Detecting Thoracic Diseases from Chest X-Ray Images

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

Abstract— Radiologists have to spend time diagnosing these chest X-ray images to find any potential lung diseases. Diagnosing X-ray require careful observation and knowledge of anatomical principles, physiology, and pathology. In this work, we applied traditional machine learning techniques for automated detection of thoracic diseases from chest X-ray images. We constructed visual bag of words using extracted SIFT descriptors. Visual bag of words is used as features for Logistic regression and SVM. Seven diseases were selected—Cardiomegaly, Edema, Emphysema, Hernia, Pneumonia, Fibrosis, Pneumothorax. We build independent binary classifier for each of these lung diseases.

INTRODUCTION

Examining chest X-ray is one of the most frequent and cost effective medical imaging examination. Radiologists have to spend time diagnosing chest X-ray images to find any potential lung diseases. Diagnosing x-rays require careful observation and knowledge of anatomical principles, physiology, and pathology. Developing automated system for such could make a huge impact to the patients, who don't have access to expert radiologists.

In our approach, we applied traditional machine learning techniques in building independent binary classifier for each of the diseases. We pre-processed image gray scale image by resizing and cropping. SIFT (Scale-invariant feature transform) computer vision algorithm was applied on pre-processed image to detect feature descriptors in the image. Visual bag of words is constructed from feature descriptors obtained from the images. Computed visual bag of words is used as a feature vector for Logistic regression and SVM. Each model's output is binary label.

RELATED WORK

Recent work [1] done on the dataset involves using state of the art Convolutional Neural Network (CNN) approach that output multi-label for the input X–ray image. However, our work is focused mainly on using traditional feature extraction techniques to produce results; instead of trying to replicate this method.

DATASET AND FEATURES

Dataset has recently released [1], that contains 112, 120 frontal-view X-ray images of 30,805 unique patients, with each image labeled with up to 14 lung diseases. Each image is a gray scale image with 1024 x 1024 in resolution (*Figure 2*).

Data split and pre-processing pipeline

Data pipeline to split data, pre-process it, and feature generation (*Figure 1)*. Each image is pre-processed by scaling from 1024 x 1024 to 224 x 224 to speed up computation. Rescaling followed by cropping to make lungs in the image focal, resulting in image of size 180 x 200. Image contrast is increased by apply histogram equalizer [2].

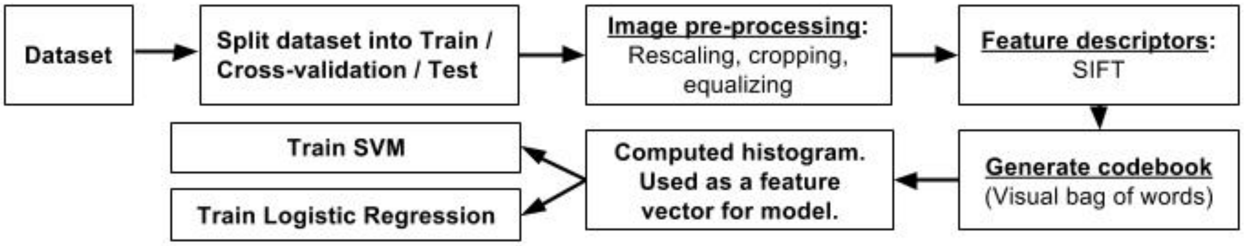

Dataset
Split dataset into Train / Cross-validation / Test
Image pre-processing: Rescaling, cropping, equalizing
Feature descriptors: SIFT
Generate codebook (Visual bag of words)
Computed histogram. Used as a feature vector for model.
Train SVM
Train Logistic Regression

Figure 1. Data preprocessing phases to generate features

Data split and sampling

We split for train / Cross-validation / Test set, by randomly selecting ~20,000 unique patients for train, ~5000 for Cross-validation, and ~5000 for Test set (*Figure 3*). Since, each disease has independent binary classifier; separate dataset is generated for each of the disease classifier. Images are randomly sampled for randomly sampled patients. For each disease classifier, data is balanced with equal number of sample for label-1 and laebl-0. Label-1 contains all the images randomly selected.

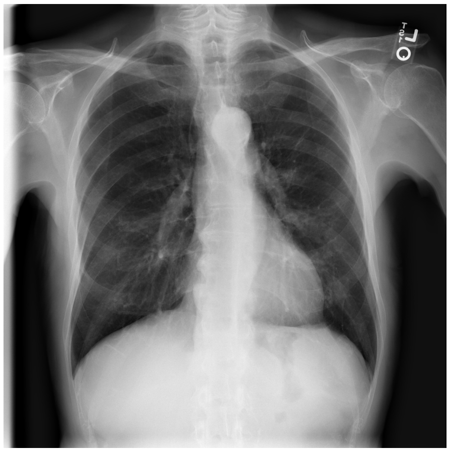


Figure 2. Raw X-ray image in dataset

Feature Extraction

For extracting features, we applied SIFT to capture local information in the image.

SIFT

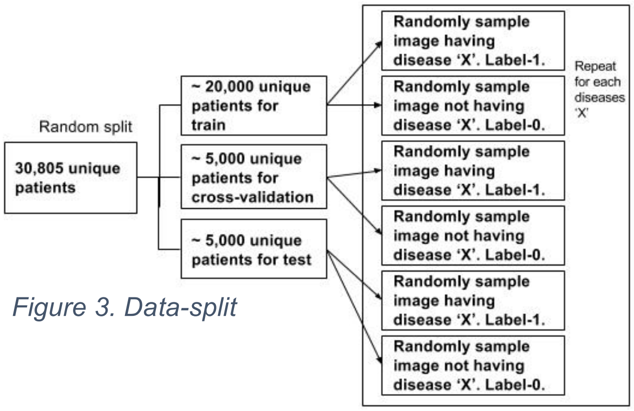

Random split
30,805 unique patients
~ 20,000 unique patients for train
~ 5,000 unique patients for cross-validation
~ 5,000 unique patients for test
Randomly sample image having disease 'X'. Label-1.
Randomly sample image not having disease 'X'. Label-0.
Randomly sample image having disease 'X'. Label-1.
Randomly sample image not having disease 'X'. Label-0.
Randomly sample image having disease 'X'. Label-1.
Randomly sample image not having disease 'X'. Label-0.
Repeat for each diseases 'X'

Figure 3. Data-split

SIFT [6] is an computer vision algorithm used to detect and describe local features in images. SIFT descriptor is invariant to translations, rotations and scaling transformations in the image and robust to moderate perspective transformations and illumination variations. SIFT first finds the keypoints (*Figure 4*) within an image and then compute descriptor vector for each keypoint. Image is convolved with Gaussian filters at different scale, and then the difference of successive Gaussian-blurred images is computed. Keypoints are the maxima/minima of the Difference of Gaussian (DoG) that occurs at multiple scales. Orientation is computed for each keypoints based on local image gradient directions. Using orientation, descriptor vector is computed for each keypoint. Descriptor returns vector of length 128 for each keypoint.

Bag of Visual Words (Codebook)

BoW [5] model constructs a large vocabulary of visual words. For BoW, features are extracted using SIFT, then codebook is generated, followed by histogram. K-means clustering is applied to extracted features from all image to generate codebook. Centroids are defined as a visual codewords. Size of the codebook is number of clusters. Each extracted feature is mapped to one of the closest centroid. Resulting histogram of for each image, which counts the number of features for each of the visual code words. Histogram is used as feature vector for training models.

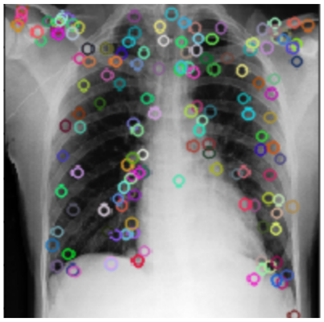


Figure 4.SIFT keypoints

Classification

For classification, we applied Logistic regression and SVM on visual bag of words feature vector.

Logistic Regression

Logistic regression uses hypothesis $h_\theta(x) = g(\theta^T x) = \frac{1}{1+e^{-\theta^T x}}$, where $g(z)$ function g(z) is called the logistic function or the sigmoid function. As $z \to \infty$, $g(z)$ tends towards 1, and as $z \to -\infty$, $g(z)$ tends towards 0. Logistic regression
returns probability of a class being 1. i.e. $p(y = 1\, |x;\, \theta) = h_\theta(x)$. Parameters /
weights (θ) , are updated using update rule $\theta = \theta + \alpha(y^{(i)} - h_\theta(x^{(i)}))x^{(i)}$, where $x^{(i)}$ is the input data and $y^{(i)}$ is the actual class / label.

Regularization parameter is fine tunes using grid search in sklearn [3].

SVM

Support vector classifier:

$$\begin{aligned} & min\frac{1}{2}||w||^2 + C\sum_i \zeta_i \\ \text{s.t. } & y_i(w^T\phi(x_i) + b) \geq 1 - \zeta_i \end{aligned} \qquad K(x,z) = \exp\Big(-\frac{1}{2\tau^2}||x-z||_2^2\Big)$$

Parameter C and τ are fine tunes using parameter grid search in sklearn [3].

EXPERIEMENTS / RESULTS / DISCUSSION

Issue Encountered (Why

Initially, we started with multi-label classifier by training one-vs-rest with data for all the diseases altogether. However, performance of multi-label was very poor. Accuracy on train was 8.14 % and on CV was 7.21% (7-8% accuracy was due to model returning all zero vector, and images with no-finding label had all zero vector. And images with no-finding labels in train & CV were 6-7%). Upon investigating we found that it was due to data imbalance, while training multi-label one-vs-rest classifier. Hence, we had to implemented independent binary classifier for each of the diseases.

Results and Discussion

Metrics used to evaluate models' performance- accuracy, precision, recall, and ROC curve. Number of cluster centroids for each of the classifier is determined using accuracy and recall. With more importance to recall, because of medical domain.

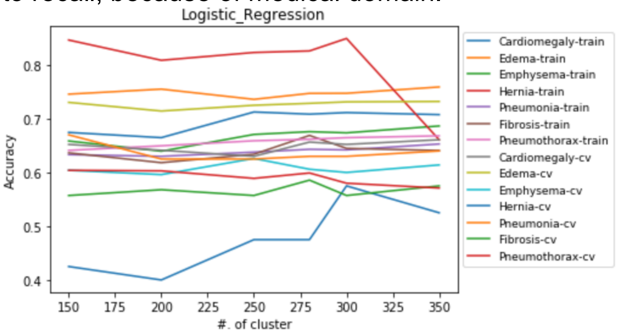


Figure 5. Disease accuracy V/S #. of centroid

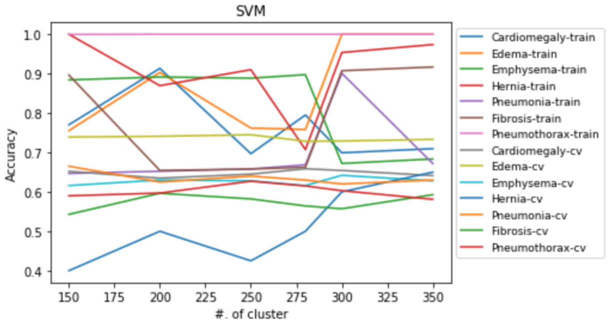


Figure 6. Disease accuracy V/S #. of centroid

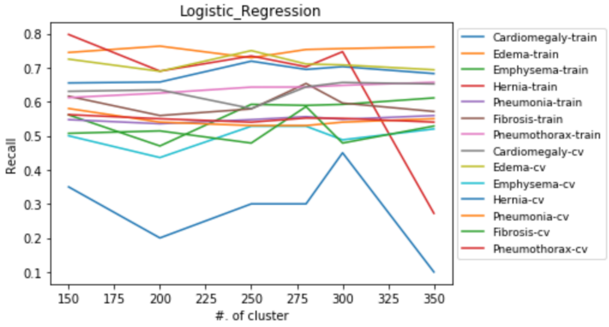


Figure 7. Disease recall V/S #. of centroid

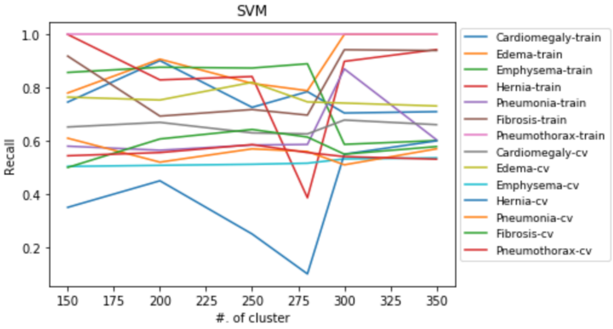
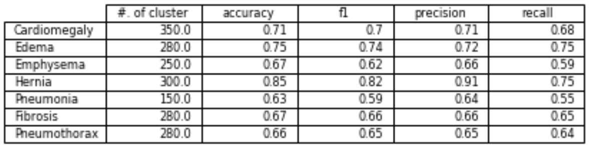
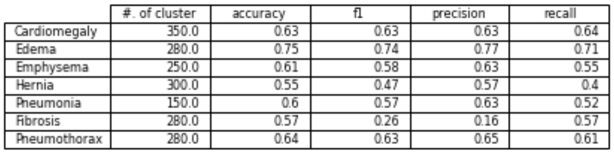


Figure 8. Disease recall V/S #. of centroid



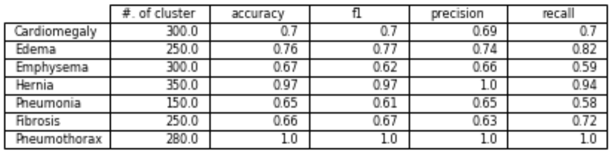
	#. of cluster	accuracy	f1	precision	recall
Cardiomegaly	350.0	0.71	0.7	0.71	0.68
Edema	280.0	0.75	0.74	0.72	0.75
Emphysema	250.0	0.67	0.62	0.66	0.59
Hernia	300.0	0.85	0.82	0.91	0.75
Pneumonia	150.0	0.63	0.59	0.64	0.55
Fibrosis	280.0	0.67	0.66	0.66	0.65
Pneumothorax	280.0	0.66	0.65	0.65	0.64

Figure 9. Logistic regression train



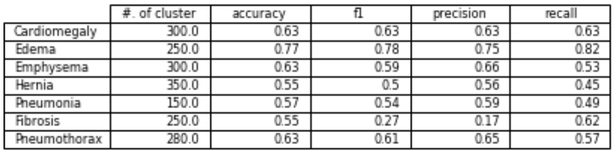
	#. of cluster	accuracy	f1	precision	recall
Cardiomegaly	350.0	0.63	0.63	0.63	0.64
Edema	280.0	0.75	0.74	0.77	0.71
Emphysema	250.0	0.61	0.58	0.63	0.55
Hernia	300.0	0.55	0.47	0.57	0.4
Pneumonia	150.0	0.6	0.57	0.63	0.52
Fibrosis	280.0	0.57	0.26	0.16	0.57
Pneumothorax	280.0	0.64	0.63	0.65	0.61

Figure 10. Logistic regression test



	#. of cluster	accuracy	f1	precision	recall
Cardiomegaly	300.0	0.7	0.7	0.69	0.7
Edema	250.0	0.76	0.77	0.74	0.82
Emphysema	300.0	0.67	0.62	0.66	0.59
Hernia	350.0	0.97	0.97	1.0	0.94
Pneumonia	150.0	0.65	0.61	0.65	0.58
Fibrosis	250.0	0.66	0.67	0.63	0.72
Pneumothorax	280.0	1.0	1.0	1.0	1.0

Figure 11. SVM train



	#. of cluster	accuracy	f1	precision	recall
Cardiomegaly	300.0	0.63	0.63	0.63	0.63
Edema	250.0	0.77	0.78	0.75	0.82
Emphysema	300.0	0.63	0.59	0.66	0.53
Hernia	350.0	0.55	0.5	0.56	0.45
Pneumonia	150.0	0.57	0.54	0.59	0.49
Fibrosis	250.0	0.55	0.27	0.17	0.62
Pneumothorax	280.0	0.63	0.61	0.65	0.57

Figure 12. SVM test

Figure 11 & 12 shows metrics accuracy / f1 / precision / recall for each of the disease for SVM regression model. Metrics are computed for the best centroid size determined from accuracy and recall curve in figure 7 and 8.

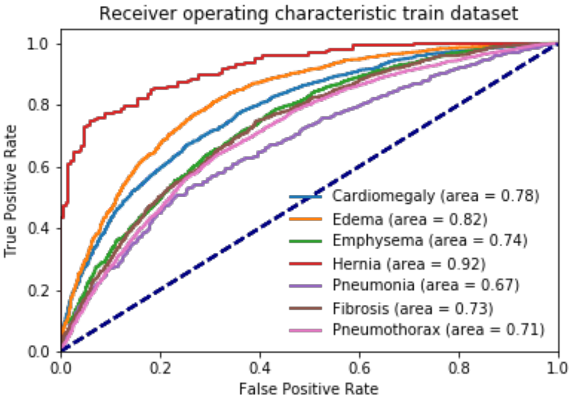


Figure 13. Logistic regression train

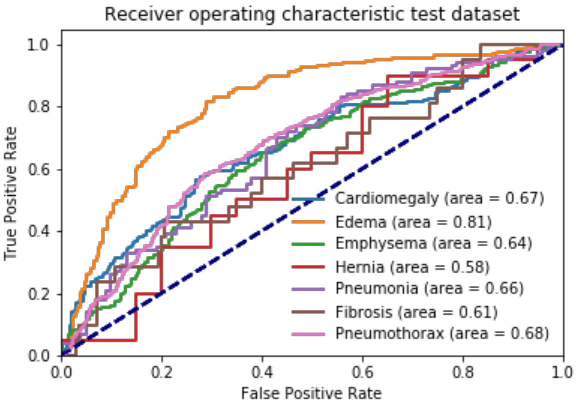


Figure 14. Logistic regression test

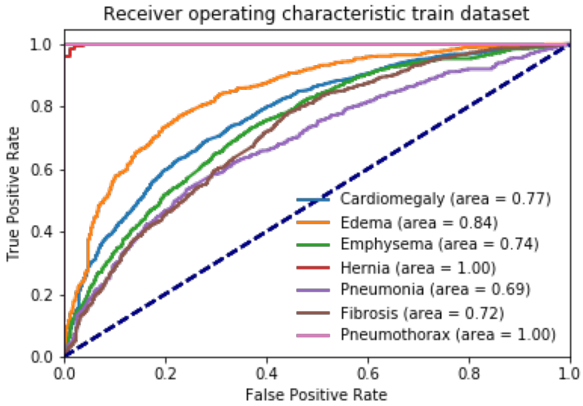


Figure 15. SVM train

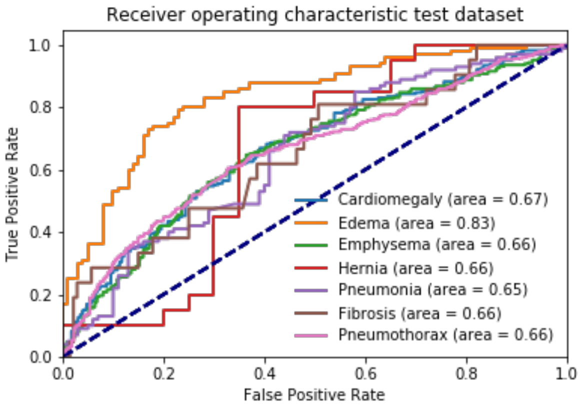
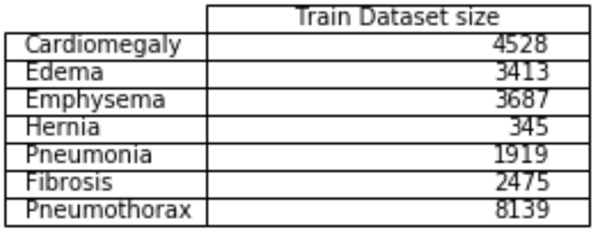


Figure 16. SVM test

Dataset size (*figure 17*) for each of the disease classifier's model. Training data contains balanced count of positive labeled data and negative labeled data.



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

Figure 17. Dataset size per disease model

Both Logistic regression and SVM (*figure 10 & 12)* are performing reasonable similar. With classifiers for disease 'Cardiomegaly', 'Edema', and 'Pneumothorax' performing great for both Logistic regression and SVM. However, SVM tends to overfitting for few disease classifier 'Hernia' and 'Pneumothorax' (*figure 15 & 16)* compared to Logistic regression (*Figure 13 & 14*). Logistic regression is generalizing much better compared to SVM.

Overall results are reasonable (*Figure 10 & 12*) with the understanding of SIFT able to find useful keypoints in the image.

CONCLUSION AND FUTURE WORK

Logistic regression performed similar to SVM. By developing classifiers using traditional machine learning technique of extracting features using computer vision technique, we were able to achieve reasonable performance.

In the future, we hope to improve our results further by applying convolutional neural networks (CNN) that has been proven to perform well for image classification task. Also, try improving pre-processing step to extract lungs from the X-ray images, and then apply SIFT technique on top of it to detect keypoints.

ACKNOWLEDGEMENTS

We would like to thank the project TA Sanyam Mehra for the initial guidance to get started on the project. Also thank the project TA Jeremy Irvin for his feedback throughout the project. Lastly, we thank Prof. Andrew Ng and Dan Boneh for the opportunity and the guidance.

CONTRIBUTIONS
(Project TA '***Lucio Mwinmaarong Dery Jr***' is already aware about it via email by Binit Topiwala)

Binit Topiwala (lead the entire project)
- Pipeline code
    - Knobs to perform operations per knob values (e.g. flag for scale image? etc.)
    - Dataset generation for varied strategy
    - Dataset split strategy
    - Scaling
    - Resizing
    - Image contrast
    - Save / load dataset to file
    - Save / load model to file
    - Logistic regression & SVM binary classifiers
    - Evaluation strategy:
        - Compute confusion matrix
        - Compute f1, precision, recall score
        - Compute & plot RoC
    - SIFT & Visual Bag of words
        - Entire implementation
- Investigation by self
    - Initially we started with multi-label using sklearn One-vs-Rest
    - But, overall accuracy of the model was just 6-7%
    - Spent 2-3 days in investigation
    - Identified the root cause (data imbalance) and tried various remedy. None worked, so had to switch to binary classifier
- Provided almost all the utilities needed by the team
- Poster
- Report

Mariam Alawadi
- HoG feature extraction (stopped contributing to project since Sunday prior to project expo)
- One v/s rest

Hari Prasad
- None

REFERENCES

[1] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, Ronald M. Summers. (2017, July 19). ChestX-ray8: Hospital-scale Chest X-ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases.

[2] Histogram equalization - wikipedia.
https://en.wikipedia.org/wiki/Histogram_equalization

[3] Gridsearch CV sklearn. http://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html

[4] sklearn machine learning algorithm library. http://scikit-learn.org/

[5] Visual bag of words. https://kushalvyas.github.io/BOV.html

[6] Opencv. https://docs.opencv.org/3.1.0/da/df5/tutorial_py_sift_intro.html