Medical Image Mining in the Era of Big Data and Deep Learning

Lawrence O. Hall

Department of Computer Science & Engineering

University of South Florida

Tampa, FL. 33620-9951

lohall@mail.usf.edu

Collaborators

Dmitry Goldgof

Bob Gillies

Mu Zhou

Henry Krewer

Steven Eschrich

Baishali Chadhury

Hailing Zhou

Bob Gatenby, M.D.

Yuhua Gu

Sam Hawkins

Ben Geiger

Yoga Balagurunathan

Matt Schabath

Renhao Lu

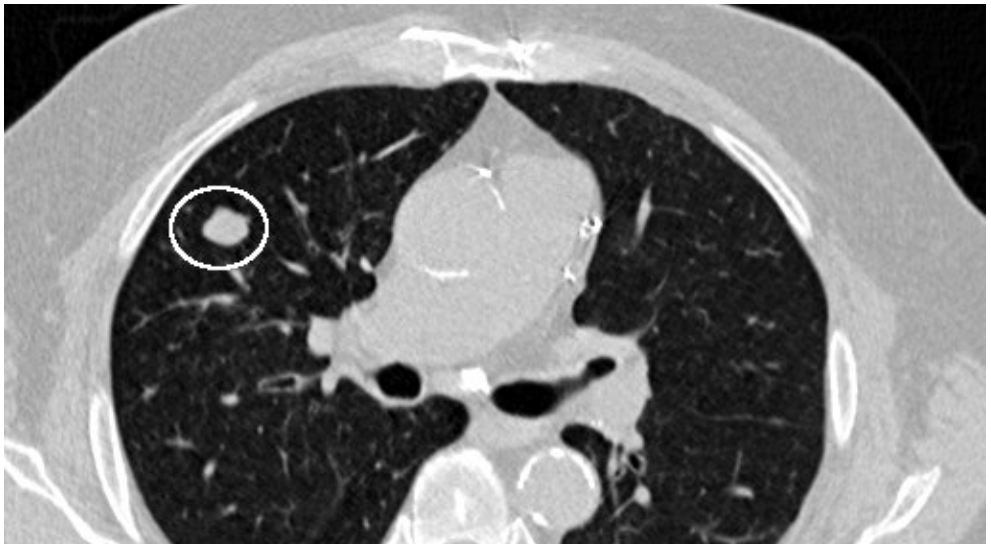
Imaging -> Personalized Treatments

- Images of the human body are non-invasive.
- Images, such as Magnetic Resonance (MR) images, may have no side effects.
- Low dose Computed Tomography (CT) has minimal effect.

Imaging -> Personalized Treatments

- If we can determine prognosis and/or treatment effects, patient-specific decisions may be made.
- Images may be used for screening so a disease can be found early and cured.
- Images can be taken at multiple time points in the treatment process.

National Lung Cancer Screening Trial (NLST) Data



Goal: Find

Benign Nodule → **Malignant Nodule**

Radiomics

- The science of using a rigorous mathematically integrated approach to evaluating radiological images.
- The problem requires computerized approaches that can analyze images and use clinical data, genetic data and more.
- A challenge, today, is that there is little image standardization in the clinic, in most cases.

Radiomics Questions

- How much can we tell, today, from just images?
- What features are useful?
  - It is easy to create more texture features than examples.
- Is there a way forward?

Radiomics Data

- For all image analysis and data mining it is critical to deeply know your data.

- Medical image data will come from different scanners (GE, Siemens, Phillips, etc.) and use different acquisition settings depending on the local M.D.'s.

Data Caveats

- We start with a “large” set of lung cancer patient images with genomic information (>500).
- Then the number with genome information is about 250.
- Of the 250 about ½ are pre-treatment and others unknown.

Data Caveats

- **Scans are repeated temporally. You want the scan from time of diagnosis. You get a (semi) random scan.**
- **The type of tumor varies and IF subtype matters it will reduce your numbers.**
- **The stage of the tumor matters. We found in one case that one type of tumor had a shorter survival time. However, it was known to have a better prognosis. Why? Later stage.**

Data Caveats

- **Treatment, of course, matters. It is in the records IF treated in the same medical center.**
- **Must stratify by treatment or, ideally, look at before treatment.**
- **No news here to those in the field, but with more medical data coming, perhaps this is helpful.**

Radiomics Examples

We will discuss:

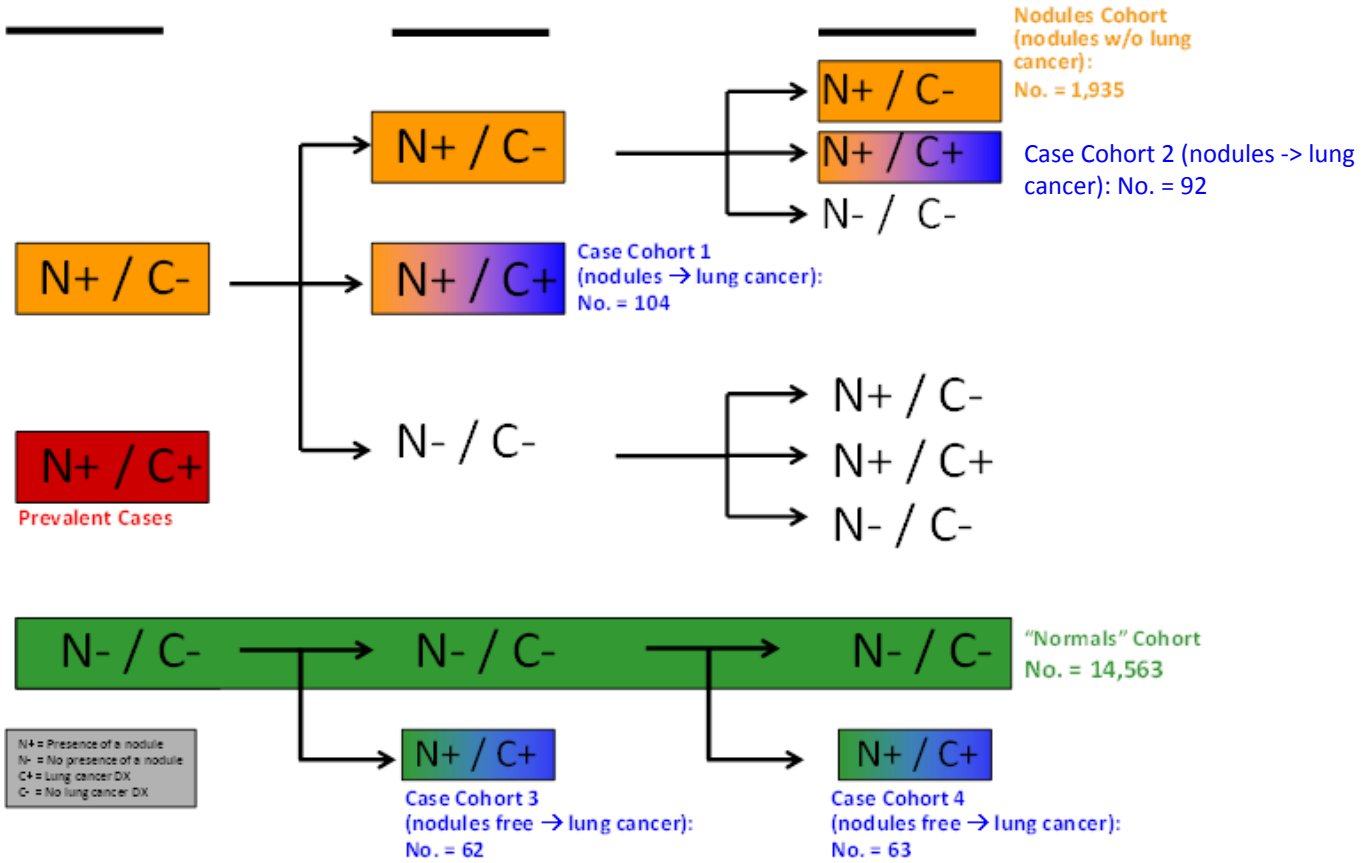
- **low dose CT imaging of the lungs for screening.**
- **MRI images of the brain for GBM prognosis.**

Lung Cancer Screening

- The National Lung Screening Trial (NLST) used low dose CT and focused on current/former smokers.
- The goal was to find nodules, especially suspicious ones that could be malignant compared to lung Xrays.
- It was so successful it was stopped early.

Data Curation Challenge

NLST dataset: **Large (~50,000 3D CT Scans, 54,000 participants), located on TCIA website, download time, all nodules not labeled. So what is Ground Truth?**

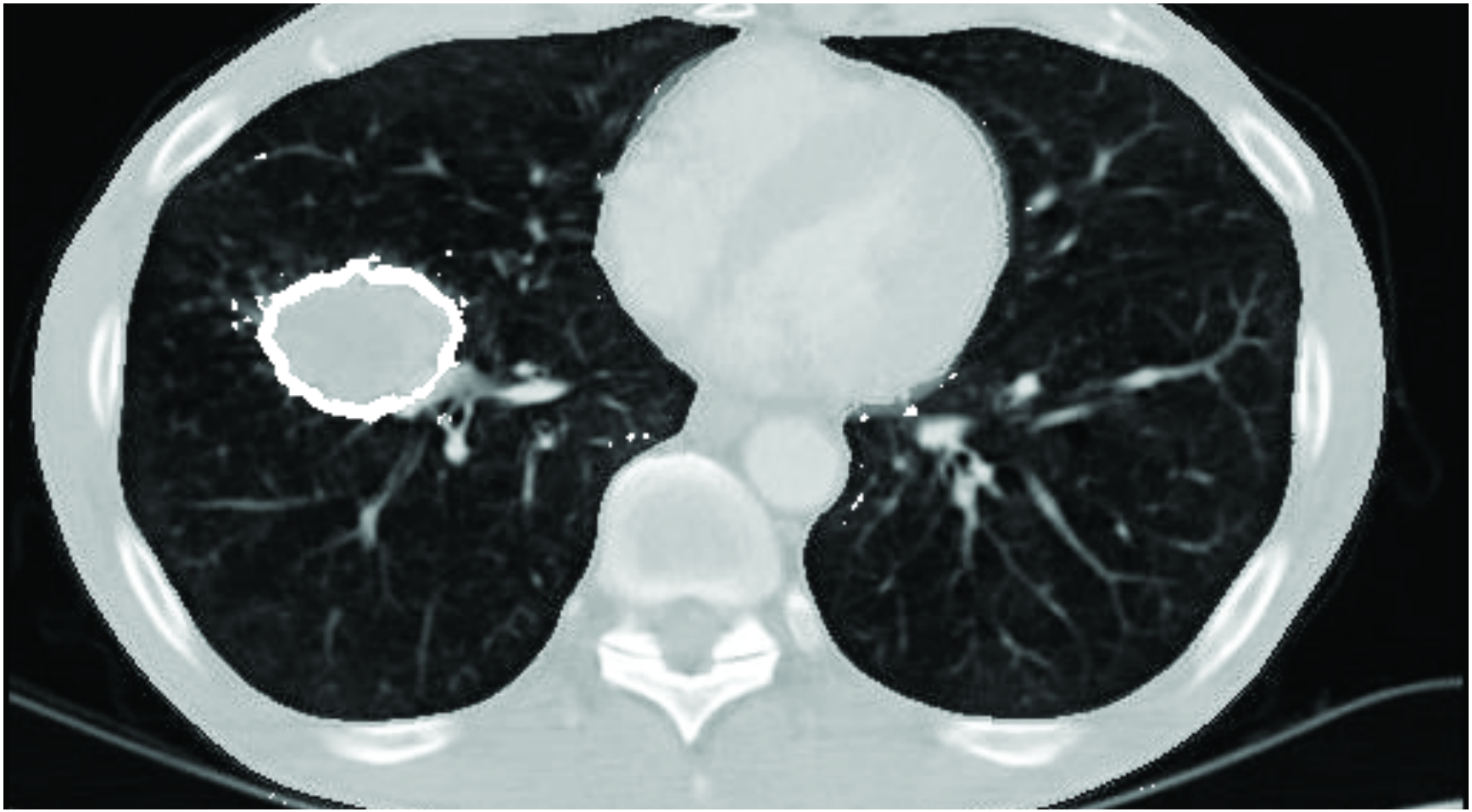


Lung Cancer Screening

- Could the prediction of nodules which are not yet a problem, but will become cancer be done with an automated imaging system?

- Radiomics is designed to a) answer the question and b) make it so.

Segmented Lung Nodule (Large)



Lung Tumor Segmentation

- Segmentation, in our work, is done with a semi-automated region growing approach.
- An expert (e.g. radiologist) clicks on the tumor.
- Our algorithm grows a region and then chooses different seed points to regrow the region.

Lung Tumor Segmentation

- Finally, an ensemble segmentation is created from the most likely intersection of the 20 segmentations.

- It is possible to tune criteria to make the region boundaries looser or tighter.

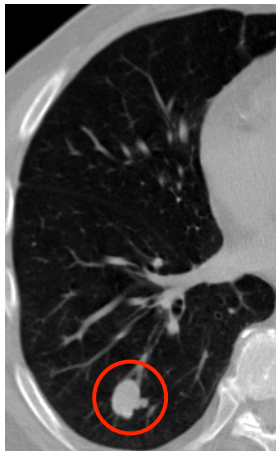
Low Dose CT Processing

- There were 47 shape features and 172 texture features extracted.
- Feature selection was done in several ways, with Correlation based Feature Subset (CFS) Selection and Relief-F (with a correlation test).
- It was also done on other data sets, just looking for stability.

Features

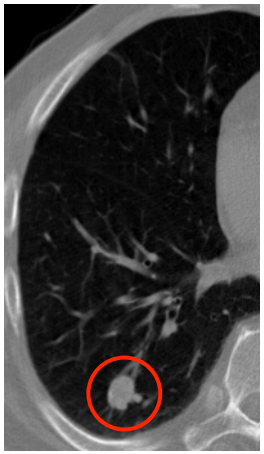
- Shape features have been the focus for describing lung tumors in the past
- The top shape features from CFS were: Longest Diameter, 3D Relative Border To Lung, 3D MIN Dist COG To Border, Elliptic Fit
- Texture features allowed for improved performance – 3D Laws L5 L5 E5 Layer 1, 3D wavelets

Example low dose nodules



Radiomics Report at Time 0:

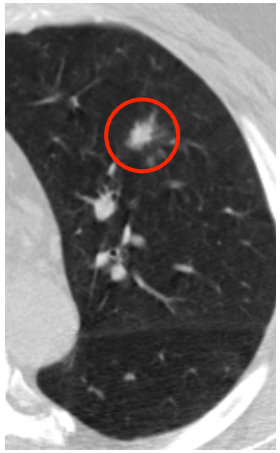
- **Volume [cm³] = 2.62**
- **Relative Volume Air Spaces = 0.074985354**
- **Mean [HU] = -92.52**



Radiomics Report at Time 1:

- **Volume [cm³] = 2.1**
- **Relative Volume Air Spaces = 0.013074205**
- **Mean [HU] = -59.28**

Benign



Radiomics Report at Time 0:

- **Volume [cm³] = 2.14**
- **Relative Volume Air Spaces = 0.341826923**
- **Mean [HU] = -307.63**

Radiomics Report at Time 1:

- **Volume [cm³] = 2.13**
- **Relative Volume Air Spaces = 0.052502453**
- **Mean [HU] = -161.48**

Malignant

Time 1 **Time 2**

NLST Data

- **Cohort 1 consists of 104 (87 viable) subjects who had an identified nodule at T0 that was benign.**
- **The same nodule was diagnosed as Cancer at the screening 1 year later.**
- **Also in the cohort are 176 subjects who had benign pulmonary nodules for both scans with appropriate age and history matching.**

NLST Data

- **Cohort 2 consists of 92 (88 viable) subjects who had an identified nodule at T0 that was benign.**
- **The same nodule was diagnosed as Cancer at the screening *2* years later.**
- **Also in the cohort are 153 subjects who had benign pulmonary nodules for both scans with appropriate age and history matching.**

NLST Data Question

- **From the first scan can we predict the nodules which will become cancerous?**

- **Can we provide a useful individual risk (probability)?**

NLST Feature selection

- **Used Rider data set (no contrast) "coffee break" scans to find stable features.**

- **Features from a manual and ensemble segmentation of the tumor that**
  - **Were quite uncorrelated (CCC > 0.95)**
  - **From both segmentations**

NLST Feature selection

- **Looked for stable features for Cohort 2 (nodules but no cancer for two scans).**

- **Features from a manual and ensemble segmentation of the tumor that**
  - **Were quite uncorrelated (CCC > 0.95)**
  - **From both segmentations**
  - **There were 38 features.**

Attributes:38

- Longest Diameter [mm]
- Short Axis * Longest Diameter [mm_]
- Short Axis [mm]
- Mean [HU]
- Volume [cm_]
- 8a_3D_Is_Attached_To_Pleural_Wall
- 8b_3D_Relative_Border_To_Lung
- 8c_3D_Relative_Border_To_PleuralWall
- 9c_3D_Compactness
- 9d_3D_AV_Dist_COG_To_Border_[mm]
- 9e_3D_SD_Dist_COG_To_Border_[mm]
- 9f_3D_MIN_Dist_COG_To_Border_[mm]
- 9g_3D_MAX_Dist_COG_To_Border_[mm]
- 10a_3D_Relative_Volume_AirSpaces
- 10b_3D_Number_AirSpaces
- 10c_3D_Av_Volume_AirSpaces_[mm_]
- Compactness
- Shape index
- Area (Pxl)
- Volume (Pxl)
- Number of pixels
- Width (Pxl)
- Thickness (Pxl)
- Length (Pxl)
- Border length (Pxl)
- avgGLN
- avgHGRE
- avgLRHGE
- avgRLN
- avgRP
- avgSRHGE
- 3D Laws features L5 L5 L5 Layer 1
- Histogram ENTROPY Layer 1
- Histogram SKEW Layer 1
- 3D Wavelet decomposition. P2 L2 C13 Layer 1
- 3D Wavelet decomposition. P2 L2 C14 Layer 1
- 3D Wavelet decomposition. P2 L2 C15 Layer 1

Experiments

- We need unseen test data to estimate the accuracy of an learned model.

- To get a general idea, we break the data into 10 disjoint subsets and train on 9 and test on the one left out (about 31 examples in test set).

NLST Results – Cohort 1

- **Average of 10, 10 fold cross validations**
- **Best Accuracy and AUC is Random Forests with RIDER stable features – 80.1% accuracy and AUC of 0.83**
- **Classifier Naïve Bayes with RIDER stable features and CFS 10 – 79.47% accuracy and AUC of 0.79**
- **Just volume gets you 75.56% accuracy with J48.**

NLST Results – Cohort 2

- **We trained on all of Cohort 1 and tested on Cohort 2.**

- **Best Accuracy and AUC is SVM-RBF and Random Forests with NLST stable features (CFS 10) and RIDER stable features (RF 10) – 74.68% accuracy and AUC of 0.72 respectively**

Risk Score Prediction

- **A recent article in the New England Journal of Medicine shows an AUC over 0.9 for a probabilistic risk score.**
- **Risk score is created by logistic regression.**
- **They use: Age, Sex, Family history of lung cancer, Emphysema, Nodule size, Nodule type - Nonsolid or with ground-glass opacity, Part-solid, Solid Nodule location, upper vs. middle or lower lobe, Nodule count per scan, Spiculation - yes vs. no**

Risk Score Prediction

- **We replicated their model on a subset of Cohort 2 for which a radiologist provided some features.**

- **For now, just accuracy: 78.9% vs. 80% for Random Forests 10 features chosen by Relief-F from the RIDER set.**

- **We will add clinical features to our model.**

Radiomics for Glioblastome Multiforme

- Brain tumors of the GBM type are deadly quite quickly

- We used 4 modalities to predict prognosis of survival <, >= 1yr.

- The modalities are T1, T2, Flair and ADM. All post contrast (gadolinium).

Brain Tumor Data Set

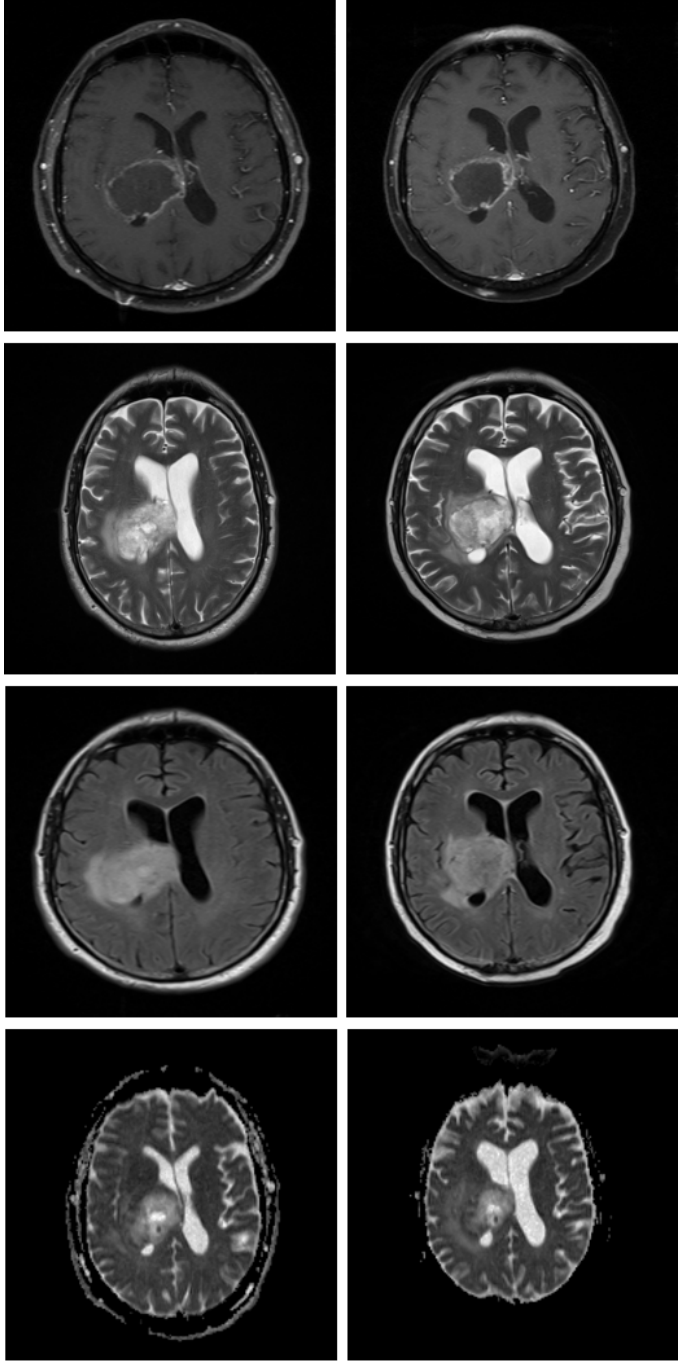
- There were 22 deidentified cases from the Moffitt Cancer Institute with images pre and post-treatment
- No resection, only radiation an chemotherapy
- A radiologist verified the tumor type and extent (manually approximately segmented the tumor bed)
- 2 classes, 12 short term cases and 10 long term cases, by choosing 12 months of survival time as division criterion.

Sample Images: T1

T2

Flair

ADM



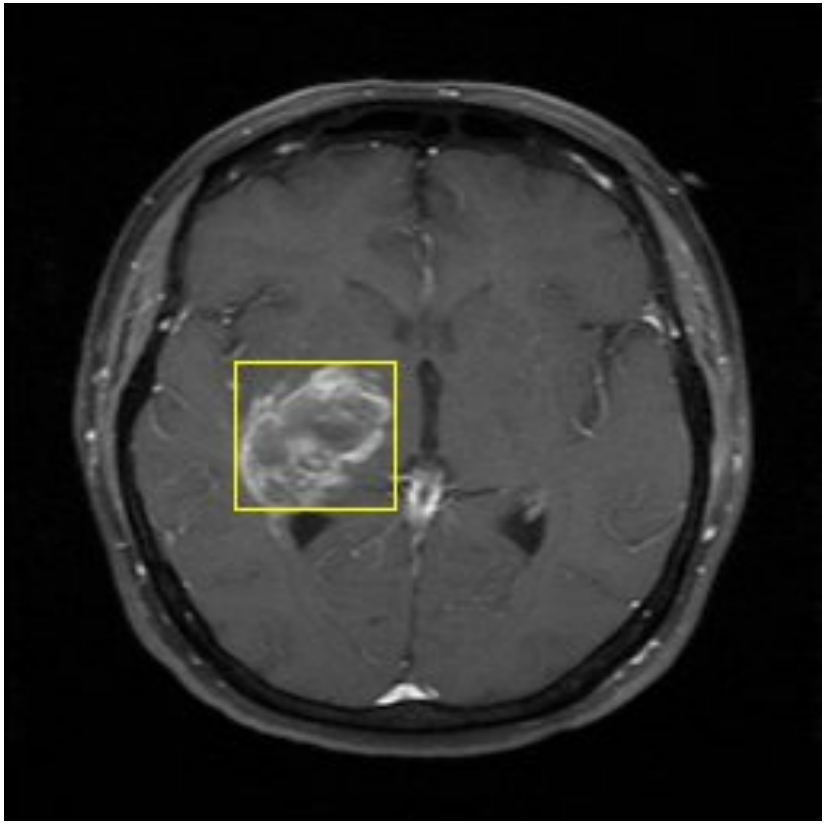
(a) Pre-treatment (b) Post-treatment

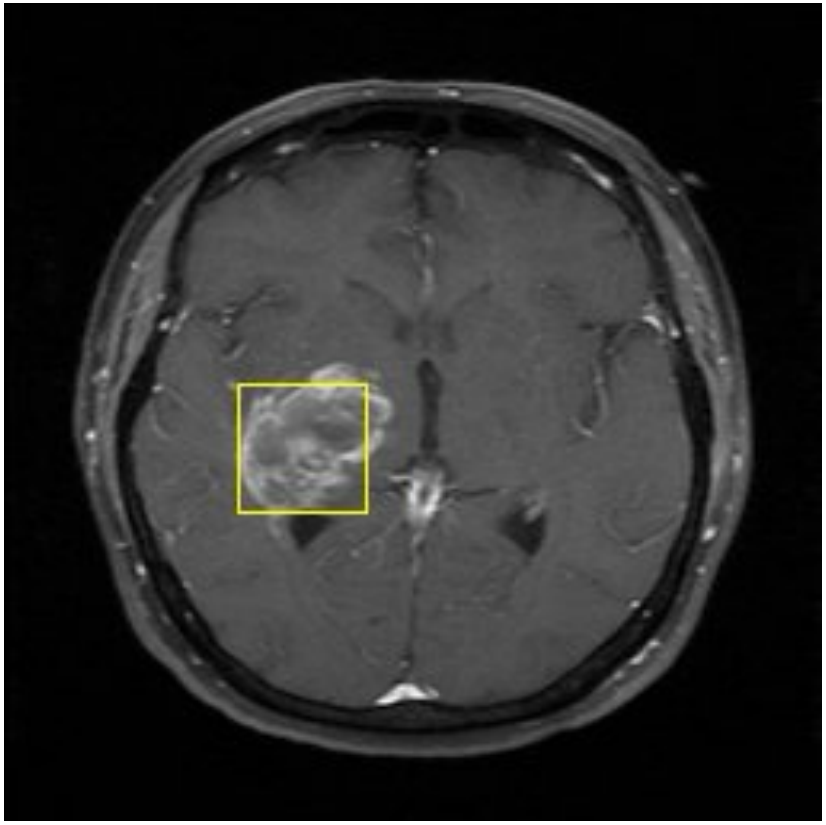
Features

- Leverage big data by using a deep neural network (CNN) trained on ImageNet
- Convert either whole tumor (warp) or 40x40 box over the tumor (cropped) to 224x224 input to pre-trained net ("CNN-F").
- Take 4096 outputs of 7^{th} layer (of 8) as features

Patch Examples

- **An example of a “Warped” Tumor Patch(left) and a “Cropped” Tumor Patch (right).**





Features

- The features were ranked by individual classification ability using a nearest neighbor approach
- Then used in a nearest class mean classifier
- We also extracted features using symmetric uncertainty

Experiments

- **We did leave-1-out cross validation with this small data set.**
- **We compared with single feature ranking and symmetric uncertainty to select features**
- **We used nearest class mean classification, nearest neighbor, J48 and Random Forests**

Conventional Feature Extraction Results

Input Pair Modalities	Accuracies on Different Modalities		
	Pre-treatment histogram features	Post-treatment histogram features	Difference histogram features
T1-weighted and FLAIR	77.27%	81.82%	77.27%
T1-weighted and T2-weighted	72.73%	77.27%	**90.91%**
FLAIR and T2-weighted	68.18%	81.82%	77.27%
T1-weighted and ADM	72.73%	77.27%	77.27%
FLAIR and ADM	72.73%	86.35%	81.82%
T2-weighted and ADM	72.13%	81.82%	77.27%

Crop Results – 1 Feature

Input Modality	Deep Feature Accuracies		
	Pre-treatment deep features	Post-treatment deep features	Difference deep features
T1-weighted	27.27%	22.73%	50%
T2-weighted	50%	**86.36%**	31.82%
FLAIR	77.27%	18.18%	45.45%
ADM	**86.36%**	18.18%	63.64%

Crop Results – 2 and 3 Features

Input Modality	Deep Feature Accuracies with nearest class mean classifier and single feature ranking feature selector	
	2 Pre-treatment deep features	3- Pre-treatment deep features
FLAIR	95.45%	100%

3 Feature Set

- **100% accuracy using nearest neighbor**
- **86.36% accuracy with Random Forests**
- **59.09% accuracy with J48**
- **Pretty variable…**

Random_Forest Warp Result

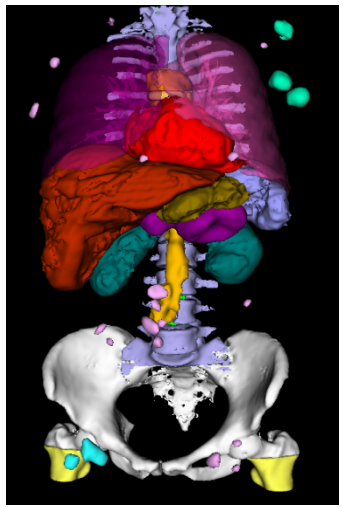
Input Modality	Pre-treatment deep features	Post-treatment deep features
T1-weighted	31.82%	9.09%
T2-weighted	**90.91%**	59.09%
FLAIR	22.73%	31.82%
ADM	72.73%	40.91%

Reading rooms of the future

annotate

AIM and Metadata

segment



extract

Radiomics DB

NBIA

Mining & decision support

Robert Gillies

Discussion

- **Features extracted from CT and MRI images can provide prognostic information.**

- **They can be used with clinical and genomic information (when available) to provide individualized prognoses with, potentially, multiple image time points.**

Summary

- Texture features (and others) can be used to build effective classifiers from medical images (CT and MRI).
- Deep Neural Networks may also provide features
- Tuning the CNN's is a next step, along with looking at other layers and channels

Summary

- Radiomics shows promise even with data variability
- Standardization of imaging, without a clinical effect, can allow for improved models
- Clean data collection is critical!

Thank you!