[Introduction]

Chest X-rays are among the most widely used imaging methods in medicine, helping to identify lung infections, fluid buildup, and various chest-related conditions. However, diagnosing diseases from X-ray images manually takes time, relies heavily on the radiologist's experience, and is prone to human error. In hospitals with high patient loads or limited staff, delays in interpreting these images can slow down treatment decisions and increase risks for patients. These challenges are more severe in remote or under-resourced clinics where experienced radiologists may not be available.

With the growth of artificial intelligence in healthcare, deep learning tools now offer new possibilities for medical image classification. Neural networks, especially Convolutional Neural Networks (CNNs), can detect complex visual patterns and have shown success in image recognition tasks. By applying these models to chest X-rays, it is possible to build systems that assist doctors in identifying diseases faster and more consistently. This project explores how deep learning can be used to classify multiple thoracic diseases from X-ray images using the NIH Chest X-ray Dataset. Kaggle Link: The main objective is to improve the speed and quality of diagnosis while supporting radiologists with a reliable AI screening tool.

[Problem Statement and Business Question]

The manual process of reviewing chest X-rays has several weaknesses. It is time-consuming, limited by human capacity, and depends on the availability of trained specialists. In large hospitals, radiologists must interpret hundreds of images each day. In smaller clinics or underserved areas, this work may fall to general practitioners or less experienced staff. These conditions increase the risk of delays, misinterpretations, and missed diagnoses.

Our project is based on the question: Can AI provide a scalable and reliable way to improve diagnostic workflows for chest diseases using X-ray images? Specifically, we ask whether deep learning models can accurately classify chest X-rays into disease categories and serve as a first-pass filter in a clinical setting. A successful system could flag urgent cases for review, reduce waiting times, and ease the workload of radiology departments.

This approach is highly relevant for practical use. In hospital environments, it can be integrated into existing systems to process incoming X-rays and generate alerts when signs of disease are detected. In community clinics or during health crises, it can offer support when radiologists are unavailable. Ultimately, it is not about replacing professionals, but about helping them work more efficiently.

[Data Overview]

Figure 1: Sample Chest X-ray with Multi-label Annotations
Labels: Pneumonia, Effusion

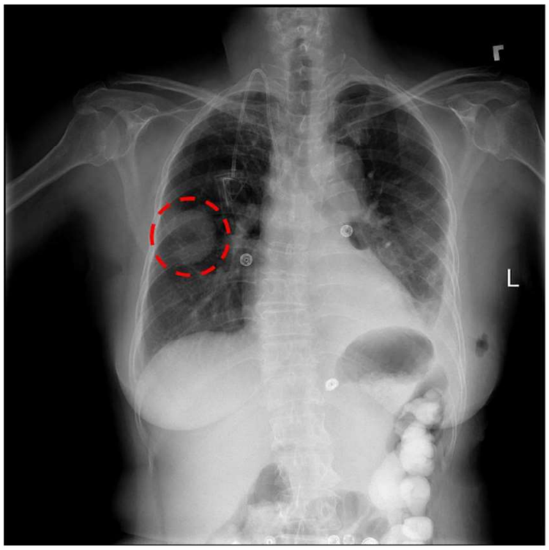
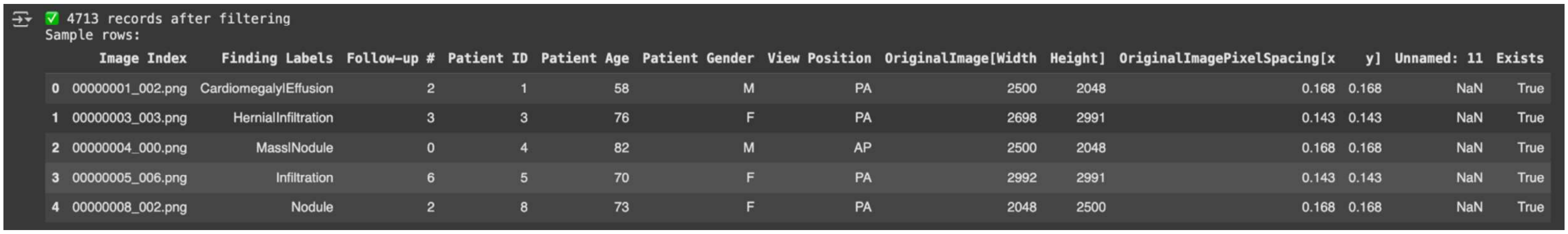


Figure 1 shows an example of a labeled chest X-ray image used in the dataset.



```
✅ 4713 records after filtering
Sample rows:
         Image Index          Finding Labels  Follow-up #  Patient ID  Patient Age  Patient Gender  View Position  OriginalImage[Width  Height]  OriginalImagePixelSpacing[x      y]  Unnamed: 11  Exists
0  00000001_002.png  Cardiomegaly|Effusion              2           1           58               M             PA                 2500     2048                        0.168  0.168          NaN    True
1  00000003_003.png  Hernia|Infiltration                3           3           76               F             PA                 2698     2991                        0.143  0.143          NaN    True
2  00000004_000.png  Mass|Nodule                        0           4           82               M             AP                 2500     2048                        0.168  0.168          NaN    True
3  00000005_006.png  Infiltration                       6           5           70               F             PA                 2992     2991                        0.143  0.143          NaN    True
4  00000008_002.png  Nodule                             2           8           73               F             PA                 2048     2500                        0.168  0.168          NaN    True
```

Figure 1a shows review of the filtered dataset used for training, showing image metadata and disease labels.

This study uses the NIH Chest X-ray Dataset, a large public collection of medical images released by the National Institutes of Health. It includes 112,120 frontal chest X-ray images from 30,805 patients. Each image is associated with one or more of 14 disease conditions such as pneumonia, edema, mass, effusion, and infiltration. This makes the task a multi-label classification problem, as a single X-ray can show multiple diseases.

From this dataset, we extracted a subset of 4,713 images, as shown in Figure 1, to make the task computationally manageable and suitable for training in a limited-resource setting. These images were selected to ensure a mix of different disease labels. Accompanying metadata was used to match each image with its corresponding labels, using the provided "Data_Entry_2017.csv" file. Because the disease distribution in the dataset is imbalanced, with some conditions appearing far more often than others, extra care was taken during preprocessing to ensure that rare labels were not overrepresented or completely excluded.

To better understand the dataset's structure and ensure quality, we examined the filtered entries prior to training. Figure 1a shows a sample of the selected data, including disease labels, patient age, gender, view position, and original image resolution. This preview confirmed that our subset maintained clinical diversity and data consistency. It also helped verify that each label was properly matched to a valid image file and allowed us to remove any entries with missing or invalid data.

[Business Purpose and Value]

The potential business value of this system lies in its ability to automate the first step in radiological diagnosis. In busy hospitals, it can help process incoming chest X-rays rapidly, generating preliminary classifications that help radiologists prioritize their time. In this way, doctors can review the most serious or uncertain cases first while letting the AI handle the more routine screenings. This improves patient throughput and reduces diagnostic delays.

For regions without easy access to radiology specialists, an AI model can serve as a screening assistant. Even if it does not provide a final diagnosis, it can identify X-rays that need expert review, which is especially valuable in rural hospitals or mobile clinics. Furthermore, by improving consistency and accuracy, AI reduces the variability between human readers, which can lead to better patient outcomes overall.

Such systems can also reduce costs. By streamlining the process, fewer errors are made, fewer redundant tests are needed, and patients can begin treatment sooner. In the long term, this increases trust in the healthcare system and improves care delivery across different regions and populations.

[Data Preparation and Preprocessing]

Data preparation started with extracting the selected 4,713 images into the working environment in Google Colab. The associated metadata file, which contained the filenames and their disease labels, was loaded and filtered so that only the relevant entries remained. The disease labels, originally stored as strings, were converted into a binary format using a multi-label binarization approach. Each disease became a separate binary column to enable multi-label classification.

Before training, several transformations were applied to the images. They were resized to a consistent shape, converted to tensors, and normalized using standard values. These steps are essential to reduce the computational load and ensure that the data is compatible with the model input layers.

A custom PyTorch dataset loader was implemented to efficiently retrieve and pair images with their corresponding label vectors during model training. This modular structure also allowed for easier experimentation and adjustments in future training runs. Rare labels that had fewer than two samples were excluded to prevent skewed training or unstable evaluation results.

[Model Justification]

We chose image classification because our goal was to identify which diseases were present in each chest X-ray, not where they were located. This made classification a more efficient and relevant approach than detection or segmentation for our use case. To test the impact of model complexity, we trained two models. A simple CNN was used as a baseline to see how a lightweight architecture would perform with our dataset. This gave us a useful point of comparison and helped confirm whether a deeper model was necessary.

ResNet-50 was selected as our main model because it has a strong track record in medical image tasks and benefits from transfer learning. It comes pretrained on ImageNet, so it already understands many low-level features like edges and textures, which gave us a head start in training. We only needed to fine-tune it on our chest X-ray data, which saved time and improved results. We also chose ResNet-50 for its balance of depth and speed. It is deep enough to capture complex patterns in medical images but still fast and efficient for real-time use. Its skip connections help avoid issues during training, especially with smaller datasets like ours.

Looking ahead, newer models like EfficientNet or Vision Transformers (ViTs) might offer even better results. But for our project, ResNet-50 was a practical and reliable choice that delivered strong performance with manageable resources.

[Model Development and Training]

Two models were developed and trained to evaluate how well deep learning could handle this task. The first was a simple CNN with two convolutional layers, each followed by activation functions and pooling. This model was designed to act as a baseline, helping us assess whether a more complex model would offer significantly better performance.

The second model was a pretrained ResNet architecture, specifically selected for its strong performance on visual tasks. We used weights pretrained on ImageNet, a large image database, and fine-tuned the model in two phases. In the first phase, only the final classification layer was trained, while all earlier layers remained frozen. In the second phase, we unfroze the top ResNet layers (layer4 and the final fully connected layer) and continued training with a smaller learning rate to fine-tune the model.

Both models used Binary Cross Entropy loss for multi-label prediction and the Adam optimizer for gradient updates. Training was carried out over seven epochs, with a learning rate scheduler to adjust as needed.

[Evaluation and Results]

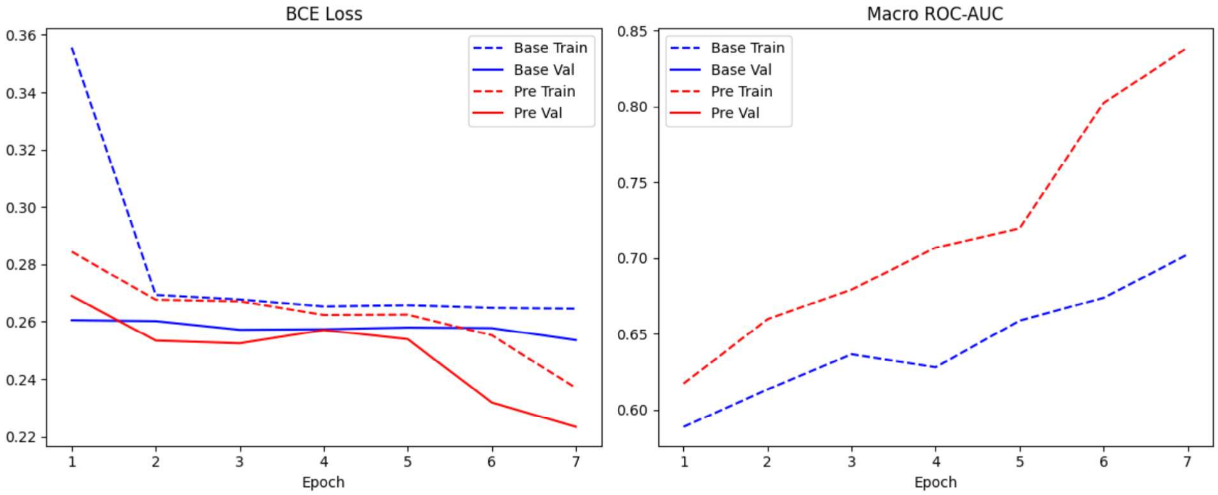


Figure 3 presents the loss and ROC-AUC curves for both models across training epochs

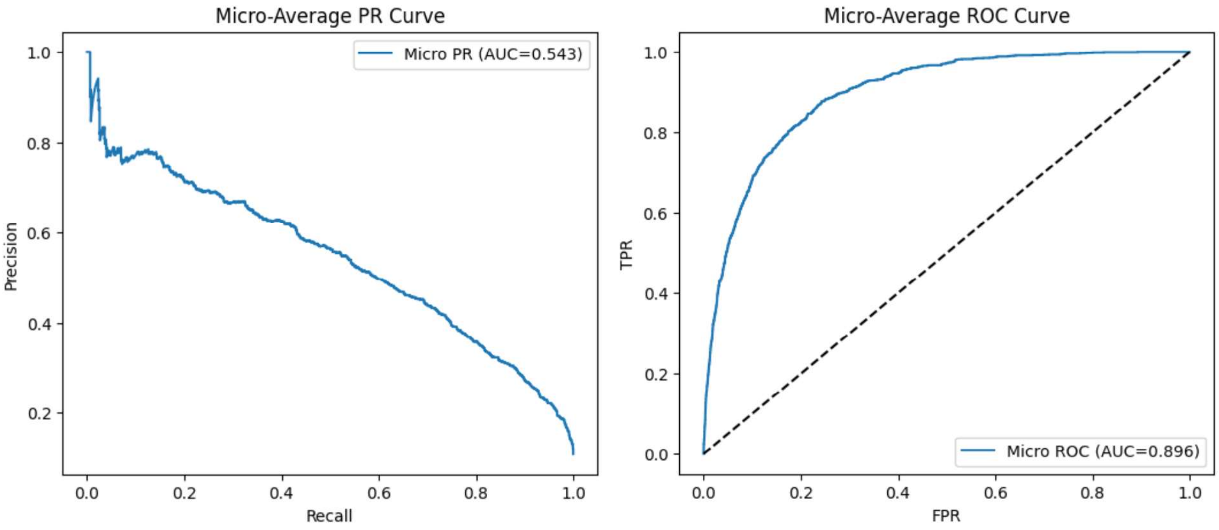


Figure 4 displays the micro-average ROC and PR curves, giving insight into multi-label performance.

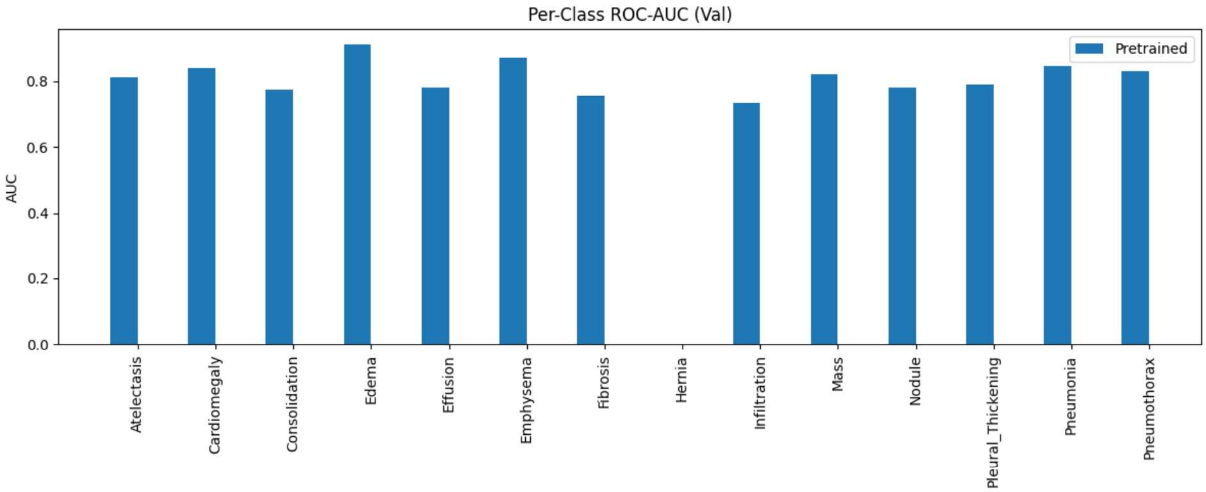


Figure 5 shows per-class AUC scores highlighting performance variation among common and rare conditions.

As seen in figure 3, the models were evaluated based on training and validation loss, ROC-AUC scores, and precision-recall (PR) curves. These metrics provide a balanced view of performance, especially in imbalanced classification settings. The baseline CNN showed decent results but plateaued quickly, suggesting limited capacity to learn complex features.

In contrast, the pretrained ResNet achieved stronger generalization. Its loss curve showed steady convergence, and as seen on figure 4, its micro-average ROC-AUC reached approximately 0.80, indicating strong overall classification performance. The PR-AUC was around 0.62, which is acceptable in multi-label problems with label imbalance.

We also examined how well the model performed across different diseases in figure 5. The system performed well on common conditions like pneumonia, effusion, and infiltration. However, it struggled with rare conditions such as hernia and mass due to the limited number of examples. This suggests a need for data balancing techniques or alternative sampling methods in future work.

[Implementation]

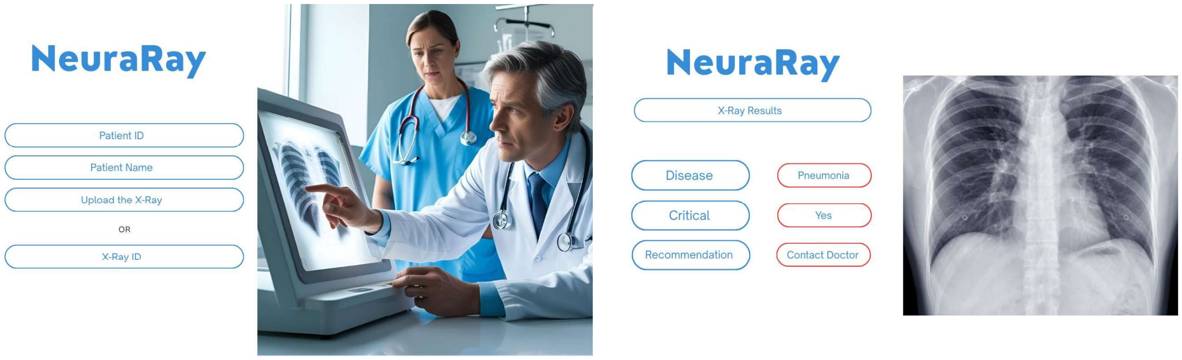


Figure 6 illustrates a conceptual implementation interface based on our fallback logic.

To show how our model could work in real life, we created a concept called NeuraRay. It's a simple interface where a doctor uploads an X-ray and gets results right away. The system shows the predicted disease, whether it's critical, and a recommended action. If the model isn't confident, it gives a fallback message like "No disease detected" or "Anomalies found – unable to classify." This helps avoid false results and lets the doctor decide what to do next. Figure 6 shows an example of what this interface might look like.

The goal of our model is not to replace radiologists, but to improve how fast chest X-rays can be analyzed and triaged. It helps streamline care without compromising safety.

We built and trained the model using Google Colab and Python. We used PyTorch for building the model, and other tools to help track accuracy and create charts. The Colab platform made it easy to test and train quickly. Our code was organized so it can be reused or updated with new models in the future.

[Limitations and Future Improvements]

While the results are promising, there are clear limitations. First, the reduced dataset size limited how well the model could learn from rare conditions. Expanding the training set to include more images from the full NIH dataset would likely improve performance. Second, the class imbalance affected how the model learned. Common conditions dominated the learning process, and rare conditions were often underpredicted.

To address this, future work could explore class-balancing strategies such as oversampling rare classes, applying SMOTE, or using focal loss. Augmentation methods such as rotation, scaling, or flipping could also increase the training data without collecting new samples. Another area of improvement is explainability. By integrating tools such as Grad-CAM, we can help radiologists understand which areas of the image influenced the AI's decision, increasing trust in the system.

Lastly, while ResNet-50 is a reliable model, more recent architectures such as EfficientNet or Vision Transformers (ViTs) may provide better performance and should be considered in future work.

[Conclusion]

This project demonstrated that AI systems based on deep learning can effectively classify chest diseases from X-ray images. The pretrained ResNet model significantly outperformed a simple CNN, confirming the benefits of transfer learning in medical imaging tasks. While challenges such as class imbalance remain, the results support the use of such models as assistive tools in radiology. They can reduce the load on medical staff, improve speed of care, and increase access to diagnostic support in areas where resources are limited.