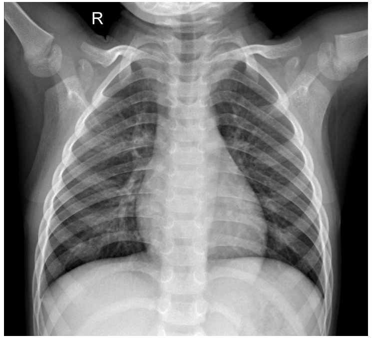
Diagnosis of Respiratory Infections from Chest X-ray Images

Alexandra Drossos, Julia Hossu, Anne Marshall, Hassan Saad

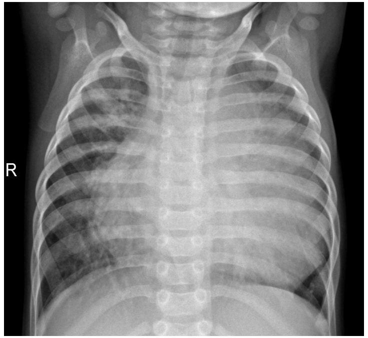
Research Question

How accurately can a machine learning model diagnose the following respiratory infections based on a chest x-ray?



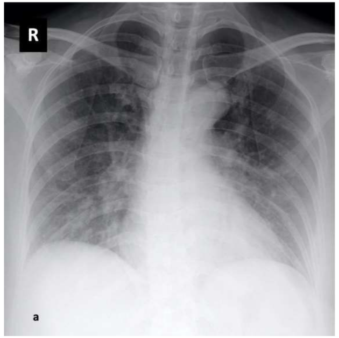
Healthy

vs.



Pneumonia

vs.



COVID-19

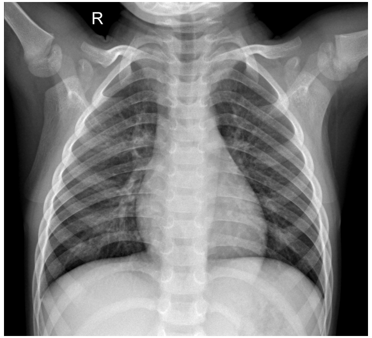
vs.

Tuberculosis

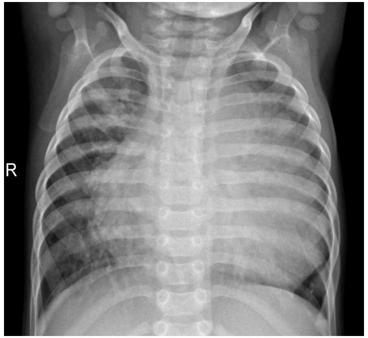
Motivation

Citations: [1] [2] [3] [4]

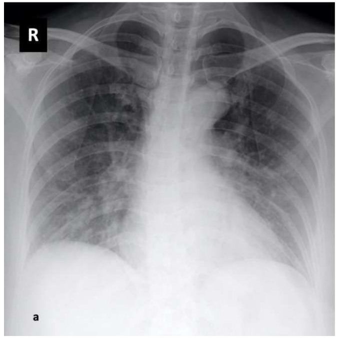
A chest X-ray exam is one of the most frequent and cost-effective medical imaging examinations. However clinical diagnosis of chest X-ray can be challenging.



vs.



vs.



vs.

Healthy

Pneumonia

13.4 deaths per 100,000 people in USA

World's leading cause of death among children under 5 years of age.

COVID-19

240.6 deaths per 100,000 people in USA

CT imaging may help detect disease with high sensitivity in asymptomatic stage

Tuberculosis

0.2 deaths per 100,000 people in USA

Preventable and typically curable disease.

Existing Work

Citations: [5] [6]

CheXNet

121-layer Dense Convolutional Neural Network

	F1 Score (95% CI)
Radiologist 1	0.383 (0.309, 0.453)
Radiologist 2	0.356 (0.282, 0.428)
Radiologist 3	0.365 (0.291, 0.435)
Radiologist 4	0.442 (0.390, 0.492)
Radiologist Avg.	0.387 (0.330, 0.442)
CheXNet	0.435 (0.387, 0.481)

Classifying 14 pathology labels (including pneumonia)

[5] Rajpurkar, Pranav, et al. "Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning." *arXiv preprint arXiv:1711.05225* (2017).
Link

COVID-Classifier

Multi-layer Neural Network

	Precision	Sensitivity	F-score	Support
COVID-19	96%	100%	0.98	25
Normal	89%	100%	0.94	31
Pneumonia	100%	82%	0.91	28

Grouped CXR images into three target classes, each containing 140 images; normal, COVID-19, non-COVID-19 pneumonia

[6] Khuzani, Abolfazl Zargari et al. "COVID-Classifier: An automated machine learning model to assist in the diagnosis of COVID-19 infection in chest x-ray images." *medRxiv : the preprint server for health sciences* 2020.05.09.20096560. 18 May. 2020, doi:10.1101/2020.05.09.20096560. Preprint.
Link

Data

Our model is running on a Kaggle CXR dataset, pulling from 3 different sources to compile 7135 photos of COVID, Pneumonia, Tuberculosis, and Normal X-Rays

Pneumonia: Sampled from 5,863 X-ray JPEGs of 2 categories (Pn, normal)

- Selected from retrospective cohorts of pediatric patients 1-5 yrs old
  - Definitely affects generalizability
- All radiographs were screened for quality control
- Diagnoses were graded by 2 expert physicians

Tuberculosis: Sampled from 6300 X-ray JPEGs of 2 categories (TB, normal)

- Compiled by a team from researchers spanning three different institutions

COVID-19: Sampled from public GitHub repository of 2 categories (COVID, normal)

- Data collected from numerous public sources, as well as indirect collection from hospitals and physicians

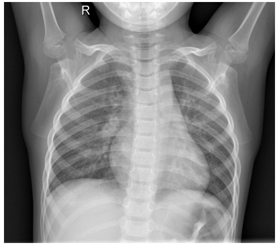
set	train	test	val
COVID	460	106	10
NORMAL	1341	234	8
PNEUMONIA	3875	390	8
TUBERCULOSIS	650	41	12
TOTAL	**6326**	**771**	**38**

train/dev split: 80/20*

* from **train** data above

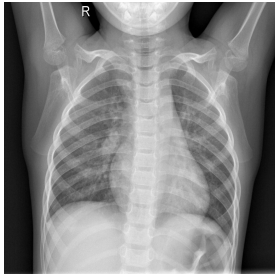
Hassan

Data Pre-Processing



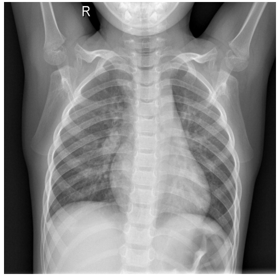
JPEG format
Vary in aspect ratio
Vary in resolution
All on scale 0-255 RGB

Pillow image processing library



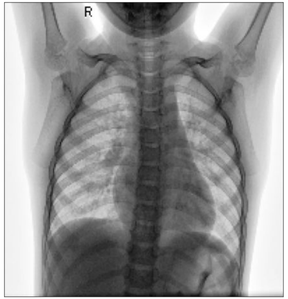
Convert to [0,1] grayscale

Subsample to 200x200



Feature vector of length 40,000

Additional Normalization



Mean: 0.49, SD: 0.249
Didn't yield any model improvement

Approach

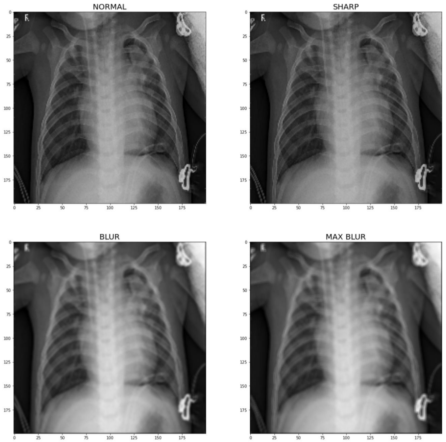
Given this specific application, our approach was to develop 4 single models with optimal parameters then combine them into an ensemble model.

Single Models

Model Type	Parameters	F1 Score (on Dev)
KNN	metric: euclidean n_neighbors: 5	95.4
Naive Bayes	alpha: 71	75.1
SVM	C: 100 gamma: 0.001 kernel: rbf	96.9
Multi-layer Perceptron	activation: logistic alpha: 10 hidden_layer_sizes: (5,) solver: lbfgs	99.9

Experiments & Exploration

Gaussian Image Blurring



Neither blurring, nor sharpening impacted accuracy

Ensemble Model

Baselines

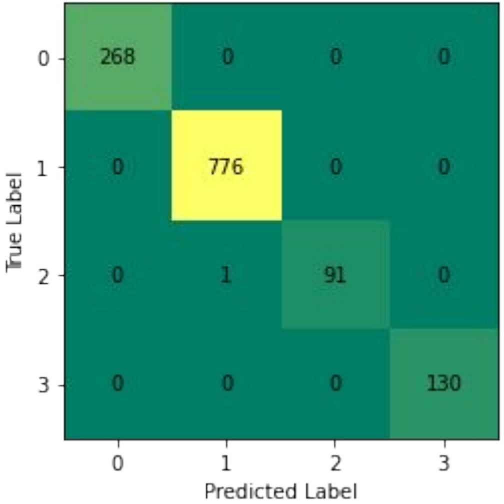
KNN
MultinomialNB
Logistic Regression
SVM

Hard Voting: 94.47
Soft Voting: 94.54
Weighted: 96.05

Additional

MLP Bagging: 92.23
Adaboost: 79.20

MLP Model Confusion Matrix



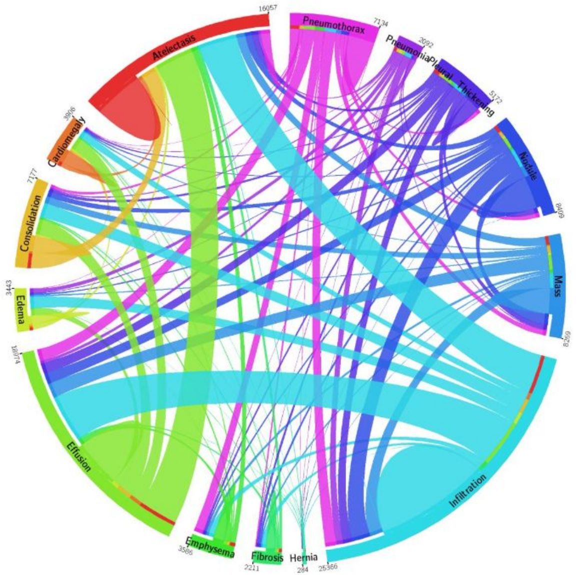
With this small of a dataset we are able to overtrain

Anne

Further Work - Applying to the NIHCC Dataset

	F1 Score (95% CI)
Radiologist 1	0.383 (0.309, 0.453)
Radiologist 2	0.356 (0.282, 0.428)
Radiologist 3	0.365 (0.291, 0.435)
Radiologist 4	0.442 (0.390, 0.492)
Radiologist Avg.	0.387 (0.330, 0.442)
CheXNet	0.435 (0.387, 0.481)

Source: https://arxiv.org/pdf/1711.05225.pdf



Why did our model do so well compared to these other systems?

Dataset Differences:

- NIHCC has ~20x the number of X ray images
- Different source data (NIHCC ChestXRays dataset includes 14 diagnoses, ours has 3)
- Different distribution of data (natural priors vs artificial category balance)
- Kaggle data has many child X rays

Tested our best (Multi Layer Perceptron) algorithm against NIHCC dataset

- Using only samples with a single diagnoses
- 10000 X rays
- Result accuracy score: **68.73**

Data Source: https://nihcc.app.box.com/v/ChestXray-NIHCC/file/220660789610

Anne

Further Work - Applying to the NIHCC Dataset

	F1 Score (95% CI)
Radiologist 1	0.383 (0.309, 0.453)
Radiologist 2	0.356 (0.282, 0.428)
Radiologist 3	0.365 (0.291, 0.435)
Radiologist 4	0.442 (0.390, 0.492)
Radiologist Avg.	0.387 (0.330, 0.442)
CheXNet	0.435 (0.387, 0.481)

Source: https://arxiv.org/pdf/1711.05225.pdf

Why did our model do so well compared to these other systems?

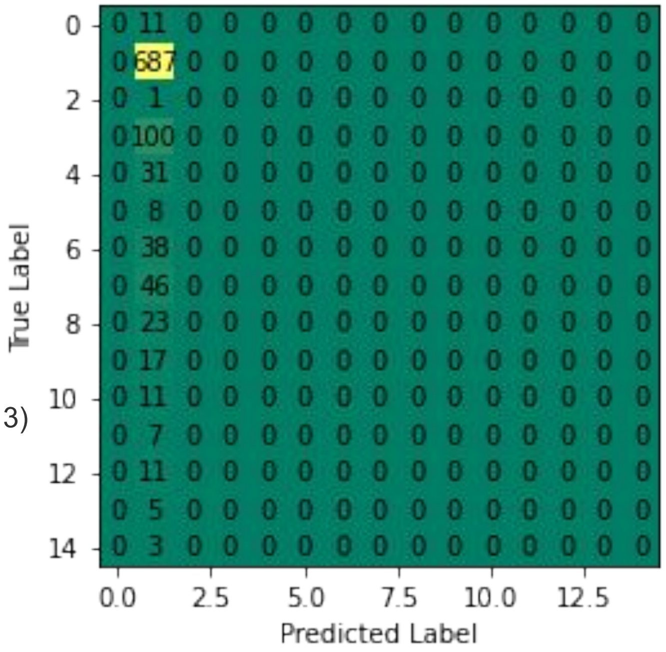
Dataset Differences:

- NIHCC has ~20x the number of X ray images
- Different source data (NIHCC ChestXRays dataset includes 14 diagnoses, ours has 3)
- Different distribution of data (natural priors vs artificial category balance)
- Kaggle data has many child X rays

Tested our best (Multi Layer Perceptron) algorithm against NIHCC dataset

- Using only samples with a single diagnoses
- 10000 X rays
- Result accuracy score: **68.73** **~ 6,878/10,000**

Confusion matrix for NIHCC dataset



Data Source: https://nihcc.app.box.com/v/ChestXray-NIHCC/file/220660789610

Alex

Final Results - Kaggle Dataset

Throughout our initial approach and experimentation, we tested our model on a larger development dataset. Once we were satisfied with the best performing model, we ran it on the test dataset once.

	Mini Train Value Counts:	Dev Set Value Counts:	Test Set Value Counts:
COVID19:	368	92	106
PNEUMONIA:	3099	776	390
NORMAL:	1073	268	234
TURBERCULOSIS:	520	130	41
TOTAL:	**5060**	**1266**	**771**
Optimal MLP Model F1 Score		99.9	75.4

The model yielding a much lower F1 score once we ran it on the test data, which could be due to distribution differences between the train and test sets, as shown above. Generalization is the main marker of success for models of this application.

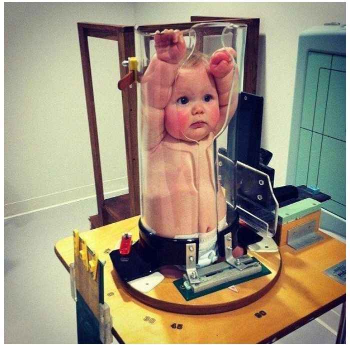
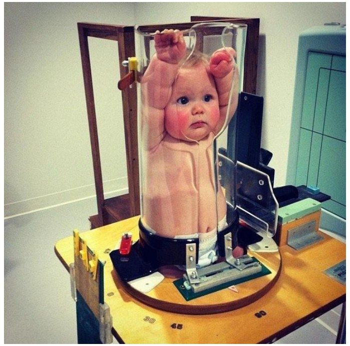
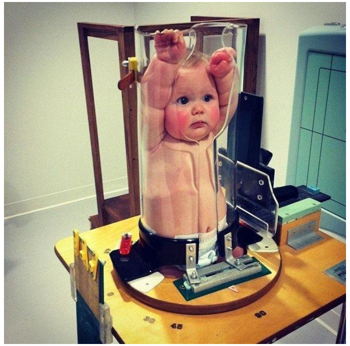
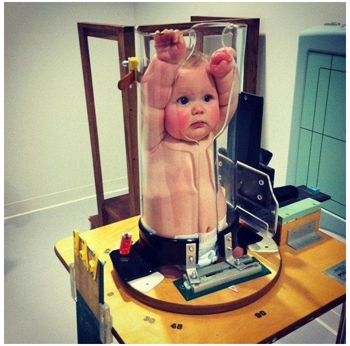
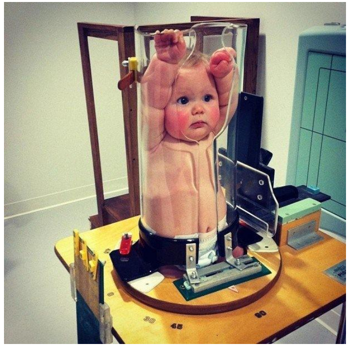
Conclusion

Citations: [7]

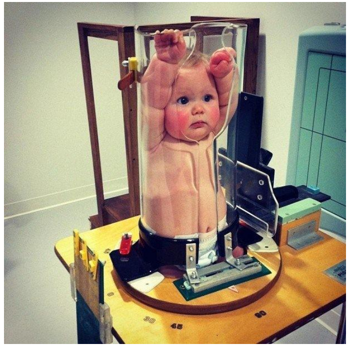
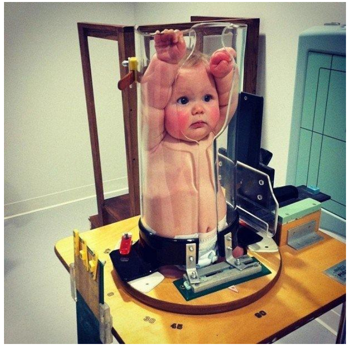
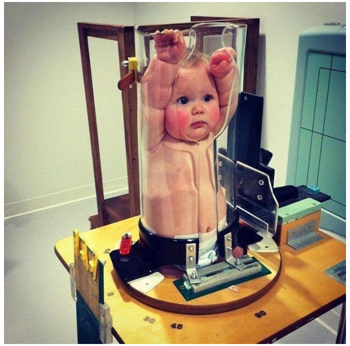
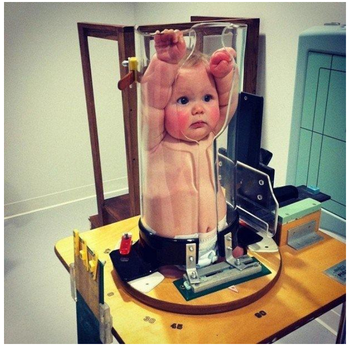
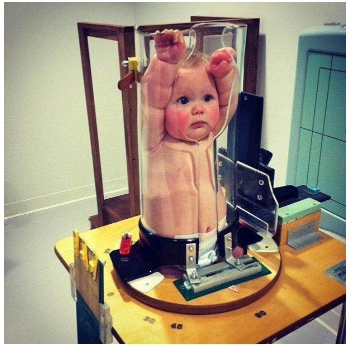
Would our model be suitable for clinical use?

- Common Practice Evaluation
    - Domain Shift Problem - University of Washington researchers audits hundreds of chest X-ray ML models and found that ~50% of them did not generalize well enough to be deployed for clinical use
    - Explainability - Evaluation of a model on external data is insufficient to ensure AI systems rely on medically relevant pathology, because the undesired ‘shortcuts’ learned by AI systems may impair performance in new hospitals.

- As a test to see if our model would generalize to other larger datasets, we ran our best model on the NIHCC dataset. We learned from this that our model that was trained on a balanced data set didn’t work as well on a more realistic dataset. It also lacks explainability, so by these standards, it would not be suitable for deployment.

Q&A

Citations

[1] “FastStats - Pneumonia.” Centers for Disease Control and Prevention, Centers for Disease Control and Prevention, 13 Sept. 2021, https://www.cdc.gov/nchs/fastats/pneumonia.htm.

[2] “Top 20 Pneumonia Facts—2019 - American Thoracic Society.” American Thoracic Society, 2019, https://www.thoracic.org/patients/patient-resources/resources/top-pneumonia-facts.pdf.

[3] Esposito, Antonio, et al. Why Is Chest CT Important for Early Diagnosis of COVID-19? Prevalence Matters. Cold Spring Harbor Laboratory, 1 Apr. 2020. Crossref, doi:10.1101/2020.03.30.20047985.

[4] “Reported Tuberculosis in the United States, 2020.” Centers for Disease Control and Prevention, 12 Oct. 2021, https://www.cdc.gov/tb/statistics/reports/2020/table1.htm.

[5] Rajpurkar, Pranav, et al. "Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning." *arXiv preprint arXiv:1711.05225* (2017).

[6] Khuzani, Abolfazl Zargari et al. “COVID-Classifier: An automated machine learning model to assist in the diagnosis of COVID-19 infection in chest x-ray images.” *medRxiv : the preprint server for health sciences* 2020.05.09.20096560. 18 May. 2020, doi:10.1101/2020.05.09.20096560. Preprint.

[7] Steffel, Catherine. “Machine-Learning Models That Detect COVID-19 on Chest X-Rays Are Not Suitable for Clinical Use.” Physics World, 29 June 2021, https://physicsworld.com/a/machine-learning-models-that-detect-covid-19-on-chest-x-rays-are-not-suitable-for-clinical-use/

Contributions

Alex

- Repository Management
- Multi-Layer Perceptron Model
- Presentation Preparation

Julia

- X-Ray Medical Research
- Image Processing / Loading Code
- KNN Model

Anne

- SVM Model
- AdaBoost & Bagging Ensembles
- NIHCC Data Processing

Hassan

- Data Loading Library
- Naive Bayes Model
- Voting Ensemble Models