# CNNs for Image Classification of Chest X-Rays

Quinlin Neuhaus
MATH 5001

May 12, 2026

## 1 Introduction

Chest X-rays are one of the most common diagnostic tools available in modern hospitals and are used in the diagnosis of countless conditions and diseases. Radiologists look at X-rays and diagnose patients everyday, but are not perfect. Misclassified X-rays can lead to dangerous unnecessary treatments or missed conditions. Human radiologists take years to learn their craft and are prone to human flaws like bias, fatigue, and ego. AI/ML, specifically CNNs, can be used in tandem or independently of human radiologists to classify and diagnose X-rays and tackle these pitfalls. In this project, I create and train CNNs for image classification of chest X-ray images and evaluate their performance against existing models in the literature.

## 2 Dataset

The dataset I am using is the NIH Chest X-Ray Dataset. It contains 112,120 black and white images of chest X-rays from 30,805 unique patients. Each image is labeled as "No Finding" or with any number of 14 possible conditions: Atelectasis, Consolidation, Infiltration, Pneumothorax, Edema, Emphysema, Fibrosis, Effusion, Pneumonia, Pleural Thickening, Cardiomegaly, Nodule Mass, and Hernia. Some tabular metadata is also available for each image including the number of follow up visits, the patient's age and gender, and the view position of the X-ray. Class labels are expected to be $> 90\%$ accurate, although some erroneous labeling is possible.

This dataset is interesting from a machine learning perspective for a number of reasons. The shear size of the dataset makes many larger architectures feasible to use, and provides plenty of training data for AI/ML purposes. This dataset is not simply multiclass, but rather

multilabeled, and many samples have multiple conditions present simultaneously. This is uncommon in traditional classification and a unique element to the dataset. The addition of tabular metadata also provides directions for ML outside of just image classification. The dataset's labels are extremely unbalanced as well, which is another challenge to tackle when modeling. This heavy imbalance is shown below.

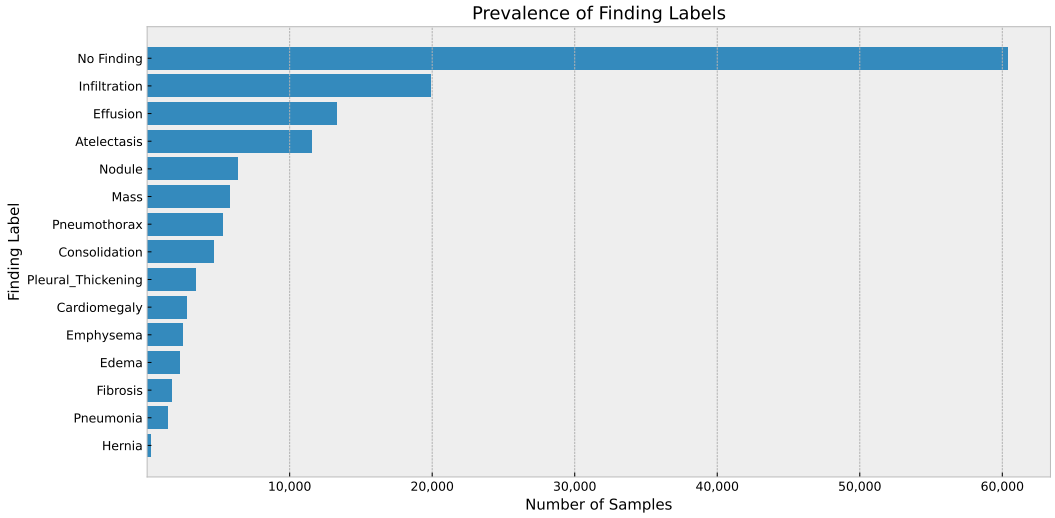


Figure 1: Distribution of Class Labels

Patients in the dataset represent a large swath of ages, with the most patients being middle aged, and few children or elderly patients.

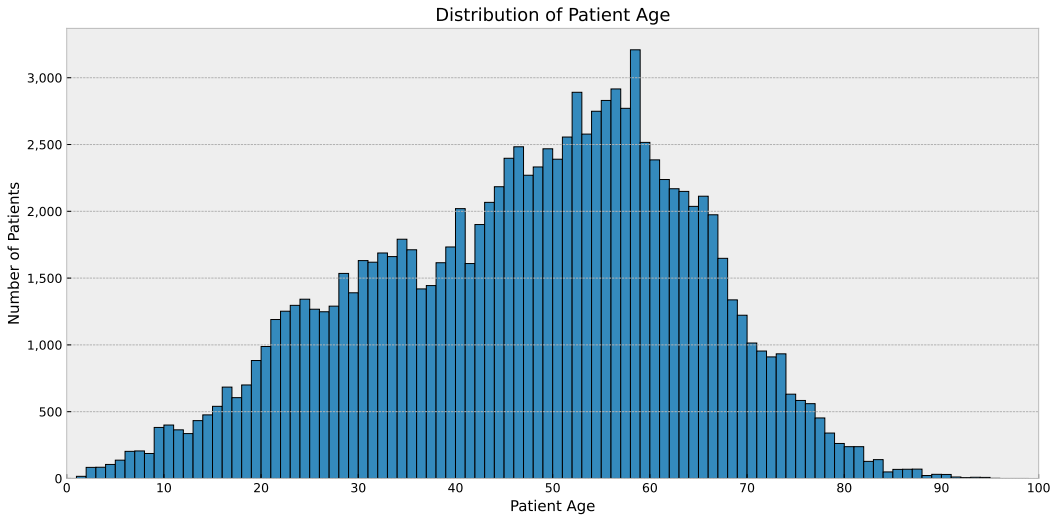


Figure 2: Patient Age Distribution

Males are overrepresented in the dataset as compared to females. Posterior anterior(PA) X-rays contribute to 60% of the images, and anterior posterior(AP) X-rays make up the remaining 40%. PA X-rays are considered the gold standard for medical imaging and are taken with the patient facing the detector, AP X-rays are taken with the back to the detector and commonly used for patients lying down.

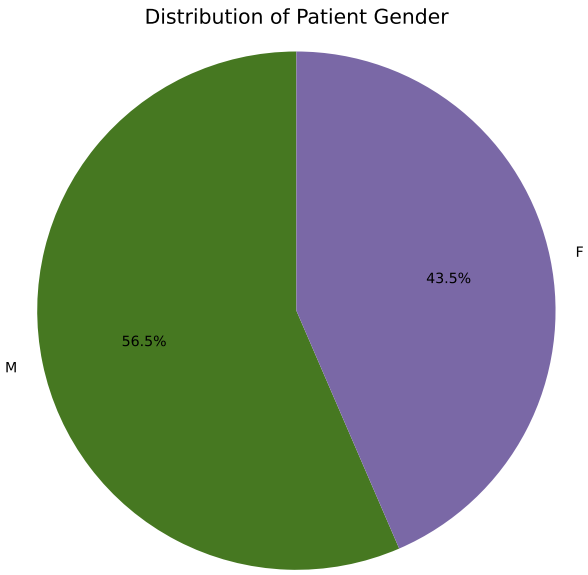


Figure 3: Patient Gender Distribution

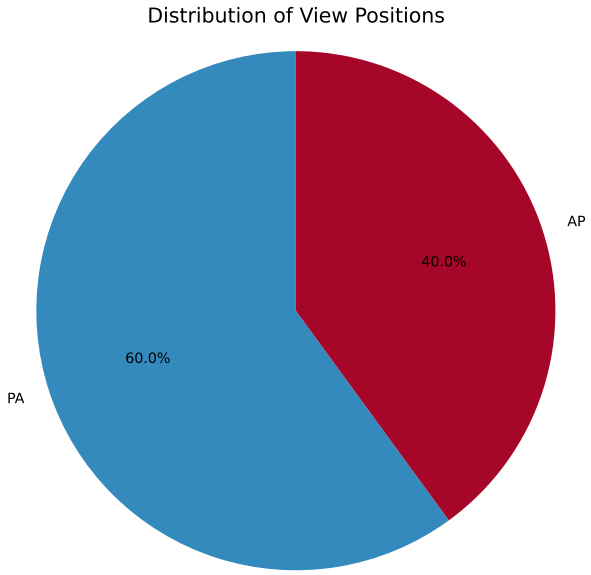


Figure 4: X-Ray View Distribution

# 3 Models

Since I am attempting to work with images, I chose CNNs as the tool to train on this data. I built two similar models: a traditional CNN that only takes in the image data, and a multimodal CNN that also uses patient metadata for classification.

## 3.1 Data Processing

The dataset I am working with is extremely large(45GB), for this reason, I will downsize my images from 1024x1024 to 224x224. This is the same size images that are used in the existing literature for this dataset. Images were also normalized according to the mean and standard deviation of this dataset instead of ImageNet statistics. I will keep patient age as a continuous predictor and standardize it, and one-hot encode the gender and view position variables. I debated augmenting the image data by adding images that were horizontally flipped or slightly rotated versions of existing images, but decided against this because of the

dataset's size and the fact that these human images are not symmetric, i.e. a horizontally flipped chest X-ray is meaningfully different than the standard view.

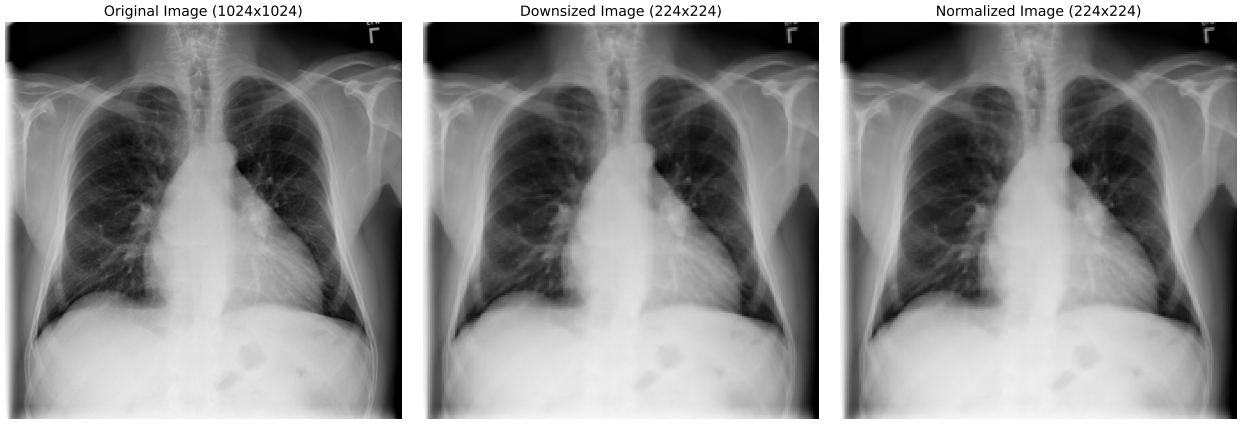


Figure 5: Image Resizing and Normalization

Because the 224x224 resolution is large, and the black and white images are all very similar, it is difficult to see the downsized and normalized transformation.

## 3.2 Class Imbalance

As previously mentioned, the dataset is very imbalanced, which if left as is could cause small classes to be ignored by my models. To alleviate this issue, the models' loss will be binary cross entropy weighted by the reciprocal of the class's prevalence in the dataset. This ensures that the models' cannot ignore small classes.

## 3.3 Architecture

The architecture of the two models was inspired by architecture in class provided code for CNNs. Both models feature four convolutional layers, each with: two rounds of convolutions of kernel size 3 and padding 1, batch norm, and ReLu activation, followed by maxpool and dropout layers. The final layer is a 256 neuron wide flat layer, which outputs to sigmoid. In the multimodal CNN, 32 additional neurons are added to the final layer for the extra metadata features. It is important to note that the output is not softmax, but rather 14 individual sigmoids. This is because the dataset is multilabeled and needs to output an individual probability for each class independently.

### 3.4 Training

I have access to the dataset through Kaggle, and used Kaggle's GPUs for training through batch jobs, avoiding downloading 45GB of data locally. Both models were trained on twin Nvidia Tesla 4s, and took about 10hrs each. No transfer learning from existing CNN model weights was used, nor transfer learning from one model to another, both models were trained from scratch.

An 80/20 Train/Test split was used, but because each patient can have multiple X-rays, it was ensured that an individual patient only ended up in one split. In total, there were 89,826 images across 24,644 patients in the training split and 22,294 images across 6,161 patients in the test split. To ensure fair comparison between the two models, the same split was used in the training and evaluation of both.

Both models were trained for 25 epochs with a batch size of 32. The Adam optimizer was chosen with learning rate 0.001. Dropout rates in the convolutional and final layers were set to .25. Training and test loss were recorded as the model trained. AUC of the ROC was also recorded per class per epoch to evaluate the models' performance on the large number of imbalanced classes.

## 4 Results

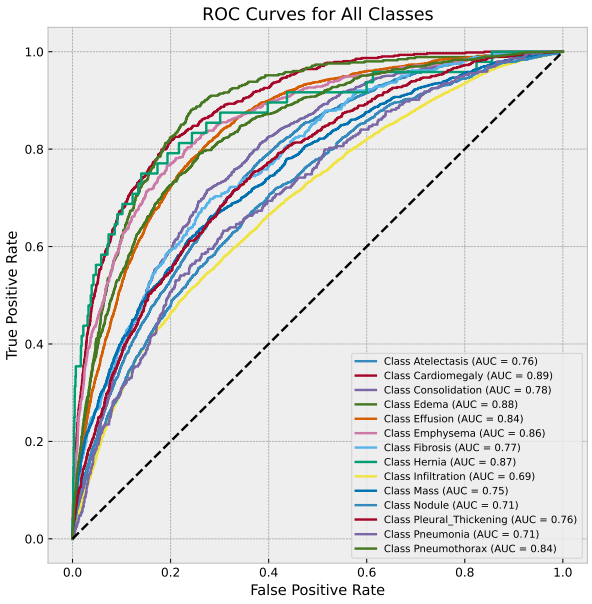


Figure 6: Pure CNN ROC by Class

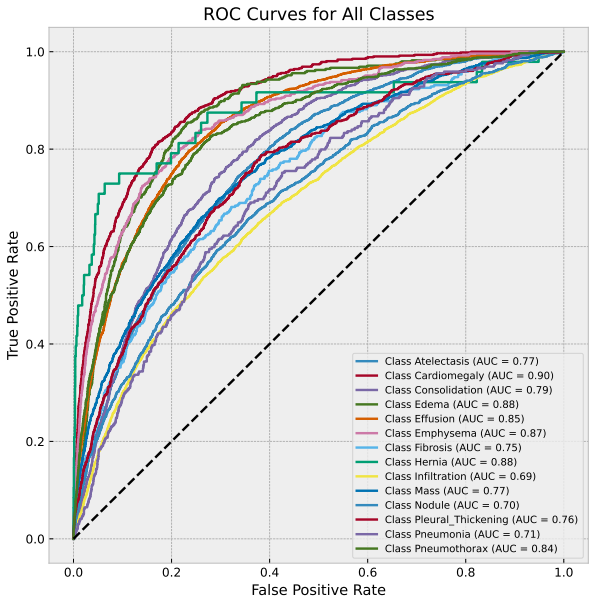


Figure 7: Multimodal CNN ROC by Class

## 4.1 Evaluation

Both models performed well, and for the models' relatively small size and training resources, greatly exceeded my expectations. The traditional and multimodal CNNs had macroaverage AUC of 0.794 and 0.799 respectively, representing good predictive power. Classes like Cardiomegaly, Edema, and Hernia had the best discrimination, while classes like Infiltration, Nodule, and Pneumonia performed worse. During training, training loss consistently decreased while test loss initially decreased and then plateaued.

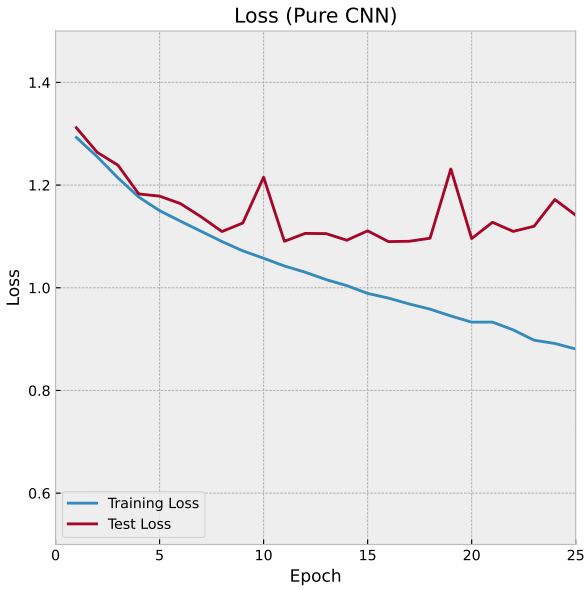


Figure 8: Pure CNN Loss

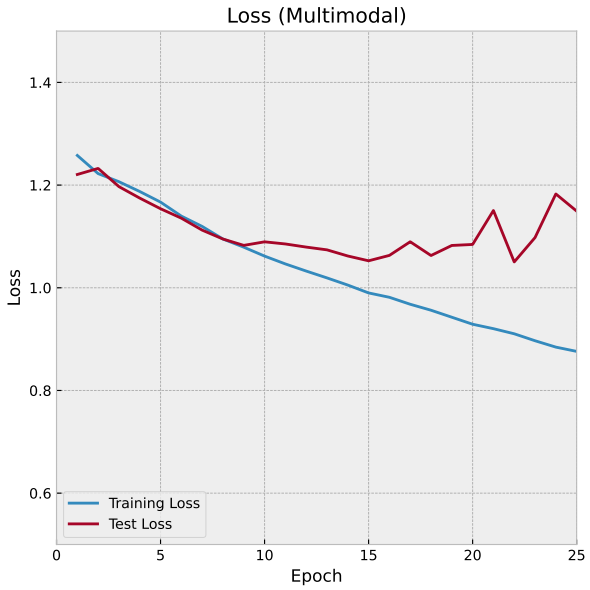


Figure 9: Multimodal CNN Loss

## 4.2 Comparison

Comparing the traditional CNN to the multimodal CNN shows that the models' performance is very similar. On average, the multimodal CNN slightly edges out the traditional one, but on a per-class level it is not clear which model wins. Neither model I created beats out all the other models in the existing literature at any specific class. However, as a whole my models are median performers among those in the literature, which represents a slew of different NN architectures created for this dataset in the last 10 years.

| Label | Yao [8] | Wang [6] | Shen [5] | Güendel [2] | Yan [7] | Baltruschat [1] | Kufel [3] | Rajpurkar [4] | Multi | CNN |
|---|---|---|---|---|---|---|---|---|---|---|
| Atelectasis | 0.733 | 0.700 | 0.766 | 0.767 | 0.792 | 0.763 | **0.817** | 0.809 | 0.772 | 0.761 |
| Cardiomegaly | 0.856 | 0.810 | 0.801 | 0.883 | 0.881 | 0.875 | 0.911 | **0.925** | 0.901 | 0.890 |
| Effusion | 0.806 | 0.759 | 0.797 | 0.828 | 0.842 | 0.822 | **0.879** | 0.864 | 0.854 | 0.842 |
| Infiltration | 0.673 | 0.661 | **0.751** | 0.709 | 0.710 | 0.694 | 0.716 | 0.735 | 0.689 | 0.689 |
| Mass | 0.777 | 0.693 | 0.760 | 0.821 | 0.847 | 0.820 | 0.853 | **0.868** | 0.767 | 0.751 |
| Nodule | 0.724 | 0.669 | 0.741 | 0.758 | **0.811** | 0.747 | 0.771 | 0.780 | 0.709 | 0.712 |
| Pneumonia | 0.684 | 0.658 | **0.778** | 0.731 | 0.740 | 0.714 | 0.769 | 0.768 | 0.714 | 0.709 |
| Pneumothorax | 0.805 | 0.799 | 0.800 | 0.846 | 0.876 | 0.840 | **0.898** | 0.889 | 0.845 | 0.838 |
| Consolidation | 0.711 | 0.703 | 0.787 | 0.745 | 0.760 | 0.749 | **0.815** | 0.790 | 0.792 | 0.783 |
| Edema | 0.806 | 0.805 | 0.820 | 0.835 | 0.848 | 0.846 | **0.908** | 0.888 | 0.876 | 0.881 |
| Emphysema | 0.842 | 0.833 | 0.773 | 0.895 | **0.942** | 0.895 | 0.935 | 0.937 | 0.868 | 0.859 |
| Fibrosis | 0.743 | 0.786 | 0.786 | 0.818 | **0.833** | 0.816 | 0.824 | 0.805 | 0.754 | 0.775 |
| Pleural Thickening | 0.724 | 0.684 | 0.759 | 0.761 | 0.808 | 0.763 | **0.812** | 0.806 | 0.758 | 0.758 |
| Hernia | 0.775 | 0.872 | 0.748 | 0.896 | 0.934 | **0.937** | 0.890 | 0.916 | 0.881 | 0.867 |
| Average | 0.761 | 0.745 | 0.775 | 0.807 | 0.830 | 0.727 | **0.843** | 0.841 | 0.799 | 0.794 |

Table 1: Comparison of AUC Values on Existing Models Across Classes

# 5 Conclusion

Overall, this project demonstrates the power that relatively small CNN architectures have in image classification. CNNs like these could improve the accuracy and speed of medical diagnoses, mitigate human problems in medicine, or allow X-rays to be performed at lower radiation levels. The models I trained exhibit good predictive power as multilabel classifiers, and hold their ground against other architectures.

Performance is limited by incorrect labels, patients' jewelry, body piercings, and image artifacts in the data. Perhaps the biggest limitation is the human labeling of the images itself. Any model created from this dataset is evaluated against the ground truth of the labels, which can only be as good as the radiologists in charge of labeling.

# References


[1] Baltruschat, I. M., et al. Comparison of deep learning approaches for multi-label chest X-ray classification. *Scientific Reports*, 9(6381), 2019.

[2] Gündel S., et al. Learning to Recognize Abnormalities in Chest X-rays with Location-Aware Dense Networks. *CIARP 2018*, 2019, pp. 757-765.

[3] Kufel J., et al. Multi-Label Classification of Chest X-ray Abnormalities Using Transfer Learning Techniques. *Journal of Personalized Medicine*, 13(10), 2023.

[4] Rajpurkar, P., et al. CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning. *arXiv preprint arXiv:1711.05225v3*, 2017.

[5] Shen, Y., & Gao, M. Dynamic routing on deep neural network for thoracic disease classification and sensitive area localization. *Machine Learning in Medical Imaging*, 2018, pp. 389-397.

[6] Wang, X., et al. ChestX-ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. *arXiv preprint arXiv:1705.02315*, 2017.

[7] Yan C., et al. Weakly Supervised Deep Learning for Thoracic Disease Classification and Localization on Chest X-rays. *Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, 2018, pp. 103-110.

[8] Yao, L., et al. Learning to diagnose from scratch by exploiting dependencies among labels. *arXiv preprint arXiv:1710.10501*, 2017.