Choose Your Own Project: Pancreatic ductal adenocarcinoma biomarkers

Daniel Flores Orozco

2022-01-24

1. Introduction

Pancreatic ductal adenocarcinoma (PDAC) is the most common pancreatic neoplasm and is considered a highly aggressive and lethal cancer (Sarantis et al., 2020). In fact, less than 20% of individuals diagnosed with PDAC survive the first year, and less than 10% survived more than five years (Debernardi et al., 2020; Sarantis et al., 2020). The poor prognosis is mainly attributed to late diagnosis and the lack of efficient treatments during an early stage. However, when the disease is detected in the early stages, the probability of surviving more than five years increases up to 70% (Debernardi et al., 2020). Therefore, the development of novel tests for the early detection of PDAC is essential to improve the patients' prognosis.

Debernardi et al. (2020) identified a group of proteins (creatinine, LYVE1, REG1A, REG1B, TFF1, and plasma CA19-9) in the urine that could potentially serve as biomarkers to detect PDAC. Debernardi et al. (2020) developed an algorithm to differentiate cancer samples from benign samples using LYVE1, REG1B, and TFF1 levels, as well as creatinine levels and age as predictors. Using logistic regression, they achieved a sensitivity and specificity of 0.963 and 0.967, respectively.

In this document, different Machine Learning algorithms, including random forest, kNN, SVM, logistic regression, and bayesian logistic regression, were used to predict if a given sample belonged to a benign or cancer group using the data provided by Debernardi et al. (2020). The ultimate objective of this project was to improve the performance of the model (logistic regression) used in the original manuscript.

2. Methodology

2.1 Database

The pancreatic cancer biomarkers data was downloaded from the database repository Kaggle (https://www.kaggle.com/johnjdavisiv/urinary-biomarkers-for-pancreatic-cancer). The original data contain 14 columns: *sample_id, patient_cohort, sample_origin, age, sex, diagnosis, stage, benign_sample_diagnosis, plasma_CA19_9, creatinine, LYVE1, REG1B, TFF1, REG1A*. A description of each column information is presented in **Table 1**. The database contained 590 entries of the same number of patients. The patients' ages range from 26-89 years, with 299 males and 291 females. Around 33.7% of the entries came from patients diagnosed with cancer, and the rest from healthy or other benign pancreatic diagnostic.

Table 1: Columns details in original data set

Column name	Details
sample_id	Unique string identifying each subject
patient_cohort	Cohort 1, previously used samples; Cohort 2, newly added samples
sample_origin	BPTB: Barts Pancreas Tissue Bank, London, UK; ESP: Spanish National Cancer Research Centre, Madrid, Spain; LIV: Liverpool University, UK; UCL: University College London, UK
age	Age in years
sex	M = male, F = female
diagnosis	1 = control (no pancreatic disease), 2 = benign hepatobiliary disease (119 of which are chronic pancreatitis); 3 = Pancreatic ductal adenocarcinoma, i.e. pancreatic cancer
stage	For those with pancratic cancer, what stage was it? One of IA, IB, IIA, IIIB, III, IV
benign_sample_diagnosis	For those with a benign, non-cancerous diagnosis, what was the diagnosis?
plasma_CA19_9	Blood plasma levels of CA 19â€“9 monoclonal antibody that is often elevated in patients with pancreatic cancer. Only assessed in 350 patients (one goal of the study was to compare various CA 19-9 cutpoints from a blood sample to the model developed using urinary samples).
creatinine	Urinary biomarker of kidney function
LYVE1	Urinary levels of Lymphatic vessel endothelial hyaluronan receptor 1, a protein that may play a role in tumor metastasis
REG1B	Urinary levels of a protein that may be associated with pancreas regeneration.
TFF1	Urinary levels of Trefoil Factor 1, which may be related to regeneration and repair of the urinary tract
REG1A	Urinary levels of a protein that may be associated with pancreas regeneration. Only assessed in 306 patients (one goal of the study was to assess REG1B vs REG1A)

Since the information *patient_cohort*, *sample_origin*, was not useful for the development of the different models, they were removed. Similarly, *stage*, and *bening_sample_diagnosis* were not included as the number of samples in each group was small.

2.2 Data exploration

The first step was to explore how the different predictors were distributed in the diagnosis groups (healthy, benign, and cancer) and determine if all of them could be good predictors.

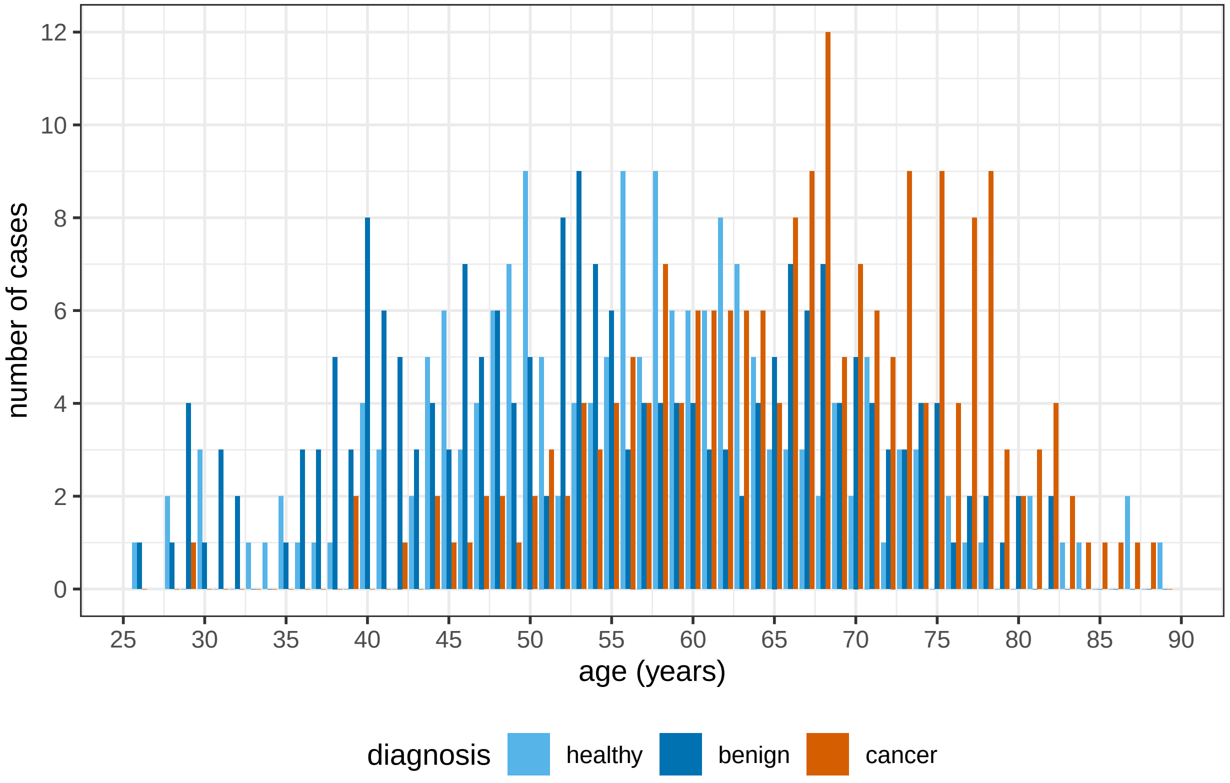
2.2.1 Diagnosis by sex

The first approach consisted of exploring PDAC diagnosis was similar in male and female patients. A contingency table combined with a Chi-squared test were used to evaluate this. The results indicated that cancer diagnosis was significantly (p-value = 0.0001) more common in males than females. For instance, 58.3% of the cancer group were males.

2.2.2 Diagnosis by age

The distribution of diagnosis by age was also evaluated (Fig.1). This analysis revealed that cancer was most common in older people. Samples from people between 65-85 were more likely to be from the cancer group. For instance, more than 93% of cancer cases came from people older than 50 years.

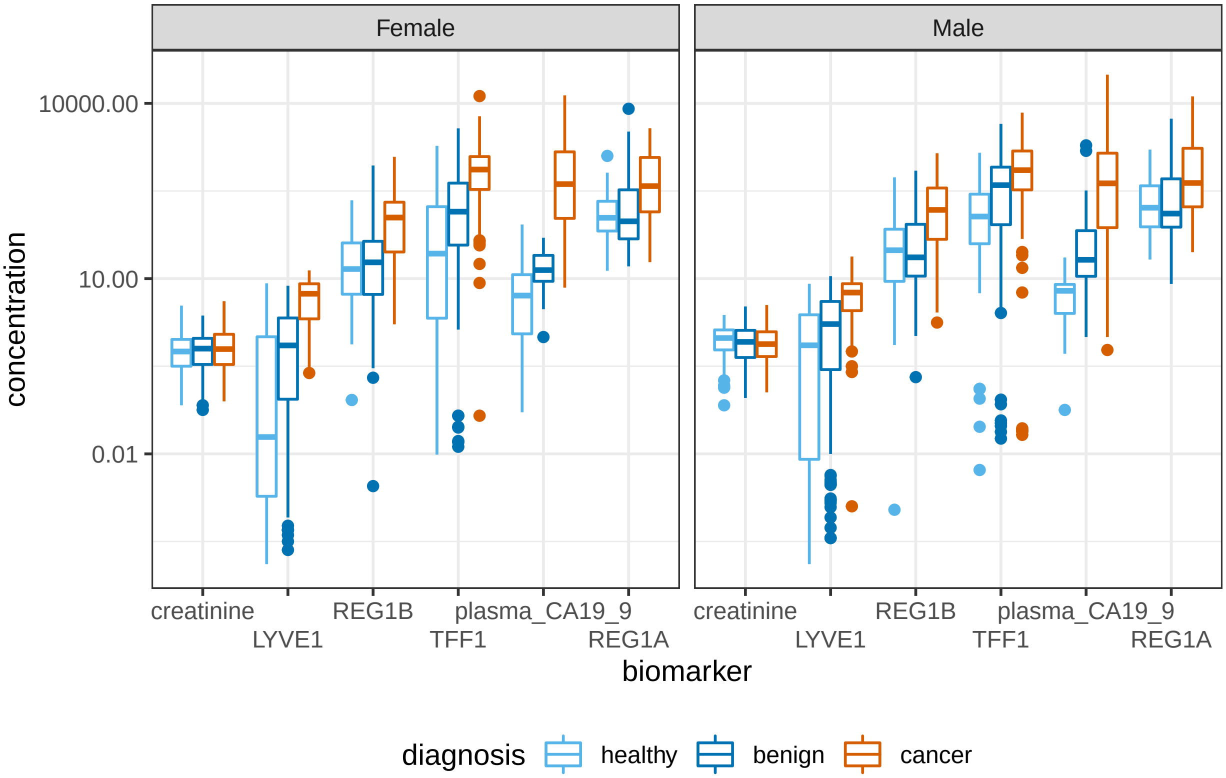
Figure 1. Diagnosis by age



2.2.3 Biomarkers by diagnosis group and sex

The evaluation of the concentration of the levels of the different biomarkers revealed that all of them, except creatinine, were overexpressed in the cancer groups compared to healthy and benign groups, as seen in **Fig. 2**. Moreover, **Fig.2** also shows that there is no noticeable difference in levels of the different biomarkers between males and females.

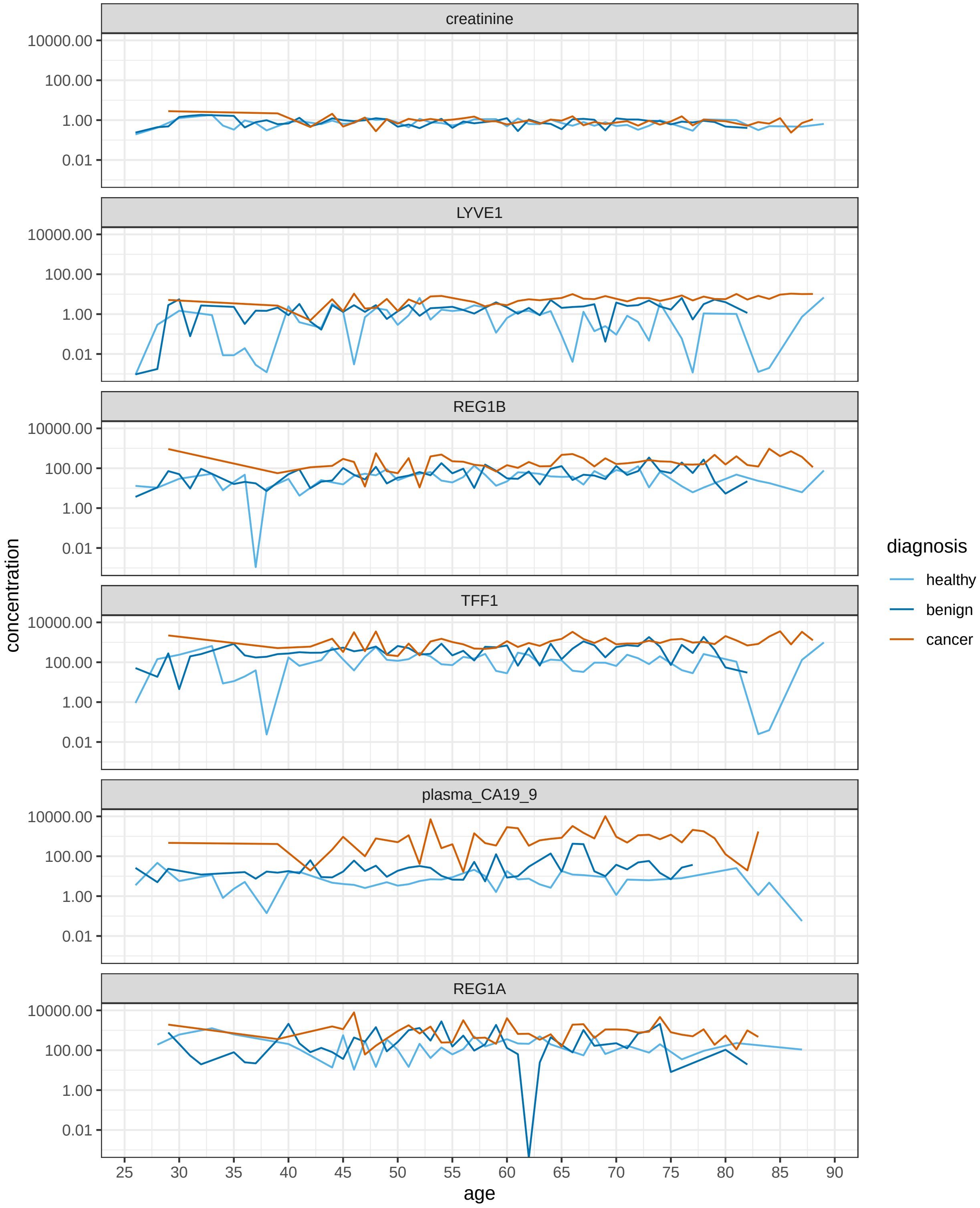
Figure 2. Biomarkers levels by diagnosis group and sex



2.2.4 Biomarkers by age and diagnosis group

Figure 3 shows the average concentration of each of the biomarkers in cancer, benign, and healthy groups at different ages. This figure shows that the creatine levels are similar and steady across all ages, which suggests that this protein may not be a good predictor. On the other hand, the levels of the other biomarkers seem to be constantly higher in cancer samples across all ages. These differences are substantially larger in older patients, which appears to result from a drop of these proteins in healthy patients rather than an increase in sick patients.

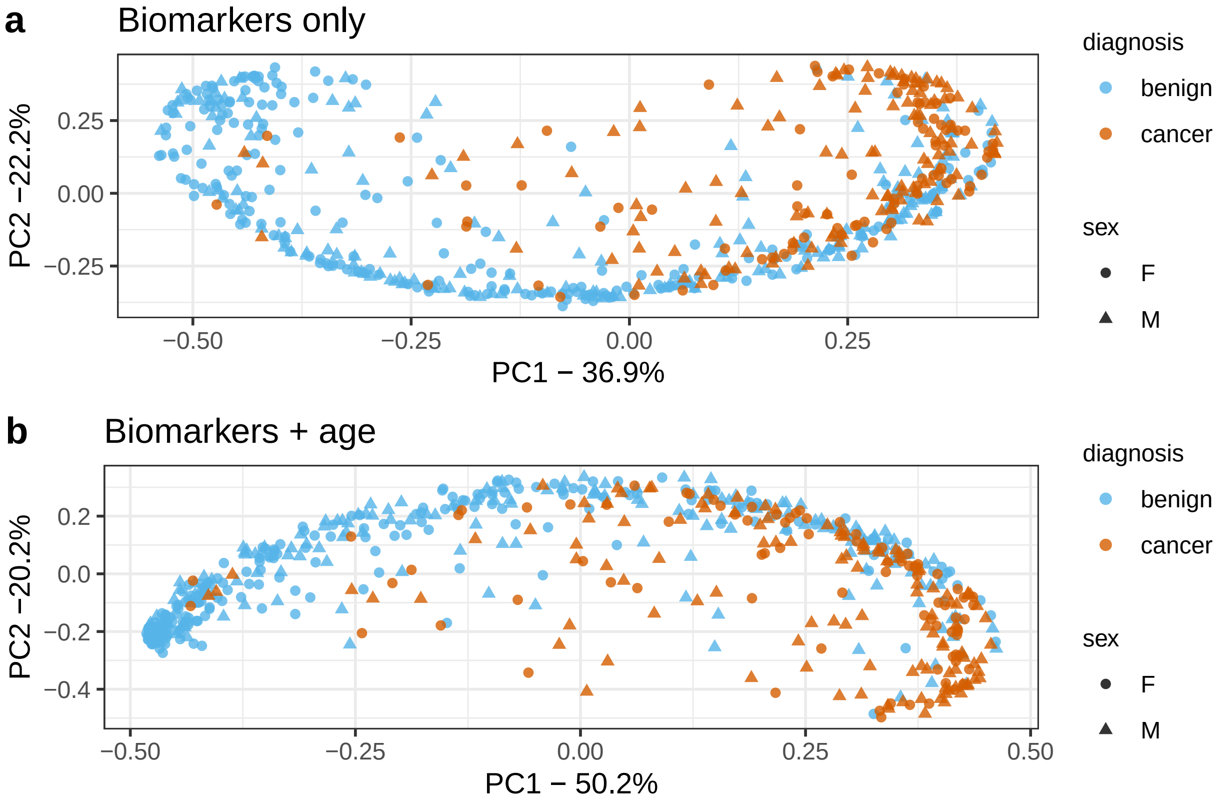
Figure 3. Average biomarker concentration by age and diagnosis group



2.2.5 Samples' distance

The final approach consisted in ordinating the samples with multidimensional scaling or principal coordinate analysis (PCoA) using the Bray-Curtis distance (Fig 4). For this, samples were classified as cancer or benign to simplify the analysis. Fig. 4a shows that when considering the biomarkers only, the two PC account for 59.1% of the variability (PC1 = 36.9%; PC2 = 22.2%). When including the patients' age (Fig.4b), the variability explained by the two PC increased to 70.4% (PC1-50.2%; PC2-22.2%), suggesting that age could be a good predictor. Although part of the variability of the data could be described for all these predictors, Fig.4 shows an overlapping of a significant fraction of cancer and benign samples. Fig.4 also shows there is no evident difference between male and female samples in the groups.

Figure 4. PCoA using Bray−Curtis distance



2.3 Loss funtion

The performance of the models was evaluated based on the proportion of cases correctly identified as patient with cancer or benign condition (overall accuracy) as well as the sensitivity and specificity of the model, which were defined as follows:

$$sensitivity : TPR = \frac{TP}{TP + FN}$$

$$specificity : TNR = \frac{TN}{TN + FP}$$

Where TPR and TNR represent the true positive rate and true negative rate, respectively; TP represents the true positive results, FN false negative results, TN true negative results, and FP false positive results.

2.4 Models

The available data was fit to several different models, including the general linear model (GLM), bayesian GLM, k-nearest neighbors (KNN), linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), random forest (rf), and a support-vector machine (SVM) model. In a first approach, only the four biomarkers present in all samples (LYVE1, REG1B, TFF1, and creatinine) plus the patient's age were used as predictors. In a second approach, all biomarkers and age were used as predictors. Thus, only samples with information on all biomarkers were included (N = 209). Since data exploration suggested that the sex of the patient was not a strong predictor, it was not included in the model developed in this project. The caret R package was used for cross-validation of all models to ensure that overfitting was avoided. A final ensemble with prediction based on the majority of votes was also developed and evaluated. The performance of the models was then compared to assess which set proteins would be better predictors of PDCA.

2.4.1 Models with 4 biomarkers

For these models, the entire data set was split into train and test sets. The train set was composed of 70% of the data, whereas the test set of the other 30%. All models were developed using the training set only and then evaluated with the test set.

2.4.2 Models with 6 biomarkers

For these models, the subset of samples with a reported concentration of all six biomarkers (N=209) was also randomly divided into training (70%) and test sets (30%).

3. Results

3.1 Three biomarkers + age models

The performance of the models using age, LYVE1, REG1B, and TFF1 as predictors is shown in Table 2. The general linear model (GLM, model used in the original study) achieved an accuracy of 0.8202 with sensitivity and specificity values of 0.667 and 0.8983, respectively. In terms of accuracy, only the KNN (0.8483) and LDA (0.8315) models performed better than the GLM. The bayesian GLM, RF and the ensemble (consensus) performed similarly to the

GLM, whereas the SVM linear kernel and QDA achieved lower accuracy. The sensitivity was comparable between all models (0.6333-0.6667), although the QDA model achieved the highest value of 0.8333. in terms of Specificity, the KNN (0.9407) was substantially better than the GLM (0.8983). In general, the results indicated that the KNN models with the three biomarkers and age as predictors was the best option to predict if a patient had cancer or not. The results also suggested that with these predictors, there was some limitation preventing most models from achieving sensitivity values greater than 0.67.

Table 2: Performance of models with three biomarkers and age as preditors

Model	Accuracy	Sensitivity	Specificity
KNN	0.8483	0.6667	0.9407
LDA	0.8315	0.6667	0.9153
GLM	0.8202	0.6667	0.8983
Bayesian GLM	0.8202	0.6667	0.8983
Random Forest	0.8202	0.6333	0.9153
Ensemble:Consensus	0.8202	0.6667	0.8983
SVM Linear Kernel	0.8146	0.6500	0.8983
QDA	0.8090	0.8333	0.7966

3.2 All biomarkers + age models

The performance of the models using age and all biomarkers are shown in Table 3. The results indicated that the use of all biomarkers and age improved the overall accuracy and sensitivity of all models. With these predictors, the QDA and RF models yielded the highest accuracy (0.9062), which was substantially higher than the obtained with the GLM (0.8750). The QDA model was also the best in terms of sensitivity (0.9512) followed by RF (0.9268) and LDA (0.9268).

Table 3: Performance of models with all biomarkers and age as predictors

Model	Accuracy	Sensitivity	Specificity
QDA	0.9062	0.9512	0.8261
Random Forest	0.9062	0.9268	0.8696
LDA	0.8906	0.9268	0.8261
GLM	0.8750	0.9024	0.8261
Bayesian GLM	0.8750	0.9024	0.8261
Ensemble:Consensus	0.8750	0.9024	0.8261
SVM Linear Kernel	0.8594	0.8780	0.8261
KNN	0.8438	0.7805	0.9565

In general, the specificity of these models is lower than the obtained with only three biomarkers as predictors. These results suggested that including all biomarkers could lead to a more significant amount of false-positive diagnoses. However, this might not be bad since false negatives could have much worse consequences, such as lack of an early diagnosis and treatment.

4. Discussion

This project aimed to improve the predictions of cancer or benign diagnosis obtained with the regular linear regression model (GLM) used by Debernardi et al. (2020) and evaluate if the use of all suggested biomarkers yielded better predictions. The results indicated that when using the LYVE1, REG1B, TFF1, and creatinine concentrations and the patient age as predictors, the kNN and LDA models could predict better if a patient had cancer or a benign condition compared to the GLM model. However, the sensitivity of these models was comparable to the GLM model. In fact, the sensitivity of all model range between 0.8333-0.6667. This relatively low sensitivity could be attributed to the high variability of the concentrations of the biomarkers in some samples. For instance, the ordination with the Bray-Curtis distance in the PCoA (Fig. 4a) showed that some cancer samples were indistinguishable from the benign samples.

When including all six biomarkers (LYVE1, REG1B, TFF1, creatinine, plasma_CA19_9, and REGG1A), the accuracy and the sensitivity of the predictions increased, which suggested that the panel with the six proteins could serve as better predictors. However, the sample size (N=209) used in these models was substantially smaller, which reduced the statistical power. It would be helpful to collect more data and increase the number of samples with concentrations of the six biomarkers, improving statistical power and making the predictions more trustable.

5. Conclusion

This study demonstrated that different machine learning algorithms could be helpful to discriminate whether a patient had pancreatic ductal adenocarcinoma (PDAC) or another benign condition based on the concentration of some urine biomarkers and age reported by Debernardi et al. (2020). The study showed that kNN, LDA, QDA, and random forest could improve the predictions obtained with the GLM model used in the original research. This study also indicated that including the biomarkers REG1A and plasma_CA19_9 as predictors could substantially increase the accuracy and sensitivity of the predictions.

6. References

Debernardi, S., O'Brien, H., Algahmdi, A.S., Malats, N., Stewart, G.D., Pljesa-Ercegovac, M., Costello, E., Greenhalf, W., Saad, A., Roberts, R., Ney, A., Pereira, S.P., Kocher, H.M., Duffy, S., Oleg Blyuss, Crnogorac-Jurcevic, T., 2020. A combination of urinary biomarker panel and PancRISK score for earlier detection of pancreatic cancer: A case-control study. PLoS Med. 17, e1003489. https://doi.org/10.1371/journal.pmed.1003489

Sarantis, P., Koustas, E., Papadimitropoulou, A., Papavassiliou, A.G., Karamouzis, M. V., 2020. Pancreatic ductal adenocarcinoma: Treatment hurdles, tumor microenvironment and immunotherapy. World J. Gastrointest. Oncol. 12, 173. https://doi.org/10.4251/WJGO.V12.I2.173