11-785 Project Proposal for Pancreatic Cancer Detection

Joshmin Ray
Department of Machine Learning
Carnegie Mellon University
Pittsburgh, PA 15213
joshminr@andrew.cmu.edu

Rohan Chawla
Department of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213
rchawla2@andrew.cmu.edu

Jesse Shen
Department of Biomedical Engineering
Carnegie Mellon University
Pittsburgh, PA 15213
jwshen@andrew.cmu.edu

Lohan Nye, MD
Department of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213
lnye@andrew.cmu.edu

1 Introduction

Pancreatic cancer is an aggressive form of cancer that is typically detected in late stage with a low long-term survival rate, ranging from 2-9% at 5 years Davide Placido [2023]. It presents as a leading cause of death among cancers in the United States and is projected to continue impacting lives in the coming decades Rahib et al. [2014], with a projected increase in incidence rate by 2030. Major research institutions such as the U.S. National Institute of Health (NIH) attribute much of pancreatic cancer's poor prognosis from late detection that leaves limited and ineffective options for treatment Ramírez-Maldonado et al. [2024]. Some of the reasons for this late stage detection include the inability to adequately visualize the pancreas given its retro-peritoneal location, its lack of specific molecular markers, and its indistinct early clinical symptoms Tripathi et al. [2024]. One approach to early detection is artificial intelligence.

According to Patel et al. [2024], there has been an exponential increase in the number of published research regarding AI and the pancreas. One of the approaches we found in the literature is the application of deep learning algorithms to predict the risk of pancreatic cancer. This will lead to possible early detection.

Our research study aims to build upon an existing deep learning algorithm for pancreatic cancer risk prediction reported in *Nature Medicine* in March 2023. We model our methodology as outlined in the baseline study as a foundation. Our goal is to enhance the predictive accuracy and clinical applicability of this AI model for pancreatic cancer prediction by leveraging the rich, multimodal patient data available in the MIMIC-IV and MIMIC-IV-Note datasets. These enhancements are intended to address potential limitations of the baseline study and incorporate alternative AI modeling techniques across diverse data types to improve early detection and personalized risk assessment of pancreatic cancer.

2 Literature Review

The field of AI-related early cancer detection uses multiple means to calculate a patient's risk. These methods take in input data from one of the following modalities: patient records, medical images, or biomarkers. With pancreatic cancer (PC), all three means have been studied as promising diagnostic indicators Tripathi et al. [2024]. In this section, we will cover these three lines of research as well as some research with the specific patient record dataset, Medical Information Mart for Intensive Care (MIMIC) - IV, we intend to use.

Biomarker-based classification offers precise detection only possible at the molecular level. The diagnostic field with most relevance to AI is liquid biopsy research, which seeks to identify markers for cancer via a sample of non-solid biological tissue– usually blood. Known biomarkers can be proteins, circulating DNA/RNA sequences, and circulating tumor cells Ramírez-Maldonado et al. [2024]. Current machine learning (ML) models applying biomarker detection exclusively analyze genetic or protein profiles of patients Patel et al. [2024]. At the same time, there remains active debate regarding biomarkers themselves. A given biomarker's predictive value needs to be generalized and effectively biochemically isolating and detecting molecules of interest continue to be active areas of research Ramírez-Maldonado et al. [2024].

Applying deep learning models to medical imaging data shows promise in both early diagnosis and characterizing subtypes of cancers for treatment. PC is a solid tissue tumor that is identified when patients report symptoms via a computed tomography (CT) scan. Other imaging modalities used with AI in pancreatic cancer diagnosis include magnetic resonance imaging (MRI) and EUS (endoscopic ultrasound). The field of radiomics specifically seeks to identify the minute details of medical images identified by machine learning models to aid medical professionals with models further able to classify cancers for optimal treatment. With an emphasis on early diagnosis for PC, AI algorithms have demonstrated high sensitivity, but pancreatic tumors with a spherical diameter of less than 2 cm prove difficult to detect Alexandra Corina Faur [2023]. While imaging remains recommended for high-risk individuals by medical professionals, this preferred method is currently impractical for the general population Klein [2021]

A "top-down" approach comes from analyzing patient data. The earlier the diagnosis the better the prognosis, which is the impetus of deep learning models analyzing large databases of patient records Muhammad et al. [2019]. Tracing the cause of a given cancer has also proven useful in characterizing it and providing targeted treatment Hu et al. [2019]. Diagnostic tests are often focused on high-risk individuals, leaving the potential for large patient record databases to detect PC early in the general population. However, the particular risk factors of interest are not clear for early PC diagnosis. Deep learning models are being researched as a solution, in conjunction with large patient databases used to study patient trajectory Davide Placido [2023].

One of the most common applications of the MIMIC-IV's patient data so far has been determining diseases diagnoses from blood-glucose data. One paper identified type 2 diabetes cases and another used the variability in blood-glucose to detect coronary artery disease Zhang W [2023], He et al. [2024]. These examples may be additional variables to consider for our project, besides the other patient data used, since diabetes emergence is a risk factor for PC Klein [2021].

3 Baseline Models

There have been several different approaches to using machine learning in the field of cancer detection, ranging from support vector machines (SVMs) and random forests to deep neural networks, natural language processing (NLP), and transformers Patel et al. [2024]. There have also been both text and imaging approaches.

The paper we chose as a baseline compares sequential neural networks, gated recurrent unit (GRU) models, and the Transformer model. These were compared against their baseline bag-of-words model, which used disease codes as the words and ignores time and order of disease events Davide Placido [2023].

They used the area under the receiving operator characteristic (AUROC) as their performance metric and obtained their best performance through using the transformer (AUROC = 0.879 (0.877-0.880)) Davide Placido [2023].

The baseline study demonstrated the feasibility of using deep learning algorithms to predict pancreatic cancer risk from disease trajectories. However, opportunities exist to enhance model performance and applicability, particularly by integrating a broader range of data types and employing more advanced AI techniques. The MIMIC-IV dataset, with its extensive collection of clinical data, including electronic health records, lab results, and clinical notes, provides an ideal resource for developing a more comprehensive and accurate pancreatic cancer prediction model.

4 Dataset Description

We plan to analyze data from the Medical Information Mart for Intensive Care (MIMIC) - IV dataset Johnson et al., 2023]. This dataset contains digitized patient care data from the Beth Israel Deaconess Medical Center (BIDMC) intensive care unit (ICU) between the years of 2008-2019. This dataset is split into three separate modules: *hosp*, *icu* and *note*. *icu* contains patient data from the ICU. This includes features such as time-series monitoring of IV treatments being administered, patient responses, any procedures and treatments documented, as well as notes made in the chart by providers. The majority of this data is organized at events. The *hosp* module contains hospital-wide electronic health record (EHR) data. This includes things like patient demographics, patient transfers, lab test results, microbiology measurements, prescribed medications and billing events. This provides longer-scale data that describes the state of patients beyond their stay in the ICU. Finally, the *note* module contains free-text unstructured data about patients: both from their discharge summaries and from radiology reports.

A concerted effort has been made to de-identify the patients in this dataset and remove any features that are too sensitive. There are around 50,000 total unique ICU patients and 180,000 unique hospital patients. The average age for these patients is around 60. An important aspect of this dataset is that it contains contemporary data from 2019 and it has been structured in a modular manner. This allows for ICU data to be linked with hospital-wide administrative records, which we expect will be useful in our model. However, this data is pretty sparse and there are some nonsensical data entries due to procedural error, so we expect some data cleaning and pre-processing will be necessary before performing any in-depth analysis.

References

Jessica X. Hjaltelin Chunlei Zheng Amalie D. Haue Piotr J. Chmura Chen Yuan Jihye Kim Renato Umeton Gregory Antell Alexander Chowdhury Alexandra Franz Lauren Brais Elizabeth Andrews Debora S. Marks Aviv Regev Siamack Ayandeh Mary T. Brophy Nhan V. Do Peter Kraft Brian M. Wolpin Michael H. Rosenthal Nathanael R. Fillmore Søren Brunak Chris Sander Davide Placido, Bo Yuan. A deep learning algorithm to predict risk of pancreatic cancer from disease trajectories. *Nature Medicine*, 2023.

Lola Rahib, Benjamin D. Smith, Rhonda Aizenberg, Allison B. Rosenzweig, Julie M. Fleshman, and Lynn M. Matrisian. Projecting cancer incidence and deaths to 2030: The unexpected burden of thyroid, liver, and pancreas cancers in the united states. *Cancer Research*, 05 2014. doi: https://doi.org/10.1158/0008-5472.CAN-14-0155.

Elena Ramírez-Maldonado, Sandra López Gordo, Rui Pedro Major Branco, Mihai-Calin Pavel, Laia Estalella, Erik Llàcer-Millán, María Alejandra Guerrero, Estrella López-Gordo, Robert Memba, and Rosa Jorba. Clinical application of liquid biopsy in pancreatic cancer: A narrative review. *International Journal of Molecular Sciences*, 2024. doi: https://doi.org/10.3390/ijms25031640.

Satvik Tripathi, Azadeh Tabari, Arian Mansur, Harika Dabbara, Christopher P. Bridge, and Dania Daye. From machine learning to patient outcomes: A comprehensive review of ai in pancreatic cancer. *Diagnostics*, 14(2), 2024. ISSN 2075-4418. doi: 10.3390/diagnostics14020174. URL `https://www.mdpi.com/2075-4418/14/2/174`.

Hardik Patel, Theodoros Zanos, and D. Brock Hewitt. Deep learning applications in pancreatic cancer. *Cancers*, 16(2), 2024. ISSN 2072-6694. doi: 10.3390/cancers16020436. URL `https://www.mdpi.com/2072-6694/16/2/436`.

Laura Andreea Ghenciu Alexandra Corina Faur, Daniela Cornelia Lazar. Artificial intelligence as a noninvasive tool for pancreatic cancer prediction and diagnosis. *World Journal of Gastroenterology*, 2023. doi: DOI: 10.3748/wjg.v29.i12.1811.

Alison P. Klein. Pancreatic cancer epidemiology: understanding the role of lifestyle and inherited risk factors. *Nature Reviews Gastroenterology Hepatology*, 2021. doi: https://doi.org/10.1038/s41575-021-00457-x.

Wazir Muhammad, Gregory R. Hart, Bradley Nartowt, James J. Farrell, Kimberly Johung, Ying Liang, and Jun Deng. Pancreatic cancer prediction through an artificial neural network. *Frontiers in Artificial Intelligence*, 2, 2019. ISSN 2624-8212. doi: 10.3389/frai.2019.00002. URL `https://www.frontiersin.org/articles/10.3389/frai.2019.00002`.

Jessica X. Hu, Marie Helleberg, Anders B. Jensen, Søren Brunak, and Jens Lundgren. A Large-Cohort, Longitudinal Study Determines Precancer Disease Routes across Different Cancer Types. *Cancer Research*, 79(4):864–872, 02 2019. ISSN 0008-5472. doi: 10.1158/0008-5472.CAN-18-1677. URL `https://doi.org/10.1158/0008-5472.CAN-18-1677`.

Cao T. Zhang W. Automated type 2 diabetes case and control identification from the mimic-iv database. *AMIA Jt Summits Transl Sci Proc.*, 2023.

Hao-Ming He, Shu-Wen Zheng, and et al. Simultaneous assessment of stress hyperglycemia ratio and glycemic variability to predict mortality in patients with coronary artery disease: a retrospective cohort study from the mimic-iv database. *Cardiovascular diabetology*, 2024. doi: https://doi.org/10.1186/s12933-024-02146-w.

Alistair Johnson, Lucas Bulgarelli, Tom Pollard, Steven Horng, Leo Anthony Celi, and Roger Mark. MIMIC-IV.

Alistair E. W. Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J. Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, Li-wei H. Lehman, Leo A. Celi, and Roger G. Mark. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific Data*, 10(1):1, January 2023. ISSN 2052-4463. doi: 10.1038/s41597-022-01899-x.