Traceable medical image de-identification pipeline for AI application (TMI):

“DICOManon”

Version 1.0

Group 0:

Yidu Guo

Ben Wilson

Nigam Lad

Table of Contents:

1. Introduction

1.1 Purpose

Medical research of the modern world has become more and more reliant on data, and processing it on large scales. AI has become an invaluable asset to the medical research field. One major concern that many researchers face is lack of access to clinical data. This project introduces a pipeline that streamlines the process of secure access to medical data for researchers within the BC medical system. The project seeks to anonymize real patient data from clinical studies, and streamlines researchers' access to the data either by themselves or by AI. .

1.2 Intended Audience

This project's immediate aim is to assist BC researchers. However, with the right steps toward implementation, the project could be beneficial to researchers or professionals in other industries who require access to clinical data.

1.3 Intended Use

The purpose of this project is to automatically de-identify DICOM files by removing sensitive PHI (Patient Health Information) from it. This server will also provide secure access to the anonymized data for BC researchers.

1.4 Scope

The project will serve as a plug-in for clinical data retrieval for AI and researchers for BC Cancer. This plug in will ensure that researchers or AI have proper access to desired clinical data without breaching patient confidentiality.

Specifically, the plug-in will be designed to anonymize DICOM files sent by clinical imaging machines such as CT scanners, and be able to communicate these anonymized data upon request by researcher or AI. The plug-in will also be developed with back-tracking in mind, such

that anonymized data is able to be grouped, sorted, and retrieved based on PHI, but without exposing patient PHI to researchers.

1.5 Definitions and Acronym

Term	**Definition**
PHI	Personal Health Information.
DICOM	International Standard file type for managing and communicating medical images and data.
TMI	Traceable Medical Image
PACS Server	Picture Archiving and Communication System Server

2. Overall Description

2.1 Data Flow Diagram

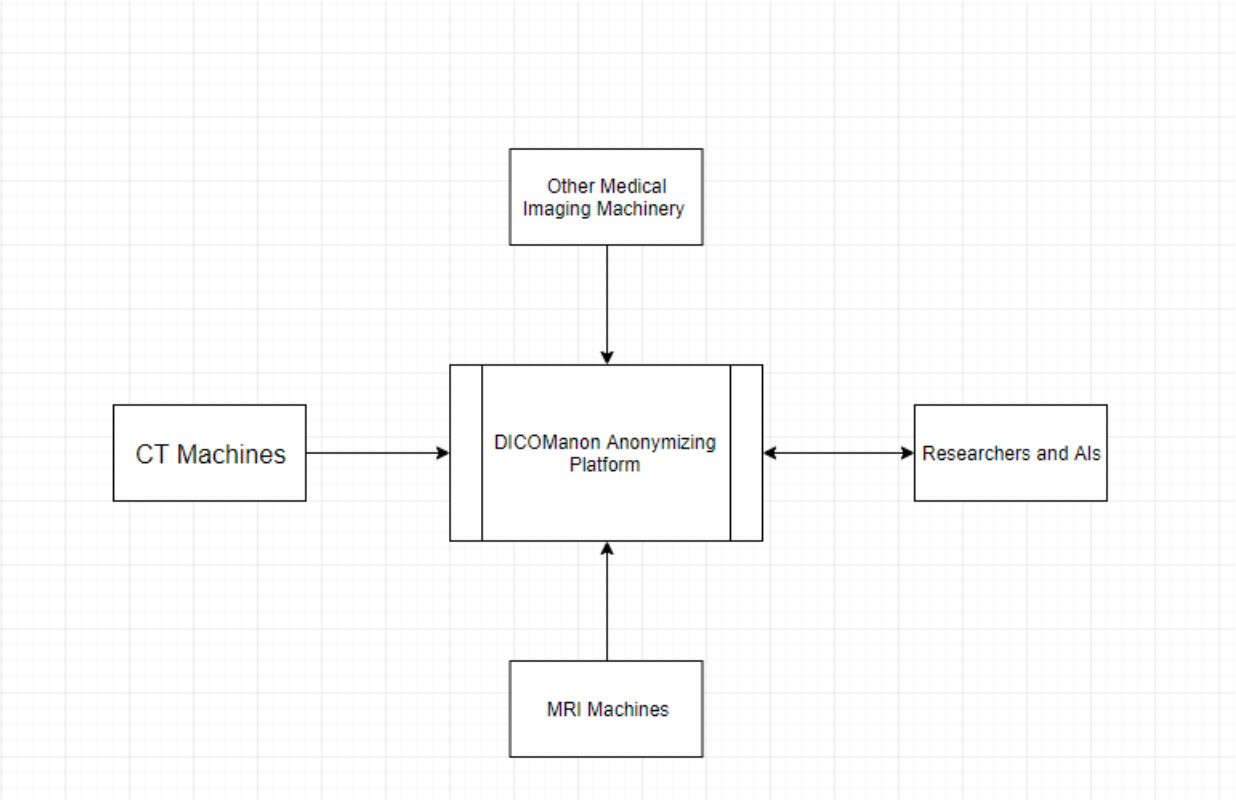

Figure 1: Level 0 DFD

Medical imaging machines such as MRI or CT machines, will send an instance of patient data along with a copy of their medical image in the form of a DICOM file to a server. The data will be processed and anonymized before presented to researchers and AIs based on their query. Researchers and AIs can also sort, group, and modify the data without being exposed to PHI Our purpose and goal is to de-identify and anonymize the dicom file. Our hope is that the plug in can be widely used by researchers throughout the province.

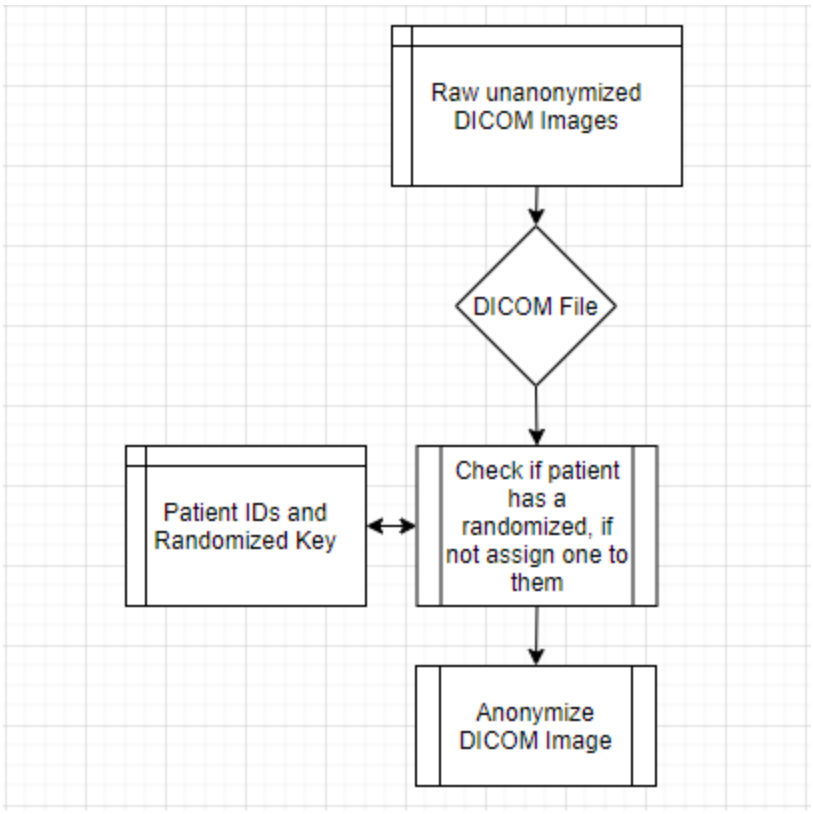

Figure 2: Level 1 DFD

The Orthanc server we will be programming will need to keep track of DICOM images coming in from MRI machines. The system will need to anonymize patient data and give them randomized IDs. Multiple DICOMs from the same patient may arrive at different times, so the server must be able to link the new images to previous patient's data and provide the same randomized ID. to it

2.2 User Needs

The functions of the plug-in aim to fulfill is currently carried out by the client manually. The client desires the process to be automated in a streamlined fashion.

3. System Features and Requirements

3.1 Functional Requirements

3.1.1 Anonymizing DICOM file:

The core purpose of this project is to automate the process of anonymizing DICOM files from its current state of manual process. The anonymizing process is a vital aspect of the entire TMI pipeline. Specifically, the function is to censor patient PHI from DICOM files sent from medical imaging machines such as CT scans. It is also important to note that the anonymized data is still used for countering mismatch, grouping, and sorting.

3.1.2 Create back-tracking cross reference table

As mentioned, the data that is being anonymized is used for organizing the DICOM file. One important aspect that the client emphasized is for the anonymized data to be able to back-tracked to the censored data. For researchers, it is important to be able to know the progress of a certain patient or a certain study without personally seeing PHI associated with them. The back-tracking cross reference table will be able to trace an anonymized DICOM file to the desired group, either by patient or study at the back end. The goal of this function is to be able to trace a DICOM file given a perimeter that may include PHI, and return desired results without exposing PHI to researcher or AI querying.

3.1.3 Store DICOM file in PACS server

Once a DICOM has gone through the anonymized process, it is then placed in the PACS server. Researchers and AIs will be able to query the database. The database will also interact with the back-tracking cross reference table mentioned before to check mis-matches or to comply with sorting or grouping requirements from researchers/AI.

3.2 System Features

- Given the project aim is to achieve a specific function of an entire TMI pipeline project, extra features or functions are difficult to predict and plan. It is up to the client to initiate more system features.

3.3 Non-Functional Requirements

- Security: DICOM file before de-identification and anonymizing contain personal health information (PHI) o Access to anonymized data must be limited and monitored
- Reliability: Robust design to prevent breaches on PHI
- Maintainability: Admin needs to able change the perimeter of which part of DICOM file is to be de-anonymize

3.4 Environmental Constraints

- Lack of access to BC cancer facility and clinical sample data

4. Testing and Milestones

4.1 Milestones

- Milestone 1
  - Complete Data Flow Diagram level 0 and 1
  - Complete Software Requirement Specification Document
  - Set up weekly meeting with client
- Milestone 2
  - Set up Orthanc environment
  - Set up PACS server
  - Determine language (Python or C)
- Milestone 3
  - Complete all functional requirement
    - Complete anonymizing function
    - Complete cross-reference table for back tracking
    - Complete processed DICOM file storage in PACS server
  - In Group Testing

- Milestone 4
    - Peer testing #1
    - Peer testing #2
    - Project Deliverance

4.2 Testing Strategy

The current testing plan includes an initial phase of unit-test, specially to test for corner cases. We plan to create our own DICOM files and feed it through the anonymizing platform. Our hope is to be able to connect the BC Cancer system and continue to integrate tests on their server. However, given we will be processing actual patient PHI, the likelihood for this testing to occur is dependent on the client.

More testing strategies will be planned and implemented as more information and access by the client is released to us.

5. Questions

What type of testing methods are you implementing? Regression? Unit-testing? Integration?

We will be performing a unit test on our main functional requirement. We must test and make sure that the DICOM file after the anonymization process contains no PHI.

What language have you decided on between Python and C++?

We have already decided on python, given some of the packages such as panda and numpy are inline with the data manipulation that we will be doing, it makes sense to develop with python with these tools already included.

What is the data scale that your application will have to run at? How large are these files and how long (worst case) would it take for your system to run through a set of these images? Do you think you'll have to consider writing portions of your code in something like C to take advantage of its speed to handle large amounts of data? Do you anticipate other bottlenecks in the speed of your system?

According to the client, our system needs to be able to process thousands of images everyday. The server will need to handle large amounts of DICOM files in the 10s of Megabytes, however this logistical problem will be handled by BC Cancer. Our task of anonymizing is a simple process or reading and writing one at a time to a database.

What security measures are in place to ensure patient information is not leaked?

According to the client, security does not need to be handled by our team. Our system will be integrated into their own secure internal network. When this system is integrated into the BC Cancer system, security and operations of the software will be handled by them.