Implementation Plan

Cancer can be defined as a disease where one’s own cells choose to grow without control (National Cancer Institute, 2021). These cells (named cancer cells) not only grow uncontrollably, but also takes resources from neighboring cells, tricks the immune system into assisting it, and finally kill the host (National Cancer Institute, 2021). For these reasons, it is crucial to discover cancerous cells as early as possible (NCI Staff, 2022). To assist in the early diagnosis of cancerous cells, an application using Artificial Intelligence to diagnose images based on patient data will be implemented.

Software Requirements

The software requirements will consist of mainly python libraries. To develop the machine learning model, Google’s Big Query and the NBIA Data Retriever will be used to collect DICOM data from The Imaging Data Commons database.

* Imaging Data Commons (IDC)
  + Tool used to allow researchers with access to data sets from multiple data coordinating centers (*Imaging Data Commons | CRDC*, n.d.).
  + Allows for some data exploration as one can observe samples from the datasets selected for querying on Google’s Big Query platform (*Imaging Data Commons | CRDC*, n.d.).
* Google’s Big Query
  + High flexibility, with multiple APIs for languages such as python, Java, JavaScript, and Go (*What Is BigQuery?*, n.d.).
  + Allows the use of SQL to query any data sets (*What Is BigQuery?*, n.d.).
* Python Programming Language
  + Contains Graphical User Interface Libraries (Sanner, n.d.).
  + High level programming language allows one to easily use complex functions while maintaining readability within code (Sanner, n.d.).
* TensorFlow library
  + Machine learning Library that allows ease of use and development of models from the concept down to the deployment of said model (TensorFlow, n.d.).
  + Allows easy visualization of machine learning model through computational graphs (Goldsborough, n.d.).
  + Allows usage of GPU to train machine learning models, thus decreasing the amount of time spent training the model (Goldsborough, n.d.).
* Pydicom library
  + Allows one to read, write, or delete DICOM files (*About Pydicom |*, n.d.).
  + Compatibility with the python tkinter library (*About Pydicom |*, n.d.).
* Matplotlib library
  + Library for creating visualization using Python (Matplotlib, 2012).
  + Compatibility with the python Tkinter library (*Embedding in Tk — Matplotlib 3.6.0 Documentation*, n.d.).

Deployment

The software will be deployed as its own independent executable file. As it is written in Python, one will not have to worry about creating a version that is compatible with windows as opposed to a version that is compatible with Linux (Aaberge, 2020). Instructions will also be provided to guide the user into downloading and installing the Python programming language from the official site and provide a version of python that is compatible for use with the software application. An online page will be open for users of the application to voice their concerns (this page may possibly be a reddit page).

Strategies, Impact, & Activities

The main strategies will be to allow the user to use any number of DICOM files for making predictions. A separate API may be used to allow the update of the model using the hospital’s own data, but this may be part of the second iteration. The user will also be allowed to view a fixed dashboard and download the dashboard itself. This software app is meant to impact the amount of time spent determining whether a tumor is malignant or benign, thus saving radiologists time.

Stakeholders

Stakeholders will be doctors, patients, and radiologists. Despite the potential of AI to assist in early cancer detection, medical experts (doctors, radiologists, and other medical professionals) may show concerns regarding the possible high number of false positives (NCI Staff, 2022). For this reason, an accuracy of at least 95% is expected of the model used for making predictions on images through the software application. Not only is this a concern for the medical experts, but also the patient who may undergone unnecessary treatments and extra tests due to the prediction made by the Artificial Intelligence (NCI Staff, 2022).

Security & Risk

In terms of security, the access of the data will be left completely to the user. Every hospital follows the same standards set by HIPAA (hhs, 2013). Based on this, interconnectivity to any established server by the hospital will be made available so that the user can access the software application locally. To prevent data breaches due to malicious code within certain libraries, one will only use reputable libraries and inspect libraries for vulnerabilities before updating the software application.

Data Governance

Data Governance will be handled through a series of data preprocessing procedures. New data introduced for training the machine learning model will undergo data balancing based on categorical values (such as age, weight, and ethnicity). By balancing the data, one prevents any possible bias against a specific age group or ethnicity. The finalized model will be tested against a random data set of similar nature to the training set and in the case that its accuracy does not meet the minimum 95% requirement, then the model will be scrapped, the data will be further examined, and a new model will be developed for training.

Source(s):

Aaberge, M. A. (2020, September 13). *Everything You Need to Get Started With Python Programming*. Medium. https://towardsdatascience.com/everything-you-need-to-get-started-with-python-programming-4a37a46e427b#:~:text=Operating%20Systems

*About Pydicom |*. (n.d.). Pydicom.github.io. https://pydicom.github.io/about

*Embedding in Tk — Matplotlib 3.6.0 documentation*. (n.d.). Matplotlib.org. Retrieved October 6, 2022, from https://matplotlib.org/stable/gallery/user\_interfaces/embedding\_in\_tk\_sgskip.html

Goldsborough, P. (n.d.). *A Tour of TensorFlow Proseminar Data Mining*. https://arxiv.org/pdf/1610.01178.pdf

hhs. (2013, July 26). *Summary of the HIPAA Security Rule*. HHS.gov. https://www.hhs.gov/hipaa/for-professionals/security/laws-regulations/index.html

*Imaging Data Commons | CRDC*. (n.d.). Datacommons.cancer.gov. Retrieved October 6, 2022, from https://datacommons.cancer.gov/repository/imaging-data-commons

Matplotlib. (2012). *Matplotlib: Python plotting — Matplotlib 3.1.1 documentation*. Matplotlib.org. https://matplotlib.org/

National Cancer Institute. (2021, May 5). *What Is Cancer?* National Cancer Institute; Cancer.gov. https://www.cancer.gov/about-cancer/understanding/what-is-cancer

NCI Staff. (2022, March 22). *Can Artificial Intelligence Help See Cancer in New Ways? - NCI*. Www.cancer.gov. https://www.cancer.gov/news-events/cancer-currents-blog/2022/artificial-intelligence-cancer-imaging#:~:text=Finding%20cancer%20early&text=Scientists%20have%20developed%20AI%20tools

Sanner, M. F. (n.d.). *Python: a programming language for software integration and development*. Retrieved October 5, 2022, from https://d1wqtxts1xzle7.cloudfront.net/25505223/10.1.1.35.6459-with-cover-page-v2.pdf?Expires=1665025244&Signature=cz2a~eRrnV1NdOEGS~77W2jyINX-Uob~wjkbVCjLP3jHHca9heKuxGUNUpXuRhZBQ04rD7430rXoXlBgBa1004KBIQVGEH58LSlAswL0y-l8CPxtm4y0c8Mr54BFLMhaH1ZOJHc0GplnLBIFyyC9AKcGLUIM4H-0c1J1X~m64glPdpqKe6mqeESj~tygnJyf3Scbu8enqJq-06cD8ULwlMtro~2srdj34eMIUjbSHrp9BBiC8~R2BN~b9nwmM1QO4mo2ArIDOQlfURUoLnpaheELjZu0DXdAz2ssxhy6xwNRq9nmYtQUZ0wr12g6WHVeylcTfZrHyyzgShiYyE2pFQ\_\_&Key-Pair-Id=APKAJLOHF5GGSLRBV4ZA

TensorFlow. (n.d.). *Why TensorFlow*. TensorFlow. https://www.tensorflow.org/about

*What is BigQuery?* (n.d.). Google Cloud. https://cloud.google.com/bigquery/docs/introduction