Master's Thesis Pitch
The Norwegian Open AI Lab And
The Norwegian Research Center for AI Innovation (NorwAI)
March 9, 2021

How can we separate medical data from personal data?

Introduction

dedeX is a med tech startup. dedeX is looking to solve the problem: how do we get the data needed to build AI algorithms that can empower medical history taking and physical examination? dedeX understands both real-world healthcare and AI and can give the student feedback from this unique perspective.

Problem Description

The value of an AI system in healthcare is related to its benefit to patients and healthcare providers and inversely related to its cost and privacy risk. Most ideas for AI solutions in healthcare are simply rejected because the benefit is unknown, the cost may be low or medium, and the privacy risk is high.

$$\textit{Value Of AI System In Healthcare} = \frac{Benefit}{Cost \times Privacy\ Risk}$$

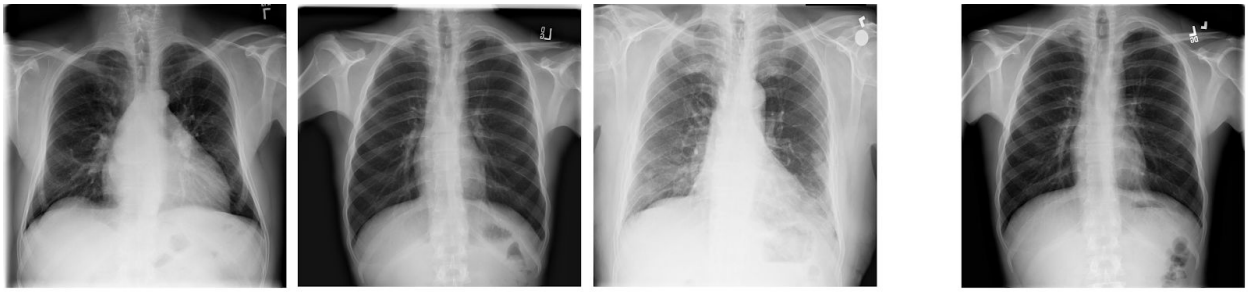
If your AI system can cure cancer then it can be valuable even if its cost and privacy risk is very high. Most AI systems however are not curing cancer so a useful strategy is to reduce the privacy risk in order to make the AI system more valuable.

To achieve this we need better methods for anonymizing data where it is mathematically possible. We are willing to spend a large amount on healthcare (around 10% of GDP) including large sums on individual solutions that are beneficial to patients. The added computational cost of anonymizing data is relatively small and will likely be outweighed by the reduction in privacy risk and value will increase.

While it is expected that using anonymization for an AI system will lower accuracy for any given dataset, in a real-use scenario an anonymized system may allow for larger datasets and could end up having a higher accuracy.

Data

To explore this problem we will look at chest x-rays. A Chest x-ray is a common radiological test performed to evaluate among others the lungs and heart of a patient.



Here are 4 chest x-rays from 3 different patients. Can you guess which of the 3 chest x-rays on the left comes from the same patient as the chest x-ray on the right?

Chest x-rays were chosen for this task because of the availability of a large open dataset of over 100,000 images that has been evaluated by other machine learning researchers. In chest x-rays the information that is medically relevant such as what is happening in the lung fields (the large black areas) is not exactly the same as the information that allows us to identify patients (the shape of the skeleton or the presence of implants). Chest x-ray datasets are considered anonymous data by many but it is more correct to refer to them as pseudonymous data.

Chest x-ray data published by National Institutes of Health in the United States
https://nihcc.app.box.com/v/ChestXray-NIHCC

Challenges

1. Investigate the possibility of applying existing state-of-the-art deep learning methods to anonymize the chest x-rays. As part of this, the student is expected to perform a state-of-the-art literature review and implement the most relevant method(s) that can solve the problem.
2. Evaluate the effectiveness of the methods using attributes of anonymous data such as k-anonymity and l-diversity and the risk for re identification.
3. Evaluate the trade-off between anonymization and classification performance when the data is used in supervised machine learning.
4. Explore whether an anonymized dataset can achieve superior performance in a machine learning classification task when it is compared to a smaller non-anonymized dataset.
5. Make the chosen method time effective, feasible and available for researchers or developers of AI technologies.

Thesis Information

Timeframe: 6 or 12 months
Supervisor: This thesis has been discussed with Professor Adil Rasheed.
dedeX contact: Jon Bekker - jon@dedex.ai