Detecting Pancreatic Ductal Adenocarcinoma (PDAC) Across Datasets

Rianna Campbell and Max Shushkovsky
02-518 Computational Medicine

Introduction

Our project is focused on the use of urinary biomarkers to detect pancreatic ductal adenocarcinoma (PDAC), the most common type of pancreatic cancer. In the paper, "A Combination of Urinary Biomarker Panel and PANCRISK Score for Earlier Detection of Pancreatic Cancer: A Case–Control Study," researchers aimed to develop an accurate diagnostic test for PDAC. They developed their own model to create a PancRISK score from the data; this led us to conclude that computational means from software could be highly effective, given that the urine panels resulted in various concentration levels for samples (Debernardi et al.). There are existing notebooks on Kaggle that use a common method—light gradient-boosting machine—for high accuracy results. However, we plan to conduct an exploration of various other classifiers to determine which models could be more accurate for this specific detection of pancreatic cancer. After conducting extensive research on this topic, we have concluded that there is limited research and quantitative evidence on using urinary biomarkers for PDAC detection, although urinary data is more easily accessible than RNA-seq data. As a result, we want to further evaluate the possibility of using models trained on urinary data to determine if they can primarily detect PDAC, but also possibly be applied to RNA-seq data.

Data

The urinary dataset has 590 total samples and uses protein concentration data from four key biomarkers associated with PDAC: creatinine, LYVE1, REG1B, and TFF1. In addition, a concatenation of various NCBI GEO DataSets of RNA-seq data across PDAC tumors and healthy pancreatic cells (from tissue) was used, including 250 total samples. Along with numerous extraneous expression data, these contain expression levels for LYVE1, TFF1, the gene GPX1 (which shows biochemical association with creatinine) and either REG1B or REG1 (has a similar primary structure to REG1B); there was no direct creatinine data, so GPX1 was used. In each of the urinary and NCBI datasets, we split the data into separate training sets and testing sets in an 80/20 ratio. Both the urinary and NCBI datasets use quantitative expression levels of biomarkers dependent on pancreatic condition, which can be easily passed into machine learning models to learn from.

Additionally, the urinary dataset diagnosis label was modified to become a simpler binary classification problem instead of having three labels; two labels indicated no pancreatic cancer, so they were joined into one label. In Figure 1 below, the 0 label corresponds to a non-cancerous label, and 1 corresponds to a cancer label.

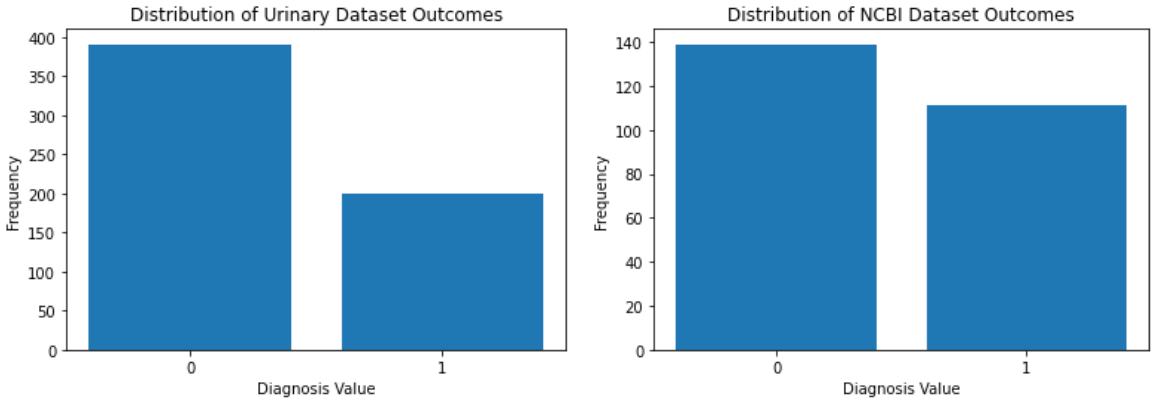


Figure 1: *Distribution of dataset outcomes. Left: Distribution of urinary dataset outcomes. There were 391 non-cancerous (0) labels and 199 cancerous (1) labels. Right: Distribution of NCBI dataset outcomes. There were 139 non-cancerous (0) labels and 111 cancerous (1) labels.*

Methods

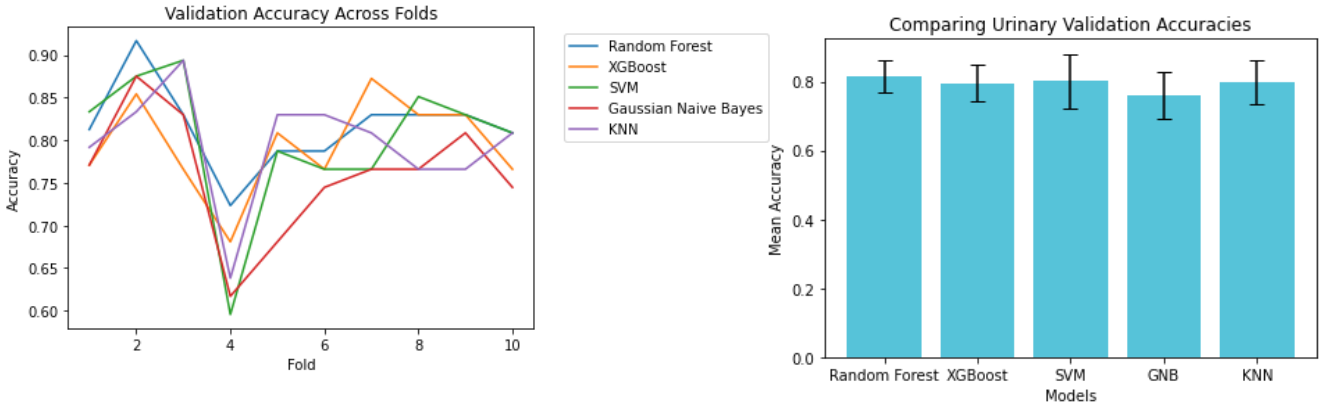
We used the following five Python supervised learning classifiers from the Sci-Kit Learn and XGBoost packages for their popularity in many common use cases: Random Forest (RF), XGBoost (XGB), Support Vector Machine (SVM), Gaussian Naive Bayes (GNB), K Nearest Neighbors (KNN) Classifiers. Feature selection was not really needed because the significant biomarkers were already presented. However, it was important to impute any missing data, as some expression levels were missing. To do so, the median expression value was imputed across the samples for a specific biomarker. Then, 10-fold cross validation training was performed to make sure training was robust and not poorly sampled.

Validating the results included examining accuracy values, precision scores, recall scores, F-1 scores, and conducting a McNemar test across the best and worst models in each dataset. Accuracy told us specifically which samples the classifiers predicted correctly on. Precision is the ratio between true positives and all the positives. Recall is the measure of identifying true positives; for all samples with X outcome, recall is how many samples correctly identified having X outcome. Having an F-1 score for each model utilized both precision and recall to determine how many times the model actually made a correct prediction instead of false positives and false negatives. Lastly, the McNemar test was used to determine if there was statistical significance in how models make errors. This is helpful to relate to the F-1 scores for each model.

Results

The results from the 10-fold cross validation on the 5 classification models trained on the urinary dataset are shown in Figures 2 and 3. Overall, the 5 models performed similarly, scoring around 80% accuracy for validation. After testing these urinary models on the urinary test dataset, we calculated metrics including test accuracy, precision, recall, and F-1 score, which are listed in Figure 4. The KNN and SVM models had the highest test accuracies at 80.5%, although all of the models scored between 77% and 80.5%. For the precision, recall, and F-1 scores, all

models performed similarly with the exception of the XGB and GNB models, which have lower recall and F-1 scores.

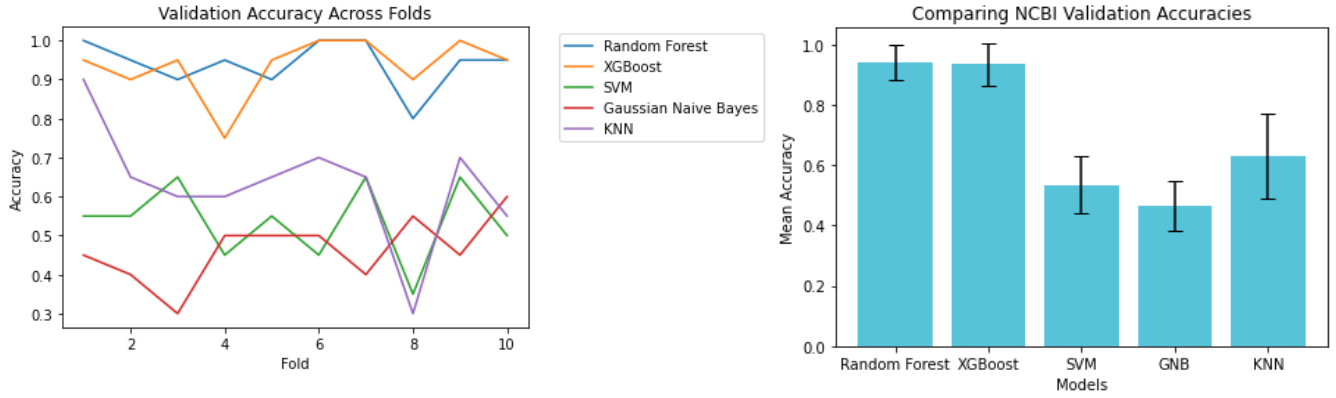


Figures 2 (left) and 3 (right). *Figure 2: A line plot of the validation accuracies for each of the 10 folds for the models trained on the urinary dataset (RF, XGB, SVM, GNB, KNN). Figure 3: A bar plot depicting the mean validation accuracy and standard deviation across all folds for each of the 5 trained and validated models.*

Model	*Random Forest*	*XGBoost*	*Support Vector Machine*	*Gaussian Naive Bayes*	*K-Nearest Neighbors*
Test Accuracy	0.797	0.771	0.805	0.771	0.805
Precision	0.7	0.686	0.743	0.76	0.707
Recall	0.7	0.6	0.65	0.475	0.725
F-1 score	0.7	0.667	0.693	0.585	0.716

Figure 4: *A table of metrics calculated for each of the 5 urinary models after testing them on the urinary test dataset. The metrics reported are test accuracy, precision, recall, and F-1 score.*

Next, we performed 10-fold cross validation on 5 separate models (with the same hyperparameters as the urinary models) on the NCBI training dataset. Figures 5 and 6 display the validation results, including the mean validation accuracies across all 10 folds. The RF and XGB models performed extremely well with validation accuracies greater than 90%, but the SVM, GNB, and KNN models performed rather poorly, with validation accuracies between 40% and 60%. In Figure 7 we report the metrics for the NCBI models tested on the NCBI test dataset, which confirm that RF and XGB are the highest performing models on the NCBI dataset due to their high precision, recall, and F-1 scores.



Figures 5 (left) and 6 (right). *Figure 5: A line plot of the validation accuracies over 10 folds for the 5 models trained on the NCBI dataset. Figure 6: A bar plot of the mean validation accuracies across the 10 folds and their standard deviations for all 5 NCBI models.*

Model	*Random Forest*	*XGBoost*	*Support Vector Machine*	*Gaussian Naive Bayes*	*K-Nearest Neighbors*
Test Accuracy	0.98	0.94	0.46	0.5	0.6
Precision	0.958	0.952	0.167	0.438	0.579
Recall	1.0	0.87	0.043	0.304	0.478
F-1 score	0.979	0.933	0.069	0.359	0.524

Figure 7: *A table of metrics calculated for each of the 5 NCBI models after testing them on the NCBI test dataset. The metrics reported are test accuracy, precision, recall, and F-1 score.*

In addition, we performed McNemar tests (Figure 8) to understand how the best and worst models for each dataset (urinary and NCBI) differed in the types of incorrect predictions (false positives and false negatives). The best and worst models were determined using a combination of the test accuracy, precision, recall, and F-1 scores. The McNemar score between the best (RF) and worst (SVM) models for the NCBI dataset was statistically significant, meaning that the RF and SVM NCBI models made errors in different ways. This could explain why there was a large division in the test accuracies between the 5 NCBI models.

	Best model	*Worst model*	*McNemar Score*
Urinary Dataset	SVM	XGBoost	0.453
NCBI Dataset	RF	SVM	9.443e-07

Figure 8: *A table of the McNemar scores for the best and worst models for each of the 2 datasets. A McNemar score of less than 0.05 is considered statistically significant.*

To determine if the urinary models and NCBI models could accurately predict PDAC outcomes from an alternative form of data (urine samples or RNA-seq data), we first tested the urinary models on the NCBI test dataset and then tested the NCBI models on the urinary test dataset. The metrics calculated for each of these tests are in Figures 9 and 10, respectively. The results for the urinary models tested on the NCBI test dataset (Figure 9) show that the test accuracies for all 5 models were fairly low, with the highest being a 54% accuracy for the KNN model. The precision, recall, and F-1 scores were low as well. The results for the NCBI models tested on the urinary test dataset (Figure 10) display a wide range of test accuracies, with the highest as 73.7% for SVM and the lowest as 35.6% for RF. The SVM model, however, had a high precision score (0.909) but a low recall score (0.25), resulting in a low F-1 score, indicating that this model may not be the best overall despite its high test accuracy.

Model	*Random Forest*	*XGBoost*	*Support Vector Machine*	*Gaussian Naive Bayes*	*K-Nearest Neighbors*
Test Accuracy	0.46	0.48	0.42	0.46	0.54
Precision	0.389	0.333	0.25	0.357	0.5
Recall	0.304	0.174	0.13	0.217	0.435
F-1 score	0.341	0.381	0.171	0.27	0.465

Figure 9: *A table of metrics (test accuracy, precision, recall, F-1 score) calculated for each of the 5 urinary models after testing them on the NCBI test dataset.*

Model	*Random Forest*	*XGBoost*	*Support Vector Machine*	*Gaussian Naive Bayes*	*K-Nearest Neighbors*
Test Accuracy	0.356	0.381	0.737	0.695	0.559
Precision	0.336	0.34	0.909	0.553	0.409
Recall	0.925	0.9	0.25	0.525	0.675
F-1 score	0.493	0.51	0.392	0.538	0.509

Figure 10: *A table of metrics (test accuracy, precision, recall, F-1 score) calculated for each of the 5 NCBI models after testing them on the urinary test dataset.*

Discussion

After evaluating our investigation, there were some potential limitations in our procedure. In the NCBI dataset, REG1B and LYVE1 were sparse across many samples; most of the expression values were 0. Since the value was reported as 0, we did not try to impute any data.

Having fewer valuable features would make it harder for the models to learn from. To mitigate this, we attempted to drop the REG1B feature entirely and rerun our procedure, but there was marginal change in accuracies. Additionally, we ended up using the GPX1 gene expression for the NCBI dataset because the protein expression of creatinine was not provided. This, along with the fact that the NCBI data reported a different form of data (gene expression) versus the urinary data (protein concentration), very likely contributed to the models not being able to accurately predict PDAC outcomes on the alternative data sources.

Ultimately, the urinary models performed quite well on predicting PDAC from urinary data. With test results close to 80%, this could become a feasible alternative to predicting PDAC from RNA-seq data. On the other hand, the NCBI models were a poor predictor of PDAC, likely owing to the method of data collection and potentially the types of models used. A future extension for this would be to apply deep learning methods with neural networks to try and improve the classification accuracy. After comparing the results of the models' prediction on the alternative data sources, results showed accuracies that were almost worse than random guessing. Thus, we cannot easily apply these models to other data sources.

References

Cameron, DP, et al. “Geo Accession Viewer.” *National Center for Biotechnology Information*, U.S. National Library of Medicine, www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE211468.

De Monte, L, et al. “Geo Accession Viewer.” *National Center for Biotechnology Information*, U.S. National Library of Medicine, www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE242916.

Debernardi, Silvana, et al. “A Combination of Urinary Biomarker Panel and PANCRISK Score for Earlier Detection of Pancreatic Cancer: A Case–Control Study.” *PLOS Medicine*, Public Library of Science, journals.plos.org/plosmedicine/article?id=10.1371%2Fjournal.pmed.1003489. Accessed 01 Dec. 2023.

Fei, L, et al. “Geo Accession Viewer.” *National Center for Biotechnology Information*, U.S. National Library of Medicine, www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE216854.

Owens, N, et al. “Geo Accession Viewer.” *National Center for Biotechnology Information*, U.S. National Library of Medicine, www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE205163.

Wittenberg, MM, et al. “Geo Accession Viewer.” *National Center for Biotechnology Information*, U.S. National Library of Medicine, www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE232860.

Xue, D, et al. “Geo Accession Viewer.” *National Center for Biotechnology Information*, U.S. National Library of Medicine, www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE228662.