Detecting Thoracic Diseases from Chest X-Ray Images

Binit Topiwala, Mariam Alawadi, Hari Prasad {topbinit, malawadi, hprasad}@stanford.edu

CS 229
Fall 2017

Introduction and motivation

- We developed classifiers to find any potential lung diseases from chest X-ray
- Built **independent binary classifiers** for each of the lung diseases
- Trained Logistic Regression and SVM models on top of **SIFT** and **HoG** features extracted from X-ray image

Dataset

- Published by National Institutes of Health (NIH) Clinical Center
- 100,000+ frontal-view X-ray images
- 32,717 unique patients, 14 lung diseases
- Each image has multi-label
- Images are gray scale of size 1024 x 1024

Features

- **Scaled** 1024 x 1024 image to 224 x 224
- **Cropped** image to make lungs a focal point, resulting image size 180 x 200
- Applied histogram equalizer to **increase contrast** of the image
- Derived **SIFT** and **HoG** features. Both techniques helps in finding robust key-points in the image

Models

[1] http://scikit-learn.org/
[2] https://kushalvyas.github.io/BOV.html
[3] https://docs.opencv.org/3.1.0/da/df5/tutorial_py_sift_intro.html

Results

SIFT

Logistic Regression

Train

	f1	precision	recall
Cardiomegaly	0.66	0.66	0.66
Edema	0.75	0.73	0.76
Emphysema	0.54	0.65	0.47
Hernia	0.77	0.87	0.69
Pneumonia	0.58	0.63	0.54
Fibrosis	0.59	0.62	0.56
Pneumothorax	0.63	0.64	0.63

Test

	f1	precision	recall
Cardiomegaly	0.59	0.6	0.58
Edema	0.74	0.78	0.7
Emphysema	0.51	0.62	0.43
Hernia	0.5	0.56	0.45
Pneumonia	0.55	0.64	0.49
Fibrosis	0.25	0.17	0.48
Pneumothorax	0.65	0.67	0.63

SVM

Train

	f1	precision	recall
Cardiomegaly	0.91	0.92	0.9
Edema	0.9	0.89	0.91
Emphysema	0.88	0.89	0.88
Hernia	0.85	0.88	0.83
Pneumonia	0.61	0.66	0.56
Fibrosis	0.66	0.63	0.69
Pneumothorax	1.0	1.0	1.0

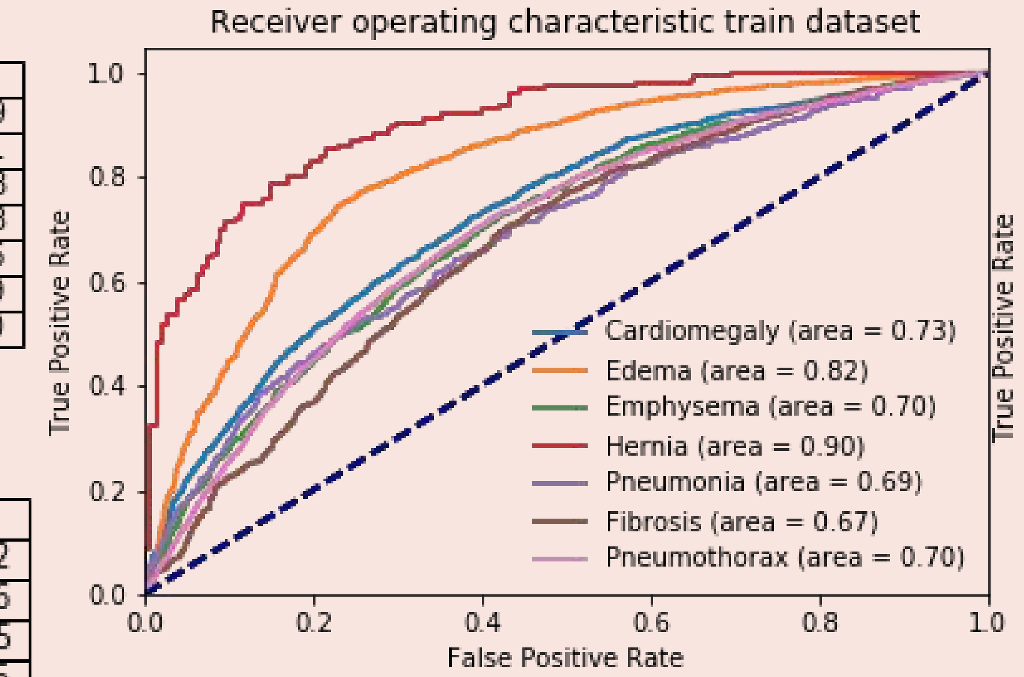
Test

	f1	precision	recall
Cardiomegaly	0.6	0.59	0.62
Edema	0.76	0.76	0.76
Emphysema	0.57	0.65	0.5
Hernia	0.59	0.65	0.55
Pneumonia	0.52	0.62	0.45
Fibrosis	0.22	0.14	0.48
Pneumothorax	0.61	0.65	0.58

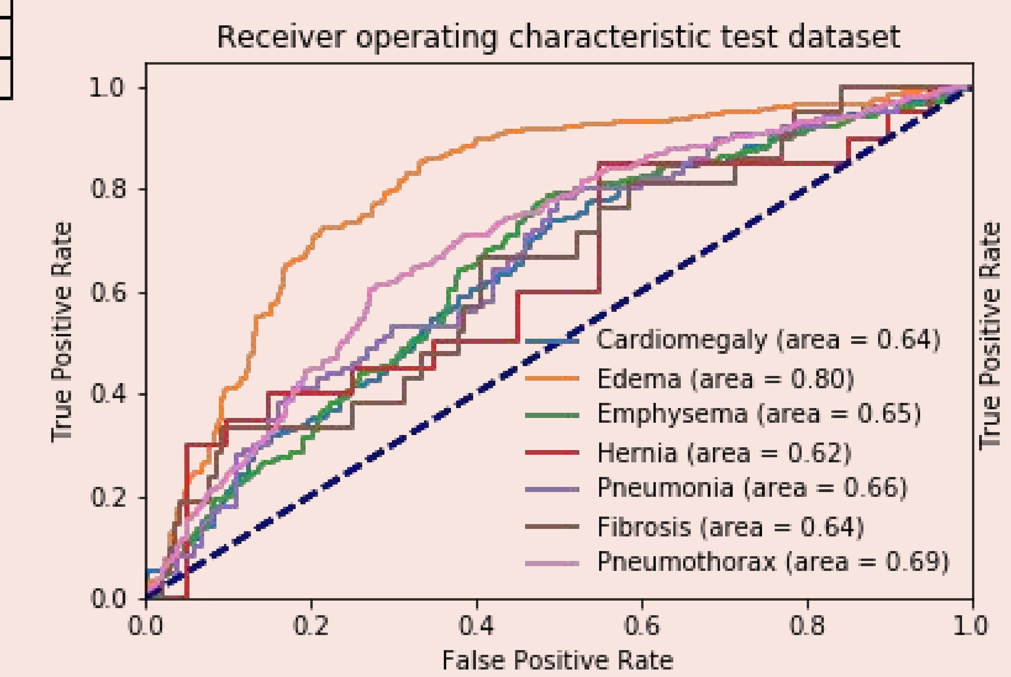
	Train Dataset size
Cardiomegaly	4528
Edema	3413
Emphysema	3687
Hernia	345
Pneumonia	1919
Fibrosis	2475
Pneumothorax	8139

Logistic Regression

Train

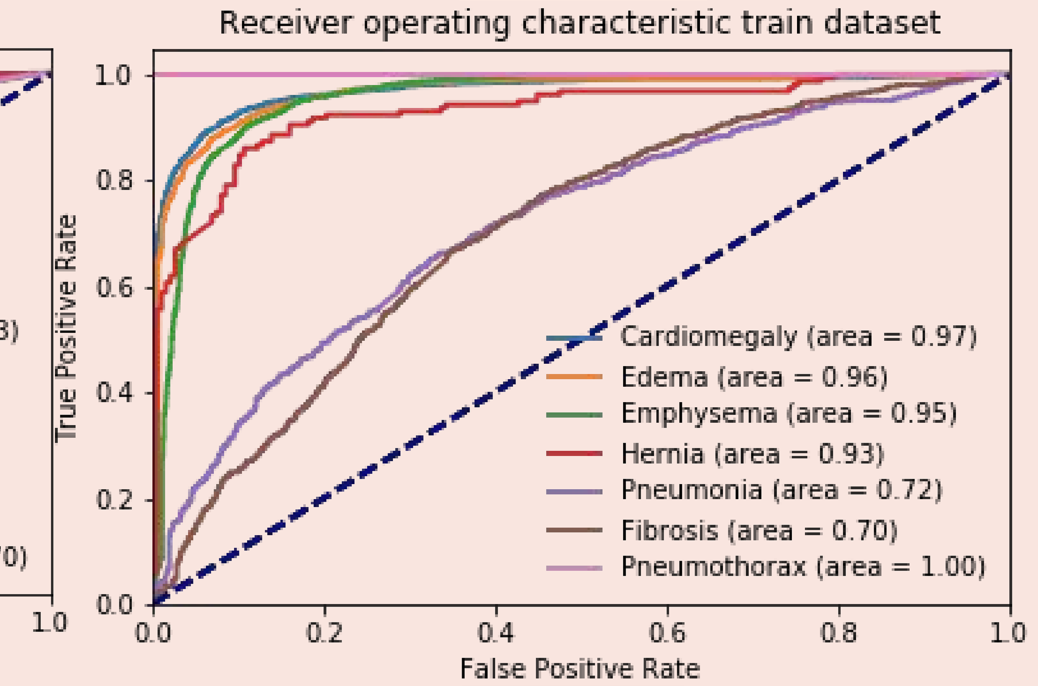


Test

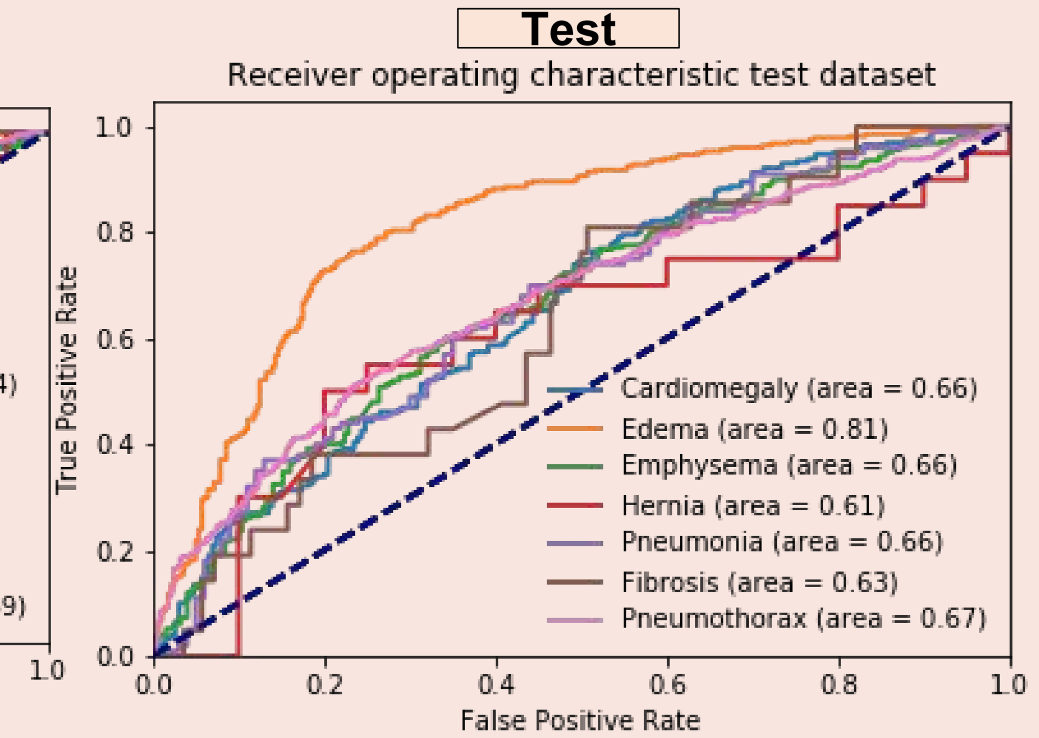


SVM

Train



Test



Models

- **Logistic Regression**
  - Cost function:

$$J(\theta) = \frac{1}{2}\sum_{i=1}^{m}\left(\frac{1}{1+e^{-\theta^T x}} - y^{(i)}\right)^2$$

- **SVM with Radial basis function**
  - Cost function:

$$J_\lambda(\alpha) = \frac{1}{m}\sum_{i=1}^{m} L(K^{(i)T}\alpha, y^{(i)}) + \frac{\lambda}{2}\alpha^T K\alpha$$

$$L(z, y) = \max\{0, 1 - yz\}$$

$$K(x, z) = \exp\left(-\frac{1}{2\tau^2}||x - z||_2^2\right)$$

Discussion

- Logistic regression seems to be generalizing much better than SVM(Kernel) for SIFT. SVM is overfitting training data
- SIFT was able to detect various descriptors on the lung in the image, that proved useful as a features

Future

- Pre-trained CNN models for feature extraction
- Min-hashing to generate feature
- HoG with PCA
- Pre-processing to make lungs more focal