-SGA) as a nutritional screening tool for cancer patients in an outpatient setting.* Nutr Cancer, 2013. **65**(2): p. 234-9.
7. Yue, M., et al., *Understanding cachexia and its impact on lung cancer and beyond.* Chin Med J Pulm Crit Care Med, 2024. **2**(2): p. 95-105.
8. Cao, Z., et al., *Biomarkers for Cancer Cachexia: A Mini Review.* Int J Mol Sci, 2021. **22**(9).
9. McMillan, D.C., *The systemic inflammation-based Glasgow Prognostic Score: a decade of experience in patients with cancer.* Cancer Treat Rev, 2013. **39**(5): p. 534-40.
10. Talbert, E.E., et al., *Circulating monocyte chemoattractant protein-1 (MCP-1) is associated with cachexia in treatment-naive pancreatic cancer patients.* J Cachexia Sarcopenia Muscle, 2018. **9**(2): p. 358-368.
11. Tsai, V.W., D.A. Brown, and S.N. Breit, *Targeting the divergent TGFbeta superfamily cytokine MIC-1/GDF15 for therapy of anorexia/cachexia syndromes.* Curr Opin Support Palliat Care, 2018. **12**(4): p. 404-409.
12. Permuth, J., et al., *Race-based differences in serum biomarkers for cancer-associated cachexia in a diverse cohort of patients with pancreatic ductal adenocarcinoma.* Communications Medicine, 2025. **in press**.
13. Pudjihartono, N., et al., *A review of feature selection methods for machine learning-based disease risk prediction.* Front. Bioinform, 2022. **2**: p. 927312.
14. Mwangi, B., T.S. Tian, and J.C. Soares, *A review of feature reduction techniques in neuroimaging.* Neuroinformatics, 2014. **12**(2): p. 229-44.
15. Permuth, J.B., et al., *The Florida Pancreas Collaborative Next-Generation Biobank: Infrastructure to Reduce Disparities and Improve Survival for a Diverse Cohort of Patients with Pancreatic Cancer.* Cancers (Basel), 2021. **13**(4).
16. Kraskov, A., H. Stogbauer, and P. Grassberger, *Estimating mutual information.* Phys Rev E Stat Nonlin Soft Matter Phys, 2004. **69**(6 Pt 2): p. 066138.
17. Loganathan, G. and M. Palanivelan, *Ovarian cancer detection from mutual information-ranked clinical biomarkers using an explainable attention-based residual multilayer perceptron.* Comput Biol Chem, 2025. **120**(Pt 2): p. 108714.
18. Aljalal, M., et al., *Selecting EEG channels and features using multi-objective optimization for accurate MCI detection: validation using leave-one-subject-out strategy (vol 14, 12483, 2024).* Scientific Reports, 2024. **14**(1).
19. Butner, J.D., et al., *Hybridizing mechanistic modeling and deep learning for personalized survival prediction after immune checkpoint inhibitor immunotherapy.* NPJ Syst Biol Appl, 2024. **10**(1): p. 88.
20. Lall, R., and Robinson, T., *Efficient Multiple Imputation for Diverse Data in Python and R: MIDASpy and rMIDAS.* Journal of Statistical Software, 2023. **107**(9): p. 1—38.

21. Bengio, Y., Yao, L., Alain, G., and Vincent, P., *Generalized denoising auto-encoders as generative models.* Proceedings of the 27th International Conference on Neural Information Processing Systems, 2013. **1**: p. 899–907.
22. Géron, A.I., *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow : concepts, tools, and techniques to build intelligent systems*. Second edition. ed. 2019, Beijing China ; Sebastopol, CA: O'Reilly Media, Inc. xxv, 819 pages.
23. Borboudakis, G. and I. Tsamardinos, *Forward-Backward Selection with Early Dropping.* Journal of Machine Learning Research, 2019. **20**.
24. Guven, E., *Decision of the Optimal Rank of a Nonnegative Matrix Factorization Model for Gene Expression Data Sets Utilizing the Unit Invariant Knee Method: Development and Evaluation of the Elbow Method for Rank Selection.* JMIR Bioinform Biotech, 2023. **4**: e43665
25. Behr, M., M. Noseworthy, and D. Kumbhare, *Feasibility of a Support Vector Machine Classifier for Myofascial Pain Syndrome: Diagnostic Case-Control Study.* J Ultrasound Med, 2019. **38**(8): p. 2119-2132.
26. Chapelle, O., P. Haffner, and V.N. Vapnik, *Support vector machines for histogram-based image classification.* IEEE Trans Neural Netw, 1999. **10**(5): p. 1055-64.
27. Guyon, I., et al., *Gene selection for cancer classification using support vector machines.* Machine Learning, 2002. **46**(1-3): p. 389-422.
28. Lee, C.P. and C.J. Lin, *A study on L2-loss (squared hinge-loss) multiclass SVM.* Neural Comput, 2013. **25**(5): p. 1302-23.
29. Boughorbel, S., F. Jarray, and M. El-Anbari, *Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric.* PLoS One, 2017. **12**(6): p. e0177678.
30. Chicco, D. and G. Jurman, *The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation.* BMC Genomics, 2020. **21**(1):6
31. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E., *Scikit-learn: Machine Learning in Python.* Journal of Machine Learning Research, 2011. **12**: p. 2825—2830.
32. Sun, L. and J. Xu, *Feature selection using mutual information based uncertainty measures for tumor classification.* Biomed Mater Eng, 2014. **24**(1): p. 763-70.
33. Brown, G., et al., *Conditional Likelihood Maximisation: A Unifying Framework for Information Theoretic Feature Selection.* Journal of Machine Learning Research, 2012. **13**: p. 27-66.
34. Breiman, L., *Random forests.* Machine Learning, 2001. **45**(1): p. 5-32.
35. Svetnik, V., et al., *Random forest: a classification and regression tool for compound classification and QSAR modeling.* J Chem Inf Comput Sci, 2003. **43**(6): p. 1947-58.
36. Menze, B.H., et al., *A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data.* BMC Bioinformatics, 2009. **10**: p. 213.
37. Chen, Y., et al., *Machine learning to identify precachexia and cachexia: a multicenter, retrospective cohort study.* Support Care Cancer, 2024. **32**(10): p. 630.
38. Yin, L.Y., et al., *Identifying cancer cachexia in patients without weight loss information: machine learning approaches to address a real-world challenge.* American Journal of Clinical Nutrition, 2022. **116**(5): p. 1229-1239.
39. Yin, L., et al., *Early identification of potentially reversible cancer cachexia using explainable machine learning driven by body weight dynamics: a multicenter cohort study.* Am J Clin Nutr, 2025. **121**(3): p. 535-547.
40. Fearon, K., et al., *Definition and classification of cancer cachexia: an international consensus.* Lancet Oncol, 2011. **12**(5): p. 489-95.
41. Martin, L., et al., *Diagnostic criteria for the classification of cancer-associated weight loss.* J Clin Oncol, 2015. **33**(1): p. 90-9.