PancreaSwift: Early-stage Pancreatic Cancer Detection through Image Classification

Nivrith Ananth Iyer, Engineering and Science University Magnet School

Abstract

Pancreatic cancer, though accounting for only 3% of all cancers in the US, is highly lethal. Early detection is critical for effective treatment, with up to 44% recovery when identified before tumor maturation. Current diagnostic tools, mainly blood tests, often lack sensitivity for early detection with accuracies of up to 87%, thus highlighting the need for a new approach to diagnostic tests. I have previously explored the alternate method of utilizing urinary biomarkers to detect early-stage pancreatic cancer, and with a 590 sample dataset, I achieved a 90% accuracy with gradient boosting models. However, plot holes and inconsistencies with the separability between the cancerous and non-cancerous groups of my previous dataset question urinary biomarkers as a reliable solution. My project aims to explore another systematic way of detecting pancreatic cancer via deep image learning. By obtaining x-ray scans of cancerous and non-cancerous pancreas samples, I want to develop and compare common image classification models in terms of general performance(accuracy, precision, F1 scores, and recall) for detecting early-stage pancreatic cancer to see if image classification is more efficient than either blood tests or urinary biomarkers. I plan to learn about and hyper-tune parameters specific to different image classification models to achieve higher results and comparisons to other novel detection mechanisms. With my PancreaSwift initiative, my project aims to compare AI image analysis to urinary biomarkers and blood test accuracies in hopes of furthering the advancement of early-stage detection to increase rates of treatability before tumor maturation.

PROJECT MOTIVATION AND APPROACH

Problem Background and Current Initiatives

Pancreatic cancer, though constituting only 3% of all cancers in the US, is exceptionally lethal [1]. Treatment is viable when the tumor is identified in its early stages [2], with up to 44% of patients recovering completely [3]. Unfortunately, symptoms often appear only after metastasis, leading to a survival span of 8-12 months [4]. Presently, blood tests, hindered by undetectable gradient levels, serve as the diagnostic tool, limiting early detection of pancreatic cancer [5].

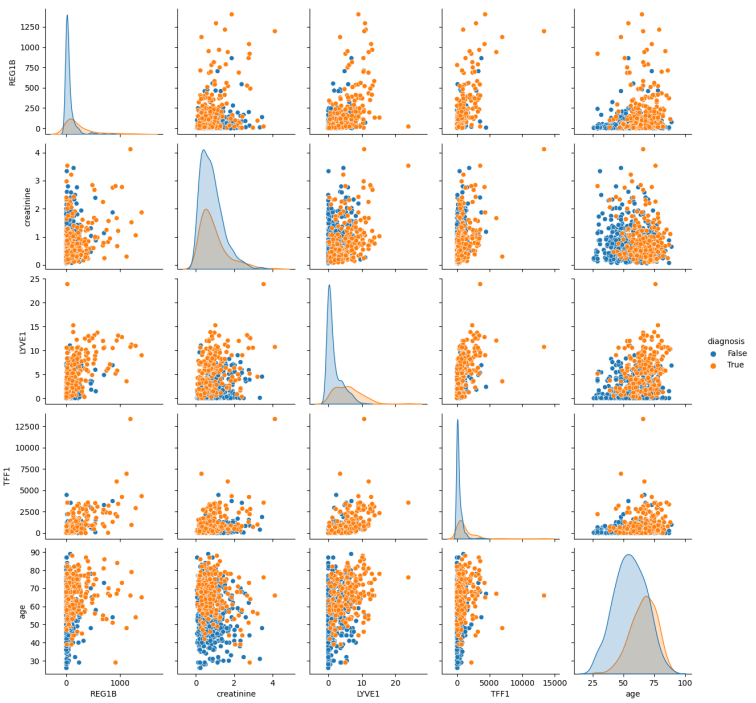
Current blood tests for pancreatic cancer, with an average accuracy of 87%, often lack conclusiveness, necessitating additional testing like imaging and biopsies [6]. The delayed results, taking two to three weeks, hinder early intervention, allowing tumors to mature beyond treatable stages [7]. A more precise and timely diagnostic approach is imperative for enhancing accuracy and enabling early detection of pancreatic cancer.

In the pursuit of novel diagnostic avenues, urinary biomarkers have emerged as a promising frontier due to cost and time effectiveness. However, their potential is hindered by a critical limitation - the lack of separability within training data samples in popular datasets [8]. The cancerous and non-cancerous data points often overlap in the datasets encompassing urinary biomarkers, sharing similar values. This lack of distinctiveness poses a formidable challenge, making it arduous to differentiate between cancer and non-cancer instances accurately. The difficulty in achieving clear separability within the training data undermines the reliability of urinary biomarkers as a robust diagnostic tool for pancreatic cancer.

A Previous Iteration, My Proposed Method, and Its Approach

After realizing all of the inconsistencies and errors that my initial research on urinary biomarkers had, I want to explore a new detection mechanism route. Instead of obtaining numerical levels of data that could go through human error thus causing significant fluctuations that affect accuracy, using an image-based dataset will serve as a more reliable basis for a machine to learn from. Using image-based data will not contain any errors in the training data since it doesn't rely on human precision but rather relies on machine precision, which is much more consistent. Because image data can not be misinterpreted to a slightly different value when used as training data, it naturally serves as a more reliable dataset. However, the training algorithms required for image processing are much different compared to numerical data processing.

When examining my urinary biomarker dataset, I first visualized the dataset in a data visualization file and then a model training file. In the data visualization file, I split up the data into two categories: "no_cancer", which is marked as false, and "cancer", which is marked as true. After that, I created pair plots, heat maps, and bar graphs to compare the abundance of urinary biomarkers per group along with any external patterns. It was in this file that I realized my data had minimal separability since cancer and non-cancer data points lined up alongside each other. Knowing this, I made a model training file where I again split up the data into two categories and initially recorded the amount of training data in "X_train." I found out that I had 472 training samples and 68 testing samples,



thus making my model training not as reliable due to the lack of testing that my models could go through.

To address the lack of separability, I coded a random generator model to generate sample sizes for the non-cancerous and cancerous groups based on the trends in the dataset. Differentiating the key differences in the two groups would allow my trained models to analyze such differences to a higher degree, thus making their produced accuracies more realistic to real-world trends. I used various models: KNN, CNN, logistic regression, random forest classifiers, and gradient boosting models. In the end, LightGBM and XGBoost ended up producing the highest accuracies of 91%, with LightGBM having a slightly higher ROC curve than XGBoost.

In this test, gradient-boosting models excelled due to their robustness to outliers, feature importance, and regularization techniques. They outperformed other models in understanding the relative importance of features in a categorical group. In contrast, CNN models performed poorly, reaching a maximum accuracy of 40%. However, these gradient-boosting models will not perform nearly as well on image data compared to convolutional networks because of many factors: spatial relationships, parameterization, and a lack of convolutional operations.

Convolutional neural networks are specifically designed to capture spatial patterns between pixels in images while gradient-boosting models lack such spatial awareness. Also, CNNs use convolutional layers which help them set specific parameters to analyze image data properly while gradient boosting models can't create convolutional networks, thus deeming them useless when trying to hyperparameter tune them because of a lack of layered parameters.

For my PancreaSwift initiative, I want to try using a layered CNN-LSTM model. This would add to the work of a research paper exploring the usage of a 1D CNN-LSTM model for

urinary biomarker detection: "Automated classification of urine biomarkers to diagnose pancreatic cancer using 1-D convolutional neural networks." The study showcases three gates: the forget gate, the input gate, and the output gate. The forget gate either passes or ignores dataflow from a dataset, as defined by the following equation:

$$F_t = \sigma(W_F \times [X_t,\ h_{t-1}] + b_F)$$

Variable Names: Output of Forget Gate = Sigmoid Activation Function multiplied by the product of the weight matrix and the current timestamp input and hidden timestamp state when added to the bias coefficient.

The forget gate also deals with presenting the updating and current timestamps of cell states, with C_t and C_{t-1} respectively.

$$C_{t-1} \times F_t = \begin{cases} 0, & F_t = 0 \\ C_{t-1}, & F_t = 1 \end{cases}$$

The input gate selects certain information to be updated using a sigmoid function, I_t, and then compresses the given input pattern using a hyperbolic tangent(tanh) function, C_t, to add an immediate state of long-term impact.

$$I_t = \sigma\left(W_I \times [x_t, h_{t-1}]\ b_I \right)$$

$$\tilde{C}_t = \tanh(W_c \times [X_t, h_{t-1}] + b_c)$$

The output gate figures out the long-term effect and updates the outputs of the current cell state, C_t, and the hidden state(h_t) using sigmoid and tanh functions via the following equations:

$$O_t = \sigma(W_O \times [X_t,\ h_{t-1}] + b_O)$$

$$h_t = O_t \times \tanh(C_t)$$

LSTM models, designed for sequential dependencies, aren't traditionally used for image classification tasks and are typically used for sequential tasks [9]. However, when combined with a layered CNN, benefits in temporal dependencies and sequential feature learning can enhance model accuracy. The combination allows the model to understand changes over time and capture both spatial and temporal features, resulting in a detailed network for distinguishing cancerous and non-cancerous samples.

This project is feasible with the layered CNN-LSTM architecture and is versatile for various image classification purposes [10]. Despite challenges in accessing datasets with pancreas images due to paywalls, available datasets online demonstrate the feasibility of researching the capabilities of the layered CNN-LSTM model on pancreas samples.

For this initiative, I only need a laptop since my proposal is entirely computational. I already have one and can thus set up a Google Colaboratory file for examining specific datasets. However, I require access to a dataset with thousands of pancreas X-ray images for proper model training. I seek mentorship from the THINK team to gain insights from mentors and professors and potentially access cancerous and non-cancerous datasets from MIT's computational biology department. Learning about different image classification models, beyond my proposed

CNN-LSTM, through the THINK team's resources would enhance my research's impact on model comparisons.

PROJECT LOGISTICS AND ORGANIZATION

With the PancreaSwift Initiative, I want to achieve three main milestones that will have a great impact on my research once completed:

1. **Create, evaluate, and successfully hyperparameter tune a layered CNN-LSTM model.** With the success that the 1D CNN-LSTM model brought researchers in the urinary biomarker dataset, I want to create a layered CNN-LSTM model that can successfully detect patterns in images and associations leading to cancer. Expanding upon the CNN-LSTM architecture can extend research into the model's capabilities, and can be further perfected to detect early-stage cancer in future studies. In addition, I want to tune different parameters in a layered CNN-LSTM model so I can dictate which parameters have more significant impacts on the model's accuracy. F1 scores, precision, and recall will also be calculated to further assess the model in producing a consistent output.
2. **Compare different image classifiers based on performance and ROC curves**. I want to test and compare different image classifiers based on different model architectures so the best kinds of architectures for such an image-based dataset can be found. Furthermore, I want to examine universal parameters across models that have the most prevalent effects on boosting accuracies. I will calculate and plot TPF against FPF across varying cut-offs in the data to generate an ROC curve to assess overall diagnostic performance.
3. **Compare the accuracies of image classifiers in general to urinary biomarker classifiers.** I want to figure out whether image data serves as a more reliable and efficient

source for a machine to learn common patterns associated with pancreatic cancer when compared to the previous urinary biomarker dataset I used. Once I figure out which dataset architecture results in a higher accuracy produced by an AI model, I want to explore that specific dataset then and test more models and parameter-tuning to approach maximum potential accuracy which will contribute to its overall performance.

The following three risks could occur to my initiative, but they can also be mitigated so that there is a minimal effect on the results of my research:

1. An unreliable dataset can feed my AI models false data or poorly compromised data. If the data is of poor quality, unrepresentative, or contains biases, the model's performance, and generalization may be compromised. By working with the THINK mentorship, I will avoid this risk by being granted a reliable and sizable dataset from MIT on which I can train my AI models.

2. The interpretability of my models' outputs can be difficult to understand. When creating a data visualization file for my previous project, I found it difficult to understand the heat maps and box plots since it seemed as if the variables depicted did not correlate with each other at all. When creating my data visualization file for my image-based dataset, my AI models can potentially create uninterpretable graphs that can prevent me from properly understanding the patterns depicted in images. However, I can avoid this by researching commonly used measurements for image-based data and only coding specific algorithms that depict understandable types of relationships through sensible graphs.

3. My AI models can undergo overfitting. Overfitting poses a risk of models learning training data intricacies too well, leading to poor real-world performance. To counter this, I will

employ proper hyperparameter tuning, ensuring model generalization, and use cross-validation to prevent learning specific patterns not applicable in real-world scenarios.

Timeline:

1. 2/15/2024: I will find a suitable dataset for machine model training with clear X-ray images of the pancreas that is sourced from a reliable database that contains both cancerous and non-cancerous samples.
2. 2/27/2024: I will start a data visualization file using Matplotlib and Seaborn to analyze image aspects such as pixel distributions. I will convert image data to a format compatible with these libraries.
3. 3/15/2024: I will begin a model training file testing the proposed layered CNN-LSTM model with parameters related to the forget gate, input gate, and output gate. I will compare its performance with other common image classification models.
4. 4/07/2024: I will start hyperparameter tuning for image classification models and CNN-LSTM in a second training iteration. I will also test various parameters to maximize performance using the same testing data.
5. 4/20/2024: I will initiate communication with local hospitals and cancer research organizations for new X-ray samples. In addition, I will verify the performance of the best-performing models on new datasets to prevent overfitting and training errors.
6. 5/15/2024: I will maximize models under new datasets to enhance the reliability of deep-learning algorithms. I will communicate findings with researchers and explore potential publication opportunities.

Current Progress and Project Budget

I recently concluded a research project titled "Evaluating Hyperparameter Tuned Machine Learning Classifiers for Pancreatic Cancer Detection via Urinary Biomarkers" [11]. This project focused on early-stage pancreatic cancer detection using urinary biomarkers in a numerical dataset, and I am actively pursuing publication. While this project did not involve image data, my current initiative aims to explore the performance of image classifiers compared to the model classifiers used in the previous research. The objective is to determine which detection method achieves higher accuracy and reliability. I have completed half of this goal, with ongoing efforts focused on developing a layered CNN-LSTM model, improving commonly used image classifiers, and acquiring a dataset with an appropriate number of samples.

Funding will go to purchasing an SSD that can fit in my computer's Thunderbolt port so that higher dataset sizes can run smoothly on my computer without any ping or a lack of storage.This will let my computer process images smoothly for my models to rely on.

Item	Amount	Cost	Link
Portable SSD X5 Thunderbolt™3 2TB	1	$599.99	https://www.samsung.com/us/computing/memory-storage/portable-solid-state-drives/portable-ssd-x5-2tb-mu-pb2t0b-am/

PERSONAL INTEREST

From an early age, I found myself drawn toward the world of technology, playing Roblox games on my iPad constantly and finding a curiosity toward the complexities of computers. My fascination led me into the computer science field through computer game development on Roblox using Lua. It was during my Freshman year of High School, particularly in AP Biology, that my fascination with computational biology took root. The idea of merging computer science and biology became increasingly appealing. Furthermore, I was inspired by my dad's friend, a data scientist, through his intriguing lectures on machine learning algorithms and their

practicality in games and detection systems whenever we met at religious gatherings. Motivated by a personal connection—my best friend's dad got pancreatic cancer—I delved into investigating pancreatic cancer detection algorithms. Frustrated by the limited research on pancreatic cancer accuracy compared to other cancers, I enrolled in a machine learning course with InspiritAI. Working with a mentor, I initiated a pneumonia detection project, which evolved into my independent venture—a pancreatic cancer detection project utilizing urinary biomarkers. Drawing from various research papers and mentorship, I now aspire to explore alternative detection methods, further advancing my contributions to pancreatic cancer detection algorithms. Currently, I understand how to train a general model to analyze image data, but I need to learn how to train image classifiers with hyperparameter tuning and I also need to learn how to train a reliable AI to detect even the slightest changes in features in a specific image for classifying purposes.

Referenced Resources

[1] *Key statistics for pancreatic cancer*. (n.d.). American Cancer Society. https://www.cancer.org/cancer/types/pancreatic-cancer/about/key-statistics.html#:~:text=Pancreatic%20cancer%20accounts%20for%20about.

[2] Tan, D. J., Mitra, M., Chiu, A., & Coller, H. A. (2020). *Intron retention is a robust marker of intertumoral heterogeneity in pancreatic ductal adenocarcinoma*. Npj Genomic Medicine, 5(1). https://doi.org/10.1038/s41525-020-00159-4

[3] Pancreatic Cancer - *statistics*. (2023, December 14). Cancer.Net. https://www.cancer.net/cancer-types/pancreatic-cancer/statistics#:~:text=If%20the%20cancer%20is%20detected,relative%20survival%20rate%20is%2015%25.

[4] UpToDate. (2023, November). S*upportive care for locally advanced or metastatic exocrine pancreatic cancer*. Retrieved December 30, 2023, from https://www.uptodate.com/contents/supportive-care-for-locally-advanced-or-metastatic-exocrine-pancreatic-cancer

[5] Hoffman, M., MD. (2009, March 6). *Pancreatic cancer diagnosis and early detection*. WebMD. https://www.webmd.com/cancer/pancreatic-cancer/pancreatic-cancer-diagnosis#:~:text=Certain%20substances%2C%20such%20as%20carcinoembryonic,is%20advanced%2C%20if%20at%20all.

[6] Post, E. (2023, November 1). *The latest developments in early detection for pancreatic cancer*. Pancreatic Cancer Action Network. https://pancan.org/news/the-latest-developments-in-early-detection-for-pancreatic-cancer/#:~:text=It's%20important%20to%20note%20that,and%20a%20biopsy%20are%20necessary.

[7] Pancreatic Cancer UK. (2021, March 16). *Tests for pancreatic cancer* - Pancreatic Cancer UK. https://www.pancreaticcancer.org.uk/information/how-is-pancreatic-cancer-diagnosed/tests-for-pancreatic-cancer/

[8] *Urinary biomarkers for pancreatic cancer*. (2020, December 10). Kaggle. https://www.kaggle.com/datasets/johnjdavisiv/urinary-biomarkers-for-pancreatic-cancer

[9] Vankdothu, R., Hameed, M. A., & Fatima, H. (2022). A *Brain Tumor Identification and Classification Using Deep Learning based on CNN-LSTM Method*. Computers & Electrical Engineering, 101, 107960. https://doi.org/10.1016/j.compeleceng.2022.1079601

[10] Farooque, G., Xiao, L., Yang, J., & Sargano, A. B. (2021). Hyperspectral image classification via a novel Spectral–Spatial 3D COnVLSTM-CNN. *Remote Sensing*, 13(21), 4348. https://doi.org/10.3390/rs13214348

[11] *Evaluating hyperparameter tuned machine learning classifiers for pancreatic cancer detection via urinary biomarkers - Nivrith Ananth Iyer*. (n.d.). Google Docs. https://docs.google.com/document/d/1kBj2HHNPkj3A6xyp4865uC1EEVgb57A723whh_jpeco/edit?usp=sharing

External Resources

- *New tests for early detection of pancreatic cancer offer significant hope - Queen Mary University of London*. (n.d.). https://www.qmul.ac.uk/research/featured-research/new-tests-for-early-detection-of-pancreatic-cancer-offer-significant-hope/#:~:text=Currently%20pancreatic%20cancer%20is%20identified,to%20perform%20in%20any%20laboratory.
- Debernardi, S., O'Brien, H., Algahmdi, A. S., Malats, N., Stewart, G. D., Plješa-Ercegovac, M., Costello, E., Greenhalf, W., Saad, A. S., Roberts, R., Ney, A., Pereira, S. P., Kocher, H. M., Duffy, S. W., Blyuss, O., & Crnogorac‐Jurčević, 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
- *New tests for early detection of pancreatic cancer offer significant hope - Queen Mary University of London*. (n.d.-b). https://www.qmul.ac.uk/research/featured-research/new-tests-for-early-detection-of-pancr

eatic-cancer-offer-significant-hope/#:~:text=The%20biomarkers%20appeared%20to%20be,potential%20for%20future%20clinical%20application.

- *Biomarkers in the diagnosis of pancreatic cancer: Are we closer to finding the golden ticket?* (2021, July 14). PubMed Central. Retrieved December 30, 2023, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8311531/
- GeeksforGeeks. (2023, May 23). *LightGBM Light Gradient Boosting Machine*. https://www.geeksforgeeks.org/lightgbm-light-gradient-boosting-machine/#.
- *What is XGBoost?* (n.d.). NVIDIA Data Science Glossary. https://www.nvidia.com/en-us/glossary/data-science/xgboost.
- Raj, A. (2022, August 6). Everything about support vector classification — above and beyond. *Medium*. https://towardsdatascience.com/everything-about-svm-classification-above-and-beyond-cc665bfd993e#:~:text=SVM%20does%20not%20perform%20very,samples%2C%20the%20SVM%20will%20underperform.
- Team, C. (2023, November 22). *Random Forest*. Corporate Finance Institute. https://corporatefinanceinstitute.com/resources/data-science/random-forest/#:~:text=Among%20all%20the%20available%20classification,other%20classes%20in%20the%20data.
- *NivrithA478 - Overview*. (n.d.). GitHub. https://github.com/NivrithA478