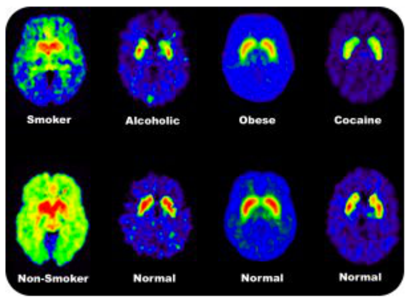
Emily Chu
Parsons School of Design - MS Data Visualization

Supervised Learning Task: Continuous image processing based on intervention instructions for health goals

I would like to address how to use intervention instructions to manipulate images of PET scans (see Figure 24.12.4 below). The particular application I have in mind is for setting and maintaining health goals. Examples could be applied to smoking cessation or weight loss goals, where someone inputs “Picture [subject] after 10 years of quitting smoking starting today” or “Picture [subject] losing 2 pounds a year for 5 years”. The PET scan is processed based on the data it has learned from the average progression over time of other relevant use cases. The output is a series of images by interval based on the duration specified for the health goal i.e. a total time frame = 5 years, intervals = 1 year would return 5 total images, the first one “taken” one year after the original assuming no smoking, and the final fifth image “taken” five years after the original assuming no smoking).

Figure 24.12.4 from Chemistry LibreTexts:
PET scan of dopamine binding in brains of normal and addicted individuals.



Data Inputs

I imagine a large amount of image data associated with progression (i.e. at least 10,000 people each with their own image series identifiable by time) by different use cases (i.e. weight, smoking) would be needed for this project. It would probably be best to identify one type of use case i.e. smoking to start with. Using this approach, each image in the dataset can be assumed as “smoker” or “non-smoker”, identifiable by the ID number of a person’s series and its position in each time series.

Obtaining data inputs

An initial search query of “public dataset PET scans” returned the following top results:

- The Reference Image Database to Evaluate Therapy Response (RIDER)
  - Measures the response to drug or radiation therapy
  - Dataset: PET scans for 244 patients, 1,349 series, representing 275 studies and totaling 269,511images
- The FDG-PET/CT and radiotherapy planning CT imaging collection
  - Measures 300 patients from four different institutions in Québec with histologically proven head-and-neck cancer
  - Dataset: PET scans for 298/300 patients, 2,413 series, representing 504 studies and totaling 117,165 images
- The most promising dataset is from The Cancer Imaging Archive
  - The data are organized as “Collections”, typically patients related by a common disease (e.g. lung cancer), image modality (MRI, CT, etc) or research focus.

- Dataset: Covers a large breadth of conditions; the largest datasets within the collection are:
  - CBIS DDSM for Breast Cancer: 6,720 subjects and 6,775 images
  - Prostatex for Prostaste Cancer: 346 subjects and 18,321 images

It would be free to use the public datasets, but it appears that there is a large amount of PET scans mostly for cancer in the public domain and less for lifestyle conditions like smoking or obesity.

Learning Model

The model would be used to learn the progression of metabolic activity *from* current state *to* future state *if* an intervention takes place *and* is followed for a specified amount of time. I imagine this to assist in health goal maintenance and intervention plan adherence. This is assuming the model accurately predicts and generalizes as detailed below.

The input used would be PET scans already separated into the following features: 1) de-identified patient ID, 2) the body part scanned, 3) the condition and 4) the position the image was taken in its time-series. Added features would be: 1) position of each colored cluster i.e. red, yellow, blue, green, 2) size of each color cluster.

Image processing would be needed to transform input images, perhaps using computer vision, to identify the size and position of each colored cluster. The labelling would be the intervention and stage of the intervention.

I would try all models available first (linear, stochastic gradient descent, Bayesian, ridge, perceptron) to test results, although I am initially more inclined towards the Bayesian. The Bayesian seems to align with the theoretical concept of determining the probability that certain clusters will be in certain places at a certain proportion *if* an intervention happens *given* current state of clusters.

Given the data available, I may resort to partitioning one fully-labelled dataset into training and test sets. The generalizability of the model for a specific condition i.e. smoking might be tested in randomized clinical trials for a small number of patients from before an intervention to the end of an intervention, with each clinical trial representing a different study or type of intervention.

My task is specific to helping participants set lifestyle goals i.e. losing weight, reducing smoking by seeing a future state of their brain or body next to normal levels. To these ends, an acceptable rate of accuracy is about 80 or 90% and fairly generalizable. However, if this was applied to decide on interventions for serious conditions like cancer, the error rate should be as close to 0% as possible. Most likely in the latter instance, a model would have to be created for each specific type of disease and generalizes well within certain patient characteristics, like history of disease, socioeconomics, demographics.