Machine Learning Model to Predict Pancreatic Ductal Adenocarcinoma Cancer

Matias Lee, Applied Mathematics and Biology, Brown University, Providence RI. GitHub

Introduction

Developing machine learning (ML) algorithms to address modern biomedical challenges is crucial for advancing medicine. One such challenge is pancreatic cancer, which in 2024 is expected to result in over 60,000 diagnoses and 50,000 deaths. Among all cancers, pancreatic cancer has the highest mortality rate, with pancreatic ductal adenocarcinoma (PDAC) being one of the most lethal forms. Only 9% of PDAC patients survive beyond five years.

Supervised ML has the potential to detect pancreatic cancer earlier, improving survival rates, as earlier detection generally leads to better outcomes. The data for developing this ML pipeline originated from Debernardi et al. (2020), where researchers conducted ELISA assays to measure protein levels in patients' urine samples. Using these biomarker data, the researchers developed an ML pipeline based on multiple logistic regression to predict whether a patient had PDAC, a benign disease, or a tumor. Their model achieved a receiver operating characteristic (ROC) AUC score of approximately 0.92 for distinguishing control versus PDAC cases, and an AUC of roughly 0.85 for other comparisons, indicating a successful algorithm.

While AUC is a useful metric, it doesn't account for "misclassification costs" in cancer diagnosis. Additionally, AUC can be hard for clinicians and patients to interpret. Therefore, I used the F1 score as an evaluation metric, which balances precision and recall, making it more relevant for this context.

In addition to using a different evaluation metric, I explored alternative ML algorithms such as XGBoost, Random Forest, KNN, and Support Vector Machines (SVM). My goal was to develop a simpler algorithm capable of predicting cancer presence. To focus on this critical biomedical question, I remapped benign disease-related diagnoses to a "non-cancer" group, simplifying the target variable into two classes: class 1 (cancer) and class 0 (non-cancer). This binary classification better aligns with the urgent need to determine whether cancer treatment is necessary.

Exploratory Data Analysis (EDA)

The PDAC dataset consists of 590 urine specimens and 14 features collected from six different biobanks. The target variable in this dataset is cancer diagnosis, categorized as either control, benign hepatobiliary disease, or PDAC. For this project, I merged the control and benign hepatobiliary disease categories into a single group designated as non-cancer to simplify the problem into a binary classification: cancer or non-cancer.

The target variable is named ‘diagnosis’. The categorical features in the dataset include ‘sex’, ‘stage’, ‘benign_sample_diagnosis’, ‘sample_origin’, ‘patient_cohort’, and ‘sample_id’. Since the goal is to predict cancer diagnosis, most of these features—except for ‘sex’—were excluded from the machine learning pipeline. Features such as ‘stage’, ‘benign_sample_diagnosis’, ‘sample_origin’, ‘patient_cohort’, and ‘sample_id’ were omitted because they are either identifiers, post-diagnostic notes, or irrelevant to the pre-diagnostic classification of cancer.

On the other hand, all continuous features were retained for the pipeline, as they are biologically relevant to the prediction of cancer. These features include ‘plasma_CA19_9’, ‘creatinine’, ‘LYVE1’, ‘REG1B’, ‘TFF1’, ‘REG1A’, and ‘age’. Below are figures highlighting key aspects of these features in the dataset.

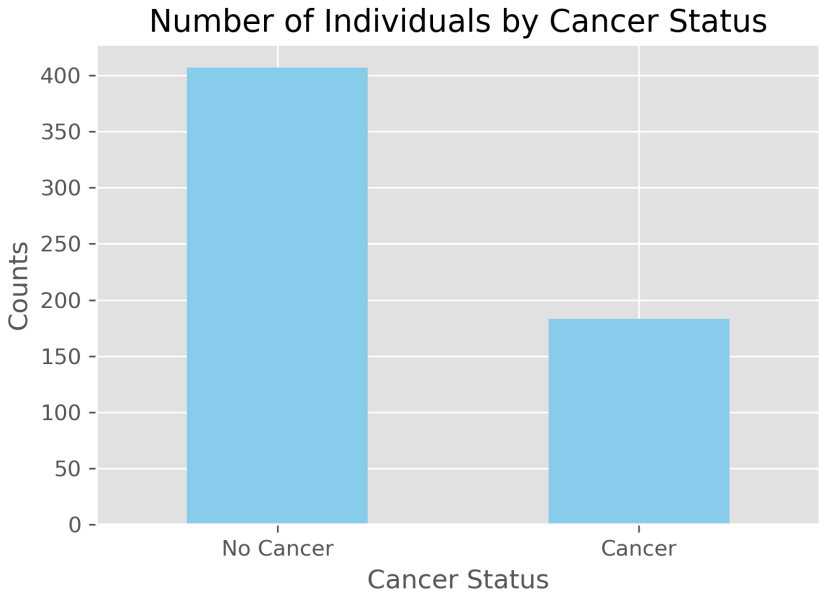


Figure 1: Distribution of Target Variable: Diagnosis

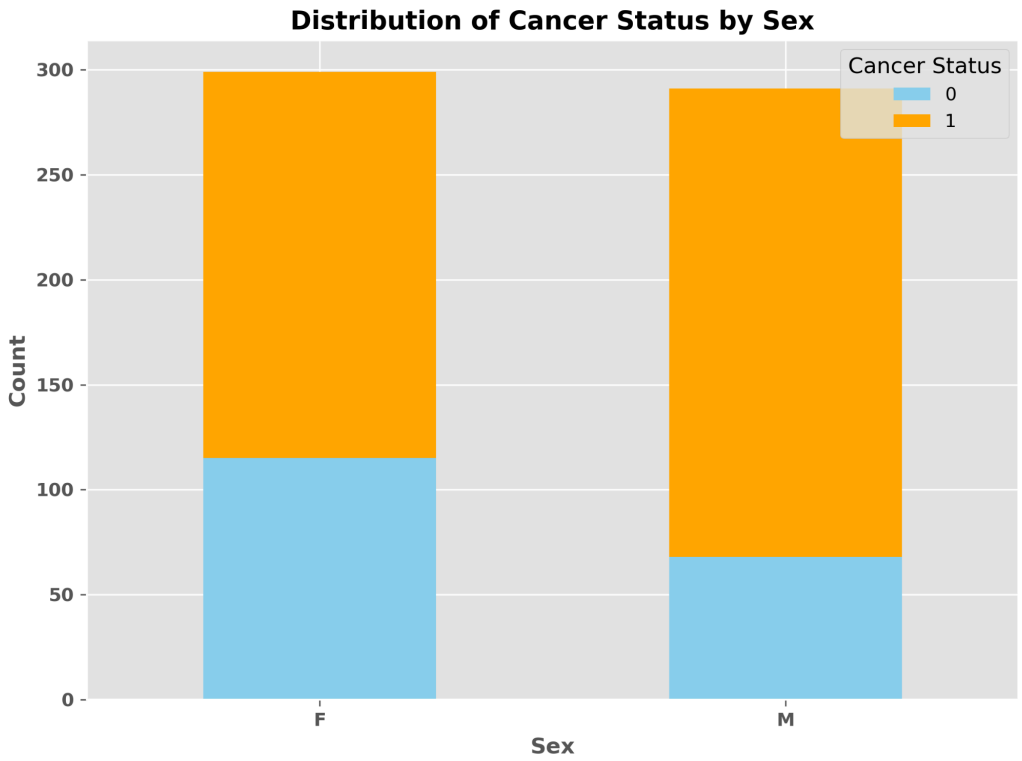


Figure 2: Distribution of Categorical Feature based on Diagnosis: Sex

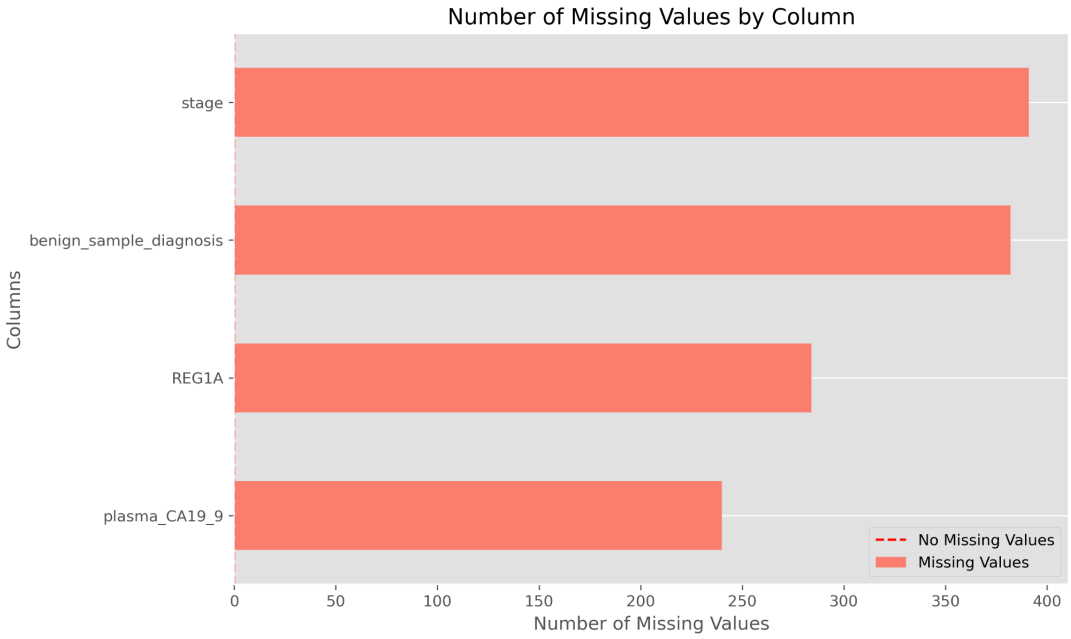


Figure 3: Number of Missing Values from PDAC Dataset

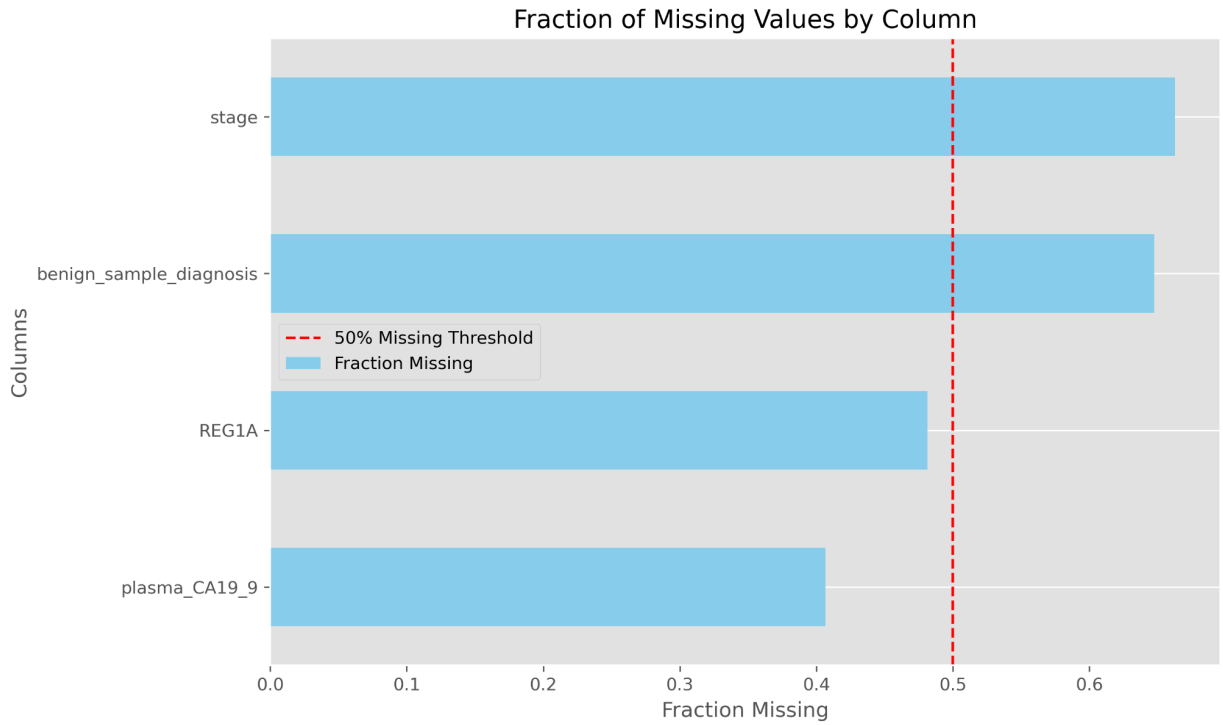


Figure 4: Percent of Missing Values from PDAC Dataset. REG1A and Plasma_CA19_9 are the two features that will be used in the dataset.

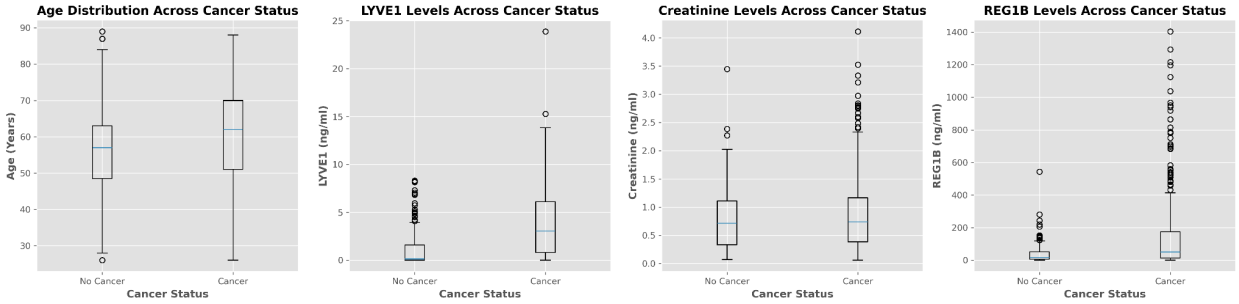


Figure 5: Boxplots of Age, LYVE1, Creatinine and REG1B depending on Cancer Status.

Figure 1 indicates that the cancer and non-cancer diagnoses are sufficiently similar to be considered balanced. Similarly, **Figure 2** demonstrates that the categorical feature 'sex' is evenly distributed across diagnoses, with no significant imbalance. **Figures 3** and **4** reveal missing values in the dataset: not all samples were assigned a stage or a benign sample diagnosis, and some lacked measurements for the REG1A and plasma_CA19_9 biomarkers. **Figure 5** presents boxplots of age and selected biomarkers (LYVE1, Creatinine, and REG1B). Both age and LYVE1 show a slight increase in mean values from non-cancer to cancer diagnoses, while Creatinine and REG1B exhibit minimal or no change in mean across diagnoses.

Methods

The PDAC dataset is generally independently and identically distributed on the target variable, ‘diagnosis’, allowing for the use of a k-fold split (k=5) with a standard 60/20/20 division for training, validation, and test data. For preprocessing, I applied a OneHotEncoder for the categorical feature ‘sex’ and used scikit-learn's StandardScaler for the continuous features. Missing values in REG1A and plasma_CA19_9 were imputed using scikit-learn’s IterativeImputer, which provides a simple yet robust method for handling missing data. Preprocessed datasets for each model’s training and test splits, across random states, are stored in the /data folder of the GitHub repository.

I evaluated five machine learning algorithms: XGBoost, Logistic Regression, Random Forest, Support Vector Classification (SVM), and K-Nearest Neighbors (KNN), optimizing their hyperparameters using GridSearchCV with k-fold cross-validation (see **Figure 7**). Logistic Regression performed best with a regularization strength (C) of 0.01 and the liblinear solver, which helped control model complexity. Random Forest achieved optimal performance with a maximum tree depth of 20, a minimum of 5 samples required to split a node, and 100 estimators, promoting deeper trees and robust ensemble learning. SVM performed best with a regularization strength (C) of 1, a gamma value of 0.1, and an RBF kernel, which enabled the model to capture non-linear decision boundaries. For KNN, the best configuration was 5 neighbors with a uniform weighting scheme, giving equal importance to all neighbors. XGBoost performed optimally with a tree depth of 1, 100 estimators, and a subsample fraction of 20%, balancing model complexity and overfitting. These parameters were at the edge of the hyperparameter range and could not be further reduced. The summary of these results can be found in **Table 1.**

Once the optimal hyperparameters were determined, I refit each model and tested it on the corresponding test data, repeating this process five times for different random states. This produced a set of five test score means and their standard deviations, along with baseline scores.

The evaluation metric for this pipeline was the F1 score, which is the weighted harmonic mean of precision and recall. The F1 score emphasizes the significance of false negatives and false positives, both of which are critical in cancer diagnosis. Misdiagnosing cancer—whether as a false positive or false negative—has equal implications for patient outcomes. I chose the F1 score over other metrics like accuracy or ROC AUC because it is more interpretable in the context of cancer diagnosis.

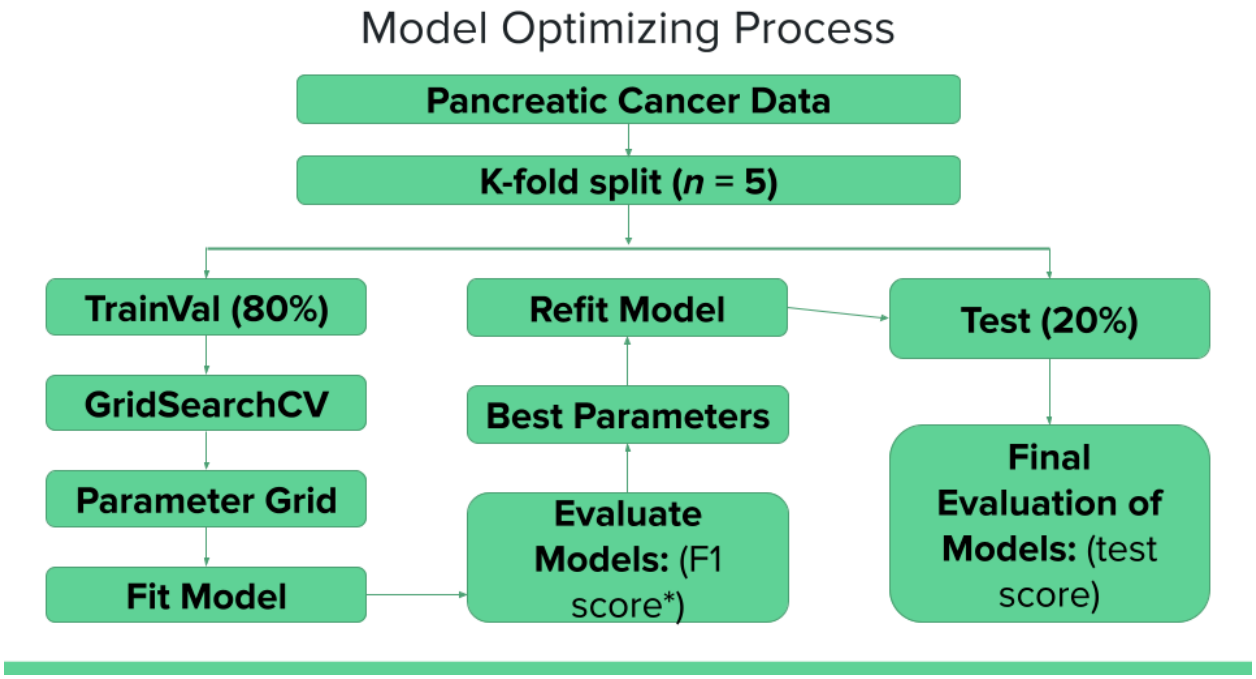


Figure 7: Schematic of cross validation and model optimizing workflow for a given random state

ML Algorithm	Tunable Hyperparameters	Optimal Values
XGBoost	model__max_depth model__subsample model__n_estimators	[**1**, 2, 3, 5, 7, 9] [**0.2**, 0.4, 0.5, 0.8, 1.0] [100]
Random Forest	model__max_depth model__min_samples_split model__n_estimators	[2, 5, 10, **20**, 30, 50, 100] [2, **5**, 7, 10, 15, 30] [100]
Logistic	model_ C model__solver	[0.0001, 0.001, **0.01**, 0.1, 1, 10, 100] [**'liblinear'**, 'lbfgs']
SVM	model__C model__kernel model__gamma	[0.0001, 0.001, 0.01, 0.1, **1**, 10, 100] ['linear', **'rbf'**] [0.0001, 0.001, 0.01, **0.1**, 1, 10, 100, 1000]
KNN	model__n_neighbors model__weights	[3, **5**, 10, 15, 30, 100] [**'uniform'**, 'distance']

Table 1: Table of ML Algorithm and corresponding optimal hyperparameter values. The optimal ML algorithm and the optimal values for each ML algorithm are in bold.

Results

The performance of each ML algorithm is summarized in **Table 2**, where the standardized difference is calculated as the difference between the mean test score and the baseline score, divided by the standard deviation of the mean test score. The baseline score was 0.5044. The best-performing model was Random Forest, with an F1 score of 0.715 and a standard deviation of 0.0243, yielding the highest standardized difference of 8.665. XGBoost and SVM followed closely, with F1 scores of 0.69 and 0.685, respectively, and standardized differences of 6.613 and 3.651. This suggests that while XGBoost and SVM performed similarly, XGBoost showed more consistency with a smaller standard deviation. The lowest-performing models were Logistic Regression and KNN, with scores of 0.67 and 0.657, respectively. These results are also depicted in **Figure 7**. In **Figure 8**, confusion matrices for the best random state show that SVM's best model slightly outperformed Random Forest's, with a total of 21 false negatives and false positives, compared to 22 for Random Forest.

Algorithm	Mean Test Score	Standard Deviation	Standardized Difference
Logistic Regression	0.67	0.046	3.555
Random Forest	0.715	0.024	8.665
SVM	0.685	0.049	3.651
KNN	0.657	0.05	3.054
XGBoost	0.69	0.028	6.613

Table 2: Table showing model's test scores and standard deviations compared to baseline

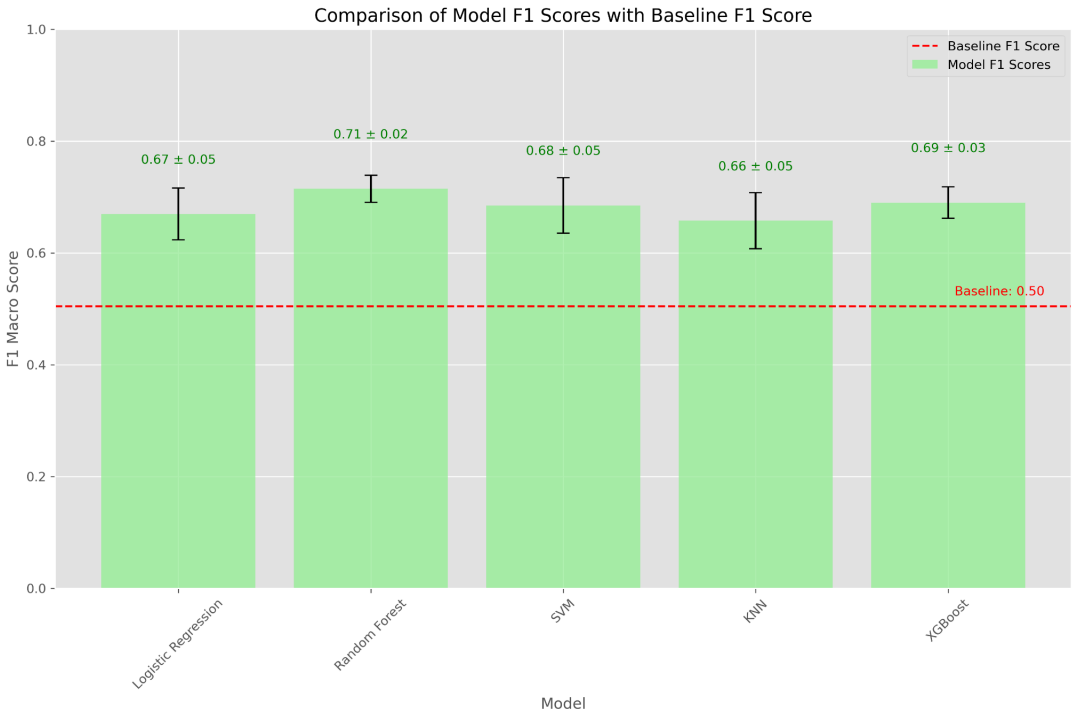
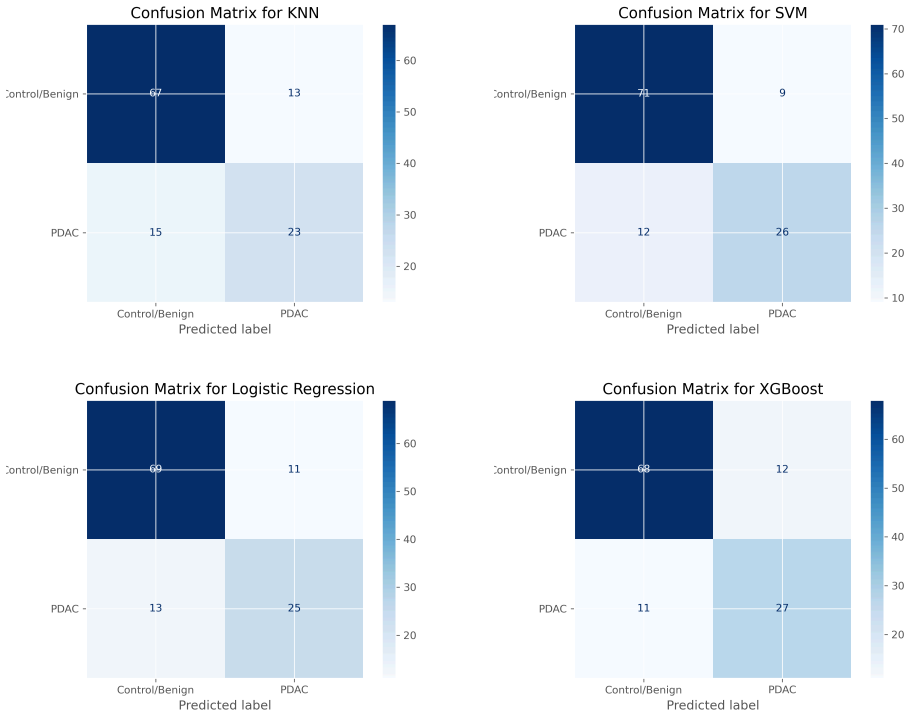


Figure 6: Bar chart for F1 Test Scores of 5 models compared to Baseline. Random Forest and XGBoost with the highest F1 scores and highest standardized differences.



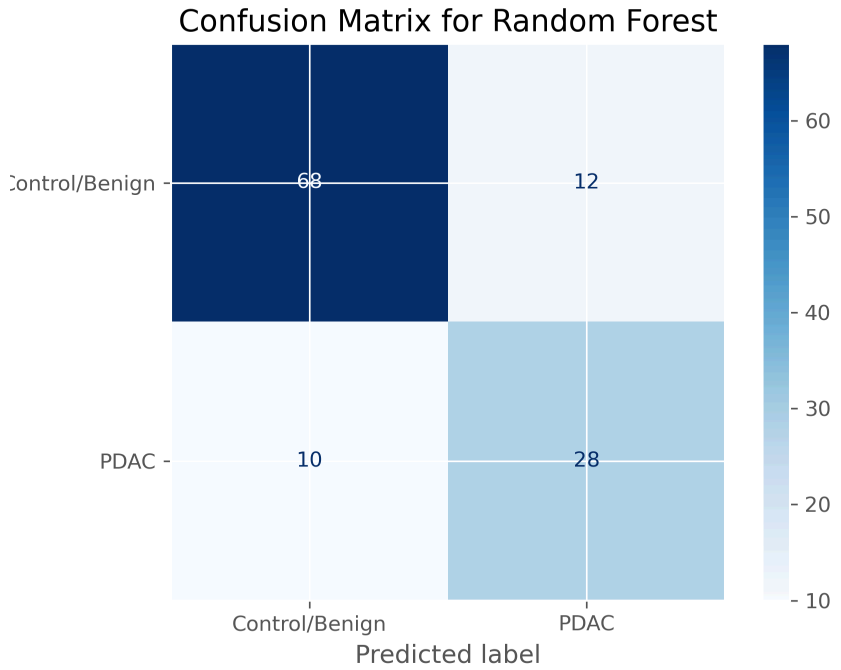


Figure 7: Confusion matrices for each of the given models' best random state with Random Forest enlarged.

Global and Local Feature Importance

Using the Random Forest model's best random state, I calculated the mean absolute SHAP values for the most important features and created a summary plot of the SHAP value distributions (**Figures 8 and 9**). Additionally, in **Figure 10**, I examined XGBoost's global feature importance metrics—Weight, Gain, and Cover—to identify any similarities or differences in the features driving predictions compared to Random Forest. From **Figures 8** and **9**, LYVE1, Age, Creatinine, and TFF1 emerge as the most influential features, with LYVE1 showing nearly double the importance of the others. Similarly, in Figure 10, these features consistently rank among the top five, aligning well with the SHAP analysis.

The prominence of LYVE1 and TFF1 is biologically meaningful, as both proteins are overexpressed in response to cancers. TFF1 is linked to increased cancer motility and is expressed in various cancers, including breast, lung, and prostate cancers, while LYVE1 is associated with lymphatic vessel growth and cancer metastasis. Age, a significant determinant of cancer, is also consistent with its role in cancer diagnoses, as the likelihood of developing cancer increases with age.

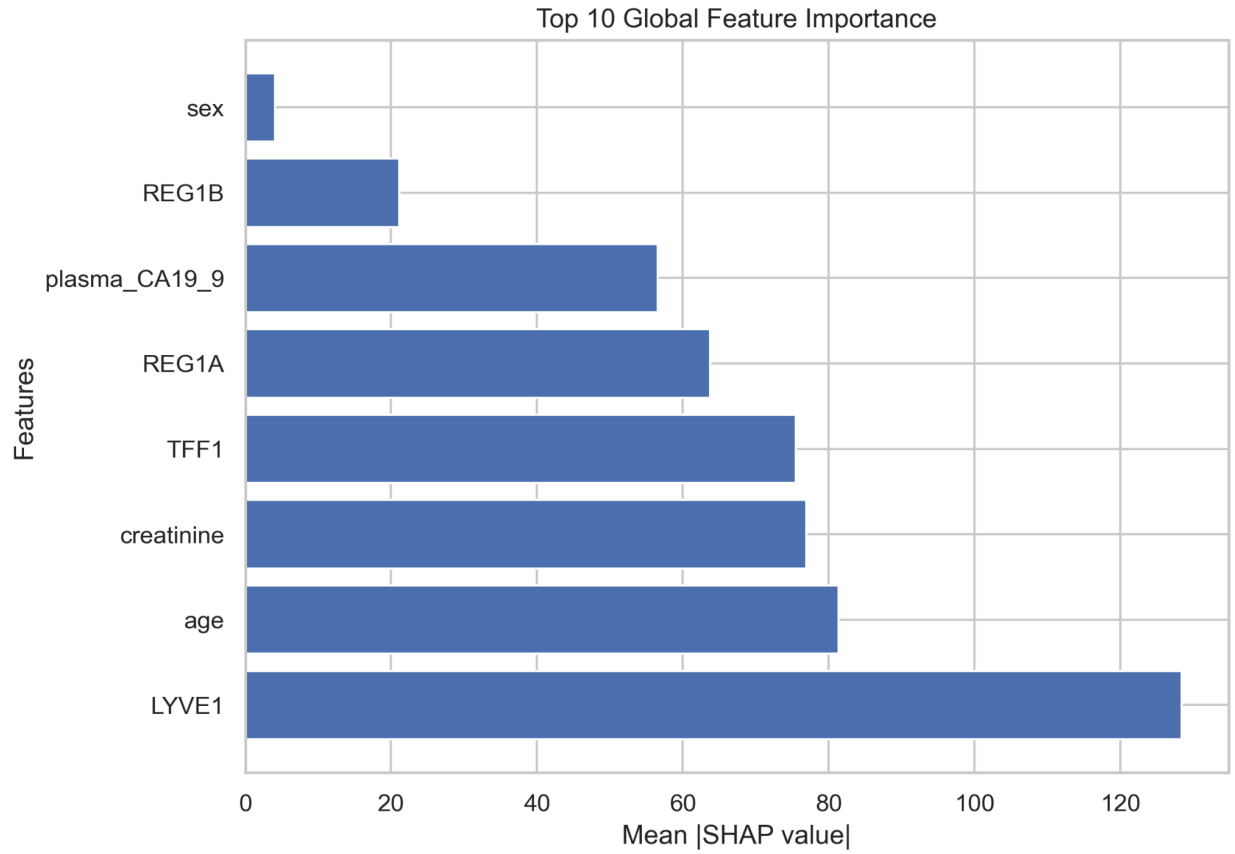


Figure 8: Mean absolute SHAP Value plot showing LYVE1, Age Creatinine and TFF1 being some of the most important features with SHAP values over 70.

Figure 9: A SHAP value summary plot that shows the distribution of the top features at high and low feature values. Most notably, low feature values of LYVE1 and Age were more likely to produce a diagnosis of no cancer and at high levels were more likely to indicate a positive diagnosis of cancer.

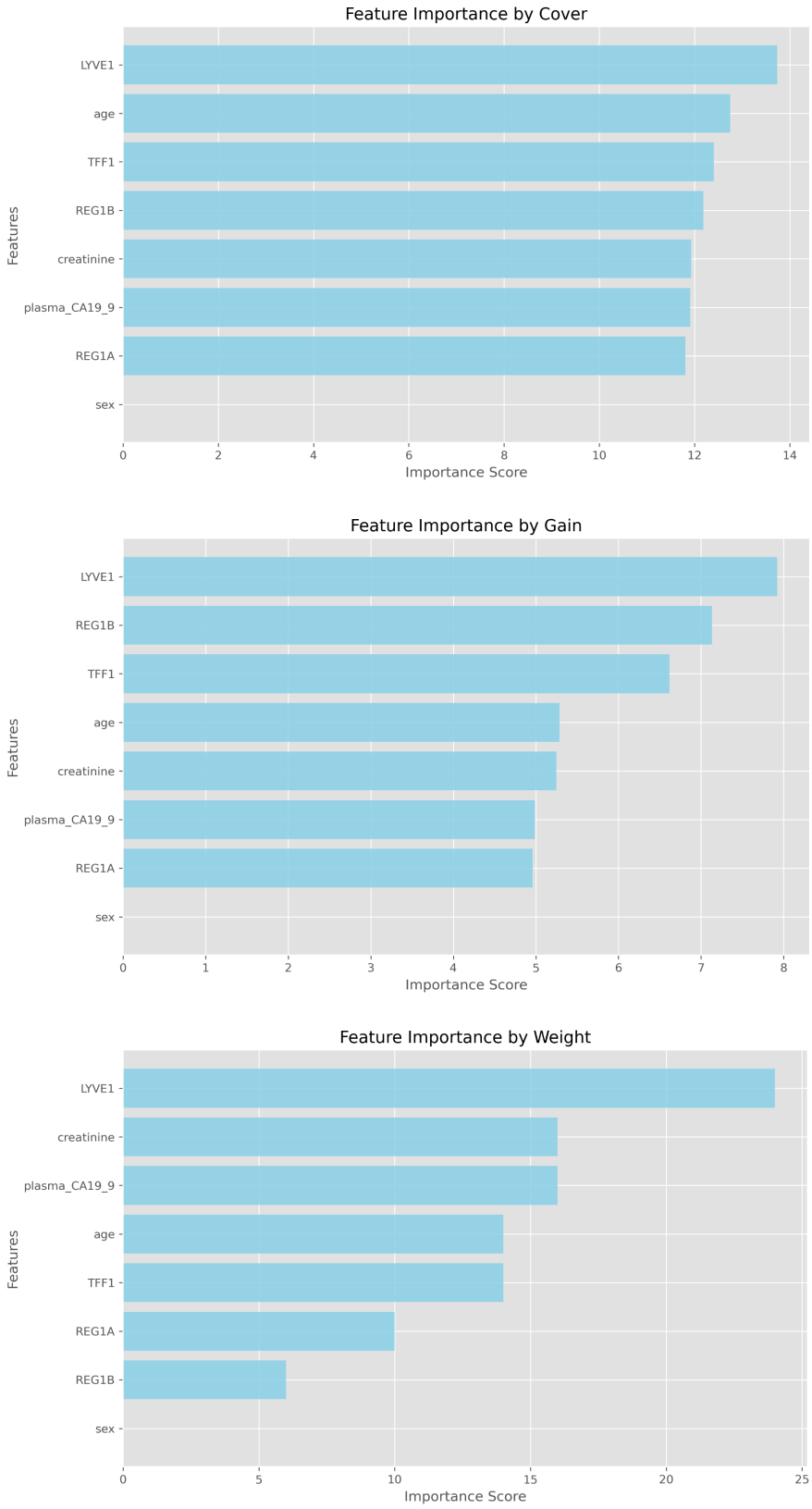


Figure 10: Feature Importance Metrics for using XGBoost's Gain, Weight and Cover

In **Figure 11**, I created three separate force plots to investigate local feature importances and assess their alignment with global feature importances. For the three randomly selected data points, TFF1, LYVE1, Creatinine, and Age consistently appear as key contributors to the predictions. Both Age and LYVE1 show a positive correlation between feature values and SHAP values, which aligns well with the global feature importance analysis. However, plasma_CA19_9 emerges as the primary driver for indices 1 and 101, despite not being prominent in the global feature importance metrics. While this may indicate outliers, it is notable that plasma_CA19_9 plays an outsized role in these specific cases.

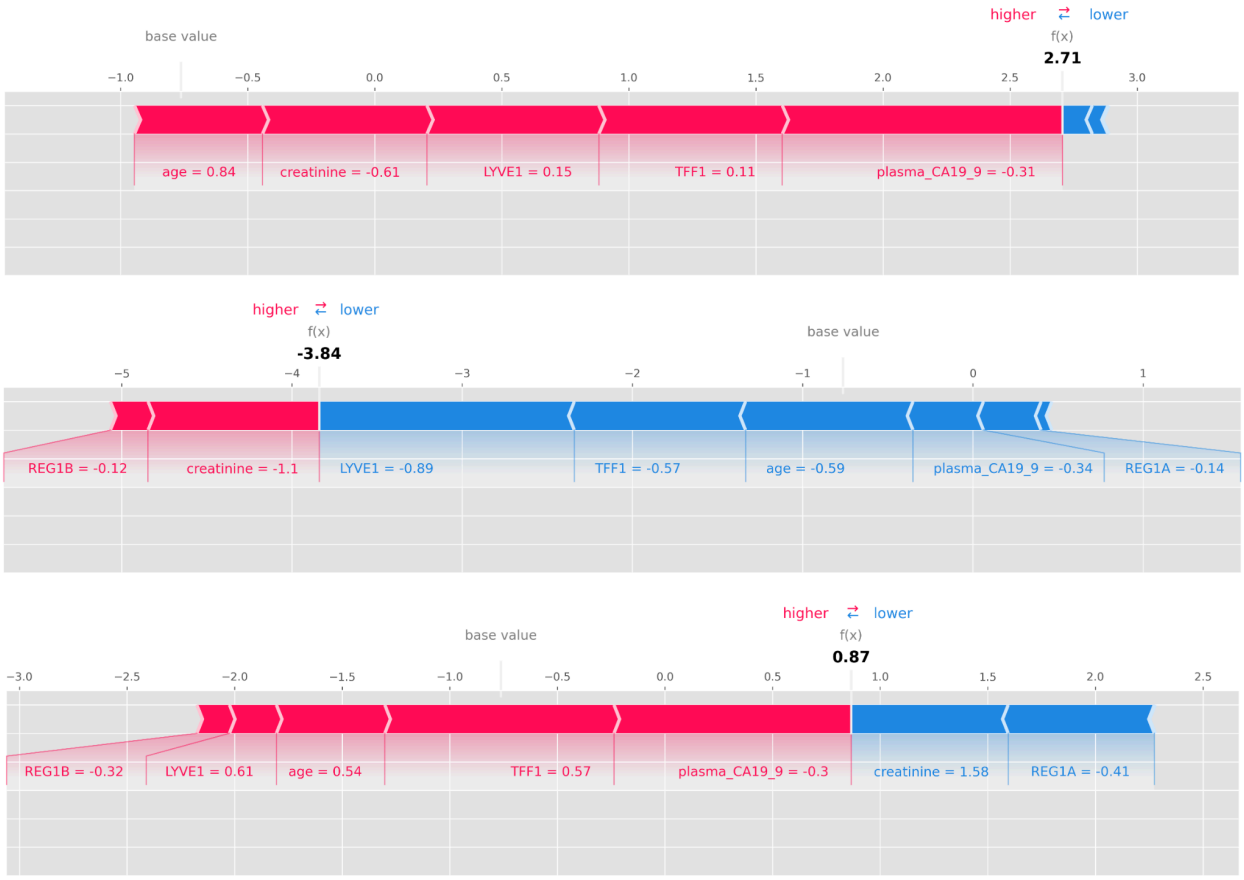


Figure 11: SHAP Force plots for data indices 1, 51, 101.

Outlook

In the future, I would like to explore improving the model by using a stratified split. A simple k-fold split may not be sufficient, especially with the remapping of diagnosis to a binary format, which could impact the balance of the target variable. Additionally, I am interested in developing a multiclass classification model as it may better reflect the complexity of the biomedical problem. Differentiating between cancer and non-cancer with a binary approach may oversimplify the challenge. Lastly, I would focus on enhancing model performance by incorporating additional features or biomarkers. Collecting new biomarkers that correlate strongly with existing ones, such as LYVE1 and TFF1, could improve the model's ability to identify positive or negative cancer diagnoses.

References

Arumugam, Thiruvengadam, et al. "Trefoil Factor 1 Stimulates Both Pancreatic Cancer and Stellate Cells and Increases Metastasis." *Pancreas*, U.S. National Library of Medicine, Aug. 2011, pmc.ncbi.nlm.nih.gov/articles/PMC4319540/.

Debernardi S;O'Brien H;Algahmdi AS;Malats N;Stewart GD;Plješa-Ercegovac M;Costello E;Greenhalf W;Saad A;Roberts R;Ney A;Pereira SP;Kocher HM;Duffy S;Blyuss O;Crnogorac-Jurcevic T; "A Combination of Urinary Biomarker Panel and PANCRISK Score for Earlier Detection of Pancreatic Cancer: A Case-Control Study." *PLoS Medicine*, U.S. National Library of Medicine, pubmed.ncbi.nlm.nih.gov/33301466/. Accessed 14 Dec. 2024.

Halligan, Steve, et al. "Disadvantages of Using the Area under the Receiver Operating Characteristic Curve to Assess Imaging Tests: A Discussion and Proposal for an Alternative Approach." *European Radiology*, U.S. National Library of Medicine, Apr. 2015, pmc.ncbi.nlm.nih.gov/articles/PMC4356897/#:~:text=ROC%20AUC%20ignores%20clinical%20differentials,slope%20equals%20one%20%5B5%5D.

Karinen, Sini, et al. "Tumour Cells Express Functional Lymphatic Endothelium-Specific Hyaluronan Receptor in Vitro and in Vivo: Lymphatic Mimicry Promotes Oral Oncogenesis?" *Nature News*, Nature Publishing Group, 5 Mar. 2021, www.nature.com/articles/s41389-021-00312-3.

Post, Erin. "Pancreatic Cancer Five-Year Survival Rate Increases to 13%." *Pancreatic Cancer Action Network*, 16 Apr. 2024, pancan.org/news/pancreatic-cancer-five-year-survival-rate-increases-to-13/#:~:text=An%20estimated%2066%2C440%20Americans%20will,of%20all%20the%20major%20cancers.