Lung Lesion Classification: A Tale of Class Imbalance and Dataset Cleanliness

Group 1: Abbie Cheng, Daniel Tan, Eric Yang, Lara Rostomian

Spring 2021

ABSTRACT

Background: The early identification of lung lesions is very important as lesions have a high likelihood of becoming cancerous. The objective of this study is to build a model for automated chest radiograph interpretation of lung lesions with high performance. Such a model could improve workflow prioritization, support decision making in clinical settings, and ultimately help progress large-scale screening and global population health initiatives.

Methods: Using the Stanford CheXpert Dataset, we employ various modeling, sampling, and data cleaning methods throughout our study. Classification models used include convolutional neural networks with VGG16 and DenseNet121 architectures and a random forest model as baseline for comparison. Sampling methods used include oversampling the minority case and undersampling the majority case with different proportions. Various data cleaning strategies were employed including redefinition of negative cases and exclusion of lateral chest x-rays in an attempt to improve model focus.

Results: We found that careful training data selection yielded the highest performance boost during testing. Our best performing model addressed the heterogeneity of the dataset by re-defining negative cases to ones that are either explicitly labeled as negative for lung lesion or labeled as “no finding”. Only frontal view images were included during training of the model. This highest performing model employed the VGG16 architecture with pre-trained weights, achieving an accuracy of 0.82, AUC of 0.82, precision of 0.76 and recall of 0.40.

Discussions: The results of the various modeling, sampling, and data cleaning approaches included in this study highlight the importance of the data generating, data labeling, and data cleaning processes. Selecting the appropriate data for the task can yield significant performance boost compared to hyperparameter tuning.

INTRODUCTION

A lung lesion is a well-defined and rounded opacity surrounded by lung tissue with a diameter less than or equal to 3 cm. **Figure 1** shows examples of chest X-rays that are positive and negative for lung lesion. These lesions can be cancerous or benign, with a 40% overall likelihood of being cancerous [1]. Lung lesions are common and are found in 1 in 500 chest X-rays and 1 in 100 chest CT scans. There has recently been an increase of lung lesion recognition, largely due to the increased application of CT scans for lung cancer screenings. Early detection and screening is particularly important in the case of lung lesions due to the high risk of a lung lesion becoming cancerous.

The objective of this study is to build a model for automated chest radiograph detection of lung lesions with high performance. Previous studies have shown the potential that artificial intelligence systems have in medicine, particularly in areas of diagnosis and prognosis predictions [2]. When deployed appropriately, artificial intelligence systems can help improve healthcare delivery in the 21st century by complementing and supporting clinical decision making. An automated classification model to detect lung lesions with a high level of accuracy could improve workflow prioritization, support decision making in clinical settings, and ultimately help progress large-scale screening and global population health initiatives. In addition, we studied the effect of dataset imbalance and cleanliness on artificial intelligence model performance and explored methods to mitigate them.

METHODS

For this study, we used the Stanford ML Group's CheXpert dataset that contains X-rays from Stanford Hospital performed between October 2002 and July 2017 [3,4]. The dataset includes 224,316 chest radiographs of 65,240 patients. The images are graded for the presence or absence of 14 pathologies including enlarged cardiomegaly, cardiomegaly, lung lesion, lung opacity, edema, consolidation, pneumonia, atelectasis, pneumothorax, pleural effusion, pleural other, fracture and support devices. The labels are derived from a natural language processing model that can extract observations from free-text radiology reports and capture uncertainties present in the reports. There are four potential labels applied to each pathology, including positive, negative, uncertain, or not applicable (N/A) if there was no mention of the pathology within the reports.

During programming, we used Python (v 3.6.7) as our choice of language. We also leveraged packages such as pandas (v 1.1.5), numpy (v 1.19.5), scikit-learn (v 0.22.2), keras (v 2.4.3) and tensorflow (v 2.4.1) to aid our development.

In all approaches, we split the images into 80% train and 20% test sets on the patient level to prevent information leakage and over estimation of our model performance. Our target pathology, lung lesion, was positive in 4% of the full dataset, negative for 0.06% of the dataset and the remaining data had no mention for lung lesions, showing clear class imbalance. We built our classification models using the VGG16 [5] and DenseNet21 [6] structures and with pre-trained weights for transfer learning. We tuned the hyperparameters for every model in each approach, including learning rate, epoch size, loss function type, and minor architectural additions. For a baseline comparison, we also built random forest models, where each pixel in the images was considered an independent feature. In addition to traditional machine learning model tuning, we also aimed for two broad approaches to boost performance of our models. First, we attempted various sampling techniques. We experimented with various ratios and raw counts of positive and negative examples during training using the imblearn package (v 0.0). Second, we made multiple attempts to redefine our training data by excluding portions of negative instances such as filtering for X-Ray view type and the presence of other pathologies labels. Each of the approaches described below contains combinations of these two broad tuning methodologies. We assessed each models' performance by accuracy, AUC, precision and recall.

Initially, we set out to use the entire dataset for classification (approach 1). We excluded patients that had uncertain labels and included all patients that were labeled either negative or NA for lung lesion as negative cases for lung lesion. Error auditing was performed to understand which part of the model was most suffering from by looking at the label distribution of false positive cases.

After feedback from error auditing, we altered our sampling techniques (approach 2). We oversampled the minority case (positive to negative = 194921:194921) for augmentation of positive cases and also undersampled the majority cases (positive to negative =8411:8411) to understand the effect of our imbalanced dataset. We also experimented with various uneven proportions of positive to negative cases.

To address continued high false positive rate, we analyzed the false positive instances that our model identified. After consulting the medical literature [4] and radiology AI experts, we re-defined our negative cases to ones that are labeled as "no findings", indicating that the image is completely normal without other diagnoses, in order to decrease the noise in the training data (approach 3). In this approach, the training set imbalance still remained (positive to negative = 7615:18731)

In an attempt to further improve model recall and precision, we then varied sampling to account for class imbalance (approach 4) on our newly modified dataset from approach 3. We again applied various sampling methods (undersampling of the majority case and oversampling of the minority case) with different proportions. Our final approach (approach 5) built upon approach 3. Here, we further focused the training data by excluding the lateral view of chest X-rays (positive to negative = 5855:14183).

Finally, in order to interpret the regions of the images the model focuses on to preliminarily study model interpretability, we implemented the Grad-CAM [7] technique to visualize high importance gradients.

Exact details of all of our methods can be found in this **Google Collaboratory notebook**: https://colab.research.google.com/drive/1HGU-q3fZgeX3oNRMZAi1k3LXU_ycEVzB#scrollTo=EJ-97cKLDoxI

RESULTS

The reported results below are the models with the highest AUC values after hyperparameter tuning for each approach. Our initial approach (approach 1), where we included heterogenous negative cases and did not deal with dataset imbalance, resulted in suboptimal performance with zero percent in both precision and recall. This pattern was evident for both our baseline random forest model and deep learning models (full results of all models are provided in **Table 1**). The deep learning approaches here were trained for 10 epochs with the Adam optimizer and a learning rate of 0.00001.

After adjusting for the imbalanced cases (approach 2), the model started to pick up some signals from the positive cases. The model that achieved the best performance in this approach was modeled on training data with an undersampling ratio of positive to negative 1:1. The model architecture used here is VGG16 with pre-trained ImageNet [8] weights. We trained the model for 10 epochs with the Adam optimizer and a learning rate of 0.00001. This model achieved 0.72, 0.08, 0.60 for AUC , precision and recall respectively (full results of all models are provided in **Table 2**).

In the third approach, the best performing model here significantly improved to 0.82,0.76,0.42 for AUC, precision and recall respectively (approach 3). Here, we filtered the heterogenous negative cases and only included cases that were negative in all other pathologies. Our model improved compared to the previous approach and achieved higher performance compared to the baseline random forest model (full results of all models are provided in **Table 3).** The best model here again has the VGG16 architecture with pre-trained ImageNet weights. We trained the model for 35 epochs with the Adam optimizer and a learning rate of 0.00003.

In the fourth approach, the best performing model improved in recall from approach 3 as we applied our sampling methods to our newly modified data (approach 4). This model yielded 0.82, 0.44, and 0.82 for AUC, precision and recall respectively (full results of all models are provided in **Table 4)**. This model utilized the VGG16 structure with pre-trained ImageNet weights, undersampling technique with positive to negative ratio 1:1, the Adam optimizer for gradient descent, and a learning rate of 0.00003 trained for 35 epochs.

Finally, in our fifth approach, we attempted to boost model performance by further focusing the training data by removing the lateral view of X-Rays in our final modeling approach (positive=7041, negative= 17729) (approach 5). Here, the model achieved an accuracy of 0.82, AUC of 0.82, precision of 0.76 and recall of 0.40. The high performing model again utilized the VGG16 structure with pre-trained ImageNet weights. We trained the model for 50 epochs with the Adam optimizer and a learning rate of 0.00003. The results of this approach was similar to approach 3.

DISCUSSION

After our first approach, we hypothesized that our model was picking up the imbalanced distribution of our data by predicting images all as negative. The results were expected as the loss function optimizes more for accuracy rather than precision. This in turn gave us high accuracy but essentially 0 for precision and recall. Therefore, with such an imbalance dataset, we believed that AUC would be a more appropriate metric to assess our model performance moving forward.

Prior to addressing the imbalanced training set issue, our initial dataset only had 4% positive cases among 194921 images. We implemented oversampling and undersampling techniques to adjust the positive to negative ratios to 1. It is important to acknowledge that oversampling methods can be prone to model overfitting because the model repeatedly samples the same positive (minority) cases that may eventually lead to the model memorizing the all positive cases. Undersampling techniques may leave out valuable information from the negative (majority) cases that make the decision boundary between the minority and majority instances harder to learn. Although we did not observe the commonly known issues in our sampling implementation, our model still did not significantly improve in performance. Most notably, the low precision after approach 2 indicated a high false positive rate.

During the process of optimizing our model, we observed that the cleanliness of the training data contributed most to the improvement in performance. According to the paper initially published for the CheXpert dataset, the labels had an hierarchical structure that grouped pathologies into subgroups. For example, under lung opacity, there are multiple pathologies including edema, consolidation, pneumonia, atelectasis and lung lesion. Upon consulting with experts and available literature [4] we found that lung lesion's phenotypes overlap considerably with those other diseases under the same subgroup, as shown in **Figure 2**. In our initial approach (approaches 1 and 2) where no data filtering was applied, we found that the top instances in our false positive cases were actually lung opacity, pleural effusion and support devices as shown in **Figure 2A**. We hypothesized that the classification boundary between our positive and negative cases is vague and provided challenges for the model to distinguish pathologies grouped under the lung opacity as they share similar image and phenotypic characteristics. This hypothesis was further validated as our model significantly improved after we filtered out the possibly ambiguous negatives cases. This process of selecting the appropriate data for training by excluding more negative instances also addressed the class imbalance problem in our dataset as a byproduct, which in concert boosted our model performance. However, data cleanliness was the driving factor of performance boost, indicated by the small performance improvement from approach 1 to 2 compared to the large performance boost from approach 2 to 3.

There are several limitations to our best performing models in terms of classification performance and clinical applicability. First, although our model performance in terms of AUC improved significantly, we still only achieved 0.46 recall, suggesting that only less than half the positive lung lesion cases were successfully identified. Therefore, this model is not yet ready for deployment in a clinical setting and needs improvements on further optimizing the recall and overall performance. Second, among the images that were "NA" for lung lesion included in our negative sample, we specifically excluded images that were positive for any other disease due to data cleanliness concerns. Because of this, we do not know how well our model will generalize to a real-life scenario where data is not as clean (e.g. patients having multiple diseases at a time). Third, the number of raw positive images in our training set was relatively small (~6000 images), especially compared to the number of negative images. Ideally, we would want more images to improve model generalizability. This smaller number of positive images could be contributing to our model's lower recall. Lastly, the labels for our dataset could have some inaccuracies as they were derived from natural language processing (NLP) rather than from manual annotation. The accuracy of NLP-based approaches can be limited by a host of different factors such as misspellings, use of acronyms/abbreviations, and other linguistic nuances.

In terms of clinical relevance and applicability, we acknowledge two limitations. First, neither the radiologists nor the model had access to patient history or previous examinations, which has been shown to decrease diagnostic performance in chest radiograph interpretation [9,10]. Second, without an experiment comparing radiologists' performances with our models, we cannot definitely conclude the clinical improvements that our models offer.

Throughout the project, we faced significant challenges highlighted throughout this report during model development and learned a tremendous amount from them. From learning to handle large data imbalance to understanding the importance of data selection, we were able to explore the nuances of deep learning. All aspects of the entire interconnected AI pipeline contribute to model performance and simple tuning of hyperparameters is not guaranteed to yield desired outcome. It is also important to select the appropriate metrics to evaluate models to maximize clinical applicability. Finally, domain knowledge is invaluable throughout the whole process. Understanding eventual use cases are important to guide design decisions and appropriate data selection.

There are several logical next steps to build onto our work. Exploration of other approaches to improve model performance and generalizability is crucial. One of such approaches relates to architecture tuning. It may be helpful to freeze the pretrained base layers of the model and add additional new fully connected layers to leverage the high level feature characteristics extracted from the original transferred model. Another direction to explore is to collect more samples, especially positive lung lesion samples to increase chances of the model to learn the lung lesion pattern. Including datasets from other institutions will also help with model generalizability. Lastly, we briefly explored model interpretability by using Grad-CAM visualizations (**Figure 3**) and found that the model was struggling to consider the abnormality regions of interest in decision making. In the future, especially for clinical deployment, it is essential to tune the model so that it mimics the decision making process of physicians during diagnoses.

Tables and Figures:

Table 1: Approach 1 - Comparison of model performance after using the entire dataset for classification.

Model	**Accuracy**	**AUC**	**Precision**	**Recall**
Random Forest	0.95	0.52	0	0
VGG16	0.96	0.61	0	0
DenseNet121	0.96	0.66	0	0

Table 2: Approach 2 - Comparison of model performance after tuning various sampling techniques for classification using the entire dataset.

Model	**Accuracy**	**AUC**	**Precision**	**Recall**
Random Forest	0.62	0.58	0.05	0.45
VGG16	0.70	0.72	0.08	0.60
DenseNet121	0.61	0.65	0.07	0.57

Table 3: Approach 3 - Comparison of model performance after filtering for “no findings” subset of data as negative instances for classification.

Model	**Accuracy**	**AUC**	**Precision**	**Recall**
Random Forest	0.77	0.62	0.41	0.13
VGG16	0.82	0.82	0.72	0.46
DenseNet121	0.80	0.78	0.69	0.38

Table 4: Approach 4 - Comparison of model performance after tuning various sampling techniques on the filtered dataset from approach 3.

Model	Accuracy	AUC	Precision	Recall
Random Forest	0.77	0.60	0.33	0.08
VGG16	0.7	0.82	0.44	0.82

Figure 1: Examples of Chest Radiographs which are (A) positive and (B) negative for lung lesions from the Stanford CheXpert dataset.

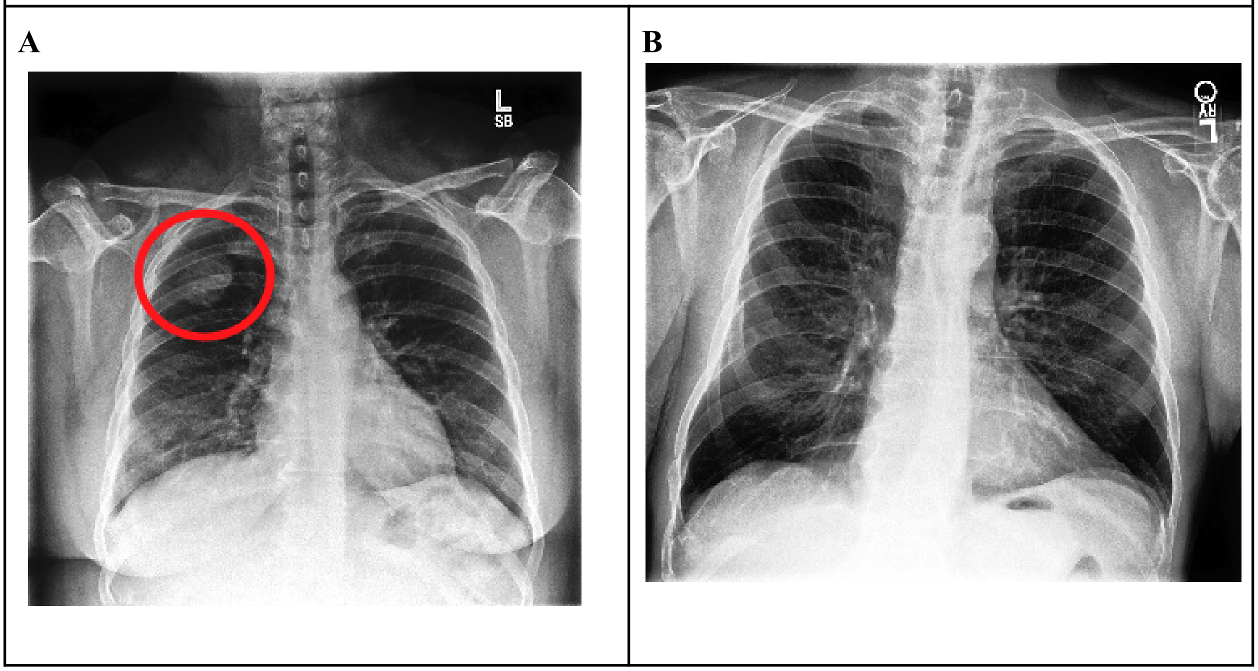


Figure 2: Error auditing and analysis of dataset heterogeneity. (A) Distribution of abnormalities that false positive cases have after approach 2. (B) Hierarchy of abnormalities in the Stanford CheXpert dataset.

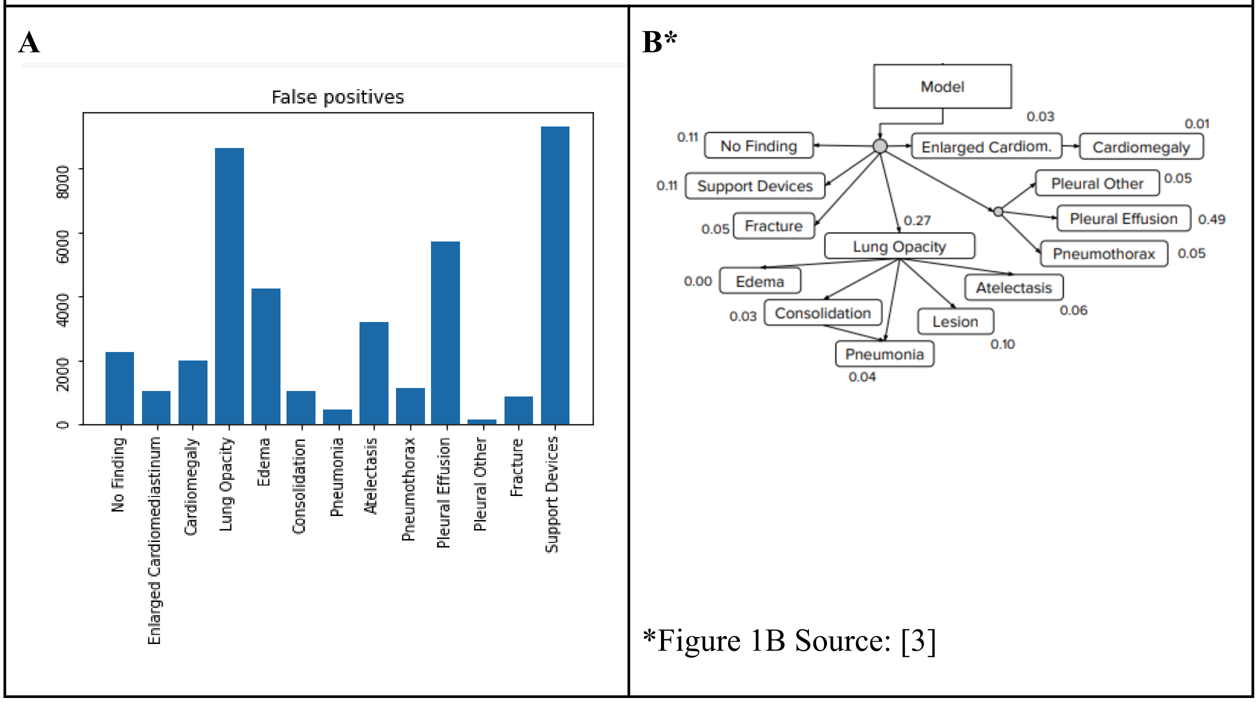
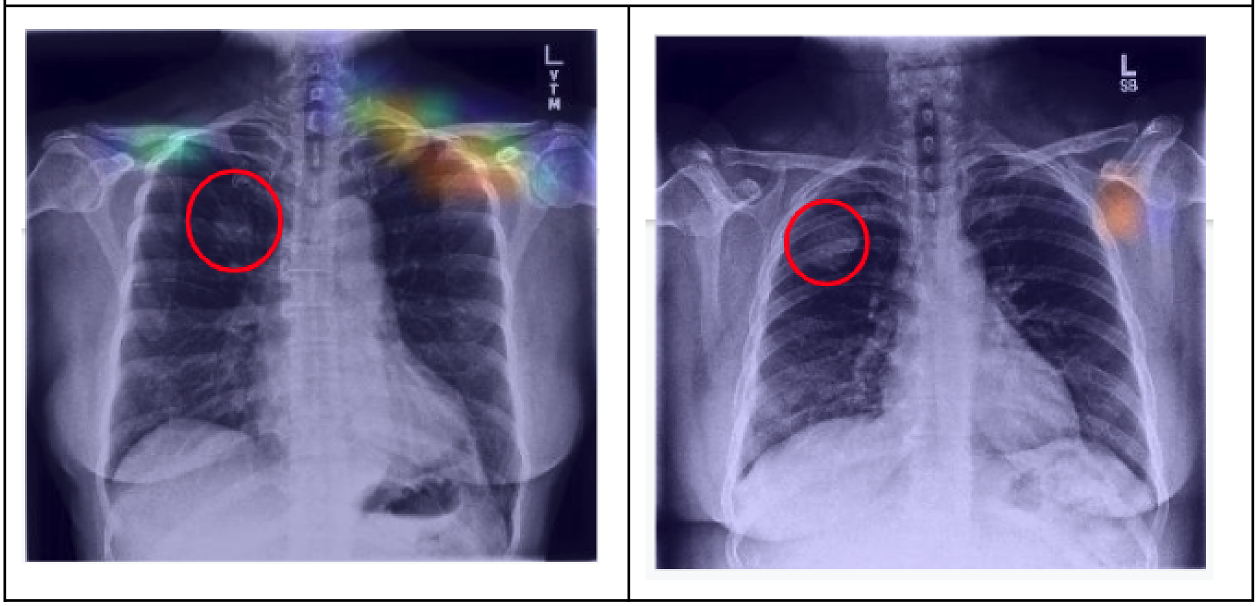


*Figure 1B Source: [3]

Figure 3: Grad-CAM visualizations to interpret model focus. The red circles indicate true areas of lung lesion while the rainbow heatmap (orange = high) indicates pixels the model used for classification.



References

[1] “Pulmonary Nodules and Lung Lesions: Condition: UT Southwestern Medical Center.” UT Southwestern Medical Center, utswmed.org/conditions-treatments/pulmonary-nodules-and-lung-lesions/.

[2] Ramesh AN, Kambhampati C, Monson JR, Drew PJ. Artificial intelligence in medicine. *Ann R Coll Surg Engl*. 2004;86(5):334-338. doi:10.1308/147870804290

[3] “CheXpert: A Large Dataset of Chest X-Rays and Competition for Automated Chest X-Ray Interpretation.” CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison, stanfordmlgroup.github.io/competitions/chexpert/.

[4] Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghgoo, B., Ball, R., Shpanskaya, K., Seekins, J., Mong, D. A., Halabi, S. S., Sandberg, J. K., Jones, R., Larson, D. B., Langlotz, C. P., Patel, B. N., Lungren, M. P., & Ng, A. Y. (2019). CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison. Proceedings of the AAAI Conference on Artificial Intelligence, 33(01), 590-597. https://doi.org/10.1609/aaai.v33i01.3301590

[5] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” arXiv e-prints, 2014

[6] Huang G, Liu Z, Van Der Maaten L, Weinberger, K. Q. Densely connected convolutional networks Proceedings of the IEEE conference on computer vision and pattern recognition, 2017: 4700–4708

[7] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 618-626, doi: 10.1109/ICCV.2017.74.

[8] J. Deng, W. Dong, R. Socher, L. Li, Kai Li and Li Fei-Fei, "ImageNet: A large-scale hierarchical image database," *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248-255, doi: 10.1109/CVPR.2009.5206848.

[9] Potchen, EJ, Gard, JW, Lazar, P, Lahaie, P, and Andary, M. Effect of clinical history data on chest film interpretation-direction or distraction. *Investigative Radiology*, 1979 pp. 404–404.

[10] Berbaum, K, Franken Jr, EA, and Smith, WL. The effect of comparison films upon resident interpretation of pediatric chest radiographs. *Investigative Radiology*, 1985, pp. 124–128. 1985.

Member Contributions

Eric: identified dataset, contributed to idea brainstorming, programmed sections of data cleaning, data subsetting, data augmentation, data sampling, training data augmentation, created random forest models, performed Grad-CAM analysis, edited final report

Lara: contributed to idea brainstorming, conducted background research on lung lesions and clinical applications of diagnostic ML programs, contributed to early attempts of programming baseline models, compiled slidedeck for presentation, participated in oral presentation, drafted and edited final report.

Daniel: Trained VGG and DenseNet models and conducted hyperparameter tuning for all sets of training data. Performed some data filtering steps i.e. for cleaning up ambiguous negative

samples for lung lesion and for removing lateral images. Assisted and editing writing final report i.e. limitations and challenges portion of final report as well as references/in-text citations.
Abbie: Contributed to idea brainstorming and hypothesis of results. Drafted the result and discussion sections of the final report. Programmed early data loading steps, preprocessing, error auditing i.e. analyzed false positive instances