Challenges in Deploying Deep Learning for Chest X-Ray Interpretation to Clinical Practice

Jin Haeng Lee[1], Ryan Gust[1], Tiga Choi[1]
Georgia Institute of Technology, Atlanta, GA

Abstract
The feasibility of using machine learning to interpret chest X-rays has been intensely discussed and researched since the mid-2010's. Researchers have been competing to build accurate classification models to convince the medical community machine learning's accuracy and viability in this area. In reality, using machine learning for X-ray interpretation is still confined to testing and has no widespread clinical adoption. To truly utilize machine learning in this area, we will build upon existing research models and identify gaps in models or researches that need to be addressed before machine learning can be relied upon in routine clinical practices.

Our video presentation can be viewed at Youtube.
Our code and supplemental documentation can be downloaded at Georgia Tech Github.
Our presentation slide can be viewed at Google Drive.

1. Introduction and Background

Amongst medical imaging technology, X-ray is the most frequently used technique for disease diagnosis mostly due to its relatively low cost and its early invention and adoption in late 1800's and Chest X-ray (CXR) is the most common form of medical X-rays in medical practices worldwide. Little has changed in its underlying technology since its inception and it is a two-dimensional projection of a three-dimensional object, and can have relatively low resolution compared to other imaging technologies such as computerized tomography (CT) or magnetic resonance imaging (MRI). Interpreting X-rays (and other medical images) requires specialized medical training in the field of radiology. Even with specialized training, radiologists' errors in X-ray readings are higher than other medical imaging technology [1]. U. S. National Institute of Health (NIH) has long recognized errors in radiology and (re)published numerous articles in this area [2][3][4]. With an emerging radiologist shortage in coming years [5] and radiologists increasingly working longer hours, errors in interpreting X-rays are likely to increase, resulting in a lower level of patient care or unnecessary adverse medical outcomes [6]. Computer-aided detection/diagnostics (CAD), especially by machine learning, can play a vital role in overcoming and alleviating this challenge.

2. Problem Formulation

In the past few years utilizing machine learning, especially deep learning, techniques to interpret X-ray images has been undergoing significant research and testing, most notably at CheXpert [7] by Stanford Machine Learning Group and by Google AI lab [8]. Both institutions extracted diagnosis labels from radiology reports using different methods; automated rule-based criteria by Stanford ML Group and natural language processing (NLP) by Google AI Lab and used corresponding X-ray image files to build multi-label classification models. In addition, numerous researchers utilizing different data pre-processing and/or deep learning techniques have published their findings include Multi-label SoftMax Loss with bilinear pooling [9], a model with feature extraction labeled as segmentation-based deep fusion network [10] and a model with Adam Optimizer and automatic learning rate finder to detect signs of COVID-19 [11]. However, in almost all prior published findings, researchers tended to address technical capabilities and accuracy of machine learning in CXR interpretation and did not evaluate non-technical factors, such as the relevant cost of a missed diagnosis or underlying data characteristics, in their considerations.

[1] [jlee3693 ,rgust3, tchoi43]@gatech.edu

CXR interpretation by machine learning remains confined to research and testing and little progress has been made to its adoption in real-world clinical practices. In this research paper, we intended to examine and leverage existing researches and models, to identify areas for improvement and to construct and evaluate one or more new model(s) by incorporating our research findings.

3. Approach and Implementation

CheXpert and NIH have independently promulgated their own datasets of anonymized CXR images and diagnosis labels to the general public for research purposes. In addition, CheXpert hosts an ongoing machine learning competition for "automated chest x-ray interpretation" based upon the highest AUC (Area Under Curve) score and numerous researchers built their machine learning models and published subsequent findings using CheXpert datasets. With comparable AUC results and existing researches, we chose to use the CheXpert dataset and diagnosis label scheme as our model input. We recognized that CXR interpretation is a machine learning multi-label classification that could be optimally solved by deep learning algorithms and our model development and research focus can be summarily described by the following three interconnected processes:

i. ***Data and Pragmatics Analysis***: During exploratory data analysis (EDA) and our preliminary review, we identified several possible data and pragmatic issues with the CheXpert dataset and methodology and devised measures to mitigate their potential effects (listed in *Table 1* below).

Table 1: Issues identified in CheXpert data and methodology

#	Issues/Concerns	Description	Measures
1	Data Quality	17 frontal image records were not labeled with 'AP' or 'PA' image projection[2]. One patient record had no gender. 12 image records had all null labels.	Excluding patients and/or image files
2	Patient Age Distribution	Older patients had more image records. For example patients aged 60 and above comprised 52% of images records while patients aged 20 and below had only 1.44%. Please see *Exhibit 1* for a graphical breakdown	Excluding patients aged below 20.
3	Label Coding	Positive, negative, uncertain and unmentioned diagnosis results were labeled as 1, 0, -1 and null values, respectively. The -1 (uncertain) and null (unmentioned) values cannot be correctly processed by most ML algorithms.	Substituting -1 and null with other numeric values
4	Label Distribution	For positive (1) labels, diagnoses had uneven distributions, ranging between lowest for "Pleural_Other" and highest for "Support Devices" with 3,523 and 116,001 images files, respectively. Please see *Exhibit 2* for a graphical breakdown.	Verifying sampling representativeness
5	Image Projection Distribution	A high majority (over 70%) of patient studies comprised only AP images. Image projection breakdown included approximately 72.3% AP, 13.2% PA and 14.5% LL files.	Separate models for image projections
6	Binary Classification	Positive and negative diagnosis results were labeled as 1 and 0, respectively, regardless of a diagnosis' seriousness/severity.	Model output with probabilities
7	Validation Data	CheXpert validation data contains 234 images for 200 patients	Utilizing sampling

The issues listed above affected almost all aspects of our model development, testing and analysis and validation and we implemented the relevant solutions at various stages. Details of these solutions are described in subsequent paragraphs below.

[2] Image projection varies depending upon a X-ray machine's position and includes AP (anterior posterior) and PA (posterior anterior) for frontal images and LL for lateral images. Please refer to https://www.med-ed.virginia.edu/courses/rad/cxr/technique3chest.html for additional information.

Exhibit 1

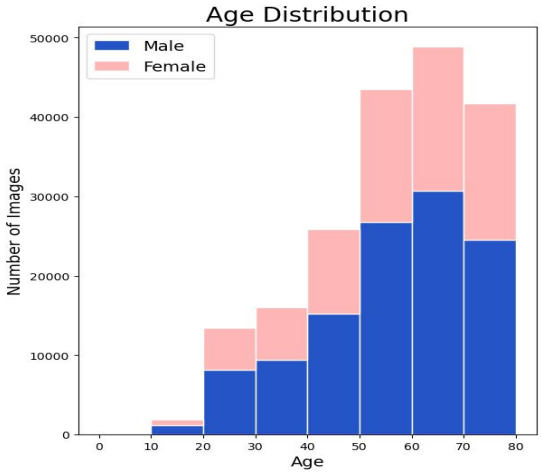
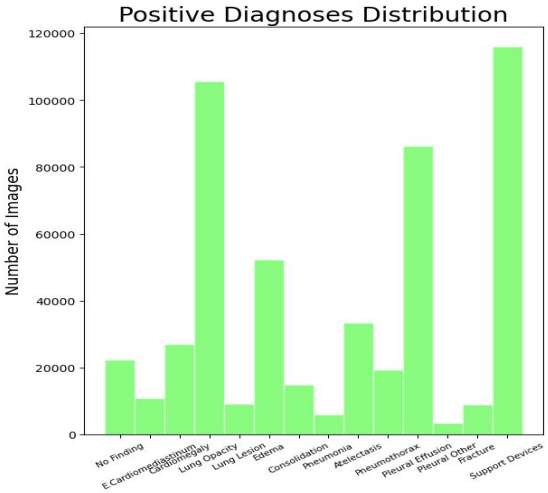


Exhibit 2



ii. ***Image Enhancement Research***: Image augmentation is commonly used technique on medical images [12]. Grayscale X-ray image files are the core data feature/input for our models, hence extensive research was conducted to improve our model by enhancing image features for our deep learning algorithm to train more effectively. These image enhancement techniques are listed in *Table 2* below.

Table 2: Procedures to enhance image features

#	Augmentation Type	Description
1	Normalization	Adjusted image pixel intensity values and equalized brightness.
2	Template Matching	Resized images to 256 x 256, unifying their sizes. Images were then cropped based on template matching with size of 244 x 244 for each lateral and frontal template using correlation coefficient method.
3	Contrast Limiting Adaptive Histogram Equalization (CLAHE)	CLAHE filter is known to improve deep learning performance [13]. Each image was subdivided into tiles with size of 8x8 by default, which in each tile, the histogram was equalized with contrast limiting threshold. CLAHE image comparisons are shown in *Exhibit 3* below.
4	Transformation	Selected images were horizontally flipped, randomly rotated 7 degrees, with 5 degrees of shear applied.
5	Random Scaling	Resized image pixels by either 2% increase or decrease for generalization purpose.

Exhibit 3: CLAHE processing comparison

Image before CLAHE processing

Image after CLAHE processing

All image files underwent the normalization technique described above and would also undergo either template matching or amalgamation of resizing with random crop techniques at equal probability (0.5). Moreover, image files could be processed by additional enhancement procedures including random scaling, CLAHE and transformation or remain unchanged at equal probability (0.5).

iii. ***Pre-processing and Model Development***: Following previous researches in this area [9][11][14], we opted for deep learning algorithms to construct the following two CXR interpretation models:

a. *Replicate DenseNet121 research model by Pham et al.[14]*: after reviewing findings by other researchers, we determined that the Pham DenseNet121 model [14] would serve as our baseline architecture for conducting all subsequent experiments. DenseNet121 required little tuning to achieve respectable performance, was robust to hyperparameter changes, and was significantly faster to train than DenseNet161 and DenseNet201. Following procedures similar to Pham et al. [14], we used a DenseNet121 model, pretrained on ImageNet, with all but the final fully-connected layer frozen for the first epoch of training. Images were passed through our aforementioned image augmentation pipeline and fed to the model in batches of size 256. Loss was calculated using a decoupled sigmoid activation function and binary cross-entropy loss on U-Ones+LSR modified labels which replace uncertainty labels (-1) with uniform random values between 0.55 and 0.85 [14]. After the first epoch, the initial learning rate of 1e-2 was reduced by a factor of 10 and the last DenseBlock was unfrozen. This pattern of gradual unfreezing was continued for the 4 remaining epochs and the learning rate was decayed after epochs 3 and 4. During inference, the trained model generated predictions on the validation set 10 times, each pass applying a modified set of augmentations in which CLAHE was removed and only one geometric transform can take place. A simple weighted average was taken across the 10 validation sets to predict the final labels.

b. *Construct projection-based model ensemble*: by utilizing separate models for image projections, our objective was to differentiate model parameter calculation and to accentuate image diagnosis features for each image projection. This model architecture required us to train three separate models and to feed/process image files for each image projection separately. During inference, we would follow a similar procedure and partition the validation set based on projection type. For each of the three projections, we used the corresponding model trained solely on that image instance to generate predictions. These prediction sets were concatenated together to produce our final set of predictions. Our expectation was that this model would yield a slightly better or similar AUC and accuracy performance as other models.

For the two models described above, our primary metric to measure their performance is AUC. AUC is a standard metric utilized by CheXpert to benchmark and evaluate model performance amongst researchers. By using AUC, we can promptly gauge our models' performance among existing research models, particularly Pham DenseNet121. Our secondary metrics to measure model performance are accuracy, false positive rate (FPR) and false negative rate (FNR). In real-world clinical practice, a missed **positive** diagnosis could pose a severe adverse effect on patient care and a serious threat to a practitioner's professional liability. To emphasize the significance of identifying true positive diagnoses, we would test probability thresholds ranging from 0.02 to 0.40 to assess the effects of changes in threshold on accuracy, FPR and FNR on five specific disease diagnoses.

4. Experimental Evaluation

From the CheXpert data, we selected two datasets for model training and testing; first dataset with filtered[3] complete CheXpert training population of 187,613 patient studies (for 223,384 images) as training data and with complete CheXpert validation population of 200 patient studies (for 234 images) as test data and second dataset with 70% and 30% of 120,000 sampled patient studies (for 142,839 images) from CheXpert training population as training and test

[3] 28 patients (for 30 images) were excluded for data quality issues described in the ***Data and Pragmatics Analysis*** paragraph.

data, respectively. We recognized that the CheXpert validation dataset is a relatively small population of 234 images and reporting metrics can be unduly affected by data variance. Consequently, we devised a much larger testing population in our second dataset based upon sampling from the CheXpert training population. Our sampling procedure included steps to validate the representative proportions of training and test populations to address the uneven distributions of positive labels amongst diagnoses (discussed in the ***Data and Pragmatics Analysis*** paragraph above). Because other researchers reported on CheXpert validation data, we would compute certain metrics on the same dataset for comparison purposes. For these two datasets, the AUCs achieved in our two models are listed in Table 3 below.

Table 3: AUC for Filtered and Sampled datasets for the two models

	AUC (All 14 Labels)	
Population (Epoch= 5)	**Repl. DenseNet121**	**Projection-based Ensemble**
Training: filtered CheXpert training data 187K patient studies Test: CheXpert validation data 200 patient studies	0.8708	0.8588
Training: 70% of sampled 120K patient studies Test: 30% of sampled 120K patient studies	0.8468	0.8531

To further analyze the models' efficacy on individual diagnosis, we calculated AUCs for five disease diagnoses identified by Pham et al. [14] and compared our model results to figures published by Pham et al. [14] in Table 4. In Exhibit 4 and 5, we displayed the receiver operating characteristic (ROC) curves and computed thresholds based upon Youden's index[4] for these five disease diagnoses under the two models.

Table 4: AUC for five disease diagnoses under the filtered dataset

	Diagnosis					
Model	**Atelectasis**	**Cardiomegaly**	**Consolidation**	**Edema**	**P. Effusion**	**Average**
Pham et al. Ensemble [14]	0.909	0.910	0.958	0.957	0.964	0.940
Pham et al. Single Model [14]	0.825	0.855	0.937	0.930	0.923	0.894
Replicated Pham DenseNet121	0.821	0.795	0.898	0.919	0.932	0.873
Projection-based Model Ensemble	0.814	0.788	0.895	0.901	0.904	0.860

Exhibit 4

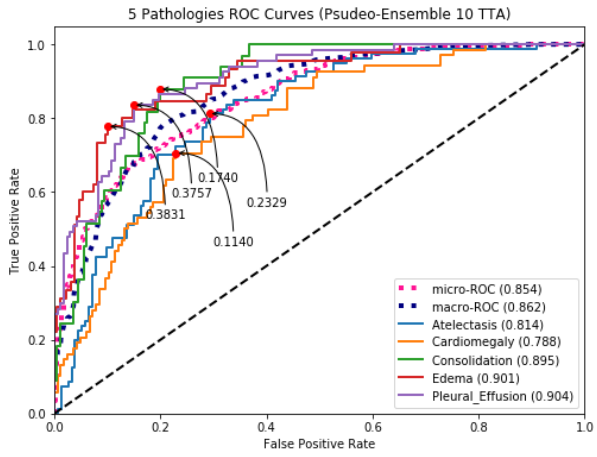
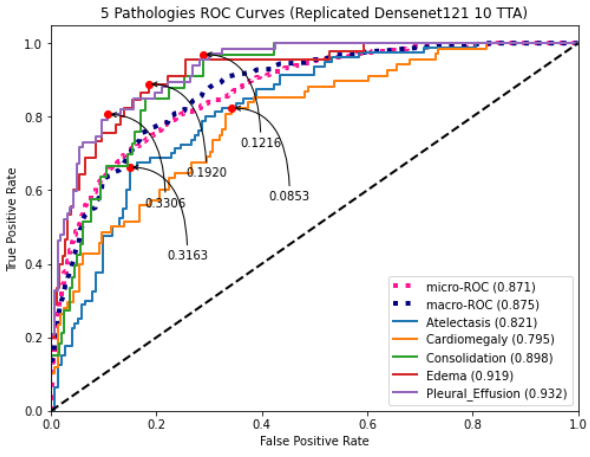


Exhibit 5



[4] https://en.wikipedia.org/wiki/Youden%27s_J_statistic

Our replicated Pham DenseNet121 model achieved an AUC of 0.8708, compared to a high of 0.894 for a single DenseNet121 model and 0.940 for an ensemble reported from Pham et al. [14]. The differences in AUC can be attributed to absence of a conditional training (CT) phase in our replicated model and the use of model ensemble by Pham et al. Using the replicated model as a baseline, our projection-based model ensemble was able to achieve an AUC of 0.8588 and 0.8531 for the CheXpert validation data and our sampled test data, respectively. We noticed the slight AUC decrease for the projection-based model ensemble (from the replicated model) and performed analyses to assess whether data partitioning caused AUC reduction for deep learning algorithms. However, our analyses did not provide a conclusive outcome. To assess whether consistent labeling within patient study would affect the AUC results, we applied label propagation and/or probability aggregation[5] (for each patient study) but these measures did not improve the AUC results.

To assess how probability threshold affects the accuracy, FPR and FNR for the five disease diagnoses, we calculated these three metrics for the CheXpert validation data under projection-based model ensemble with thresholds incremented by 0.02 and from a low of 0.02 to a high of 0.40 and charted how the metrics fluctuate as threshold changes in Exhibit 6 and 7.

Exhibit 6

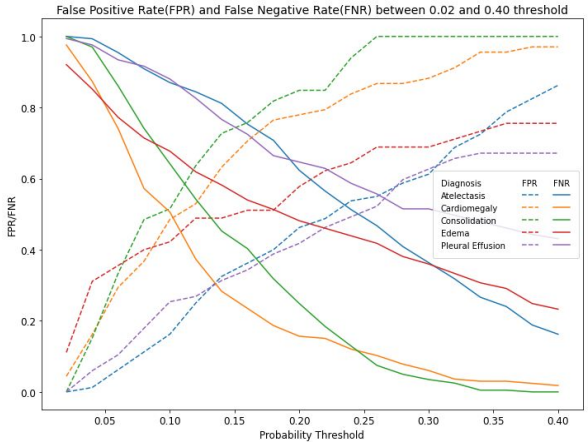
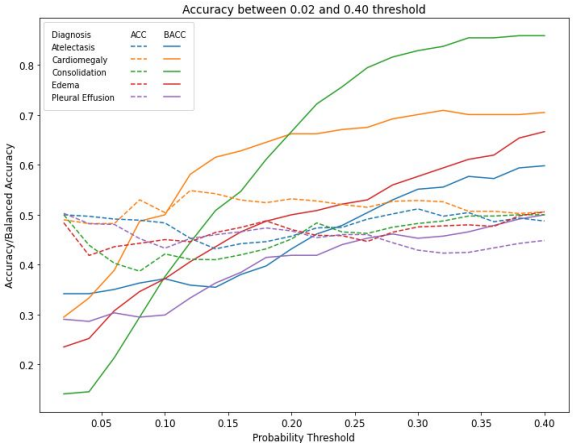


Exhibit 7



As shown in Exhibit 6 and 7, the FNR, FPR and accuracy five disease diagnoses illustrated different progression patterns under various threshold values. For example, for consolidation diagnosis, accuracy and FNR increased at a similar rate as threshold increases and FNR increased at a steep rate (from almost zero to around 0.3) when threshold increased from 0.02 to 0.05. This indicates consolidation would require a low threshold (at around 0.2) to capture most true positive diagnoses. In contrast, FNR for Atelectasis increased at a lower rate as threshold increases. Depending upon a practitioner's role, experience and specialty and diseases' severity, (s)he may choose to use different thresholds to identify potentially positive diagnoses.

5. Conclusion

With our two trained models able to process/predict on 36,000 test image files in less than 35 seconds and with AUC at 0.8708 and 0.8588 for replicated Pham DenseNet121 and projection-based model, respectively, neither speed nor efficacy should be the prohibiting factors in utilizing machine learning algorithms for CXR interpretation. In evaluating prior researches for our project, we noticed that researchers primarily focused on developing more

[5] Label propagation and probability aggregation ensured images in each patient study have consistent labels and probabilities, respectively, by standardizing their values for each diagnosis in a patient study. In CheXpert, this similar technique is called "mention aggregation".

accurate (as measured by higher AUC) models and did not discuss techniques to deploy and/or incorporate machine learning algorithms to clinical practice. As more successful and accurate models and findings were published in the past few years, deep learning algorithms were discussed as an emerging threat to the radiologist profession [15]. The absence of discussion on proper mechanisms to incorporate machine learning into CXR interpretation and the discussion of machine learning to replace radiologists could only spawn resistance for its adoption and cast doubts on its practical uses. We viewed these two factors as main drivers for the lack of progress in clinical implementation. Similar to autonomous driving technology which is also driven by machine learning algorithms, CXR machine learning interpretation should function in a very comparable role; as an ancillary tool to **help** practitioners perform their routine duties in a more efficient and effective manner. Analogous to requiring a driver in the driver seat of a self-driving car, a radiologist should be the final decision maker in CXR machine learning interpretations and can correct or override any machine learning decisions. We would recommend that CXR interpretation machine learning models allow radiologists to adjust diagnosis probability threshold according to his/her preference and provide probability readings in addition to binary outcome. When incorporated with other technology such as queueing and messaging, CXR machine learning interpretation can arrange patient priorities based upon diagnosis' seriousness/severity and promptly alert radiologists of such occurrences.

As mentioned briefly above, NIH also promulgated its own dataset for research purposes and it comprises 112,120 frontal-only chest X-ray images files for 30,805 patients under a different 14-label[6] scheme [16]. The NIH data does not have age-skewness or image project distribution issues that were embedded in the CheXpert data (discussed in the ***Data and Pragmatics Analysis*** paragraph above). When analyzing the NIH and CheXpert datasets, we noticed that the two datasets used different diagnosis labels to categorize CXR images and the shared and mutually exclusive labels are listed in Table 5 below:

Table 5: Labels used by NIH and CheXpert

	NIH	CheXPert
1	Atelectasis	
2	Cardiomegaly	
3	Consolidation	
4	Edema	
5	Pneumonia	
6	Pneumothorax	
7	No Finding[7]	
8	Effusion	Enlarged Cardiomediastinum
9	Emphysema	Fracture
10	Fibrosis	Lung Lesion
11	Hernia	Lung Opacity
12	Infiltration	Pleural Effusion
13	Mass	Pleural Other
14	Nodule	Support Devices
15	Pleural_thickening	

This difference in labeling scheme would likely render models trained using one dataset, for example CheXpert, useless and/or less accurate to report on the other dataset, for example NIH, and would require researchers to develop a new model for each dataset. For machine learning to achieve its objective of producing accurate predictions, a more comprehensive and complete labeling scheme should be used and agreed upon by the medical community. For future research activities, we would like to re-train our models on the NIH dataset under a different

[6] NIH's 14 labels could be considered as a 15-label scheme because although used in the "Label" field, "No Finding" was NOT considered as a diagnosis label for NIH but was one for CheXpert. For both NIH and CheXpert data, "No Finding" label could be viewed as a derived/filled-in value; populated when none of the other diagnosis (except for Support Devices in CheXpert) labels were detected.

[7] "No Findings" and "Support Devices" may not be considered as disease labels.

labeling scheme to evaluate accuracy and performance differences and to provide further evidence on deep learning algorithms' effectiveness and accuracy in CXR interpretation.

References

1. Brady AP. Error and discrepancy in radiology: inevitable or avoidable?. Insights into imaging. 2017 Feb 1;8(1):171-82.
2. Berlin L. Radiologic errors, past, present and future. Diagnosis. 2014 Jan 1;1(1):79-84.
3. Maskell G. Error in radiology—where are we now?. The British journal of radiology. 2019 Apr;92(1096):20180845.
4. Pinto A, Brunese L. Spectrum of diagnostic errors in radiology. World journal of radiology. 2010 Oct 28;2(10):377.
5. Bender CE, Bansal S, Wolfman D, Parikh JR. 2018 ACR Commission on Human Resources workforce survey. Journal of the American College of Radiology. 2019 Apr 1;16(4):508-12.
6. Krupinski EA, Berbaum KS, Caldwell RT, Schartz KM, Kim J. Long radiology workdays reduce detection and accommodation accuracy. Journal of the American College of Radiology. 2010 Sep 1;7(9):698-704.
7. Irvin J, Rajpurkar P, Ko M, Yu Y, Ciurea-Ilcus S, Chute C, Marklund H, Haghgoo B, Ball R, Shpanskaya K, Seekins J. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. InProceedings of the AAAI Conference on Artificial Intelligence 2019 Jul 17 (Vol. 33, pp. 590-597).
8. Maskell G. Error in radiology—where are we now?. The British journal of radiology. 2019 Apr;92(1096):20180845.
9. Ge Z, Mahapatra D, Sedai S, Garnavi R, Chakrovorty R. Chest X-rays Classification: A Multi-Label and Fine-Grained Problem. 2018.
10. Liu H, Wang L, Nan Y, Jin F, Wang Q, Pu J. SDFN: Segmentation-based deep fusion network for thoracic disease classification in chest X-ray images. Computerized Medical Imaging and Graphics. 2019 Jul 1;75:66-73.
11. Shorfuzzaman M, Masud M. On the Detection of COVID-19 from Chest X-Ray Images Using CNN-Based Transfer Learning. CMC-COMPUTERS MATERIALS & CONTINUA. 2020 Jan 1;64(3):1359-81.
12. Tang Z, Chen K, Pan M, Wang M, Song Z. An augmentation strategy for medical image processing based on Statistical Shape Model and 3D Thin Plate Spline for deep learning. IEEE Access. 2019 Sep 12;7:133111-21.
13. Vidyarthi A, Shad J, Sharma S, Agarwal P. Classification of Breast Microscopic Imaging using Hybrid CLAHE-CNN Deep Architecture. In2019 Twelfth International Conference on Contemporary Computing (IC3) 2019 Aug 8 (pp. 1-5). IEEE.
14. Pham HH, Le TT, Ngo DT, Tran DQ, Nguyen HQ. Interpreting Chest X-rays via CNNs that Exploit Hierarchical Disease Dependencies and Uncertainty Labels. arXiv preprint arXiv:2005.12734. 2020 May 25.
15. Chan S, Siegel EL. Will machine learning end the viability of radiology as a thriving medical specialty?. The British Journal of Radiology. 2019 Feb;92(1094):20180416.
16. Wang X, Peng Y, Lu L, Lu Z, Bagheri M, Summers RM. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases.

Team Contribution

Based upon our approximation, the table below depicts contribution by each team member and the workload was distributed based upon members' skillset and availability:

Phase: Task	Start Date	End Date	John	Ryan	Tiga	Total
1: Data Collection and Exploration	9/28	11/15	15	10	18	43
Data Analysis and Pre-processing	10/1	12/1	10	15	10	35
Data Exploration	11/1	11/30	10	10	15	35
Sampling Validation	11/15	11/27	15	5	15	35
2. Research and reporting	10/15	12/5	12	10	10	32
Image Enhancement	10/15	11/15	10	5	10	25
Metrics research and reporting	11/15	12/5	11	10	15	36
3: Model Development	10/12	12/5	10	20	5	35
Model Analysis	10/12	12/5	10	20	10	40
Model Testing	10/15	12/5	10	15	10	35
Model Validation	10/15	12/5	12	15	6	33
4. Project Documentation	10/15	12/6	10	5	10	25
Proposal	9/30	10/9	5	10	5	20
Draft	10/11	11/15	7	5	7	19
Final Report	11/15	12/6	5	5	10	20
Presentation and Submission Files	11/20	12/6	10	5	5	20
			162	165	161	488