Deep Learning Approaches for NIH Chest X-Ray Classification

Aysha Allahverdiyeva aa7943@nyu.edu
Courant, Center for Data Science
New York University

Melina Tsotras mmt515@nyu.edu
Center for Data Science
New York University

Moujan Saderi ms15516@nyu.edu
NYU Langone Health
New York University

Abstract

This project explores deep learning approaches for multi-label classification of thoracic pathologies using the NIH ChestX-ray14 dataset, which comprises over 120,000 chest radiographs across 14 disease categories and one "No Finding" label. To establish a strong baseline, we trained a convolutional neural network (CNN) from scratch and incorporated patient metadata (age, sex, and view position) alongside image features. We compared this model against fine-tuning approaches using ResNet-50 and a Vision Transformer (BEiT) pretrained with BERT-style masked image modeling. Performance was evaluated using metrics such as F1-score, AUC-ROC, and exact match ratio. Results indicated that incorporating metadata, applying log-scaled loss weights to address class imbalance, and tuning prediction thresholds substantially improved classification performance, especially for rare conditions. While the BEiT model showed promise in leveraging self-supervised pretraining, ResNet-50 with loss reweighting and threshold tuning yielded the best overall results. We conclude with recommendations for future work, including using more balanced dataset splits, medically-informed augmentations and advanced regularization techniques.

Keywords: Chest X-ray, Deep learning, Medical imaging, CNN, VIT, ResNet

1 Introduction

Chest radiography is one of the most frequently utilized and cost-effective diagnostic imaging techniques in clinical medicine. However, interpreting chest X-rays remains a challenging task even for experienced radiologists due to subtle visual features associated with many thoracic conditions. The complexity of diagnosis is further compounded when multiple pathologies co-occur, necessitating robust methods for multi-label classification. Reviews on chest X-ray diagnoses estimate that errors made by radiologists may occur in approximately 4–30% of cases (Gefter et al., 2023). Additionally, as many physicians have large workloads and are often overworked, finding ways to speed up the diagnosis process is critical for timely intervention in life-threatening conditions, such as lung cancer (Gefter et al., 2023).

To address these challenges, deep learning, especially using convolutional neural networks (CNNs), has become a leading approach. CNNs can learn and extract important features from raw image data, enabling accurate detection and classification of multiple thoracic conditions simultaneously. These models have demonstrated high performance in chest X-ray analysis (Baltruschat et al., 2019), offering the potential to enhance diagnostic accuracy, speed up diagnosis timing and reduce clinician workload. Their scalability also makes them particularly valuable in low-resource settings or developing countries where expert radiologists are scarce (Zuhair et al., 2024).

Progress in this area has been accelerated by the release of large public datasets such as NIH ChestX-ray14 (Wang et al., 2017), CheXpert (Irvin et al., 2019), and MIMIC-CXR (Johnson et al., 2019). These resources provide hundreds of thousands of labeled chest X-rays, making it easier to train and evaluate deep learning models. CNNs have long been the standard for chest X-ray classification, with models like ResNet-50 leveraging residual blocks to allow training of deeper models (He et al., 2015) (Baltruschat et al., 2019).

However, recent developments in vision modeling have introduced transformer-based architectures, which model global image relationships using self-attention (Dosovitskiy et al., 2021). One such model is BEiT (Bidirectional Encoder representation from Image Transformers), (Bao et al., 2021), which uses BERT-style masked image modeling and has outperformed ResNet-50 on natural image benchmarks such as ImageNet(Deng et al., 2009). While other vision transformers have been successfully applied to medical image classification, our aim is to evaluate whether BEiT's robust pretraining method, originally developed for natural language processing, can effectively generalize to the highly specialized task of multi-label thoracic disease classification from chest radiographs.

In this work, we compare the performance of ResNet-50 and BEiT against a baseline CNN on the NIH ChestX-ray14 dataset to assess the effectiveness of vision transformers for medical image classification. Both models will be fine-tuned on the dataset, and we will explore a range of training strategies, hyperparameter settings, and optimization techniques to improve performance. Our goal is to understand how these newer models perform on multi-label diagnostic tasks and whether they offer meaningful advantages over traditional CNNs in clinical imaging contexts. We hypothesize that BEiT may ultimately achieve superior performance in classifying chest X-rays, due to its larger number of parameters and use of an attention mechanism.

2 Methods

2.1 Data

The NIH ChestX-ray14 dataset is a publicly available collection of 112,120 anonymized chest X-ray images sourced from over 30,000 unique patients (Wang et al., 2017), available at `https://nihcc.app.box.com/v/ChestXray-NIHCC`. Each image is annotated with either 'No Finding' or one or more of 14 common thoracic disease labels, mined from radiology reports using natural language processing techniques. The dataset includes essential metadata such as patient age, gender, view position, and follow-up numbers, facilitating comprehensive analysis. It significantly extended prior datasets by including more disease categories and more samples, making it a valuable benchmark for evaluating computer-aided diagnostic models. Although the image labels are text-mined from radiology reports, with a reported labeling accuracy of over 90%, the dataset remains one of the most realistic and representative resources available for thoracic pathology classification research.

We used the published splits of a test and val/train set within our work. Following this, we split the val/train set into separate train and validation sets, at the patient level to prevent data leakage. We chose to use the established test set to enable comparison with prior work, although we discovered later that the test set may be suboptimal due to a higher prevalence of all diseases compared to the training set (Appendix D, Table 4). The proportion of the data in each set was as follows: Train (65.67%), Validation (11.50%), Test (22.83%).

For all models, several image augmentation strategies were employed in the training set only, which were applying small-angle rotations within ± 10 degrees, random brightness and contrast jittering. Age, Sex and View Position metadata was incorporated into each model. For metadata pre-processing, age was z-scored based on the train set mean and standard deviation. Sex and View Position were one-hot encoded.

2.2 Model Selection

Baseline CNN: To establish a baseline against more complex models, we implemented a CNN from scratch in PyTorch, trained without pretrained weights and tailored to the NIH ChestX-ray14 dataset. The architecture included five convolutional blocks with increasing channels (32→512) and adaptive average pooling. Metadata features (n=3) were concatenated with the 512-D feature vector prior to classification. The model was trained with BCELoss and Adam (lr=0.001) for 20 epochs.

This design allowed us to isolate the contribution of architectural depth and metadata fusion without the influence of pretrained representations. The increasing number of filters in the convolutional layers aimed to capture hierarchical image features, from low-level edges to high-level pathology patterns, while batch normalization and ReLU activations after each convolution stabilized training and encouraged non-linearity. Max pooling layers progressively reduced spatial resolution to control computational complexity and increase receptive field. Dropout before the final layer was added to mitigate overfitting, especially given the model's relatively small parameter count compared to pretrained architectures. Including patient metadata directly into the classifier was intended to test whether simple feature concatenation would enhance prediction, without needing more sophisticated mechanisms.

ResNet-50: The second model was built around a ResNet-50 model, pretrained on the ImageNet dataset, as proposed in (Hsu et al., 2017). To address the imbalanced classes, two options for weighting the Binary-Cross Entropy loss were tested: linear-scaled and log-scaled inverse class frequency (Phan and Yamamoto, 2020). We hypothesized that the raw weights would lead to improved recall at the cost of lower precision (due to the model over-prediction of positive cases), while the log-scaled version would better balance these two metrics, and be less sensitive to distribution shifts in the prevalence of positive cases.

The utility of threshold tuning in addressing the class imbalance was also tested. This method was of interest as it has been proposed for cost-sensitive classification, and incorporating real-world misclassification cost is critical in AI-diagnostic applications (Araf et al., 2024). The specific tuning method used was label-wise optimal threshold choice (Alotaibi and Flach, 2021): the default probability threshold of 0.5 for binary classification was tuned for each label on the validation dataset to maximize F1-score. These tuned thresholds were then used during inference.

Options for preparing input images for processing by the pretrained ResNet model were also compared: for optimal performance, images needed to be resized to 224×224 pixels, with values in 3 channels similar to the RGB values in the ImageNet dataset. The first method was direct resizing/downsampling of images, before replicating grayscale pixel values across three channels, and normalizing to the expected mean and standard deviation for each channel. The second method was designed to capture high-resolution information from the 1024×1024 images, given the hypothesis that this information may improve detection of different pathologies. In this version of the model, three convolutional layers were added at the input, to produce an embedding of size 224×224 with 3 channels.

BEiT (BERT Pre-training for Image Transformers): BEiT is a vision transformer pretrained using masked image modeling, similar to BERT (Bao et al., 2021). During pretraining, the image undergoes two transformations: one processed by a pretrained discrete variational encoder to generate an 'image vocabulary' and another divided into flattened patches with 1D positional embeddings to be used as input. The encoder learns to predict the image vocabulary from the input flattened patches, enabling robust feature learning. For this project, we use the BEiT architecture that has been pretrained on ImageNet (Deng et al., 2009). The standard BEiT processor was used to transform the images into input that can be accepted by BEiT, by down-sampling, normalizing, extracting and flattening image patches, and adding a positional embedding. BEiT uses standard image transformer architecure to encode each reduced image patch (Bao et al., 2021). Mean pooling was performed on the hidden layer and this was inputted, along with metadata representations, into a classification layer for multi-label classification.

Several strategies were employed to boost BEiT performance. While all BEiT models used the Adam optimizer, we compared the effects of using a learning rate scheduler with linear warmup (10% of total training steps) with learning rate starting at 3e-4 versus no scheduler with learning rate at 1e-4. We also evaluated the impact of log-weighting the loss compared to using unweighted loss. Transfer learning was used as a baseline to confirm that fine-tuning contributed to performance gains. Additionally, we experimented with different approaches for incorporating metadata: (1) directly appending the processed metadata to the mean-pooled hidden state, and (2) passing metadata through a multilayer perceptron (MLP) before appending it, as we hypothesized that it may allow potentially richer feature representations. Finally, we applied threshold tuning on the validation set as a post-processing step to adjust class probability cut-offs and further enhance performance.

2.3 Model Evaluation

In order to evaluate and compare model performances, the following metrics were calculated on classifier results for the test set: macro-averaged recall, precision, F1-scores, and AUROC (Area Under the Receiver Operating Characteristic curve, calculated per label), and the Exact Match Ratio (capturing the ratio of cases where all labels were correctly predicted.)

3 Results

Baseline CNN

The performance of the baseline CNN classifier on the test set is summarized in Table 3, and the label-wise performance is illustrated in the ROC curves in Figure 1a. We found that the CNN was predicting all zeros for the labels Consolidation, Fibrosis, Hernia, Pleural Thickening, and Pneumonia.

ResNet-50

A summary of comparison experiments using the ResNet-50 backbone are shown in Table 1, with the resulting ROC curves for the best model based on macro-AUC (Log Weights + Conv Input Layers) illustrated in Figure 1b. All results are after threshold tuning (see Appendix E).

Metric	Linear Weights	Log Weights	Log Weights + Conv Input Layers
Recall (macro)	0.3875	0.4150	**0.4266**
Precision (macro)	0.1821	**0.3049**	0.2866
F1-score (macro)	0.2236	0.3103	**0.3253**
AUC (macro)	0.6912	0.7745	**0.7800**
Exact Match Ratio	0.0360	**0.2130**	0.1969

Table 1: Evaluation metrics for best ResNet-50 model

BEiT

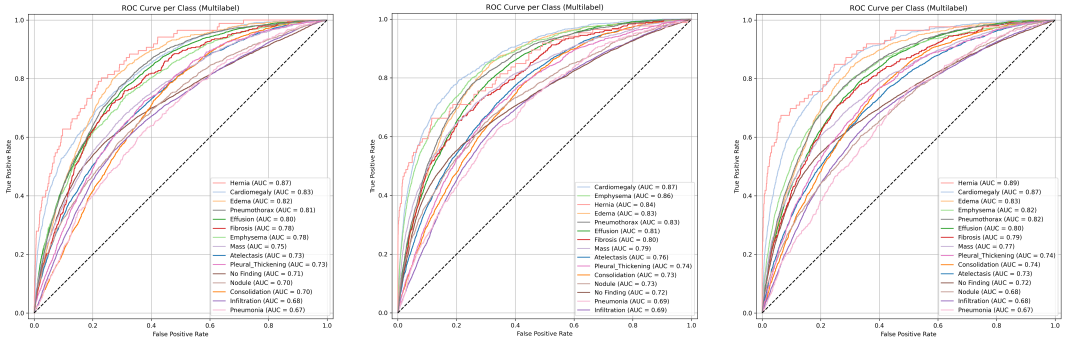
Results of BEiT optimization techniques are shown in Table 2 and ROC curves of all labels are shown in Figure 1c.

Parameters/Metric	Model 1	Model 2	Model 3	Model 4	Model 5
Scheduler with Linear Warmup	No	Yes	No	Yes	Yes
Finetune (FT) or Transfer Learning (TL)	TL	FT	FT	FT	FT
Metadata (CD=Concated Directly, MLP=Encoded First)	CD	CD	CD	CD	MLP
Loss Weighting Method	Log	Unweighted	Log	Log	Log
Recall (macro)	0.5263	0.0334	0.4978	**0.6974**	0.6664
Precision (macro)	0.0967	**0.3749**	0.0959	0.1799	0.1748
F1 Score (macro)	0.0960	0.0545	0.0817	**0.2678**	0.2624
AUC-ROC (macro)	0.5018	0.7500	0.4941	**0.7701**	0.7657
Exact Match Ratio	0.0000	0.0012	0.0000	0.0684	**0.0780**

Table 2: Evaluation of different BEiT model configurations and their performance, before threshold-tuning.

Model Comparison

Table 3 summarizes the performance of the best ResNet and BEiT models following the experiments presented above, as compared to the baseline CNN.



(a) Baseline CNN (b) Best ResNet Model (c) Best BEiT Model

Figure 1: Per-label ROC curves and AUC values for best performing models of each type.

Metric	Baseline (CNN)	ResNet	BEiT (Vision Transformer)
Recall (macro)	0.0753	**0.4266**	0.4236
Precision (macro)	**0.3110**	0.2866	0.2707
F1-score (macro)	0.1029	**0.3253**	0.3122
AUC (macro)	0.7581	**0.7800**	0.7701
Exact Match Ratio	**0.2204**	0.1969	0.1801

Table 3: Test Set Performance Metrics. ResNet and BEiT metrics are shown after threshold tuning was applied. The CNN has no threshold tuning applied.

4 Discussion

We hypothesized that BEiT would achieve superior performance, due to its larger number of parameters and use of an attention mechanism. However, our results demonstrate that given time and resource constraints, the longer training time associated with this model is a limitation. In our experiments, the fine-tuned ResNet-50 was able to be trained for 20 epochs at the time of evaluation, as compared to 9 for the vision transformer, and yielded slightly better classification performance.

4.1 Model Optimization

For ResNet-50, from the comparison of alternative model designs, the results demonstrate that the best classifier uses log-scaled weights, convolutional layers at the input prior to the pretrained ResNet, and label-wise optimal threshold choice tuning. As expected, log-scaled weights achieved a better balance between recall and precision. However, the addition of input convolutional layers to capture high-resolution image features did not significantly improve performance.

Out of BEiT training strategies, we found that using log-weights, a scheduler with linear warmup, and metadata being appended directly to the hidden state, increased performance the most. Due to time and resource constraints, we could not train BEiT to the extent that we hoped to, which potentially explains its lower than expected performance. This is particularly true in comparing metadata appending techniques, where we expected that having an MLP learn a metadata embedding would produce the strongest model, but did not.

4.2 Model Comparison

The CNN trained from scratch served as a baseline for this study, highlighting the challenges inherent in multi-label chest X-ray classification without leveraging pretrained features. Notably, the baseline

CNN had the best precision (0.3109) and exact match ratio (0.2204) across all the models, even without any threshold-tuning being applied. However, we believe that this is due to the CNN not learning how to predict labels with low prevalence well, as we see that the CNN predicted all zeros for the labels Consolidation, Fibrosis, Hernia, Pleural Thickening, and Pneumonia. The low Recall (0.0753) confirms this, showing that the model had poor performance in discovering all positive cases. These results reflect both class imbalance in the dataset and the limitations of training deep models from scratch without access to large-scale supervision or advanced regularization techniques.

Overall, the fine-tuned ResNet-50 outperformed the baseline CNN and fine-tuned BEiT, although BEiT's performance was pretty comparable. We believe that this is because ResNet is able to take advantage of a deeper model with residual blocks, allowing it to be tuned efficiently for the multi-label classification problem, without having too many trainable parameters like BEiT that require longer training times to fit well. ResNet-50 had the greatest Recall (0.4266), F1-score (0.3253) and AUC (macro) (0.7800). Both BEiT and ResNet-50 had higher recall than the CNN, indicating that they were better at identifying positive classes, but had lower precision than the CNN which indicates they often predicted False cases as True. Of note, neither the BEiT nor ResNet-50 predicted all zeroes for any single label. Their higher macro AUC shows they predicted more true positives over false positives in comparison to the CNN on average. All models showed they predicted better than random chance, indicating that training did occur.

4.3 Limitations and Future Work

First, we believe with more time and greater resources, we may see that the vision transformer will be better able to fit its much larger amount of trainable parameters (BEiT: 86M vs. ResNet-50: 25M) for the classification task. Second, the use of input convolutional layers to create smaller feature embeddings would benefit from pretraining using an autoencoder. Another limitation in our work came from the NIH ChestX-ray14 test dataset, which had proportionally more disease cases in comparison to the train/val sets (Appendix D). Given that the observed proportion of positive cases was used to inform training loss weights, the distribution shift reduces performance on the held-out test set. This could be why the baseline CNN, which has fewer parameters and may be less prone to overfitting and did not use log-weighted loss, was able to perform well. This is especially in regards to the baseline CNN's Exact Match Ratio, which outperformed ResNet and BEiT without using threshold tuning. To continue this work, we could explore using other newer chest x-ray datasets, such as CheXpert or MIMIC-CXR, or establishing our own more balanced test/val/train split. We can also aim to incorporate more regularization techniques such as domain-specific data augmentation strategies (Kora Venu and Ravula, 2021).

Finally, some labels occurred rarely, and likely weren't learned well. We may consider using focal loss (Lin et al., 2020), which would down-weight the loss of easily classified images and focus more on the harder images, or upsampling positive cases in these classes during training, while applying stronger data augmentation. It would also be important to incorporate domain knowledge to understand whether the few positive examples are representative of the typical disease presentation, or whether images features indicative of the condition can vary greatly (in which case, it would be advantageous to include more data).

5 Conclusion

Overall, this work provides a strong foundation for understanding the capabilities and limitations of different deep learning architectures and training strategies for performing multi-label image classification of thoracic pathologies, using the NIH ChestX-ray14 dataset. It also provides several avenues for continued model refinement, to address challenges arising from class imbalance and distribution shifts in diagnostic radiology applications.

Appendix A. Contribution Statements

Aysha Allahverdiyeva implemented the Baseline-CNN model, and wrote the initial draft of the paper. Melina Tsotras implemented and evaluated the BEiT models, wrote corresponding portions of the paper, and contributed to the introduction and discussion sections. Moujan Saderi implemented and evaluated the ResNet-50 models, wrote corresponding portions of the paper, and contributed to the discussion section.

Appendix B. Data Availability

The ChestX-ray14 dataset used is publicly available at `https://nihcc.app.box.com/v/ChestXray-NIHCC`.

Appendix C. Code Availability

All training/evaluation scripts and saved fine-tuned models are available on Github at `https://github.com/mtsotras/dl4med-chestxray`

Appendix D. Dataset Exploration

Condition	**Percentage of positive samples in Training Dataset**	**Percentage of positive samples in Validation Dataset**	**Percentage of positive samples in Testing Dataset**
Atelectasis	9.43	10.34	12.81
Cardiomegaly	1.99	1.90	4.18
Consolidation	3.26	3.51	7.09
Edema	1.61	1.47	3.61
Effusion	9.95	10.33	18.20
Emphysema	1.63	1.74	4.27
Fibrosis	1.44	1.50	1.70
Hernia	0.17	0.12	0.34
Infiltration	15.94	15.84	23.88
Mass	4.68	4.56	6.83
No Finding	58.48	57.74	38.53
Nodule	5.46	5.33	6.34
Pleural Thickening	2.58	2.67	4.47
Pneumonia	1.02	0.96	2.17
Pneumothorax	3.02	3.21	10.41

Table 4: Precentage of positive samples for each label, in training, validation, and test datasets

As illustrated in the above tables, the prevalence of different conditions varied between the training and the test set (which were provided in the original dataset, not created by us).

Appendix E. ResNet-50 Model: Threshold Tuning Experiments

The use of threshold tuning was found to improve F1-scores for all models, by better balancing Recall and Precision. An example of the thresholds calculated for the final model (Table 7) are as follows; Atelectasis: 0.3782, Cardiomegaly: 0.2143, Consolidation: 0.2607, Edema: 0.3528, Effusion: 0.4476, Emphysema: 0.3746, Fibrosis: 0.2050, Hernia: 0.4538, Infiltration: 0.2348, Mass: 0.4285, No Finding: 0.2398, Nodule: 0.4444, Pleural Thickening: 0.3631, Pneumonia: 0.1632, Pneumothorax: 0.3587.

Metric	Default >0.5 probability threshold	Tuned Thresholds
Recall (macro)	0.7848	0.3875
Precision (macro)	0.1373	0.1821
F1-score (macro)	0.1904	0.2236

Table 5: Test set performance metrics for ResNet-50 model using Linear-scaled inverse class frequency weighting

Metric	Default >0.5 probability threshold	Tuned Thresholds
Recall (macro)	0.2571	0.4150
Precision (macro)	0.3106	0.3049
F1-score (macro)	0.2689	0.3103

Table 6: Test set performance metrics for ResNet-50 model using Log-scaled inverse class frequency weighting

Metric	Default >0.5 probability threshold	Tuned Thresholds
Recall (macro)	0.2802	0.4266
Precision (macro)	0.3219	0.2866
F1-score (macro)	0.2883	0.3253

Table 7: Test set performance metrics for ResNet-50 model using Log-scaled inverse class frequency weighting and convolutional input layers

References

Reem Alotaibi and Peter Flach. Multi-label thresholding for cost-sensitive classification. *Neurocomputing*, 436:232–247, May 2021. ISSN 0925-2312. doi: 10.1016/j.neucom.2020.12.004. URL https://www.sciencedirect.com/science/article/pii/S0925231220318853.

Imane Araf, Ali Idri, and Ikram Chairi. Cost-sensitive learning for imbalanced medical data: a review. *Artificial Intelligence Review*, 57(4):80, March 2024. ISSN 1573-7462. doi: 10.1007/s10462-023-10652-8. URL https://doi.org/10.1007/s10462-023-10652-8.

Ivo M. Baltruschat, Hannes Nickisch, Michael Grass, Tobias Knopp, and Axel Saalbach. Comparison of Deep Learning Approaches for Multi-Label Chest X-Ray Classification. *Scientific Reports*, 9 (1):6381, April 2019. ISSN 2045-2322. doi: 10.1038/s41598-019-42294-8. URL https://www.nature.com/articles/s41598-019-42294-8. Publisher: Nature Publishing Group.

Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. BEiT: BERT Pre-Training of Image Transformers, 2021. URL https://arxiv.org/abs/2106.08254. Version Number: 2.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021. URL https://arxiv.org/abs/2010.11929.

Warren B. Gefter, Benjamin A. Post, and Hiroto Hatabu. Commonly Missed Findings on Chest Radiographs: Causes and Consequences. *Chest*, 163(3):650–661, 2023. ISSN 0012-3692. doi: https://doi.org/10.1016/j.chest.2022.10.039. URL https://www.sciencedirect.com/science/article/pii/S0012369222040818.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition, 2015. URL https://arxiv.org/abs/1512.03385. Version Number: 1.

Joy Hsu, Peter Lu, and Kush Khosla. Predicting Thorax Diseases with NIH Chest X-Rays. Technical report, Department of Computer Science and Department of Mathematics Stanford University, 2017. URL https://cs229.stanford.edu/proj2017/final-reports/5224573.pdf.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, Jayne Seekins, David A. Mong, Safwan S. Halabi, Jesse K. Sandberg, Ricky Jones, David B. Larson, Curtis P. Langlotz, Bhavik N. Patel, Matthew P. Lungren, and Andrew Y. Ng. CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison, 2019. URL https://arxiv.org/abs/1901.07031. Version Number: 1.

Alistair E. W. Johnson, Tom J. Pollard, Seth J. Berkowitz, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih-ying Deng, Roger G. Mark, and Steven Horng. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data*, 6(1):317, December 2019. ISSN 2052-4463. doi: 10.1038/s41597-019-0322-0. URL https://www.nature.com/articles/s41597-019-0322-0. Publisher: Nature Publishing Group.

Sagar Kora Venu and Sridhar Ravula. Evaluation of Deep Convolutional Generative Adversarial Networks for Data Augmentation of Chest X-ray Images. *Future Internet*, 13(1):8, January 2021. ISSN 1999-5903. doi: 10.3390/fi13010008. URL https://www.mdpi.com/1999-5903/13/1/8. Number: 1 Publisher: Multidisciplinary Digital Publishing Institute.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal Loss for Dense Object Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(2): 318–327, February 2020. ISSN 1939-3539. doi: 10.1109/TPAMI.2018.2858826. URL https://ieeexplore.ieee.org/document/8417976.

Trong Huy Phan and Kazuma Yamamoto. Resolving Class Imbalance in Object Detection with Weighted Cross Entropy Losses, June 2020. URL http://arxiv.org/abs/2006.01413. arXiv:2006.01413 [cs].

Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M. Summers. ChestX-Ray8: Hospital-Scale Chest X-Ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3462–3471, July 2017. doi: 10.1109/CVPR.2017.369. URL https://ieeexplore.ieee.org/document/8099852. ISSN: 1063-6919.

Varisha Zuhair, Areesha Babar, Rabbiya Ali, Malik Olatunde Oduoye, Zainab Noor, Kitumaini Chris, Inibehe Ime Okon, and Latif Ur Rehman. Exploring the Impact of Artificial Intelligence on Global Health and Enhancing Healthcare in Developing Nations. *Journal of Primary Care & Community Health*, 15:21501319241245847, 2024. doi: 10.1177/21501319241245847. URL https://doi.org/10.1177/21501319241245847. _eprint: https://doi.org/10.1177/21501319241245847.