De-identifying Medical Images – Made Easy

Initiating a Literature Review and Identifying Relevant Data Sources

---

Student Details

Name: Likhith Ravula

NUID: 002293027

Course Details

Course ID: IE 7945

Faculty: Prof Kirankumar Trivedi

Submission Details

Date: May 20th, 2024

By: Likhith Ravula

Contribution: 100%

Contents:

1. Introduction

1.1. De-identification of Medical Images

To protect people's privacy, data can be anonymized by taking out or changing any details that could be used to identify them. This makes it safe to share the data with others for various purposes across the industry/ organization.

Regulations like HIPAA exist to anonymize data, especially in healthcare. This anonymization process ensures information can't be linked back to specific individuals. For instance, the human subject research data needs to be analyzed but privacy for the participants must be a top priority. De-identification helps achieve this balance.

Direct identifiers and demographics, also known as Protected Health Information (PHI), includes a patient's name, address, gender, etc., and convey a patient's physical or mental health condition, or diagnosis related to that individual, as well as financial data related to healthcare like, medical records, bills, and lab results. These must be de-identified before the data is stored/ shared across serves.

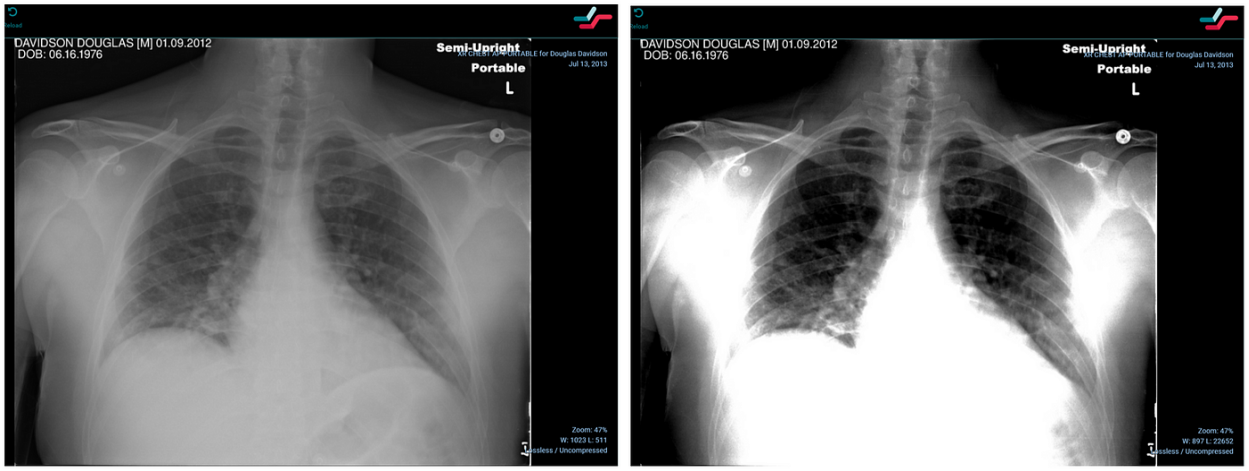
1.2. Existing methods and limitations

The existing methods to de-identify images have a lot of limitations including file formats, etc. The major limitation is that the existing methods are built for traditional file formats such as PNG, JPG, etc. Medical images (MR, CT, and others) are transmitted and stored and processed in a file type called DICOM (Digital Imaging and Communication in Medicine), which is a standard built to adhere to the HIPAA norms.

Converting medical images (DICOM) to common formats (e.g., JPEG) works for feeding data to ML models (Computer Vision & NLP) that remove sensitive text. However, this conversion can degrade image quality. Medical images use specific pixel data formats (photometric interpretations) that differ from common image formats. While using the converted images in ML models might be possible, converting the anonymized results back to DICOM format

becomes difficult as the conversion process compresses the image and loses crucial data embedded in the original DICOM file metadata.

Original DICOM image | After converting to PNG and back to DICOM



source

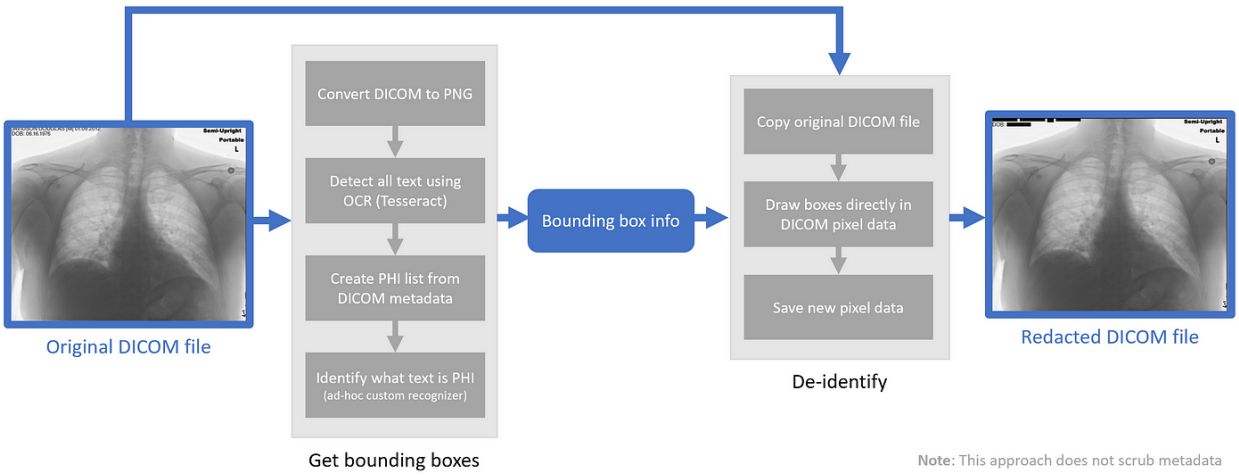
1.3. Proposed Approach

An open-source python module, Presidio DICOM image redactor, avoids the problem of losing image quality. It modifies the pixels directly within the original DICOM file.

Working: To identify which pixels to change, it uses a two-step process:

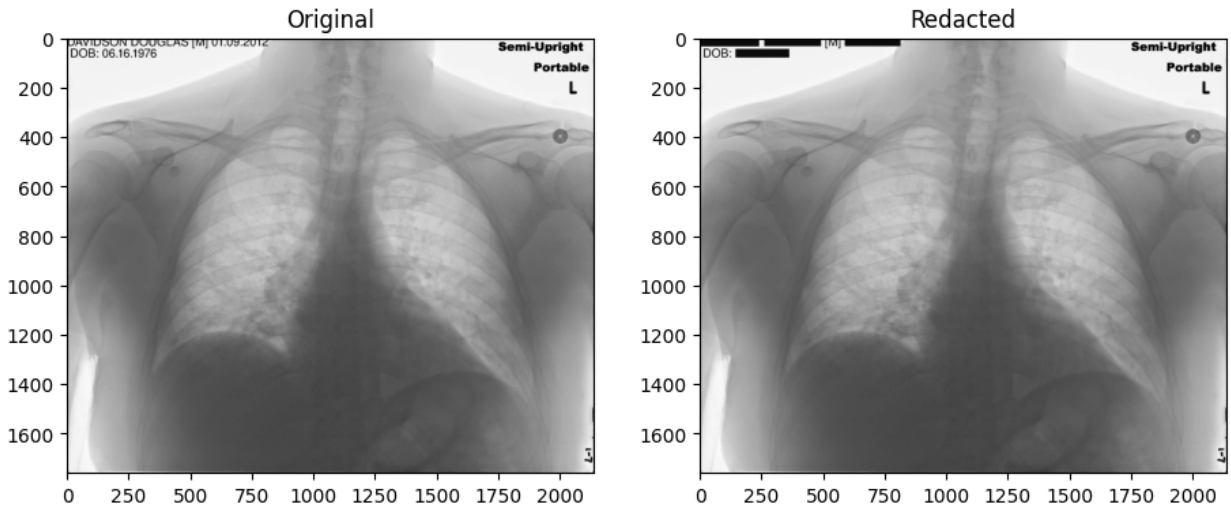
- OCR (Optical Character Recognition) detects all text in the image.
- NER (Named Entity Recognition) then pinpoints only the sensitive text containing personal health information (PHI).

Only the pixels corresponding to this identified PHI in the original DICOM file are then modified so that the de-identification is performed without any compression or loss.



source

Below is an example of an original DICOM image and a redacted DICOM image.



source

2. Literature and citations

[1] Macdonald JA, Morgan KR, Konkel B, Abdullah K, Martin M, Ennis C, Lo JY, Stroo M, Snyder DC, Bashir MR. A Method for Efficient De-identification of DICOM Metadata and Burned-in Pixel Text. J Imaging Inform Med. 2024 Apr 8. doi: 10.1007/s10278-024-01098-7. Epub ahead of print. PMID: 38587767.

[2] Vcelak P, Kryl M, Kratochvil M, Kleckova J. Identification and classification of DICOM files with burned-in text content. Int J Med Inform. 2019 Jun;126:128-137. doi: 10.1016/j.ijmedinf.2019.02.011. Epub 2019 Mar 1. PMID: 31029254.

[3] Monteiro E, Costa C, Oliveira JL. A De-Identification Pipeline for Ultrasound Medical Images in DICOM Format. J Med Syst. 2017 May;41(5):89. doi: 10.1007/s10916-017-0736-1. Epub 2017 Apr 13. PMID: 28405948.

[4] Presidio Image Redactor, Documentation

[5] Presidio Redactor sample python implementation, Notebook

[6] Redacting sensitive text from DICOM medical images in Python, Article

[7] De-identification in Medical Imaging, Article

3. Data Source

The DICOM images are extracted from The Cancer Imaging Archive (TCIA) for the development and evaluation of medical image de-identification. Data citation: Rutherford, M., Mun, S.K., Levine, B., Bennett, W.C., Smith, K., Farmer, P., Jarosz, J., Wagner, U., Farahani, K., Prior, F. (2021). A DICOM dataset for evaluation of medical image de-identification (Pseudo-PHI-DICOM-Data) [Data set]. The Cancer Imaging Archive. DOI: https://doi.org/10.7937/s17z-r072