Can a simple urine test detect pancreatic cancer?

Link Yang
10/20/2023

Outline

1. Problem Identification
2. Executive Summary
3. Modeling Preparation and Results
4. Conclusion and Future Work

Problem Identification

Background

- Pancreatic ductal adenocarcinoma (PDAC)
- Median survival: 5-6 months
- 5-year survival rate: only ~9%

5-year survival rate > 70%

if PDAC was detected at an early stage

Problem Identification

Gap in Knowledge

- No useful biomarkers for earlier detection
- Serum CA19-9 is not specific or sensitive enough

1. Healthy controls
2. Patients with non-cancerous pancreatic conditions
3. Patients with PDAC

PLOS MEDICINE

RESEARCH ARTICLE

A combination of urinary biomarker panel and PancRISK score for earlier detection of pancreatic cancer: A case–control study

Silvana Debernardi[1], Harrison O'Brien[1], Asma S. Algahmdi[1], Nuria Malats[2,3], Grant D. Stewart[4], Marija Plješa-Ercegovac[5], Eithne Costello[6], William Greenhalf[6], Amina Saad[7], Rhiannon Roberts[7], Alexander Ney[8], Stephen P. Pereira[8], Hemant M. Kocher[7], Stephen Duffy[9], Oleg Blyuss[10,11,12], Tatjana Crnogorac-Jurcevic[1]*

1 Centre for Cancer Biomarkers and Biotherapeutics, Barts Cancer Institute, Queen Mary University of London, London, United Kingdom, **2** Centro Nacional de Investigaciones Oncológicas, Madrid, Spain, **3** Centro de Investigación Biomédica en Red de Cáncer, Madrid Spain, **4** Department of Surgery, University of Cambridge, Cambridge, United Kingdom, **5** Institute of Medical and Clinical Biochemistry, Faculty of Medicine, University of Belgrade, Belgrade, Serbia, **6** Molecular and Clinical Cancer Medicine, University of Liverpool, Liverpool, United Kingdom, **7** Centre for Tumour Biology, Barts Cancer Institute, Queen Mary University of London, London, United Kingdom, **8** Institute for Liver and Digestive Health, University College London, London, United Kingdom, **9** Centre for Cancer Prevention, Wolfson Institute of Preventive Medicine, Queen Mary University of London, London, United Kingdom, **10** School of Physics, Astronomy and Mathematics, University of Hertfordshire, Hatfield, United Kingdom, **11** Department of Paediatrics and Paediatric Infectious Diseases, Institute of Child Health, Sechenov First Moscow State Medical University, Moscow, Russia, **12** Department of Applied Mathematics, Lobachevsky State University of Nizhny Novgorod, Nizhny Novgorod, Russia

* t.c.jurcevic@qmul.ac.uk

OPEN ACCESS

Citation: Debernardi S, O'Brien H, Algahmdi AS, Malats N, Stewart GD, Plješa-Ercegovac M, et al. (2020) A combination of urinary biomarker panel and PancRISK score for earlier detection of pancreatic cancer: A case–control study. PLoS Med 17(12): e1003489. https://doi.org/10.1371/journal.pmed.1003489

Academic Editor: Michele T. Yip-Schneider, Indiana University School of Medicine, UNITED STATES

Received: April 30, 2020

Accepted: November 19, 2020

Published: December 10, 2020

Data Availability Statement: All relevant data are within the manuscript and its Supporting Information files.

Abstract

Background

Pancreatic ductal adenocarcinoma (PDAC) is one of the deadliest cancers, with around 9% of patients surviving >5 years. Asymptomatic in its initial stages, PDAC is mostly diagnosed late, when already a locally advanced or metastatic disease, as there are no useful biomarkers for detection in its early stages, when surgery can be curative. We have previously described a promising biomarker panel (LYVE1, REG1A, and TFF1) for earlier detection of PDAC in urine. Here, we aimed to establish the accuracy of an improved panel, including REG1B instead of REG1A, and an algorithm for data interpretation, the PancRISK score, in additional retrospectively collected urine specimens. We also assessed the complementarity of this panel with CA19-9 and explored the daily variation and stability of the biomarkers and their performance in common urinary tract cancers.

Methods and findings

Clinical specimens were obtained from multiple centres: Barts Pancreas Tissue Bank, University College London, University of Liverpool, Spanish National Cancer Research Center

Develop an accurate model to identify patients with pancreatic cancer

to differentiate between PDAC versus non-cancerous pancreas condition and healthy

Executive Summary

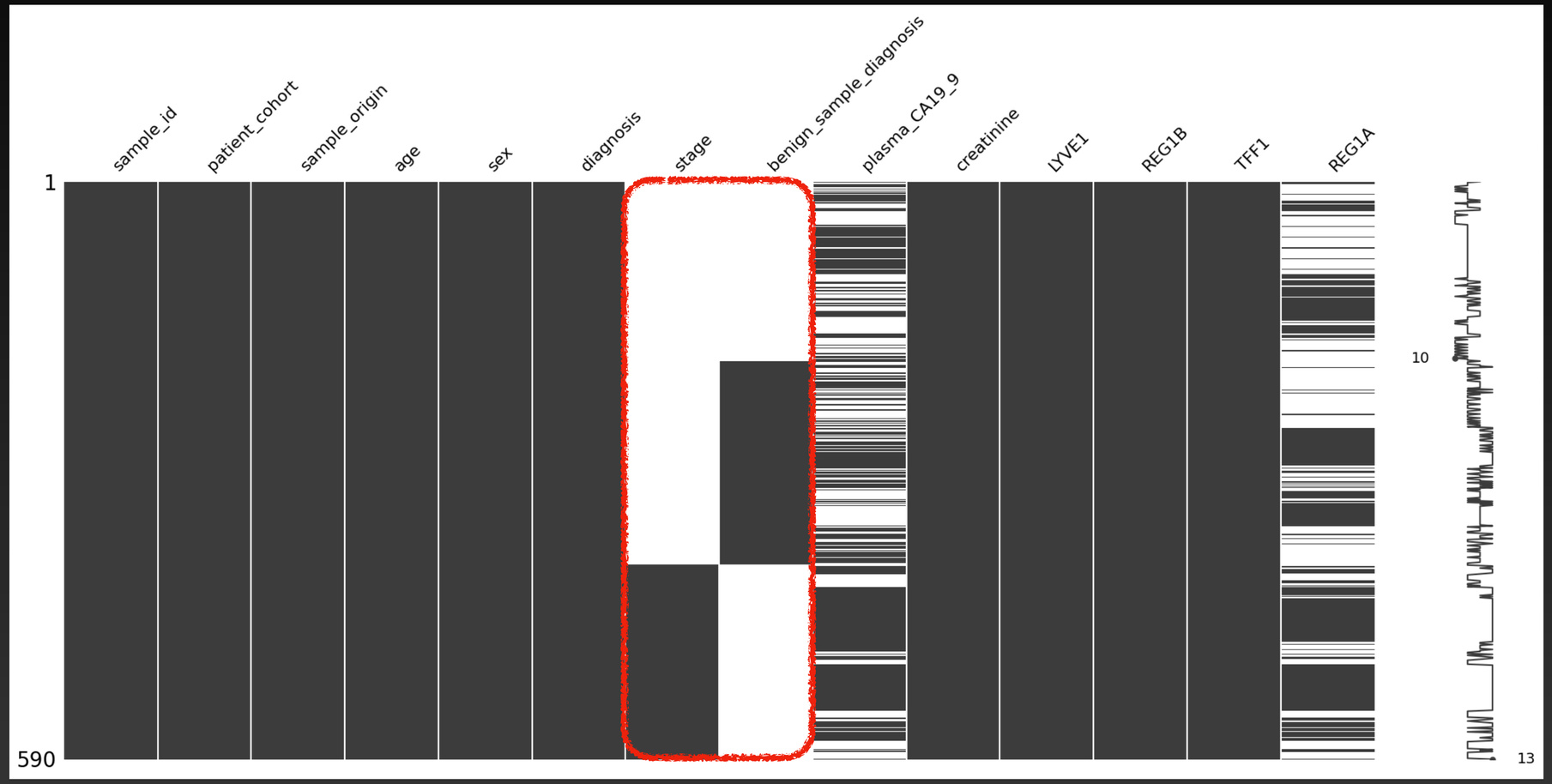
- We successfully developed a supervised machine learning model with decent performance metrics.
    - Accuracy: 78.8%
    - F1-score: 78.7%
    - Log-loss: 58.2%
- Support Vector Machine classifier outperformed the other 5 algorithms that were tested.

Modeling Preparation

Data Wrangling

- A single CSV file with 590 rows and 14 columns
    - Creatinine
    - YVLE1
    - REG1B
    - TFF1
    - REG1A
    - Plasma CA19-9

Modeling Preparation



These two should never be treated as predictor features.

Modeling Preparation

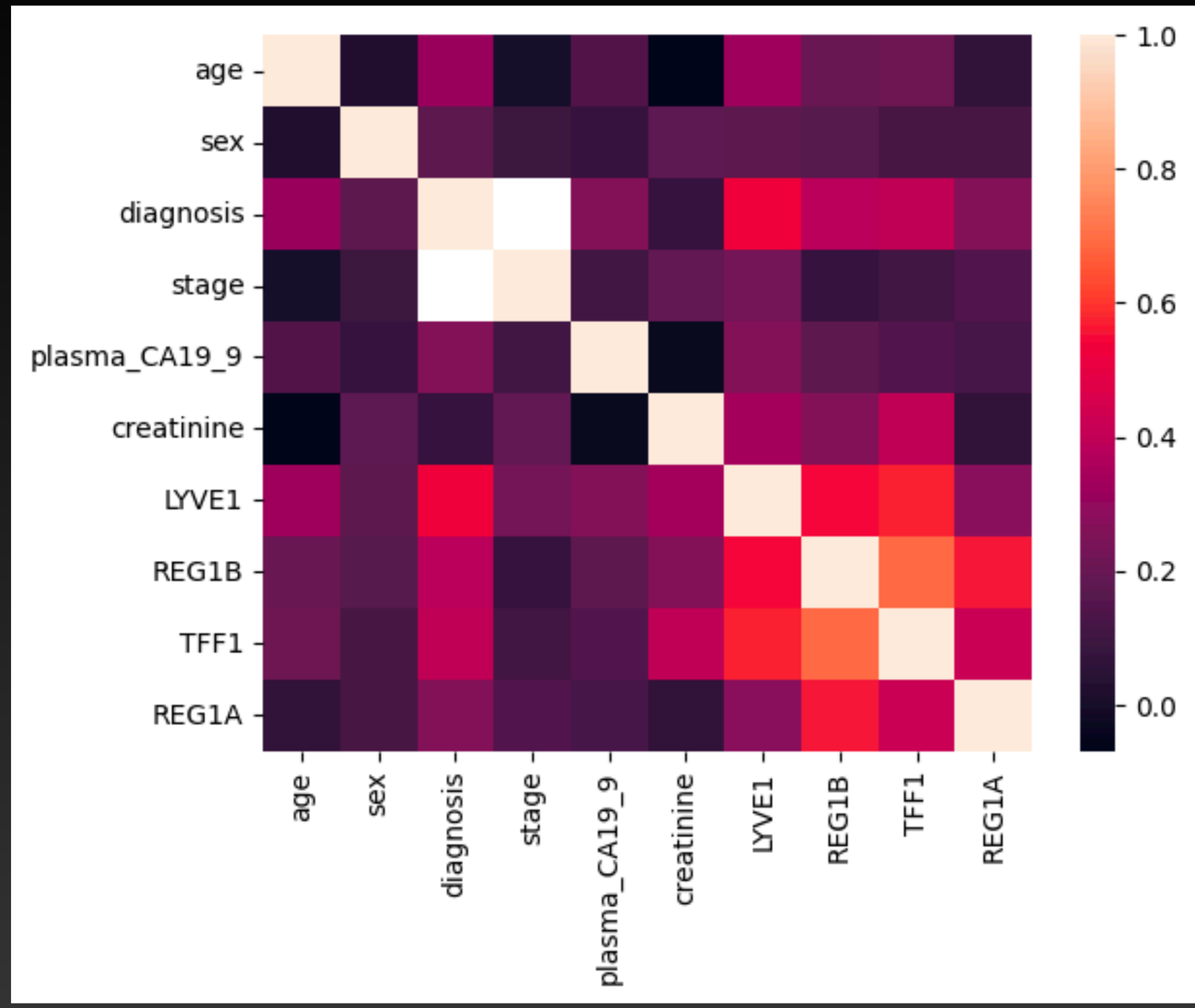
Data Wrangling & EDA

- Plasma_CA19_9 & REG1A with minimum values equal to zero
    - Replace these 0s with NaN first, and impute these missing values later
- Elderly people are more vulnerable to PDAC.
- Men are more susceptible to PDAC and non-cancerous pancreas conditions than women.
- Several urinary biomarkers are highly elevated in PDAC and non-cancerous pancreas conditions compared with healthy controls.

Modeling Preparation

EDA

- The .components_ attribute of the fitted PCA object showed that TFF1, REG1B, LYVE1, and REG1A played an important role for PC1.
- There was a relatively strong correlation between our target feature and LYVE1, followed by other predictor features.



Modeling Preparation

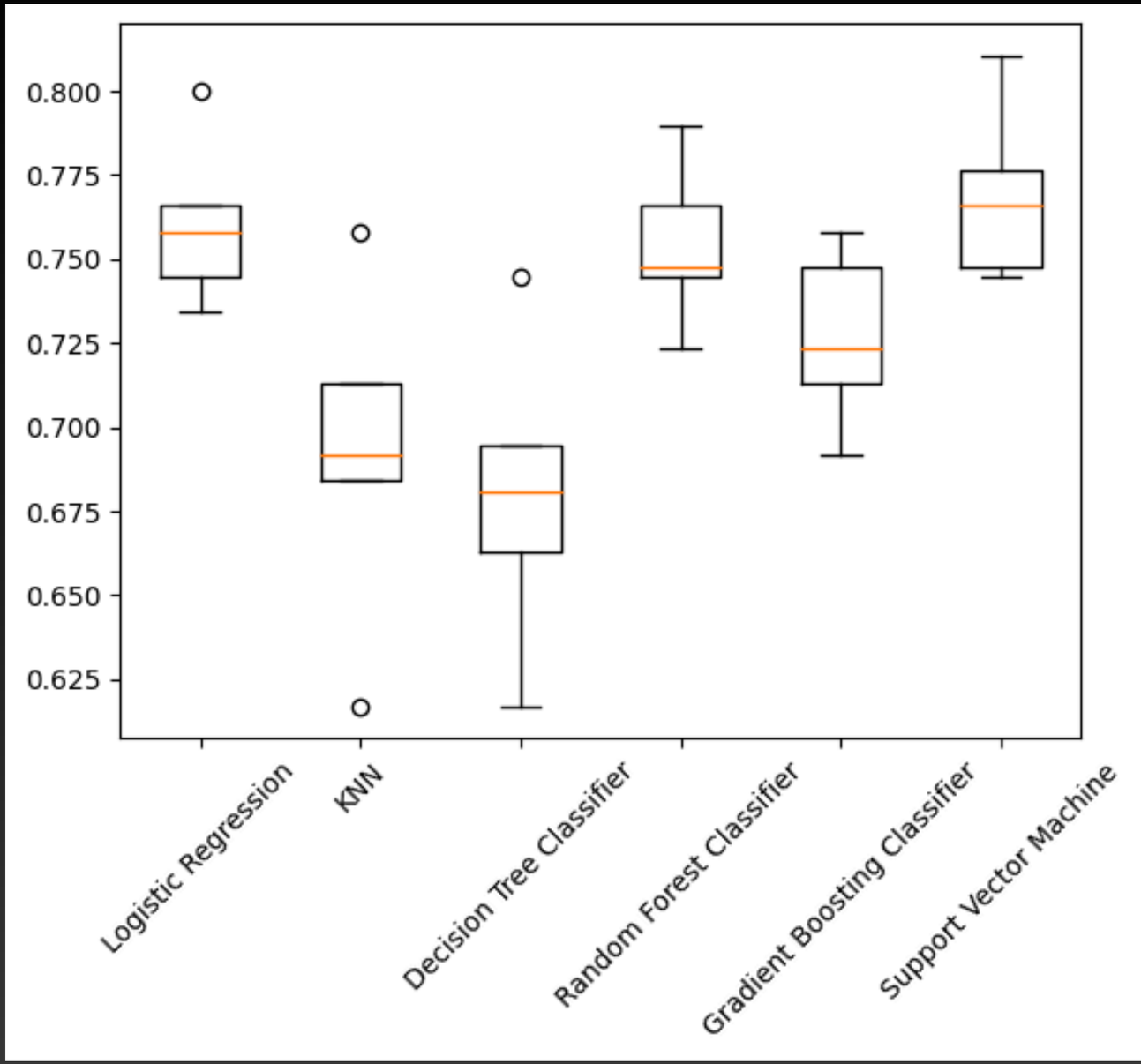
Pre-processing & Training Data Development

- Categorical Variable Encoding
    - Categorical columns containing nominal data were either one-hot or dummy encoded.
- Imputation
    - Missing numerical data was filled with median.
- Transformation
    - The distribution of all the six biomarkers was highly skewed towards the left.
- Standardization

To avoid data leakage, data was split into train and test sets before any aforementioned preprocessing.

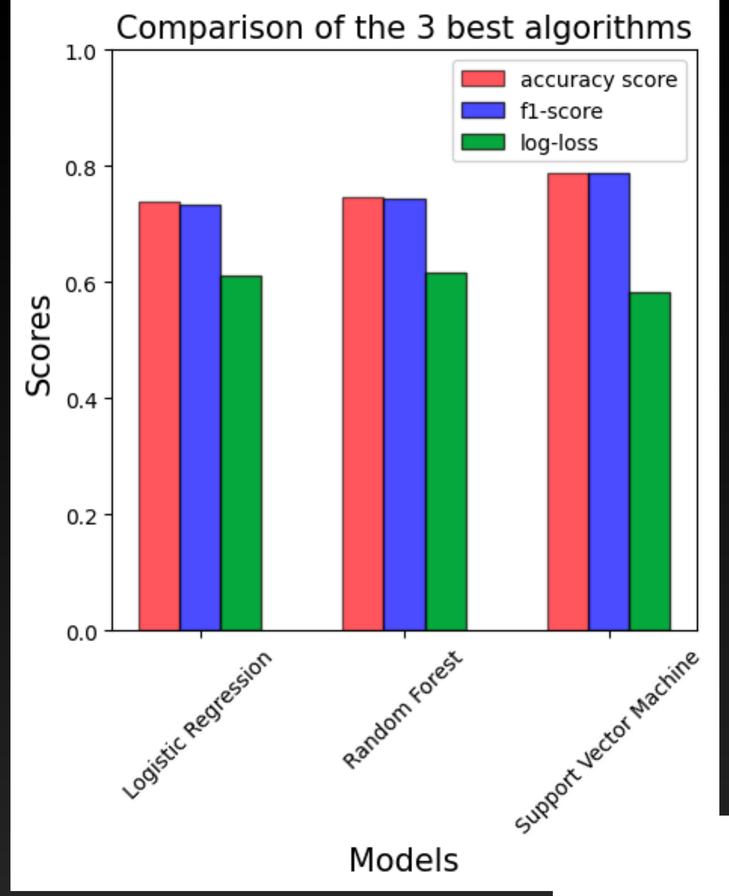
Modeling Result
without tuning

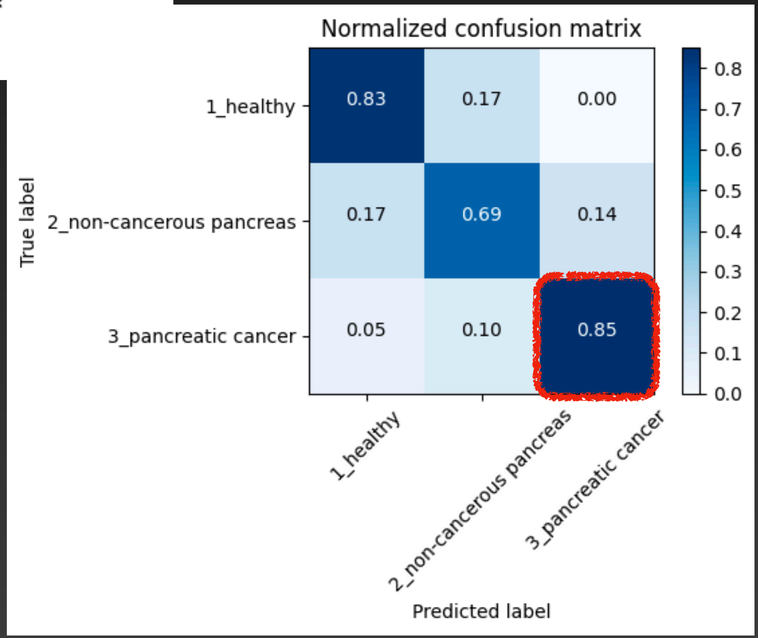
- Without any optimization, Logistic Regression, Random Forest and Support Vector Machine classifier had the average of accuracy around 75%.
- K-nearest Neighbors and Decision Tree Classifier had a significantly lower score compared with the other competitors.



Modeling Result
after hyper parameter tuning

- Three were only three models, Logistic Regression, Random Forest and Support Vector Machine classifier, with an accuracy score larger than 76%.
- Support Vector Machine classifier was the final winner:
    - Highest accuracy 78.8%
    - Highest f1-score 78.7%
    - Lowest log-loss 58.2%.
    - Decent recall score 85% for PDAC

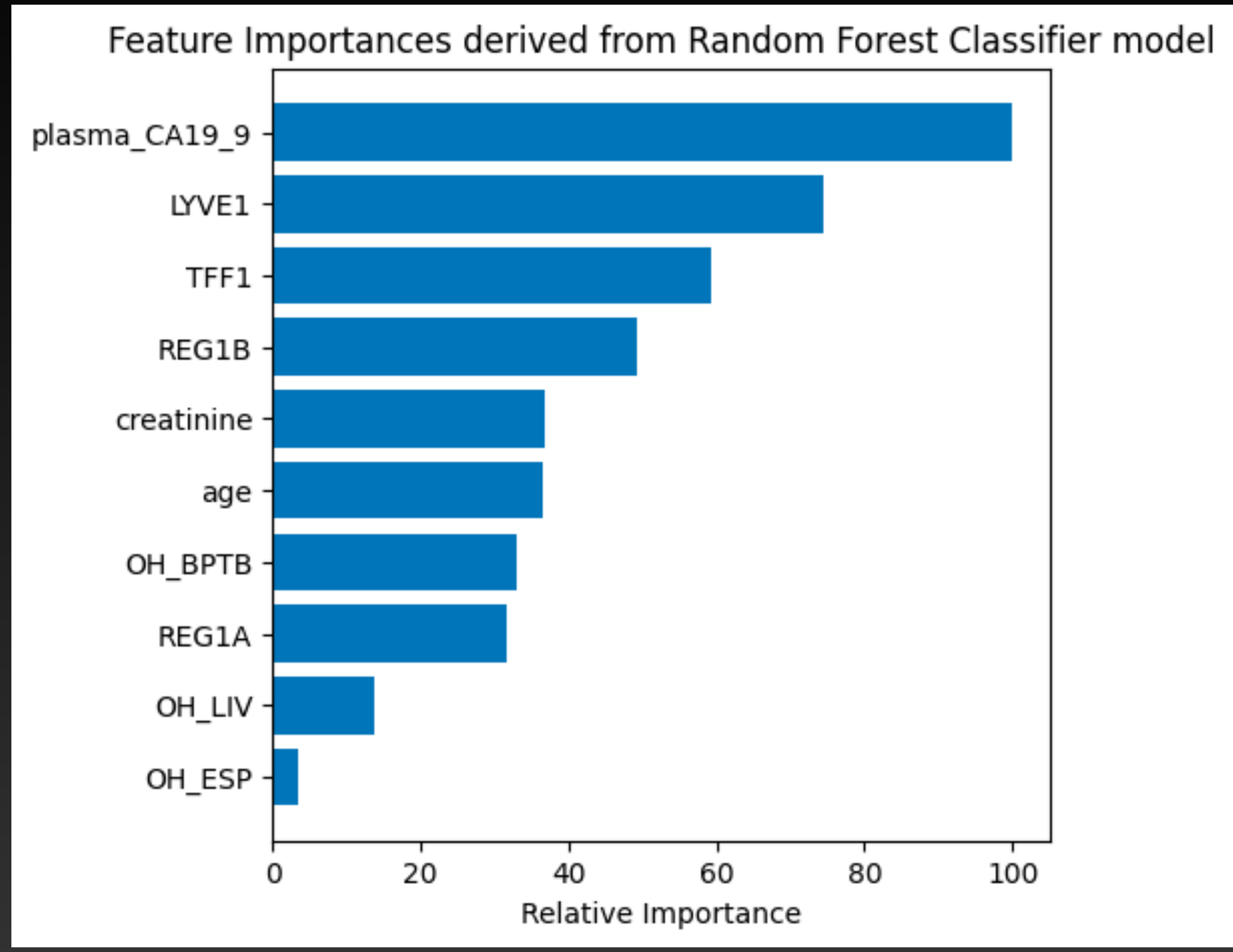




Modeling Result

Feature Importance

- Urine REG1B outperforms REG1A in detecting early stage PDAC.
- We should consider prioritizing REG1B over REG1A if we aim to limit sample collections and reduce cost.



Conclusion and Future Work

Take Home Message

- We successfully developed an accurate machine learning model which presents a potential approach to detect PDAC in a non-invasive manner.

Next Steps

- Can we augment the predictive power for non-cancerous pancreas conditions?
- Can we further improve the overall performance of our model?