Assignment #4: Computer Vision with NIH Chest X-rays

MSDS 458 Artificial Intelligence and Deep Learning

Alison Au

August 27, 2022

Abstract

The National Institutes of Health (NIH) released a sizeable dataset of 112,120 labeled chest X-ray images to help researchers discover machine learning techniques for diagnosis. Ultimately, this can assist clinicians to better detect illness and treat patients. In this first phase of the project, eleven convolutional neural networks (CNN) are tested to identify pathologies in the X-rays using multi-label classification. They vary in terms of number of filters, layers, regularization, early stopping metrics, batch sizes, and pre-trained networks. The impact of these various levers is investigated in order to find the most accurate multilabel classification neural network.

Introduction

Chest X-ray exams are one of the most common and cost-effective medical image examinations. However, diagnosis from these tests can be challenging, especially compared to chest CT imaging. It requires careful observation, complex reasoning, and knowledge of anatomical principles from trained medical professionals. To help develop computational methods for aiding radiologists, the NIH made publicly available a dataset of 112,120 X-ray images with disease labels from 30,805 unique patients in 2017. The diseases span 14 classes: atelectasis, consolidation, infiltration, pneumothorax, edema, emphysema, fibrosis, effusion, pneumonia, pleural thickening, cardiomegaly, nodule, mass, and hernia. The labels were created using natural language processing that text-mined radiological reports but are expected to be >90% accurate.

This intent of this project is to build a model that could help radiologists and other physicians detect or confirm abnormalities in the scans. This first phase focuses on CNNs of

varying architecture and pre-trained models to maximize the AUC and F1 score in classifying pathologies.

Literature Review

The NIH chest X-ray dataset is widely available for machine learning both by researchers and students alike. There are few publications of models using this data thus far, though CNN and transfer learning have been the most popular techniques. These have been proven for image classification, thus are evaluated in this project as well.

Methods

The NIH chest X-ray dataset consists of 112,120 chest X-ray images of size 1024x1024. They are labeled across 15 classes (14 diseases and one of "no findings"). A sample of images are shown in Appendix Figure 1. An exploratory analysis shows the patients skew male (57%) and middle-aged with a mean of 47. Almost half of the X-rays have no findings. Infiltration is the most common pathology (18%) followed by effusion (12%) and atelectasis (10%) (Figure 2). This class imbalance was not adjusted in this phase of the project, though this will be investigated in future iterations as this can potentially lead to much better predictions. Additionally, an X-ray can have multiple pathologies making this a multi-label classification task. In fact, the three most common illness typically show in combination with other diseases. This can be challenging for the model given that they may overlap when manifested in X-rays.

Data preparation and model building were done using Tensorflow's Keras in the Google Colab Pro environment with GPU enabled. The dataset was split 80/20 into training (89,696) and test (22,424). 10% (8,970) of the training set was carved out for validation, leaving 80,726 for training the model. ImageDataGenerator from Keras was used to augment the images including sizing down the grayscale images to 128x128 pixels to help with processing times and

standardizing the values. Random horizontal flipping, height shifting, rotation, distortion, and reflecting the fill were also applied. This helps the model train across a wider variety of image versions. The training set was separated into batches of 32 or 64 (both were tested). For some pre-trained models that require this input, the generator was used to convert the single channel images (grayscale) to three-channels by repeating the values in the image across all channels.

Eleven CNN models were evaluated. All of them utilized binary cross-entropy as the loss function and sigmoid as the output layer's activation function with 14 nodes given that this is a multi-label classification. Each class is modeled independently to produce a probability between 0 and 1. This contrasts softmax which forces the probabilities across all classes to sum to 1 and categorical cross-entropy which is suited for single-label classification.

Relating to binary cross-entropy, binary accuracy was monitored as a metric along with recall. Note that categorical accuracy was not applicable since it assigns an image to a single class with the highest probability. Binary accuracy was also not very beneficial given the sparsity of the class matrix, as X-rays mostly had 0's encoded across the 14 diseases. This resulted in unmeaningfully high accuracies. Consequently, AUC and F1-score were viewed with more importance as well as recall since our goal is to identify diseases when present. The probabilities outputted were relatively low, indicating difficulty in detecting disease. Rather than a 50% threshold, a lower 20% threshold was used to classify whether a pathology was present and for calculating F1-score, recall, and precision.

All models used Adam as the optimizer and ran for 20 epochs with 100 steps or batches per epoch. This is much fewer than the training size divided by the batch size, meaning not all data ran through the training. However, this dramatically helped with training times, dropping

from 10 minutes per epoch to 2-3 minutes. All CNNs used a kernel size of 3x3 with stride 1 and max pooling of 2x2. The experiments can be broken down as follows:

- Experiments 1-2 were 3-layer CNNs that tested using binary accuracy vs. recall for early stopping, given that binary accuracy was already so high at close to 90%
- Experiments 3-4 tested a larger batch size of 64 vs. 32 in the earlier models since those showed sharp fluctuations in the training and validation curves
- Experiment 5 added on regularization
- Experiment 6 increased the number of filters to 128, 256, and 512 as compared to 32, 64, and 128 in prior models
- Experiments 7-8 evaluated 2-layer CNN models with varying filters
- Experiments 9-11 tested pre-trained models MobileNet and DenseNet121. The last model tested training DenseNet121 beyond 20 epochs to 40

Results

A performance summary table of the eleven experiments is in Figure 3, and findings are detailed below.

Early Stopping, Batch Size, Regularization

Initial models tested various performance metrics and configuration given the difficulty of a multi-label classification with a sparse matrix. While the CNNs optimize towards binary cross-entropy, binary accuracy was first tried for early stopping. Since accuracy is already high at around 88%, inflated by the many 0's across the classes, the model ran for only a few epochs with small fluctuations. Stopping was somewhat arbitrary, and training was not sufficient. Early stopping with the Keras recall metric was also tried but not functioning correctly, stopping before the maximum was achieved within the set patience. As a result, all subsequent models

were run for 20 epochs with no early stopping. This helped showcase how accuracy is not a good indicator of how well the model performs. In the next phase, early stopping with recall or F1-score can be attempted further.

The initial models showed training and validation accuracy and loss curves with many spikes throughout 20 epochs. This suggested a potentially less stable model. This led to testing a higher batch size from 32 to 64, increasing the number of images the model trains on each time. The curves did have slightly less fluctuations, but performance was not better across the board. Test AUC remained at around 68%, meaning the model has a 68% chance of correctly identifying each disease on X-rays. This came at a cost of more training time of about 1 hour vs. 45 minutes. The extra processing time did not seem valuable enough, so all subsequent models reverted to a batch size of 32. Other factors that could help the fluctuating loss and accuracy are running more epochs until they stabilize and increasing the step size. However, this would require significantly more training time. Increasing step size from 100 to 1,164, the number of batches to run through all the data, caused each epoch's training time to increase from 2 minutes to 10.

Lastly, regularization was tested with 20% dropout applied to each of the 3 CNN layers. This led to 3% lower AUC and 1% lower F1-score. Recall did increase 7%. The training and validation curves still fluctuated, though all models thus far did not show overfitting. The performance improvements were inconsistent, and overfitting did not seem to be a large issue prior. However, regularization is still employed in subsequent models as best practice since CNN models tend to overfit. It was not a detriment to training time, which remained at 44 minutes.

Overall, the various levers changed in the first 5 models did not result in much difference. The networks potentially may have needed to be trained for longer, something that can be investigated in the future with more time.

Number of Filters and layers

The number of filters in the 3-layer CNN were increased from 32, 64, and 128 to 128, 256, and 512. Training time unexpectedly dropped to 35 minutes. AUC decreased 5% to 64%, and F1-score decreased 2% to 38%. Recall was the only improved metric by 5% to 55%. This inconsistency indicates that the large increase in filters does not necessarily result in better performance. The performance metrics also remain low, proving the model unusable for real application.

2-layer CNNs were also tested, one with 64 and 128 filters and another with 256 and 512 filters. Again, training time and AUC dropped when more filters were used. F1-score remained steady. They also performed in line with the 3-layer networks. Given the complexity of multi-label identification on these X-rays, modifying the layers between 2 to 3 or the number of filters does not drive significant change, contrary to what was seen with the MNIST digits data.

Pre-Trained Models

Network architectures MobileNet and DenseNet121with pretrained weights were evaluated next. The former contains 28 layers including depthwise and pointwise convolutions. DenseNet has 120 convolution layers and 4 average pooling layers. MobileNet did not perform better than prior models with just 2 to 3 layers and had similar training time of 48 minutes. It's AUC was the lowest at 61% and had an average F1-score of 40%. It did attain relatively high recall at 56%. This illustrated that pretrained network architectures may not necessarily perform better.

DenseNet121 did achieve significantly better performance. It garnered the highest AUC at 71%, 2% more than the next best model. F1-score was also the highest at 42%, though recall was relatively low at 44% compared to 58% of the best model. The network took the longest to train at one hour and 22 minutes but was valuable for its better classification performance. This showcases that pretrained models with over a hundred layers is beneficial for a complex classification task like medical diagnoses from X-rays.

Given the relative success of DenseNet121, it was run for 20 more epochs. The training and validation performance curves showed more stabilization (see Figure 4). Additionally, AUC jumped by 6% to 76%, recall increased 7% to 51%, and F1-score increased 1% to 43%. This further supports that the earlier models could have run for more epochs, though this did take an additional hour and 17 minutes.

Final Model

The detailed results of the best model, DenseNet121 trained for 40 epochs, shows that the diseases are not equal in terms of ease of classification (see Figure 5-6). Infiltration, effusion, and atelectasis had the best F1-scores ranging from 41% to 59%. They also had the best recall with atelectasis at a high 90%. These are also the most common pathologies, suggesting that class imbalance skewed the model to better detect these diseases. Micro metrics were also significantly higher compared to macro, further supporting that smaller volume labels were more poorly classified. Hernia had such limited sample (227 cases in the entire dataset) that the model could not be effectively trained to detect it. Edema and pneumonia also had 0% F-score. The model as it stands can be better to detect a subset of diseases but not all 14 included in the data. A sample of predictions are also shown in Figure 7, illustrating the difficulty in classification.

The probability of a disease identified in the X-rays is low, hence why the threshold for classification was lowered from the default 50% to 20%.

Conclusions

Overall, multi-label classification of several diseases in X-rays proved difficult. Most modifications did not impact performance significantly. Larger network architectures with pre-trained weights were most fruitful, as they can uncover more sets of patterns in the X-rays, especially important if multiple diseases can present themselves in one image. Even then, the best model's recall and F1-score hovered at or below 50%, unusable for clinical applications. The high 89% accuracy is a misleading indicator of actual performance. The sparsity of the classification matrix made this more challenging as well as the unbalanced classes. Also difficult was the large size of the dataset given the amount of time required for training.

Multiple improvements can be explored in the next phase of this study to make the model usable. Class imbalance should be addressed in data preparation, such as by resampling or SMOTE. This alone may vastly improve performance. With more time, the models should be run for more epochs until degradation is seen and more steps per epoch. It could also be beneficial to find a way to use recall or F1-score as an early stopping metric. Other pre-trained models with more layers could be evaluated as well as other techniques. This includes attention mechanism, which can turn on and off pixels, hybrid CNN-transformer models, and ensemble models. The final model may not necessarily be able to detect all 14 diseases with high recall and precision, though may be usable for a subset of them.

References

Kaggle. Retrieved August 14, 2022, from https://www.kaggle.com/datasets/nih-chest-xrays/data.

NIH. *NIH Clinical Center provides one of the largest publicly available chest x-ray datasets to scientific community*. https://www.nih.gov/news-events/news-releases/nih-clinical-center-provides-one-largest-publicly-available-chest-x-ray-datasets-scientific-community, September 27, 2017.

Brownlee, Jason. *Multi-Label Classification with Deep Learning.* Machine Learning Mastery. https://machinelearningmastery.com/multi-label-classification-with-deep-learning/, August 31, 2020.

Appendix

Figure 1. Sample of NIH Chest X-ray Images.

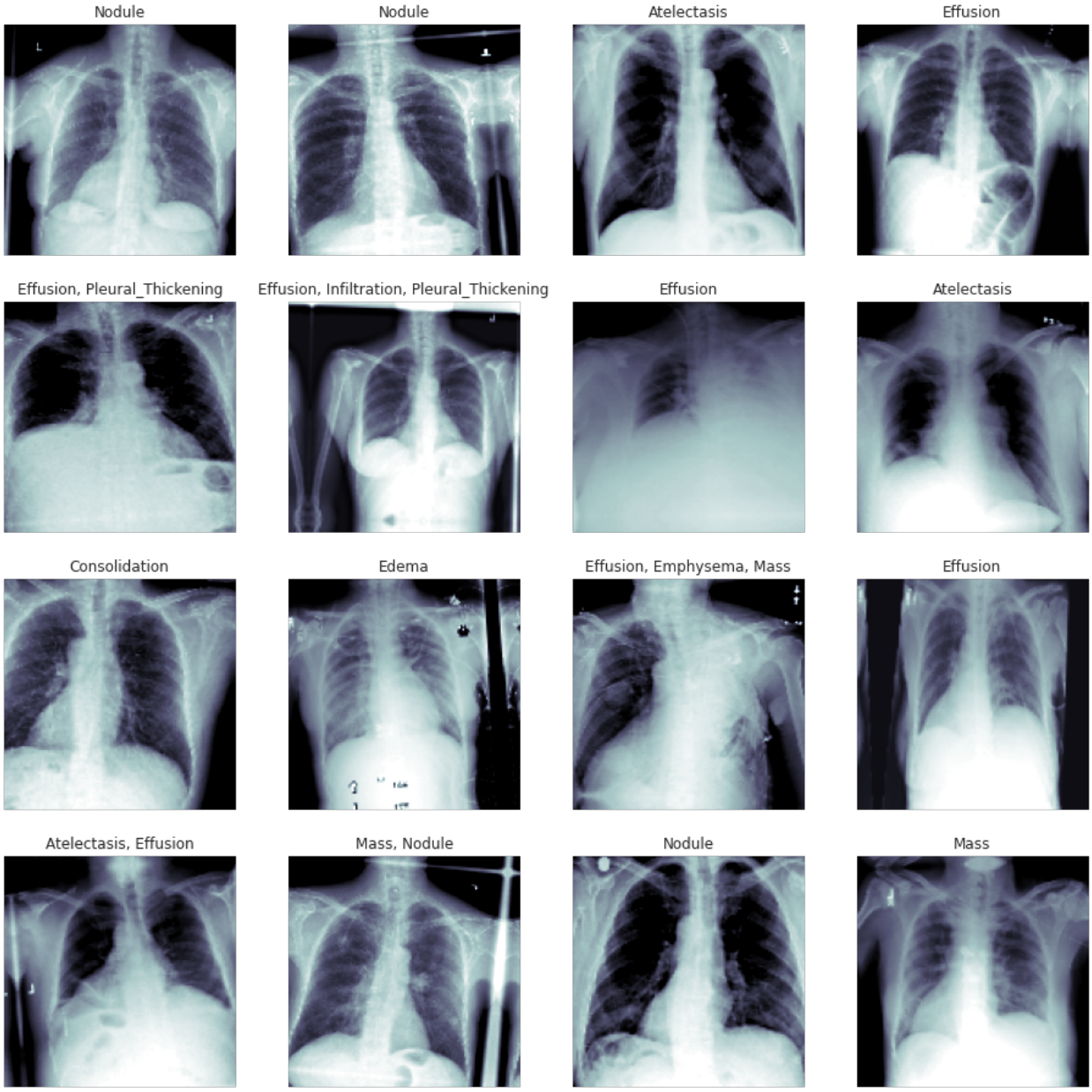


Figure 2. Number of X-rays by Pathology.

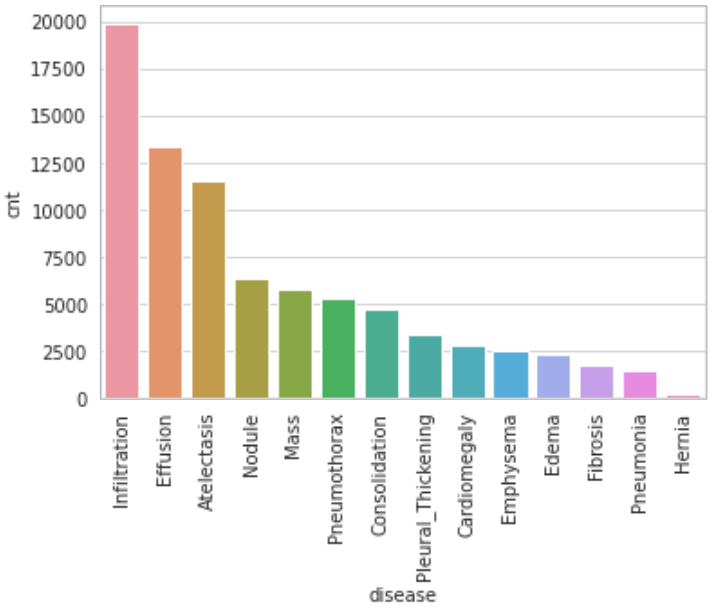


Figure 3. Summary Table of Model Results.

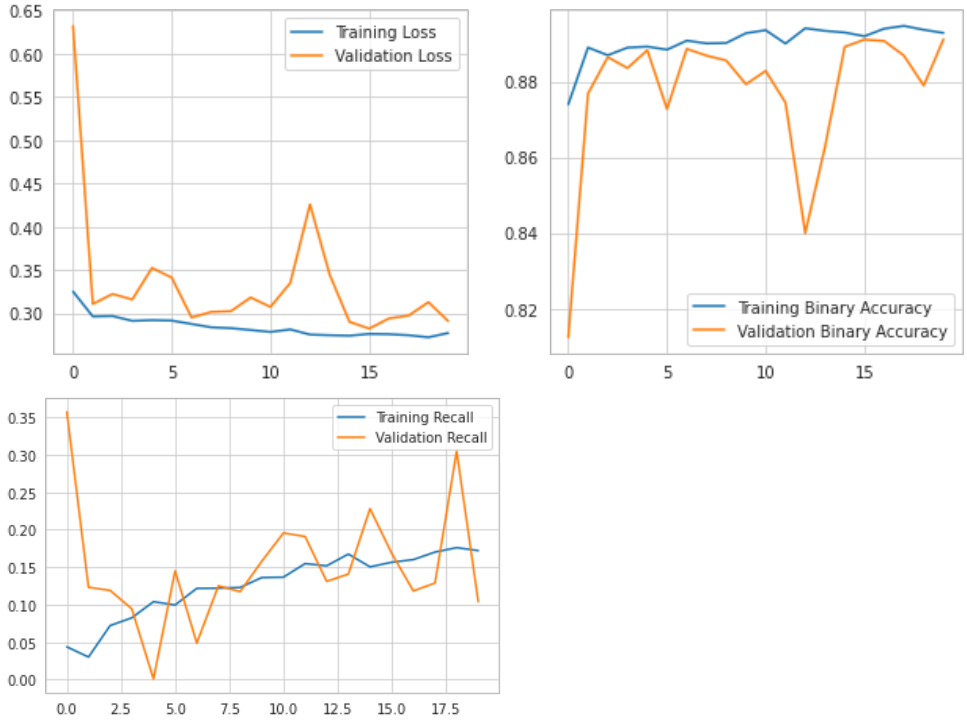
CNN Models

Evaluation	Early Stopping Metric		Batch Size		Regulariza-tion	# of Filters	# of Layers		Pre-Trained Models		More Epochs
Experiment	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**
# of Layers	3	3	3	3	3	3	2	2	14 (MobileNet)	120 (DenseNet121)	120 (DenseNet121)
# of Filters	32, 64, 128	32, 64, 128	32, 64, 128	32, 64, 128	32, 64, 128	128, 256, 512	64, 128	256, 512	32-1024	6-48	6-48
Batch Size	32	32	64	64	32	32	32	32	32	32	32
Early Stop Metric	Accuracy	Recall	Accuracy	Recall	Accuracy	Accuracy	Accuracy	Accuracy	Accuracy	Accuracy	Accuracy
Regularization	None	None	None	None	20% dropout	30% dropout	20% dropout	30% dropout	50% dropout	Batch Normalization	Batch Normalization
Training Loss	0.2910	0.2907	0.2821	0.2873	0.2923	0.2968	0.2930	0.2949	0.3033	0.2774	0.2676
Training Binary Acc	0.8877	0.8876	0.8910	0.8900	0.8879	0.8883	0.8877	0.8887	0.8853	0.8928	0.8966
Val Loss	0.2892	0.2894	0.2851	0.2869	0.2950	0.3011	0.2969	0.2982	0.3001	0.2910	0.2833
Val Binary Acc	0.8895	0.8893	0.8907	0.8900	0.8891	0.8883	0.8884	0.8883	0.8886	0.8911	0.8917
Testing Binary Acc	0.8925	0.8924	0.8922	0.8857	0.8862	0.8858	0.8848	0.8870	0.8870	0.8915	0.8907
Training Time	00:44:20	00:44:58	01:07:14	00:57:45	00:44:13	00:35:13	00:41:43	00:36:23	00:48:55	01:22:15	01:17:16
AUC	0.6886	0.6758	0.6782	0.6746	0.6562	0.6371	0.6517	0.6427	0.6098	0.7070	**0.7629**
Precision - Micro	0.33	0.33	0.34	0.29	0.30	0.29	0.30	0.31	0.31	0.40	0.37
Precision - Macro	0.21	0.16	0.20	0.12	0.14	0.14	0.15	0.14	0.10	0.24	0.25
Recall - Micro	0.50	0.53	0.55	0.43	0.57	0.55	**0.58**	0.51	0.56	0.44	0.51
Recall - Macro	0.21	0.25	0.26	0.17	0.27	0.27	0.29	0.25	0.22	0.24	0.30
F1 Score - Micro	0.40	0.40	0.42	0.35	0.39	0.38	0.39	0.39	0.40	0.42	**0.43**
F1 Score - Macro	0.14	0.17	0.18	0.11	0.17	0.17	0.17	0.16	0.11	0.21	0.25

Precision, Recall, F1 Score at 20% threshold

Figure 4. Training and Validation Curves of Best Model (DenseNet121).

Epochs 1-20



Epochs 21-40

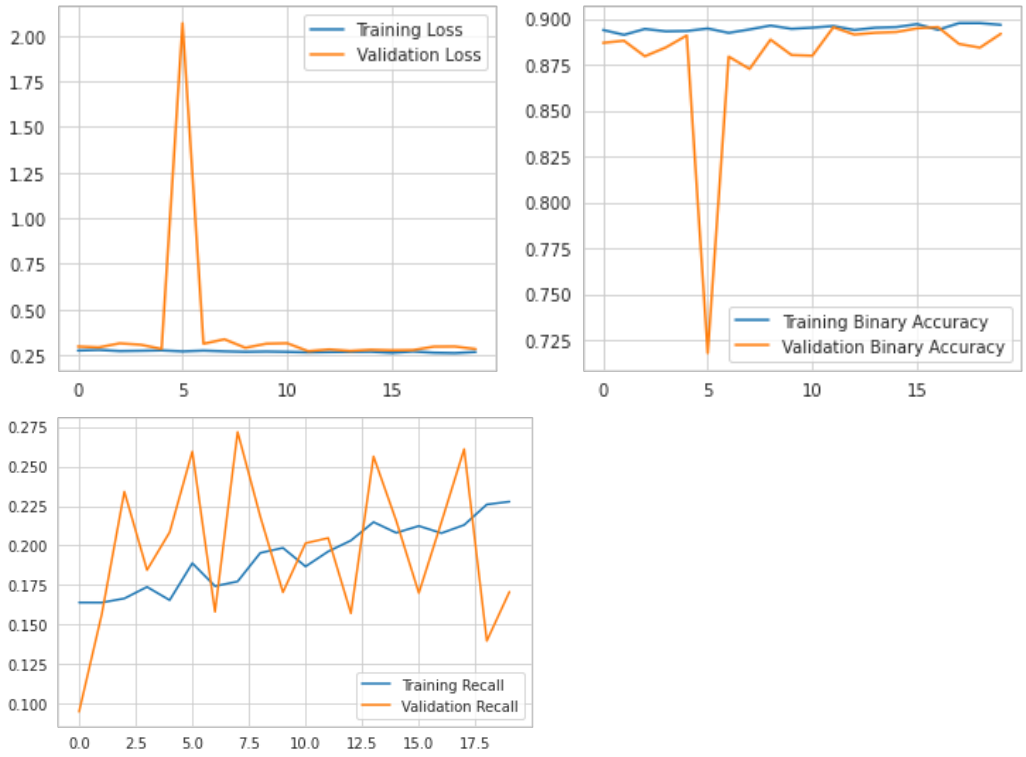


Figure 5. ROC-AUC of Best Model (DenseNet121).

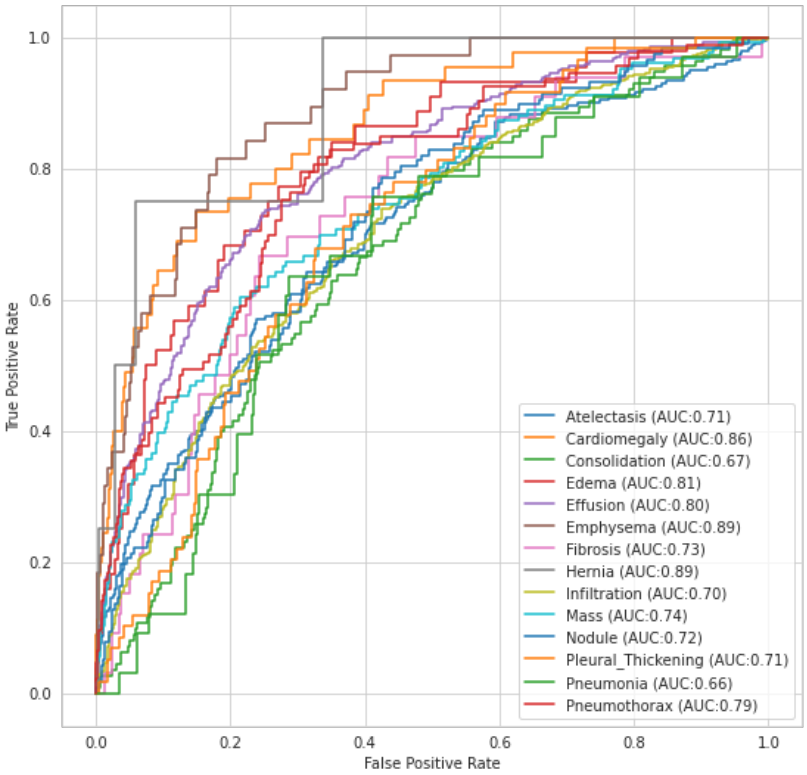
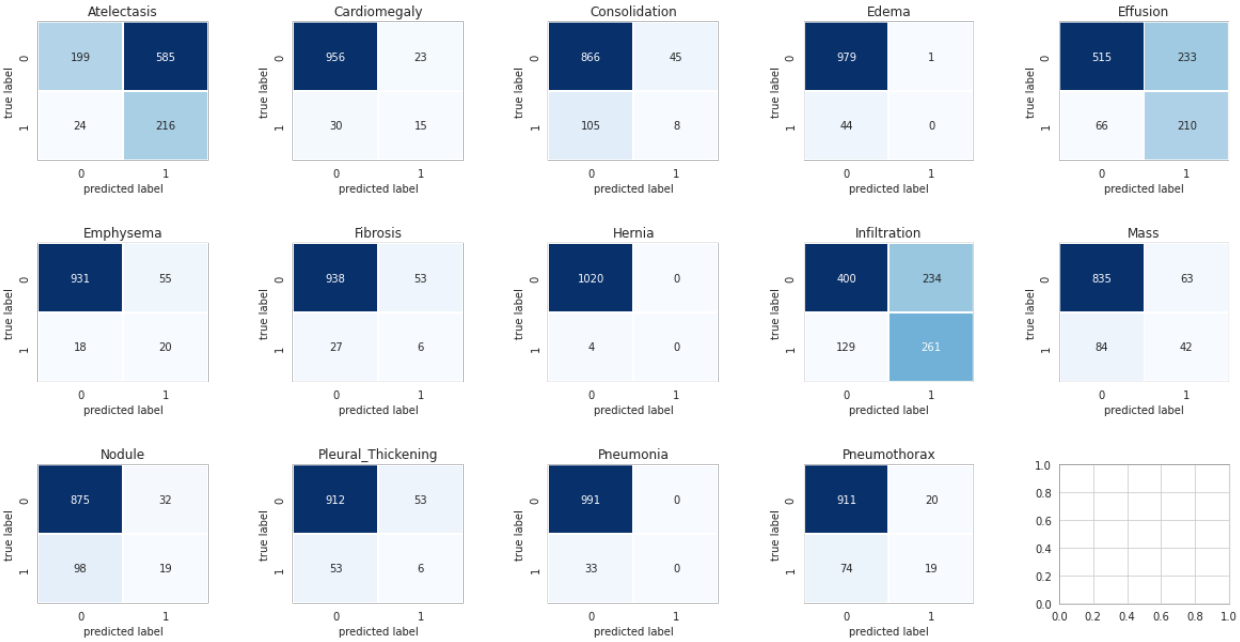


Figure 6. Confusion Matrices and Classification Report of Best Model (DenseNet121).



	precision	recall	f1-score	support
Atelectasis	0.43	0.45	0.44	240
Cardiomegaly	0.58	0.16	0.25	45
Consolidation	0.00	0.00	0.00	113
Edema	0.00	0.00	0.00	44
Effusion	0.68	0.38	0.48	276
Emphysema	0.57	0.21	0.31	38
Fibrosis	0.00	0.00	0.00	33
Hernia	0.00	0.00	0.00	4
Infiltration	0.71	0.08	0.14	390
Mass	0.56	0.07	0.13	126
Nodule	1.00	0.01	0.02	117
Pleural_Thickening	0.00	0.00	0.00	59
Pneumonia	0.00	0.00	0.00	33
Pneumothorax	1.00	0.02	0.04	93
micro avg	0.54	0.17	0.26	1611
macro avg	0.40	0.10	0.13	1611
weighted avg	0.56	0.17	0.21	1611
samples avg	0.24	0.17	0.19	1611

Figure 7. Sample Predictions from Best Model (DenseNet121).