Lumen Prize Application 2023

Name: Archie Tan
Major: Computer Science
Mentor: Dr. Spurlock

Project Title

Improving Early Diagnosis of Pancreatic Cancer with Synthetic Data

Abstract

Pancreatic cancer is a highly lethal disease, primarily because it is difficult to diagnose in its early stages using traditional methods. Using imaging and biopsy techniques is significantly challenging due to the pancreas being deep inside the human and its asymptotic characteristics. As a result, machine learning became necessary for early diagnosis to identify patterns and features associated with pancreatic cancer from medical images. However, accurate diagnosis using machine-learning models is challenging due to the inability of the current medical datasets, which are often biased and sensitive, making them difficult to share and use for research purposes. Therefore, this research focuses on improving the accuracy of pancreatic cancer diagnosis in the early stages by using artificial intelligence to generate synthetic data that closely replicate the real pancreatic cancer medical image to reduce the bias of the data and enhance the model's performance in early diagnosis.

Personal Statement

My journey in computer science started with unlimited problem-solving in life and expanded to the world.

Growing up, my family didn't have access to a computer since it was always cost prohibitive, and subsequently, we have never learned to use them. However, my passion for technology was unstoppably born from the tons of science fiction movies, especially Iron man. I adored how Tony Stark used technology to change the world and save lives as a human, and I've always dreamed of developing those machines and AI assistants. Fortunately, I received an old computer from my parent's friend when I was ten; I was so excited about this first computer in my life. I spent hours on my living room floor fiddling with the computer, trying to figure out how to connect the wires. The computer was very laggy, but I was still pleased with the opportunity to explore its hardware and software.

However, It wasn't until the day my grandfather passed away in the hospital. It was the first time I felt the kind of pain that was engraved into my bones and dug into my heart. I realized my beloved had gone forever, and I understand I'm not the only one that went through it. There are many people who also couldn't survive due to the limited technologies in healthcare. Therefore, I

firmly decided to contribute to technology and healthcare, positively impacting the world and saving lives. I began to learn Python independently and build a command line-based program to help my mom store the data from her preschool. I was surprised by the potential of the technology as it enables me to use imagination and logic in the programming language to create whatever I want, and I'm eager to enhance my skills.

Undoubtedly, I majored in computer science with my passion and dream in mind. However, my interest in computer science was always limited by the prerequisite of the classes. I was so eager to learn and dreamed that my skills could eventually develop a piece of technology that could positively impact the world. Therefore, I took outside courses like software engineering, Android development, and machine learning. I also participated in coding contests. Those experiences opened a new perspective for me on how to view the world without human abilities, such as how computers work, see, learn, or understand. I began questioning the logic behind everything in our lives, such as how we can tell what a disease is. It will involve a lot of analysis of the micro patterns of the organs, cells, or tissues. Therefore, I wish to learn more and enhance my intellectual development, to solve meaningful, real-world problems.

Therefore, I'm applying for the Lumen Prize as it can allow me to achieve my ambitious goal and extend my intellectual journey. The Lumen Prize can supply me with the necessary machine to further develop my computer vision, machine learning, and synthetic data expertise. It also allows me to take advanced computer science classes beyond those offered at Elon. With the Lumen Prize, I can travel to state-of-the-art technology and healthcare conferences. I will be able to explore the most recent cutting-edge technology and learn about the challenges in the world. I will also be able to make connections and communicate with other engaging experts in the field to enhance my knowledge, perspective, and professional development. This experience would help me become a better researcher and computer scientist that continually challenges the world's difficulties.

Project Description

Focus:

Pancreatic cancer is a devastating and aggressive cancer with a high mortality rate. The American Cancer Society estimates about 50,550 people will die of pancreatic cancer in the United States in 2023 [1]. Pancreatic cancer is often detected at advanced stages, by which time it can't be removed by surgery and doesn't respond well to chemotherapy and radiation [8]. The challenges of early diagnosis are its asymptomatic nature and the location of tumors [4]. Traditional imaging techniques prove insufficient in detecting pancreatic tumors, especially those smaller than 2 cm; abdominal CT scans can miss up to 40% of these tumors [2].

However, recent advances in artificial intelligence, specifically in deep learning, have enabled relatively accurate detection of pancreatic cancer on CT scans, even for tumors smaller than 2 cm. This development presents new hope for the early detection of pancreatic cancer [2]. As a computer science student living in a rapidly evolving world, I was fascinated by the potential of cutting-edge technologies like deep learning to transform healthcare, and I wish to use my skills to make a positive impact on the world. Pancreatic cancer has affected the lives of so many

people, including those close to me, and I believe there is a pressing need for better diagnostic tools to catch the disease in its early stages. By proposing this project, I hope to contribute to the fight against pancreatic cancer and make a difference in the lives of patients and their families.

This project expands upon the recent advancements in pancreatic cancer diagnosis using deep learning [2]. More specifically, our goal aims to enhance the accuracy of the current state-of-the-art pancreatic cancer diagnosis model by using synthetic data to overcome the challenges of the inability of the medical image datasets and emphasize its early-stage images.

Medical image diagnosis has been an active area of research for many years. However, recent studies have highlighted the challenge of class imbalance and the lack of suitable datasets in medical imaging research [5, 9]. For instance, studies that focused on the COVID-19 dataset found that the variability in data sources affects the performance of machine learning models [10, 12]. In addition, two well-known publicly available X-ray image datasets also found gender imbalance conditions [6]. Given such context, the pancreatic cancer image dataset might also contain bias.

This research will investigate the available medical image datasets of pancreatic cancer to determine the potential bias, such as the imbalance between the advanced and early stages. The class imbalance will cause the poor performance of the machine learning model in the underrepresented groups in the dataset [5]. To overcome this challenge, we will generate synthetic medical images to balance various groups. We will also generate images with the early stage feature of pancreatic cancer, such as the size of the tumors, to improve the accuracy of early diagnosis [4].

Therefore, this research will use Generative Adversarial Networks (GAN) to closely mimic real-world data to synthesize data for machine learning models. GAN uses a generator to create data and a discriminator to identify synthetic data; through this process, the generator will eventually synthesize data that is realistic enough that the discriminator can't identify. A recent study on synthetic abnormal MRI images with brain tumors using GAN shows significantly higher sensitivity and specificity in the machine-learning model when diagnosing brain tumors [3]. Therefore, the goal of this research aims to use GAN and answer the following questions:

1. Do the current pancreatic cancer medical image datasets contain bias?
2. Can we develop a model that can generate realistic pancreatic cancer images?
3. Does augmenting existing, biased datasets with synthetic images generated by our model reduce bias?
4. Do synthetically augmented datasets lead to improved performance of early-stage pancreatic cancer diagnosis systems?

The early diagnosis of pancreatic cancer remains a significant challenge. Ongoing research and advanced techniques are necessary to overcome this challenge [5]. Therefore, This project provides an innovative approach to advance the early diagnosis of pancreatic cancer. It has the potential to create impacts on the current medical and technological field through the application of synthetic data. It will first provide an insightful analysis of the current pancreatic cancer imaging dataset, which would help future researchers to determine problems in the pancreatic

cancer medical dataset. Through this research, we can also provide invaluable performance analysis of synthetic pancreatic cancer images and the potential challenges of using GAN to synthesize data. It will be an important reference for future research. Overall, this research has the potential to help the healthcare professional to diagnose pancreatic cancer in the early stage, thereby saving lives and enhancing patient outcomes, and providing a significant new approach to overcome the challenge with synthetic data.

Scholarly Process:

This research will involve leveraging frameworks from computer science and data science to guide the development. Expertise in Python and Jupyter Notebook can be especially valuable for data processing and analysis. Essential skills include machine learning, data visualization, synthetic data generation, and computer vision. We will follow a rigorous data science process that iterates over several key steps: data collection, cleaning, data analysis, training model, running experiments, optimizing, and analyzing the performances.

To initiate the research process, it is crucial to acquire relevant knowledge by thoroughly studying existing research papers. Then, it is imperative to gather and organize the relevant pancreatic cancer image dataset. I will utilize Python and Jupyter Notebook to clean and analyze the available datasets. Subsequently, I will apply computer vision and machine learning techniques to identify patterns in medical images and record the model's initial performance in pancreatic cancer diagnosis without balancing the data. Afterward, I will analyze the datasets to identify any potential bias by using the aggregate functions for each category, then categorize the images based on the cancer stage. Then I will use GAN to synthesize medical images of pancreatic cancer to enhance the diversity level of the dataset. The GAN training process will seek to produce realistic synthetic images similar to the actual images from the dataset. Then, we will use the generated data to train and test the machine learning model and record its initial performance. Finally, we will combine the synthesized and real data to train the model and test the model's performance on real data to estimate the accuracy of the diagnosis in the real world. Through continuously experimenting and optimizing the model and synthetic data, I will monitor the performance of the machine learning model and produce the most accurate diagnosis system.

Developing the model and synthesizing data will be challenging and requires knowledge in Python, Jupyter Notebook, data science, and machine learning. My coursework in database systems, data visualization, data mining, machine learning, and real-world research project experience in transportation system planning in NYC using publicly available and NYC transportation department data provides me with proficiency in those skills. Additionally, I have completed and am progressing with relevant coursework, including NYU AI school, Machine Learning and Data Science course from ZTM, and Data Mining and Machine Learning course at Elon University. Those experiences and knowledge provided me with a strong foundation for this research. To develop expertise in computer vision and GAN, I will take the relevant course at Elon university and certification programs outside of school since I have the required math and coding foundation. By leveraging my skills and experiences and continuously expanding my knowledge, I am confident in my ability to contribute meaningfully to pancreatic cancer diagnosis and develop a model that will make a real difference in the fight against this disease.

Through this research, I will gain invaluable knowledge and expertise in various disciplines, such as computer science, AI, data science, and healthcare. This research will be an excellent opportunity for me to explore cutting-edge technologies in the medical field. I can participate in state-of-the-art conferences on AI and pancreatic cancer to explore the most advanced research, technologies, and applications in the industries. It will also allow me to gain a deeper understanding of the integration of technology and healthcare, which I have been eager to explore. I can advance my skill in cutting-edge AI technologies throughout my research. I can publish my findings and advance my communication and professional presentation skills in my domain. Additionally, I will gain invaluable experience in manipulating and analyzing large real-world image data, which can enhance my critical thinking skills and develop the ability to determine and address complex problems effectively. Ultimately, the research will significantly enhance my intellectual, academic, and professional growth and allow me to contribute to society in a meaningful way and make a positive impact on people's lives.

Proposed Product:

1. A research paper submitted for publication
2. Presentation at relevant conferences.
3. A model that can accurately diagnose pancreatic cancer from medical images
4. Code that published in an open-source repository for future research access
5. Powerful computer that can run experiments with massive image datasets

Feasibility

Despite the potential challenges, logistical issues, and requirements for advanced technical skills that might exist in the research, we are confident about addressing these challenges and acquiring the required skill sets effectively. For the next two years, I will be able to take 12 credits of classes and 2 credits of LUM 4998 per semester and produce high-quality work in my classes, graduation plans, and research.

Potential Challenges:

One of the main challenges of this project will be acquiring sufficient and diverse pancreatic cancer medical image datasets from various ages, genders, and stages of pancreatic cancer. Additionally, there could be poor-quality images, such as containing artifacts.

However, we determined a few datasets that can provide enough medical images for our research, such as NIH-NCI EDRN, UK Biobank, Danish National Medical Record, and Kaggle [4].

Key Logistical Issues:

A key logistical issue is a need for a high-performance computer to process training, analysis, and experiments with large volumes of medical image data. This computer may require some investment and time to set up. Due to the tuition support from Elon University and outside

scholarships, I can utilize the funding to purchase the computer required for this research.

Another issue is to ensure the project complies with ethical and regulatory requirements. We will contact the data providers or legal and regulatory experts to ensure the project complies with privacy and data protection regulations.

Undeveloped Skills:
The undeveloped skills include: computer vision and synthetic data. The plan to develop those skills includes taking Deep Learning and Computer Vision at Carnegie Mellon University, Computer Vision and Introduction to Artificial Intelligence at Elon University, and Machine Learning Specialization and Generative Adversarial Networks Specialization from Coursera. My mentor will also provide support for my learning.

Budget

New York University Pancreatic Cancer Center **(Total: $1160)**

- Consultation for a better understanding of pancreatic cancer and its medical images
- Flight $600
- Hotel ($200/night, 3nights) $600

Professional Development **(Total: $3600)**

- Computer Vision at Carnegie Mellon University $1600
- Deep Learning at Carnegie Mellon University $1600
    - (Both courses are online certificate programs, 10 weeks)
- Machine Learning Specialization $200
- Generative Adversarial Networks(GANs) Specialization $200
    - (Both Coursera are from Coursera 49/month, 4 months)

Scholarly Materials **(Total: $600)**

- Pancreatic Cancer Detection on CT Scans with Deep Learning: A Nationwide Population-based Study $200
- Artificial Intelligence–Assisted Diagnostic Approaches for Pancreatic Disease $200
- Potential Fee for additional access to research papers and datasets $200

Conference **(Total: $3560)**

- Computer Vision and Pattern Recognition 2023
    - Virtual Ticket $100

- Pancreas Cancer: Update on Cancer Biology, Treatment, and Outcome
    - Ticket $250
    - Flight $550
    - Hotel and Food (200/night, 3night) $600

- DSS San Francisco: AI and Machine Learning in the Enterprise
  - Virtual Ticket $300

- International Conference on Artificial Intelligence Applications in Software Engineering
  - Ticket $600
  - Flight $600
  - Hotel and Food (280/night, 2 nights） $560

Dissemination of Results **(Total: $1300)**
- Publication Fee $500
- Presentation of the finding in Relevant conference and Associated Cost $800

The project will require high computation power to run experiments with machine learning, synthetic data, and medical images. Therefore, A high-performance computer is required. The PC will be built to save budgets and customized based on the focus of this project. There following are the approximated price. Pre-build Deep Learning PC ($8000 - $10,000)

Hardware Components **(Total: $8180)**
- 2 * Nvidia GeForce RTX 4090 (GPU) $4000
  The core component that affects the performance of the experiments
- SAMSUNG 34-Inch SJ55W (Monitor) $300
- AMD RYZEN™ 9 7900X3D Processor(CPU) $650
- SSD (2T GB) $250
- RAM (4*32GB = 128GB) $480
- EVGA SuperNOVA 1600 GT (Power supply) $550
- MSI PRO Z790-A WIFI ATX LGA1700 (Motherboard) $600
- Computer case $200
- Others (Wire and Fan)
  $350
- Keyboard and Mouse $200
- Potential additional cost $600

Graduate school support **(Total: $1600)**
- Graduate School Applications $1,000
  (Approximately 10 applications, each approximately $100)
- Trips to visit schools and interview $600

Total = $20,000

Bibliography

[1] American Cancer Society. “Key Statistics for Pancreatic Cancer.” (n.d.). (2023, January 12). https://www.cancer.org/cancer/pancreatic-cancer/about/key-statistics.html

[2] Chen, P. T., Wu, T., Wang, P., Chang, D., Liu, K., Wu, M., Roth, H. R., Lee, P., Liao, W., & Wang, W. (2022, Sep 13). “Pancreatic Cancer Detection on CT Scans with Deep Learning: A

Nationwide Population-based Study Radiology." 2023 306:1, 172-182 https://pubs.rsna.org/doi/10.1148/radiol.220152

[3] Huijuan Zhang, Zongrun Huang, and Zhongwei Lv. 2020. "Medical Image Synthetic Data Augmentation Using GAN." In Proceedings of the 4th International Conference on Computer Science and Application Engineering (CSAE '20). Association for Computing Machinery, New York, NY, USA, Article 133, 1–6. https://link.springer.com/chapter/10.1007/978-3-030-00536-8_1

[4] Johns Hopkins Medicine. "Pancreatic Cancer Diagnosis." (2019, June 3). [Online]. https://www.hopkinsmedicine.org/health/conditions-and-diseases/pancreatic-cancer/pancreatic-cancer-diagnosis

[5] Kenner, B., Chari, S. T., Kelsen, D., Klimstra, D. S., Pandol, S. J., Rosenthal, M., Rustgi, A. K., Taylor, J. A., Yala, A., Abul-Husn, N., Andersen, D. K., Bernstein, D., Brunak, S., Canto, M. I., Eldar, Y. C., Fishman, E. K., Fleshman, J., Go, V. L. W., Holt, J. M., Field, B., … Wolpin, B. (2021). "Artificial Intelligence and Early Detection of Pancreatic Cancer: 2020 Summative Review." Pancreas, 50(3), 251–279. https://doi.org/10.1097/MPA.0000000000001762

[6] Larrazabal, A. J., Nieto, N., Peterson, V., Milone, D. H., & Ferrante, E. (2020, May 26). "Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis." https://doi.org/10.1073/pnas.1919012117

[7] Liu, Y., Le, Y., Wang, T., Fu, Y., Tang, X., Curran, W. J., Liu, T., Patel, P., & Yang, X. (2020, March 28). "CBCT-based synthetic CT generation using deep-attention cycleGAN for pancreatic adaptive radiotherapy." PubMed Central (PMC). https://doi.org/10.1002/mp.14121

[8] National Cancer Institute. "Test Shows Promise for Pancreatic Cancer Early Detection," (2017, August 9). https://www.cancer.gov/news-events/cancer-currents-blog/2017/blood-test-pancreatic-cancer

[9] Razzak, M. I., Naz, S., & Zaib, A. (2017, November 14). "Deep Learning for Medical Image Processing: Overview, Challenges and the Future." SpringerLink. https://doi.org/10.1007/978-3-319-65981-7_12

[10] Sáez, C., Romero, N., Conejero, J. A., & García-Gómez, J. M. (2020, October 7). "Potential limitations in COVID-19 machine learning due to data source variability: A case study in the nCov2019 dataset." PubMed Central (PMC). https://doi.org/10.1093/jamia/ocaa258

[11] Shin, H. C., Tenenholtz, N. A., Rogers, J. K., Schwarz, C. G., Senjem, M. L., Gunter, J. L., Andriole, K. P., & Michalski, M. (2018, September 12). "Medical Image Synthesis for Data Augmentation and Anonymization Using Generative Adversarial Networks." SpringerLink. https://doi.org/10.1007/978-3-030-00536-8_1

[12] Tejo Catalá, O. D., Igual, I. S., Pérez-Benito, F. J., Escrivá, D. M., Castelló, V. O., Llobet, R., & Peréz-Cortés, J. C. (2021, March 10). "Bias Analysis on Public X-Ray Image Datasets of

Pneumonia and COVID-19 Patients." PubMed Central (PMC). https://doi.org/10.1109/ACCESS.2021.3065456

[13] Michael Hollingsworth, (2014). "Multiplex-Immunoflourescent Staining of Rapid Autopsy Samples from Human Pancreatic Cancer at the Primary and Metastatic Sites," National Cancer Institute, [Online]. https://doi.org/10.26252/s18j-1h24

[14] Berryman, Sean. (Sep 16, 2020). "Pancreas-CT," Kaggle, [Online]. https://www.kaggle.com/datasets/salihayesilyurt/pancreas-ct

[15] UK Biobank, "Pancreas MRI images and derived data," [Online]. https://biobank.ndph.ox.ac.uk/showcase/label.cgi?id=131

Timeline

Summer 2023	- Relevant background reading - Enroll in Machine Learning Specialization Course	- Machine Learning Specialization Certificate
Fall 2023	- Install required Computer - Enroll in Computer Vision at Carnegie Mellon University - Complete SURE Application	- A high-performance computer for later research experiments - Computer Vision Certificate
Winter 2024	- Enroll in Deep Learning at Carnegie Mellon University - Running experiment with the real medical image with deep learning and computer vision to classify the medical image of pancreatic cancer record the performance of the model - Specialist consultation for better understanding the medical image of Pancreatic Cancer and its early stage features	- Deep Learning Certificate - Analysis of the performance of the deep learning model without the augment of synthetic data
Spring 2024	- Enroll Generative Adversarial Networks(GANs) Specialization	- GAN Certificate

	- Start generating medical images of pancreatic cancer	- A model that can generate realistic pancreatic cancer medical image data.
Summer 2024	- Analyze the existing public or private image dataset of pancreatic cancer to identify the diversity level. - Gather more data if needed - Use Synthetic data to balance the datasets if needed - SURE	- A summary of the potential problems or biases of the pancreatic cancer datasets - Synthetic pancreatic cancer medical images dataset - SURF Presentation
Fall 2024	- Using GAN to generate the medical image of pancreatic cancer in the early stage - Train model with synthetic data and record its performance. Apply for NCUR	- A detailed analysis of the performance of synthetic data on the machine learning model
Winter 2025	- Final optimization of the machine model and GAN (model to generate data)	- Final version of GAN that can generate realistic pancreatic cancer medical images - Final version of machine learning model that can diagnose pancreatic cancer in the early stage
Spring 2025	- Write the research paper - Prepare for presentation on my findings	- Publish the research finding - Presentation of the research in NCUR - Publish the deep learning models on GitHub