Team Microsoft

Kaggle Machine Learning Competition

A Project to Learn Machine Learning

What We Know Now

- Project will be in C/C++
    - Possibly Cuda and OpenMPI
- Create an algorithm to help in early detection of lung cancer
- Dataset of thousands of high-resolution lung scans
    - DICOM file format
- ANN(Artificial Neural Network) most likely to be most accurate.
- Output is simply a csv file with patient name(hex string) and probability prediction
- Each image contains a series with multiple axial slices of the chest cavity. Each image has a variable number of 2D slices, which can vary based on the machine taking the scan and patient.

Additional Libraries

- Imebra for DICOM
  - This will be used to transform Dicom formatted files into more readable formats for image and header data
- Tensor Flow
  - An open source library for machine learning
- Thrust
  - From nVidia for Cuda acceleration
- OpenMPI
  - The defacto standard for parallel processing on supercomputers
  - Can be used on any size cluster of networked computers

Development environments

- Specific IDE’s / setup is unimportant, down to user preference as long as IDE specific files aren’t pushed to repo
- G++ or similar form of compilation required for C/C++ code
- Libraries must be individually compiled on each users machine
- Linux

Algorithm Selection

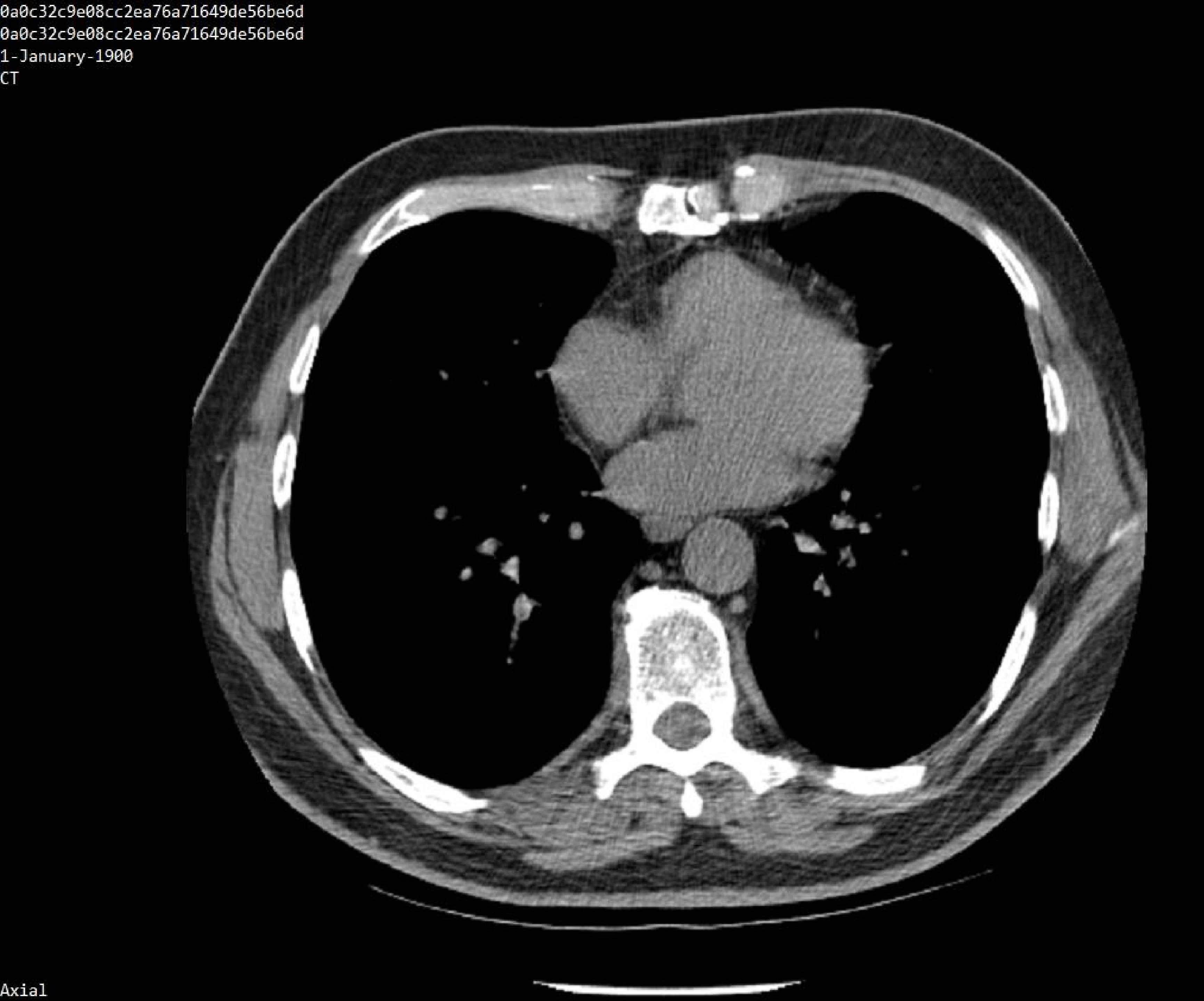
- A matter of experiment - trial and error
- Part of the learning process
- Goal is to match algorithm output with results from training set
- ANNs can be over-trained
- Anticipate most time spent working on fine-tuning the algorithm

May or may not need a simple GUI for testing and progress indicators. We can determine later whether that will be worth the time or not.

Division of labor will become more evident as our algorithm evolves.

Repo Usage

- Can be used to post individual files of weights resulting in training back-propagation neural networks.
- C++ classes can be defined in separate headers and therefore can be separated in the repo
- Result csv files from training can be posted
- Don’t keep dataset (>66 GB) in repo
  - Figure out where the best place to have a common store of the data would be
    - FTP on personal computers?
    - Azure, EOS?
- Further use of the repo will become evident as the project evolves


0a0c32c9e08cc2ea76a71649de56be6d
0a0c32c9e08cc2ea76a71649de56be6d
1-January-1900
CT
Axial