MACHINE LEARNING LAB PROJECT

TEAM 28 :

PUNNETH RANJAN -121CS0213

BODDEDA NEHITHA RATAHN- 121CS0215

SAMBIT KUMAR ROUT-121CS0216

SUBMITTED TO :

RATNAKAR DASH

Department of Computer Science and Engineering

NATIONAL INSTITUTE OF TECHNOLOGY ROURKELA

November 2024

INDEX PAGE

1.Introduction

Medical imaging plays a crucial role in modern healthcare, enabling accurate diagnosis and treatment planning for a wide range of diseases. Among various imaging modalities, chest X-rays are one of the most common and essential diagnostic tools used for assessing thoracic diseases. They are particularly valuable in detecting conditions such as pneumonia, pleural effusion, lung cancer, and other lung pathologies, which require prompt and accurate diagnosis for effective patient care. However, the interpretation of chest X-rays is highly challenging and requires expert radiologists to analyze subtle patterns. The limited availability of radiologists, coupled with increasing healthcare demands, has driven the need for automated systems that can assist in the diagnosis of chest diseases.

1.1 Overview of the Problem

Classifying diseases from chest X-rays presents several challenges. First, medical image datasets are often imbalanced, meaning that some diseases are over-represented while others are under-represented. This imbalance can lead to biased models that perform well on more common diseases but poorly on rare conditions, potentially resulting in missed diagnoses for critical conditions. Additionally, chest X-rays contain complex and subtle visual features that require advanced image recognition techniques to accurately classify. Unlike everyday images, chest X-rays often lack clear distinctions between normal and pathological states, making it difficult for traditional image processing methods to identify disease-specific features reliably.

Furthermore, chest X-rays often show overlapping structures, such as ribs, muscles, and organs, which can obscure disease patterns. To address these challenges, we will employ pre-trained convolutional neural networks (CNNs) such as ResNet and DenseNet, which have been shown to perform well on complex image classification tasks by learning deep hierarchical features that can capture subtle patterns in the data.

2.Data Exploration

2.1 Dataset Overview

The CheXpert dataset is a large and diverse collection of labeled chest X-rays that is widely used for research in medical imaging. Developed by Stanford University, the dataset consists of over 224,000 chest radiographs from more than 65,000 patients, making it one of the largest publicly available medical image datasets. Each image in CheXpert is labeled with one or more disease categories, including pneumonia, pleural effusion, cardiomegaly, and others, making it suitable for multi-label classification tasks. The dataset also includes uncertainty labels, acknowledging that some diagnoses may not be conclusive due to overlapping symptoms or image artifacts.

The CheXpert dataset offers a realistic and challenging benchmark for developing and testing automated diagnostic models. However, like many medical datasets, CheXpert suffers from class imbalance, where common diseases like no-finding or cardiomegaly are over-represented, and rarer conditions like pneumothorax are under-represented. This imbalance necessitates the use of advanced techniques to prevent the model from being biased toward more frequent classes and ensure reliable performance across all disease categories

2.2 Statistics and Visualizations

The table you provided shows the distribution of labels for various diseases in the CheXpert dataset, with each disease having three possible label values:

- 1: Indicates the presence of the disease.
- 0: Indicates the absence of the disease.
- -1: Represents uncertainty, where it is unclear if the disease is present or absent.

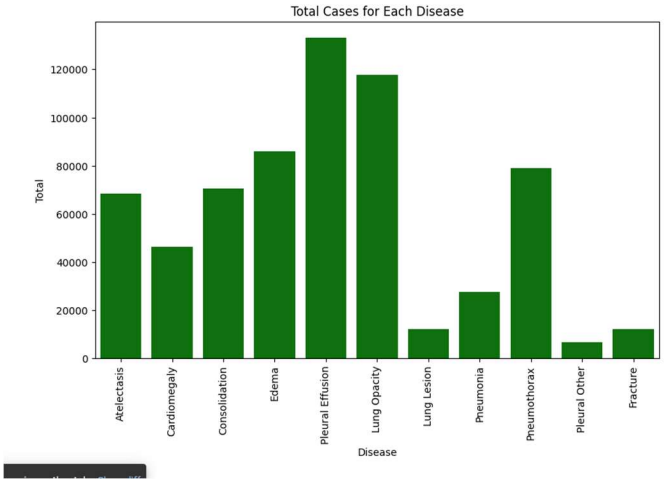
Key Observations:

1. Class Distribution: Each row represents a specific disease category, displaying the counts for each label (1, 0, -1) as well as the total number of instances for that disease. The total column sums up the instances across all labels for each disease.

High Occurrence of Certain Diseases:

- Atelectasis has a total of 68,443 instances, with 33,376 labeled as 1 (presence), 1,328 as 0 (absence), and 33,739 as -1 (uncertain). This disease has a fairly balanced distribution between certain presence and uncertain cases, with relatively few cases marked as absence.
- Cardiomegaly has 46,203 instances, with 27,000 labeled as 1 (presence) and 11,116 as 0 (absence), indicating that most cases are either labeled as present or uncertain.

	Disease	1s	0s	-1s	Total
0	Atelectasis	33376	1328	33739	68443
1	Cardiomegaly	27000	11116	8087	46203
2	Consolidation	14783	28097	27742	70622
3	Edema	52246	20726	12984	85956
4	Pleural Effusion	86187	35396	11628	133211
5	Lung Opacity	105581	6599	5598	117778
6	Lung Lesion	9186	1270	1488	11944
7	Pneumonia	6039	2799	18770	27608
8	Pneumothorax	19448	56341	3145	78934
9	Pleural Other	3523	316	2653	6492
10	Fracture	9040	2512	642	12194



- Atelectasis has a high number of cases labeled as present (1) and a similar number labeled as uncertain (-1). This suggests that Atelectasis is frequently diagnosed but also has a significant amount of ambiguity, which could be due to overlapping visual features in X-rays or the subtlety of symptoms.
- Cardiomegaly has more cases labeled as present than absent or uncertain, indicating a higher prevalence of positive diagnoses. However, there is still a notable amount of uncertainty, which may be due to variations in heart size due to patient-specific factors.
- Consolidation has a relatively low count of positive cases (1) compared to absent (0) and uncertain (-1) labels. This suggests that it may be harder to diagnose with confidence, possibly due to overlapping

symptoms with other lung conditions, making this condition one of the more challenging to identify clearly in X-rays.

- Edema has a high count of positive cases, indicating it is relatively common in this dataset. However, there are also significant numbers of absent and uncertain labels, suggesting that while edema is commonly diagnosed, there are cases where it is not clearly identifiable in X-rays.
- Pleural Effusion is one of the most common findings in the dataset, with a substantial number of positive cases (1). The lower proportion of uncertain labels suggests that it may be easier to identify in chest X-rays compared to some other conditions, likely due to the distinct fluid buildup that characterizes this disease.

3 .Dataset Prepossessing

The dataset preparation begins by filtering the data to include only frontal chest X-ray images using the 'Frontal/Lateral' column, which is set to 'Frontal'. After filtering, the dataset is shuffled to ensure randomness in the selection of samples. The mixed policy is used to assign binary labels to conditions based on predefined categories. Specifically, for classes listed in class_ones (such as Atelectasis and Cardiomegaly), the condition is marked as present (1). For all other classes, the condition is marked as absent (0). This allows for selective labelling where certain conditions are treated as positive (present) based on the predefined list, while others are always considered negative (absent). This approach helps apply different labelling strategies based on the nature of the medical conditions in the dataset.

Since the total dataset is large, a balanced selection of 2,000 positive samples (where the condition is present) is made for each class to ensure sufficient representation of each condition in the training set. These positive samples are then combined with the corresponding negative samples (where the condition is absent). Afterward, the dataset is shuffled and reset to maintain randomness and ensure that the final training set is both balanced and representative. This step is crucial to prevent class imbalance, which could negatively impact model performance, especially in cases where certain conditions are less frequent.

4.Model Development and Training

Detailed Model Architecture:

1. Base Model - DenseNet121:
    - DenseNet121: A CNN architecture known for efficient feature reuse and fewer parameters due to its “dense” connections between layers. Each layer receives inputs from all preceding layers, which helps in gradient flow and reduces the vanishing gradient problem. This characteristic is advantageous for medical imaging tasks, as it allows for deeper networks with lower risk of overfitting.
    - Pre-trained Weights: The model uses weights pre-trained on ImageNet, which are repurposed here to recognize medical imaging features after transfer learning. The ImageNet weights are beneficial for initial feature extraction, although additional training layers will be tailored to the specific needs of the CheXpert dataset.
    - Layer Freezing: The for layer freezes the weights in the DenseNet121 layers, effectively preserving the learned weights from ImageNet. This is particularly useful when working with smaller datasets, as it reduces the number of parameters to train and mitigates overfitting.
2. Global Average Pooling:
    - The GlobalAveragePooling2D layer replaces the fully connected layers often used in CNNs, drastically reducing the number of parameters. This pooling layer averages the spatial dimensions, which helps in capturing global features across the entire image, an essential characteristic for medical imaging tasks.

3. Prediction Layer:

    - The final Dense layer outputs a number of nodes equal to the number of classes in class_names and uses a sigmoid activation function. Unlike softmax (which is used for mutually exclusive classes), the sigmoid activation allows for multi-label classification, crucial for CheXpert, where a single X-ray image might indicate multiple pathologies.

4.1 Block diagram of model proposed

5.Results

5.1 Performance Metrics :

Accuracy: Calculated as (TP +TN) / (TP + TN +FN + FP), where TP is true positives and FN is false negatives, TN is true negatives and FP is false positives.

Sensitivity/Recall: Calculated as TP / (TP + FN), where TP is true positives and FN is false negatives. Specificity: Calculated as TN / (TN + FP), where TN is true negatives and FP is false positives.

Precision: Calculated as TP / (TP + FP), where TP is true positives and FP is false positives.

F1 Score: Calculated as hormonic mean of precision and recall

Confusion Matrix: Shows the distribution of true positive, true negative, false positive, and false negative predictions.

Using densenet121:

Test Accuracy: The model achieved an accuracy of 72.3

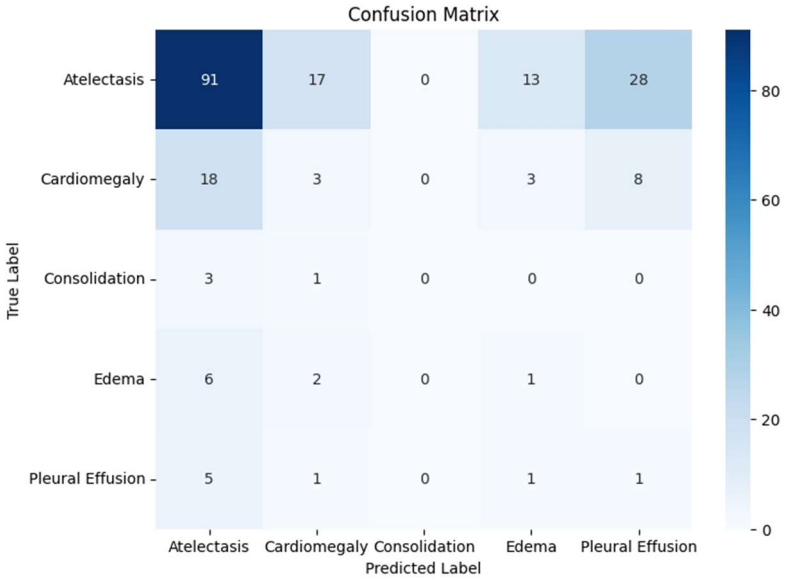
- The model achieved an accuracy of 0.84 for **Consolidation**.
- The model achieved an accuracy of 0.81 for **Edema**.
- The model achieved an accuracy of 0.71 for **Pleural Effusion**.
- The model achieved an accuracy of 0.64 for **Cardiomegaly**.
- The model achieved an accuracy of 0.61 for **Atelectasis**.

Loss : The model achieved a validation loss of 0.5711.

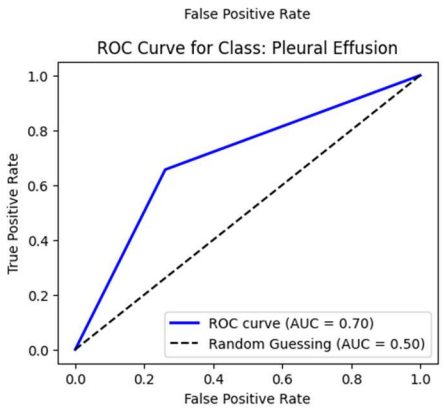
Precision: The model achieved a validation precision of 0.4508.

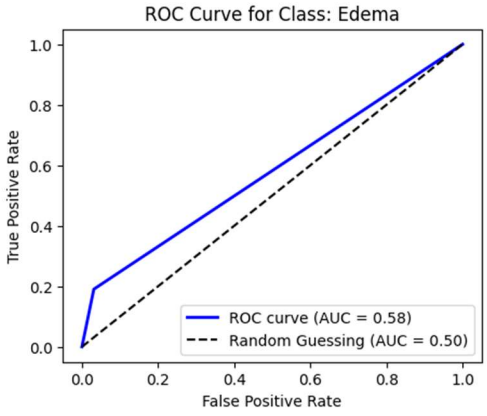
Recall: The model achieved a validation recall of 0.4265.

Confusion matrix:



ROC CURVES:





Conclusion

In this project, I worked with a limited subset of the CheXpert dataset, which likely impacted the accuracy of my model. Given the smaller number of samples, the model had fewer examples to learn from, which can result in underfitting and lower generalization ability. If I had access to the complete dataset, the model would likely achieve higher accuracy, as it would have more diverse examples to capture the complex patterns required for reliable predictions.

GUI Development

I have developed a Flask web application that allows users to upload an image, which is then processed by a machine learning model to predict medical conditions such as Atelectasis, Cardiomegaly, etc. The model's predictions are displayed as class labels below the uploaded image on the web interface. The app supports image file uploads (PNG, JPG, JPEG, GIF) and has basic error handling for unsupported file types or no file being selected.

Steps to Run:

1. Set up the environment:
    - Ensure Python and Flask are installed. You can install Flask via pip install flask.
    - Ensure all necessary dependencies are installed, including any deep learning libraries (like TensorFlow or Keras) used for model inference.
    - Run the Flask app:
    - Navigate to the project directory in the terminal and run the Flask application with python app.py.
    - The app should be available at http://127.0.0.1:5000/ in your browser.
2. Upload an image:
    - Visit the home page, upload an image, and check the results displayed after the model processes it.

This will launch the app and allow you to upload images for classification. Make sure the model inference code (inside validate()) is correctly set up for processing the uploaded images**.**

References

1. Original CheXpert Paper Irvin, Jeremy, et al. "CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison." *Proceedings of the AAAI Conference on Artificial Intelligence.* Vol. 33. 2019.
2. Deep Learning for Chest Radiographs Using CheXpert Rajpurkar, Pranav, et al. "Deep learning for chest radiograph diagnosis: A retrospective comparison of the CheXNeXt algorithm to practicing radiologists." *PLoS Medicine* 15.11 (2018): e1002686.
3. Benchmarking Robustness with CheXpert Seyyed-Kalantari, Laleh, et al. "CheXclusion: Fairness gaps in deep chest X-ray classifiers." *arXiv preprint arXiv:2003.00827* (2020).
4. Large-scale Robust Deep AUC Maximization: A New Surrogate Loss and Empirical Studies on Medical Image Classification by Zhuoning Yuan, Yan Yan, Milan Sonka, Tianbao Yang .