College of Integrative Sciences and Arts
Arizona State University

Preparation of Pseudodata for Application of Machine Learning Techniques in Medical Imaging

Randy Montoya, Dr. Michael Dugger, and Dr. Wenwei Zheng
College of Integrative Sciences and Arts, Arizona State University, Polytechnic

Introduction

Medical imaging has many uses, one of which is the diagnosis of cancers. Medical imaging can determine if a cancerous tumor is present, the location of the tumor and how much the tumor has grown. These factors are considered when planning a treatment for a patient. Our key focus when conducting the research contained in this poster is to use Machine Learning (ML) techniques for the evaluation of Computerized Tomography (CT) scans of a brain, to help aid in the diagnosis of whether a tumor is present or not. Prior to the utilization of the ML techniques, pseudodata representing brains ranging from healthy to having differing sizes and locations of tumorous growth were generated. The construction of the pseudodata was made using a C++ software library named GEANT4. A Python program named TomoPy was utilized for the tomographic reconstruction. The pseudodata is prepared for unsupervised learning techniques called isometric mapping (Isomap) or principal component analysis (PCA) to sort through our high dimensional data. It is hoped that by showing that machine learning techniques can make a clear distinction between a CT-scans, with and without cancerous tumors, that ML techniques can be constructed to find tumors that would not be caught by human inspection.

Geant4 and Tomopy

GEANT4 is software that is utilized to simulate the passage of particles through matter using Monte Carlo methods. We simulate photons being shot at a phantom head. The phantom head is made of 3 spheres: the first sphere is the outer shell which represents a skull made of bone; the second sphere fills the skull and is made of brain matter; the third sphere is a tumor that is made of tissue and iodine as a contrast agent. The beam of photons has an energy of 0.15 MeV which is enough energy to produce a scan that shows the tumor well. The detector plane is utilized to count the photons that traverse the the phantom head. The way we acquire the high dimensional data is by rotating the beam and the detector around our phantom head while the detector plane records the intensity of the photons at every degree from 0 to 180. After retrieving the high-dimensional dataset, we utilized TomoPy which uses Fourier analysis to reconstruct our data and produce a CT-scan of the phantom head.

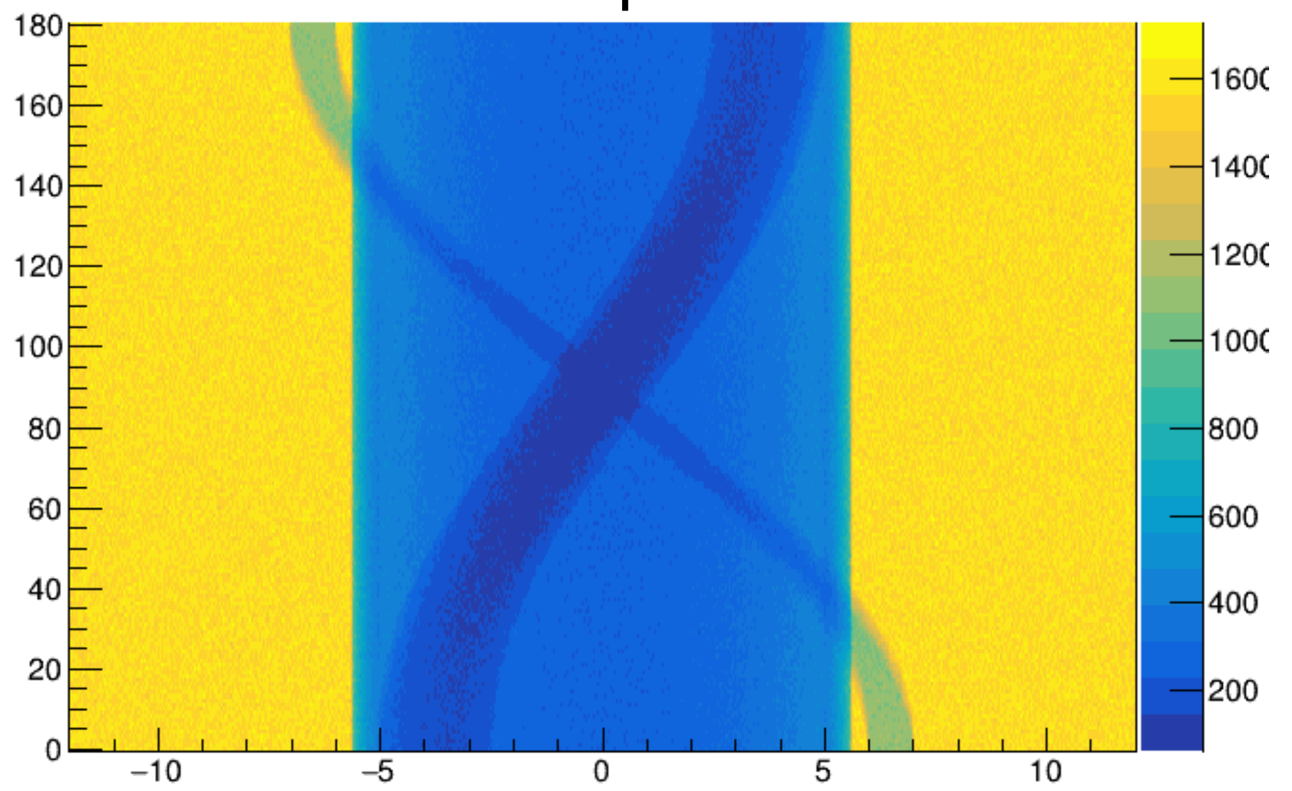

Figure 1: A sinogram which is a collection of the data in terms of angle/degree versus detector position/cm.

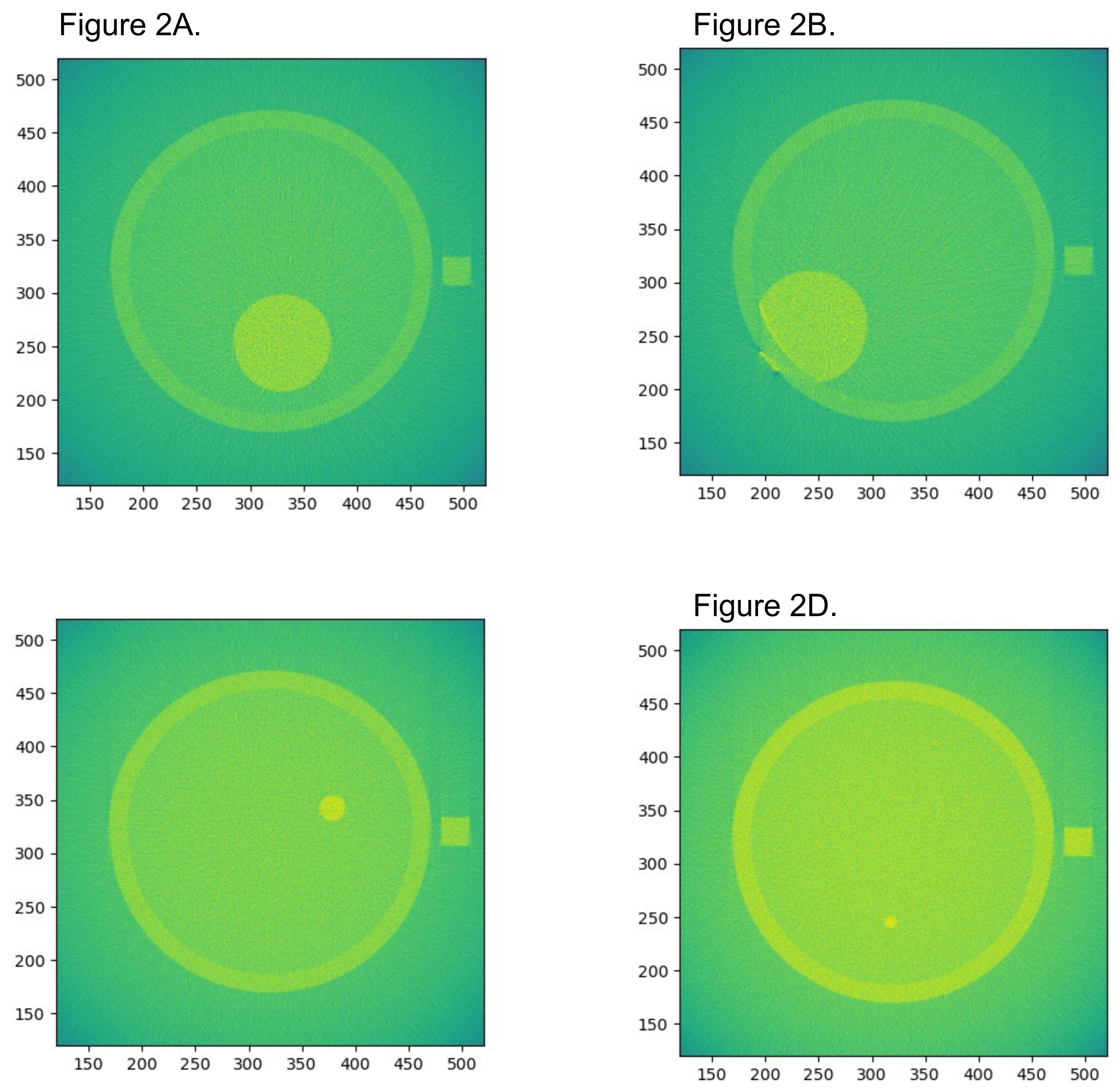
Figure 2A.

Figure 2B.

Figure 2D.

Figure 2A,2B,2C,2D show an example of the images that were produced and used to create our high dimensional data. In total, 100 tumor examples were created by 12 instances of the simulation running on two nodes of a cluster (each node having 6 processing cores). Each image has a unique tumor size and position.

Machine Learning

For the machine learning aspect, we are hoping to use an unsupervised learning technique called manifold learning. Unsupervised learning is typically utilized to identify patterns and trends in raw datasets, and clusters similar data into specific groups. By testing our data with two different techniques, we want to show that Manifold learning techniques outperform principal component analysis, more specifically isometric mapping. Before testing our data, we believe that Isomap is the optimal algorithm that could deal with my simulation of real-world data that would come from a brain scan with a tumor. The reason principal component analysis or PCA can fail when analyzing my data is because we have simulated several scans that show a tumor with varying degrees of movement and varying size. It is hard to tell before testing, but this could fluster the PCA algorithm because the difference between tumor images are not necessarily in a linear space. PCA could produce bad classifications for the reasons stated above. Isometric mapping specifically seeks a lower dimensional representation of our data that maintains geodesic distances or distance for a curved surface between my images. The geodesic distance used by Isomap might more effectively capture the difference between these images. Essentially PCA will try to flatten my data and find similarities in this way which could fail for my data and create errors.

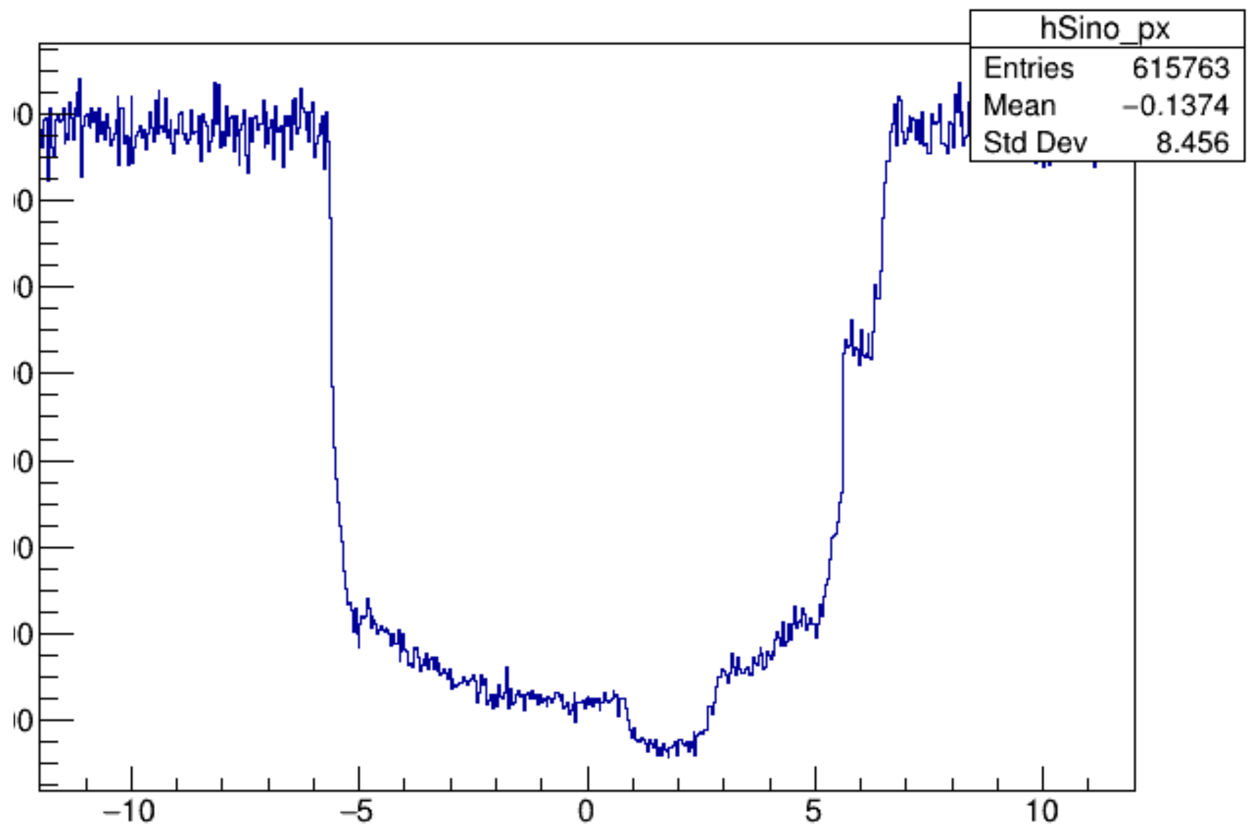

Figure 3. shows a projection of the sinogram below onto the x-axis. The divot shown is the tumor.

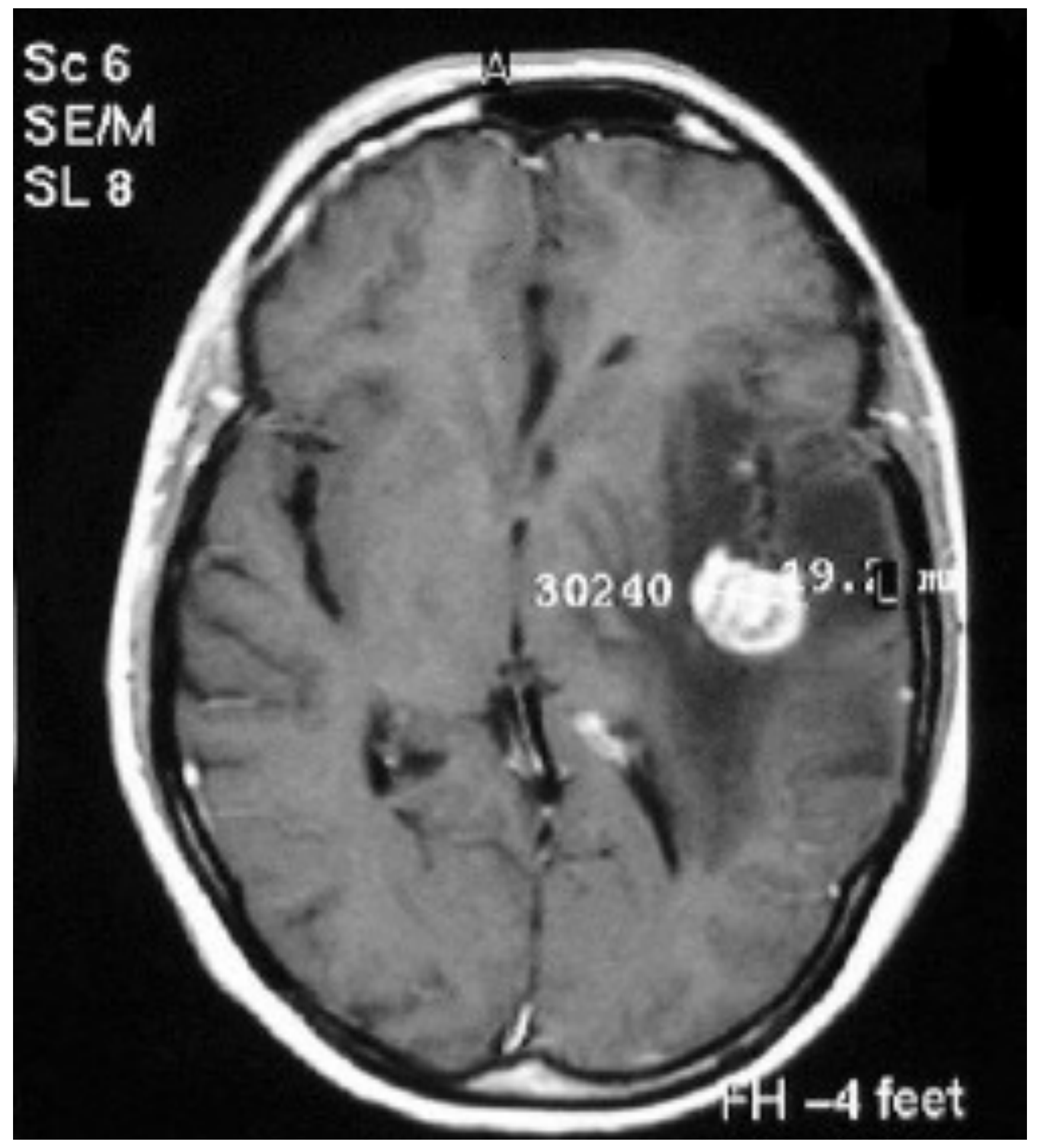

Figure 4. shows a real brain scan with a tumor present

Summary

To prepare our high dimensional pseudodata we decided to use python to open our CT-scans and convert the images into a dataset that can be implemented into two algorithms. By generating simulated CT-scans, preparing this data, and using machine learning techniques we hope to show that isomap can analyze our high dimensional data. Which will prove that these ML techniques can find tumors that can't be seen by human inspection.

Future directions

The next step is to take our prepared pseudodata and test PCA and Isomap to find which algorithm produce the least errors. Then write a program that takes the images that are flagged for having a tumor and have that program analyze the data within those pictures alone to tell the exact location of the tumor by way of Convolutional Neural Networks.