Contents:

Introduction:

This work explores the development of a diagnostic classification model for the purposes of early diagnosis of pancreatic cancer. In order to not overwhelm the length of the Jupyter notebook, Clinical Evidence, Explanations of Model Development and Explainable Analysis can be found in this readme. This file can be read in tandem with the associated Jupyter notebook.

1.Clinical Evidence

Pancreatic cancer stands at the 7th highest cause of cancer related deaths.(Rawla et al., 2019) While it is less prevalent than other cancers, its high mortality rate, which is as low as 2% in some countries, is alarming.(McGuigan et al., 2018) This is largely due late diagnosis of the cancer. (Rawla et al., 2019) In 85% of cases, the cancerous tumor are no longer surgically removable at the point of detection.(eBioMedicine, 2022) Majority of pancreatic cancer cases are pancreatic ductal adenocarcinomas, which will henceforth be used interchangeably with pancreatic cancer in this work, for simplicity. (McGuigan et al., 2018) At early stages, the disease is largely asymptomatic, or displays less concerning symptoms such as dorsal pain or nausea, contributing to its low rate of early detection.(Partyka et al., 2023; Rawla et al., 2019) Early detection of pancreatic cancer is tricky, many viable techniques are invasive and expensive to conduct for screening purposes.(Partyka et al., 2023) Thus, developing a viable, easy to conduct and inexpensive mode of detection is crucial.

Researchers have investigated various biofluid biomarkers to determine relationships between biomarker expressions and presence of pancreatic cancer. A blood biomarkers, CA19-9 is the only known biomarker for pancreatic cancer, but are not sensitive or specific enough to detect pancreatic cancer at earlier stages.(eBioMedicine, 2022; Tatjana Crnogorac-Jurcevic, n.d.) However, findings from proteomic studies using mass spectrometry have distinguished three key biomarkers: LYVE1, REG1B and TFF1 as highly probable discriminators between healthy patients, patients with benign conditions and patients with pancreatic cancer.(Tatjana Crnogorac-Jurcevic, n.d.)(Radon et al., 2015) Given the symptomatic similarity between non-cancerous ailments and pancreatic cancer, distinguishing between benign conditions with similar symptoms and pancreatic cancer is also important. As such a key area for the development of an early detection tool are urinary biomarkers. (eBioMedicine, 2022)

The dataset chosen for this work is obtained from a study by Debernandi and colleagues, “A combination of urinary biomarker panel and PancRISK score for earlier detection of pancreatic cancer: A case–control study”.(Debernardi et al., 2020) It contains 590 retrospectively obtained urine samples from 3 groups: Healthy Controls, Hepatobiliary diseases and Patients with Pancreatic Cancer. The samples were obtained in 2 cohorts, from multiple centers. The dataset was varied, to include other common malignancies in order to represent the variations in urinary biomarkers.(Debernardi et al., 2020) Below is a table explaining the data available in the dataset:

|  |  |
| --- | --- |
| Sample\_ID | Sample ID |
| Patient Cohort | Cohort the patient sample came from. There are two cohorts in which the samples were obtained. This data should not have links with the diagnosis and will be dropped further down. |
| Sample Origin | The center in which the sample was obtained from. There are 4 different centers, whereby the samples were obtained unequally. This will be dropped. |
| Age | The sample age range follows a normal distribution, between 20-90 years. |
| Sex | The gender of the patients from which the sample was obtained. The distribution is fairly equal with female samples being slightly higher than male. |
| Diagnosis | Diagnosis is split into 3:   1. Healthy Controls 2. Patients with hepatobiliary diseases 3. Patients with Pancreatic Cancer   This was changed into 0,1,2 to align with XGBoost classifiers. |
| Stage | Stage of cancer, this was removed for analysis as it would have caused data leakage since all pancreatic cancer samples would have a stage. |
| Benign Sample Diagnosis | The specific conditions in which patients with hepatobiliary diseases have been diagnosed with. This was removed as it did not have relevance to pancreatic cancer. |
| Plasma CA19\_9 (U/ml) | A blood biomarker that can be indicative of pancreatic cancer. This was removed as this was a blood plasma sample, only available in half the data points. |
| Creatinine (mg/ml) | An animo acid byproduct that is commonly used to measure kidney function.(Asif et al., n.d.) Because urine concentration can vary, when measuring urine biomarkers, creatinine is often used for normalization of biomarker values.(Tang et al., 2015) |
| Lymphatic Vessel Endothelial HA Receptor – LYVE1(ng/ml) | A cell surface receptor that is heavily expressed on lymphatic endothelial cells. Some evidence has shown decreased expression in the pancreas as compared to normal pancreatic tissues.(Samir et al., 2024) |
| Regenerating Islet-derived 1 Beta – REG1B (ng/ml) | A protein highly similar to REG1A, with its expressions highly associated. (Radon et al., 2015; van Beelen Granlund et al., 2013) This study also uses REG1B in place of REG1A due to their similar expression. |
| Trefoil Factor Family 1 – TFF1(ng/ml) | A secretory protein that maintains the integrity of gastrointestinal mucus membranes. 10 Differed expression is heavily associated with multiple cancers. (Samir et al., 2024) |
| Regenerating Islet-derived 1 Alpha – REG1A (ng/ml) | A protein expressed in the pancreas that regulates insulin secretion and regulates pancreas function. (Samir et al., 2024) Increased expression is often associated with pancreatic cancer or bowel inflammation.(Radon et al., 2015; van Beelen Granlund et al., 2013) Its expression levels is often associated with REG1B and other proteins in the REG1 family. (van Beelen Granlund et al., 2013) |

This dataset was downloaded from Kaggle. However, the raw data is available directly from the authors, through the published journal. Both files contain the same values, apart from slight changes in the headings, such as the availability of measurement units. For example: ‘Sample ID’ in the original as opposed to ‘sample\_id’ from Kaggle, and ‘REG1A (ng/ml)’ as opposed to ‘REG1A’. As it does not change any content in the cells, the dataset uploaded within the .zip file is the dataset downloaded from Kaggle.

Data Preprocessing:

To prepare the data to be input into the model, several steps were taken.

Model Development

When developing a classifier model that will distinguish between healthy controls, benign controls, and pancreatic cancer cases, the most important metric to consider is the recall when predicting cases of pancreatic cancer.(Hicks et al., 2022) In the clinical scenario, it is far more dangerous to produce a false negative than a false positive.(Hicks et al., 2022) If a pancreatic cancer patient is wrongfully classified as healthy, the potential likelihood of detecting the cancer early goes down, further minimizing survival chances.(Bartlett et al., 2021)

Distinguishing between non-pancreatic cancer related cases and pancreatic cancer would be one of the main aims of this classifier. Due to the symptomatic similarities between pancreatic cancer and other hepatobiliary diseases, it is important to be able to determine the difference between the two classes of patients. Minimizing misdiagnosis and delay of appropriate treatment in both cases will be crucial to improve patient outcomes as well as reduce healthcare burden.

In order to develop a classification model that can achieve these aims, several methods were considered:

Baseline Model: Logistic regression

Log regression is supervised learning algorithm that works as a discriminative classifier in both binary and multi-class classifications.(Bisong, 2019) It determines the probability of a sample being in a particular class (diagnosis) by mapping and applying it to the log function.(Kumar, 2023)

Advanced Model: Random Forest

Random Forest is a supervised learning algorithm that can be used in classification and regression models.(Shafi, 2024). It combines the outputs of multiple decision trees to generate a more accurate output.(Te Beest et al., 2017) Random Forest is generally effective in high dimensionality datasets, as the many trees can break down the complex tabular structure of a dataset.(Te Beest et al., 2017)

Advanced Model: XGBoost

XGBoost is a s scalable machine-learning algorithm meant for supervised learning models. Similar to Random Forest, it is an ensemble of decision trees.(Tarwidi et al., 2023) For this dataset, XGBoost may have an advantage in its ability to focus on certain outcomes, such as distinguishing between the cancer class and other hepatobiliary disease class.(Chen & Guestrin, 2016)

Model Comparison

To compare the models, classification reports and confusion matrices were generated for each of the models developed, with the most focus placed on Recall in class 2 (Pancreatic cancer prediction). Both random forest and XGBoost had outperformed log regression in terms of accuracy, by 2-3%. However, logistic regression outperformed both advanced models in class 2 recall by 10%. Based on this metric, the best model selected was the baseline model using logistic regression.

This differs from the initial assumption that XGBoost would yield the best results, given that it is the more complex and powerful algorithm. However, there are some possible reasons why log regression outperformed the more advanced models. Firstly, the comparatively small dataset could work better with more simplistic models like logistic regression. Meanwhile, more advanced models may have overfit the models, given the limited features. Additionally, the SHAP analysis showed a strong linear relationship between the variables in class 2, which could explain why logistic regression was able to perform better with making predictions for class 2.

Explainable AI

Feature Importance, LIME and SHAP analysis was conducted to understand how the model worked.

SHAP analysis indicated that all values with the exception of creatinine and sex had a strong linear relationship with pushing up the SHAP values, showcasing their importance and relevance in determining a predictor.

LIME shows an approximation of how a datapoint was likely estimated. The output in the Jupyter notebook displays the thresholds which determine the classes in which the prediction is likely to go into. In this case, the TFF1 and age value was particularly influential in the decision-making process. Meanwhile creatinine and LYVE1 were less influential.

However, the Feature Importance graph showed creatinine has the largest influence on the predictive process, while age is the least important. This makes clinical decision making based on these features tougher, as it highlights the importance of a combination of different biomarkers to make an accurate assessment.

Clinical Decision Making

While the model had a high recall, the contradictions within the explainable analysis suggest that the model needs further tuning. It is likely that a larger dataset is required to generate a more stable set of relationships to develop a solid classification model. Furthermore, a combined interpretation of multiple biomarkers is necessary to predicting pancreatic cancer.

Future work can focus on refining the model using a larger database. Additionally, developing a model that can also succeed in differentiating between stage of cancer by using slight differences between biomarker variations can make a large clinical impact on patient outcomes.

Asif, A. A., Hussain, H., & Chatterjee, T. (n.d.). Extraordinary Creatinine Level: A Case Report. *Cureus*, *12*(7), e9076. https://doi.org/10.7759/cureus.9076

Bartlett, E. C., Silva, M., Callister, M. E., & Devaraj, A. (2021). False-Negative Results in Lung Cancer Screening—Evidence and Controversies. *Journal of Thoracic Oncology*, *16*(6), 912–921. https://doi.org/10.1016/j.jtho.2021.01.1607

Bisong, E. (2019). Logistic Regression. In E. Bisong (Ed.), *Building Machine Learning and Deep Learning Models on Google Cloud Platform: A Comprehensive Guide for Beginners* (pp. 243–250). Apress. https://doi.org/10.1007/978-1-4842-4470-8\_20

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. https://doi.org/10.1145/2939672.2939785

Debernardi, S., O’Brien, H., Algahmdi, A. S., Malats, N., Stewart, G. D., Plješa-Ercegovac, M., Costello, E., Greenhalf, W., Saad, A., Roberts, R., Ney, A., Pereira, S. P., Kocher, H. M., Duffy, S., Blyuss, O., & Crnogorac-Jurcevic, T. (2020). A combination of urinary biomarker panel and PancRISK score for earlier detection of pancreatic cancer: A case-control study. *PLoS Medicine*, *17*(12), e1003489. https://doi.org/10.1371/journal.pmed.1003489

eBioMedicine. (2022). Emerging biomarkers for early diagnosis of pancreatic cancer. *eBioMedicine*, *79*. https://doi.org/10.1016/j.ebiom.2022.104064

Hicks, S. A., Strümke, I., Thambawita, V., Hammou, M., Riegler, M. A., Halvorsen, P., & Parasa, S. (2022). On evaluation metrics for medical applications of artificial intelligence. *Scientific Reports*, *12*, 5979. https://doi.org/10.1038/s41598-022-09954-8

Kumar, S. (2023, November 14). *4. Assumptions and Limitations of Logistic Regression: Navigating the Nuances*. Medium. https://medium.com/@skme20417/4-assumptions-and-limitations-of-logistic-regression-navigating-the-nuances-8ef249cc7a01

McGuigan, A., Kelly, P., Turkington, R. C., Jones, C., Coleman, H. G., & McCain, R. S. (2018). Pancreatic cancer: A review of clinical diagnosis, epidemiology, treatment and outcomes. *World Journal of Gastroenterology*, *24*(43), 4846–4861. https://doi.org/10.3748/wjg.v24.i43.4846

Partyka, O., Pajewska, M., Kwaśniewska, D., Czerw, A., Deptała, A., Budzik, M., Cipora, E., Gąska, I., Gazdowicz, L., Mielnik, A., Sygit, K., Sygit, M., Krzych-Fałta, E., Schneider-Matyka, D., Grochans, S., Cybulska, A. M., Drobnik, J., Bandurska, E., Ciećko, W., … Kozłowski, R. (2023). Overview of Pancreatic Cancer Epidemiology in Europe and Recommendations for Screening in High-Risk Populations. *Cancers*, *15*(14), 3634. https://doi.org/10.3390/cancers15143634

Radon, T. P., Massat, N. J., Jones, R., Alrawashdeh, W., Dumartin, L., Ennis, D., Duffy, S. W., Kocher, H. M., Pereira, S. P., Guarner posthumous, L., Murta-Nascimento, C., Real, F. X., Malats, N., Neoptolemos, J., Costello, E., Greenhalf, W., Lemoine, N. R., & Crnogorac-Jurcevic, T. (2015). Identification of a Three-Biomarker Panel in Urine for Early Detection of Pancreatic Adenocarcinoma. *Clinical Cancer Research: An Official Journal of the American Association for Cancer Research*, *21*(15), 3512–3521. https://doi.org/10.1158/1078-0432.CCR-14-2467

Rawla, P., Sunkara, T., & Gaduputi, V. (2019). Epidemiology of Pancreatic Cancer: Global Trends, Etiology and Risk Factors. *World Journal of Oncology*, *10*(1), 10–27. https://doi.org/10.14740/wjon1166

Samir, S., El-Ashry, M., Soliman, W., & Hassan, M. (2024). Urinary biomarkers analysis as a diagnostic tool for early detection of pancreatic adenocarcinoma: Molecular quantification approach. *Computational Biology and Chemistry*, *112*, 108171. https://doi.org/10.1016/j.compbiolchem.2024.108171

Shafi, A. (2024, October 1). *Random Forest Classifier Tutorial: How to Build a Random Forest in Python*. DataCamp. https://www.datacamp.com/tutorial/random-forests-classifier-python

Tang, K. W. A., Toh, Q. C., & Teo, B. W. (2015). Normalisation of urinary biomarkers to creatinine for clinical practice and research – when and why. *Singapore Medical Journal*, *56*(1), 7–10. https://doi.org/10.11622/smedj.2015003

Tarwidi, D., Pudjaprasetya, S. R., Adytia, D., & Apri, M. (2023). An optimized XGBoost-based machine learning method for predicting wave run-up on a sloping beach. *MethodsX*, *10*, 102119. https://doi.org/10.1016/j.mex.2023.102119

Tatjana Crnogorac-Jurcevic. (n.d.). *New tests for early detection of pancreatic cancer offer significant hope—Queen Mary University of London*. Queen Mary University of London. Retrieved 30 April 2025, from https://www.qmul.ac.uk/research/featured-research/new-tests-for-early-detection-of-pancreatic-cancer-offer-significant-hope/

Te Beest, D. E., Mes, S. W., Wilting, S. M., Brakenhoff, R. H., & Van De Wiel, M. A. (2017). Improved high-dimensional prediction with Random Forests by the use of co-data. *BMC Bioinformatics*, *18*(1), 584. https://doi.org/10.1186/s12859-017-1993-1

van Beelen Granlund, A., Østvik, A. E., Brenna, Ø., Torp, S. H., Gustafsson, B. I., & Sandvik, A. K. (2013). REG gene expression in inflamed and healthy colon mucosa explored by in situ hybridisation. *Cell and Tissue Research*, *352*(3), 639–646. https://doi.org/10.1007/s00441-013-1592-z