Project Report

By Aarij H. (Developer)

(Github - https://github.com/aarij-h/XRay_DeepLearning.git)

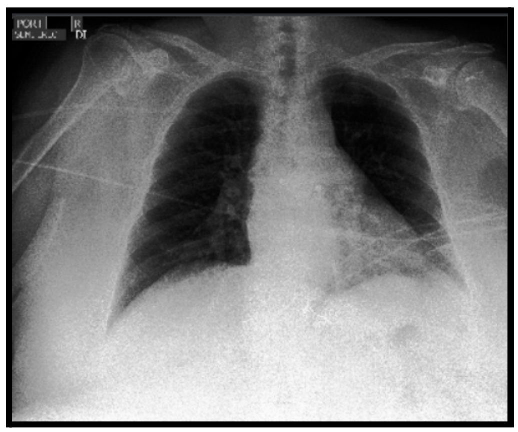
A Deep Learning Approach to Medical X-ray Analysis

Version 1.1
(Updates to be made as model improves)

Table of Contents

1.0 Aim

The aim of this project is to develop an automated system for analyzing chest X-ray images using deep learning techniques, specifically convolutional neural networks (CNNs) (2). The goal is to accurately detect and classify various medical conditions present in the X-rays, enabling faster and more efficient diagnostic support for healthcare professionals. By leveraging large datasets and advanced machine learning algorithms, this project aims to improve the accuracy and reliability of chest X-ray interpretation, ultimately contributing to better patient outcomes.



Classified as - Edema/ Pleural Effusion/ Other

2.0 Introduction

The aim of this project is to develop a deep learning model capable of accurately analyzing and interpreting chest X-ray images to detect various medical conditions. Specifically, we focus on utilizing convolutional neural networks (CNNs) for multi-label classification of chest X-rays, where each image can be classified into multiple possible conditions, such as pneumonia, lung opacities, and other abnormalities.

The primary objective is to create a model that can automatically classify chest X-ray images based on the presence or absence of these conditions. To achieve this, we use the CheXpert dataset (1), which contains a large number of labeled chest X-ray images. The dataset provides valuable annotations that allow us to train the model to recognize patterns associated with specific conditions, such as the presence of lung infections or other abnormalities.

The model architecture is built using a CNN, which is particularly well-suited for image classification tasks. The network is designed to handle the multi-class, multi-label nature of the problem, allowing it to predict the probability of each condition being present in an image. The model's performance is evaluated using various metrics, including AUROC (Area Under the Receiver Operating Characteristic Curve), sensitivity, and specificity, to ensure its ability to provide reliable and accurate predictions.

The goal of this project is to provide an automated system that can assist healthcare professionals by accurately identifying key medical conditions from chest X-ray images, thus helping to speed up the diagnostic process and improve overall efficiency in clinical settings.

3.0 Literature Review

In recent years, deep learning has transformed the field of medical image analysis, particularly in the classification of chest X-rays. Traditionally, image classification relied on handcrafted features and basic machine learning models like decision trees and support vector machines. However, these methods were limited in their ability to handle complex image data.

The introduction of Convolutional Neural Networks (CNNs) marked a significant leap, enabling models to automatically extract features from images and greatly improving performance. CNN-based architectures like AlexNet, VGG, and ResNet have been particularly successful in general image classification tasks and have been adapted for medical imaging, including chest X-rays.

Datasets like CheXpert have fueled advancements in the classification of multiple diseases from chest X-rays, moving beyond single-condition diagnosis to multi-label classification models. These models can now detect multiple coexisting conditions in a single X-ray image, making them more applicable to real-world medical scenarios.

Despite these advancements, challenges remain, such as the scarcity of large, labeled datasets, the difficulty of generalizing models across different hospital datasets, and the lack of interpretability in deep learning models. While models perform well on benchmark datasets, their application in clinical settings is hindered by these issues.

In conclusion, while deep learning has made significant strides in medical image classification, further research is needed to address issues related to dataset limitations, model generalization, and explainability to make AI-based tools more viable for clinical use.

4.0 Objectives

- **Data Preprocessing**: Clean and preprocess the CheXpert dataset ensuring proper image resizing and augmentation for model training.
- **Model Selection and Training**: Implement a Convolutional Neural Network (CNN) for multi-label classification of chest X-rays, leveraging transfer learning techniques.
- **Model Evaluation**: Assess the model's performance using appropriate metrics such as AUROC scores.
- **Hyperparameter Tuning**: Fine-tune the model's hyperparameters to optimize performance, using grid search or random search techniques.
- **Visualization**: Visualize the ROC curves for each class, highlighting the model's effectiveness in differentiating between positive and negative cases.
- **Analysis and Reporting**: Analyze the results, identify strengths and weaknesses, and document the findings in a detailed report.
- **Final Model Evaluation**: Evaluate the final model on the test set and compare it with baseline models or previous state-of-the-art methods in medical image classification.

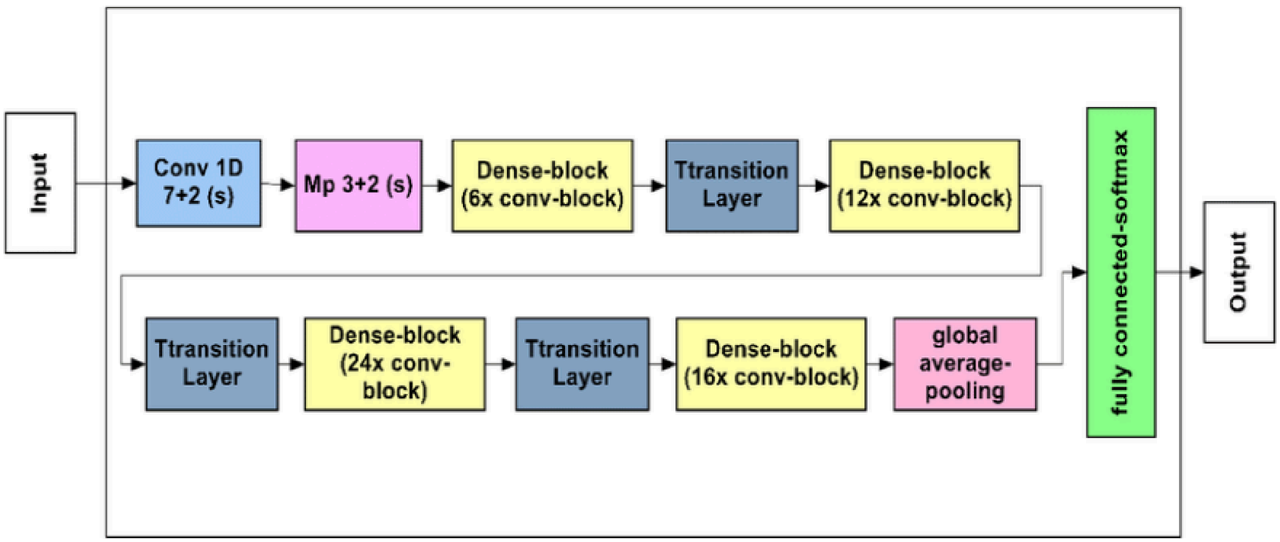
5.0 Approach

1- Dataset Selection and Preprocessing:

- The CheXpert dataset, which contains labeled chest X-ray images with multiple diagnoses per image, was selected for this task. The dataset is widely used for training and evaluating models in the medical image classification domain.
- We first performed data preprocessing, including resizing images to a consistent dimension suitable for the CNN model.

2- Model Selection - DenseNet:

- Given the complexity of the dataset and the need for efficient feature extraction, we employed DenseNet (Densely Connected Convolutional Networks) for model training. DenseNet is particularly suitable for medical image classification because it facilitates better gradient flow and encourages feature reuse by connecting each layer to every other layer in a dense block.
- We do not use a pretrained model since a pre-trained model may not hold a lot of valuable information due to the difference in type of images used here (X rays).



Densenet121 architecture

3- Multi-label Classification:

- The project aimed to classify images into multiple categories simultaneously (e.g., identifying conditions like pneumonia, lung opacity, etc.). Therefore, we used a multi-label classification approach with a sigmoid activation function for each class to output probabilities. This allowed the model to predict the likelihood of each condition independently, rather than assigning a single class label to each image.

4- Loss Function and Optimization:

- Cross-entropy loss was employed to handle multi-class classification, and the Adam optimizer was used to ensure stable and efficient convergence during training.

5- Federated Learning:

- To explore privacy-preserving and decentralized training, federated learning was integrated into the workflow (3). This approach allows the model to be trained across multiple devices or servers without sharing raw data, ensuring data privacy and security.

6- Model Tuning and Optimization:

- We performed hyperparameter tuning to optimize the performance of the model. This included adjusting learning rates, batch sizes, and other training parameters to ensure the model achieved the best possible results on the validation set.

7- Visualization and Interpretation

- The primary evaluation metric used was **AUROC (Area Under the Receiver Operating Characteristic Curve)**, as it is particularly useful for imbalanced datasets, which is common in medical imaging.
- For better interpretability and to assess the model's decision-making process, we visualized the ROC curves for each class, helping to understand the trade-offs between true positive and false positive rates for different threshold values

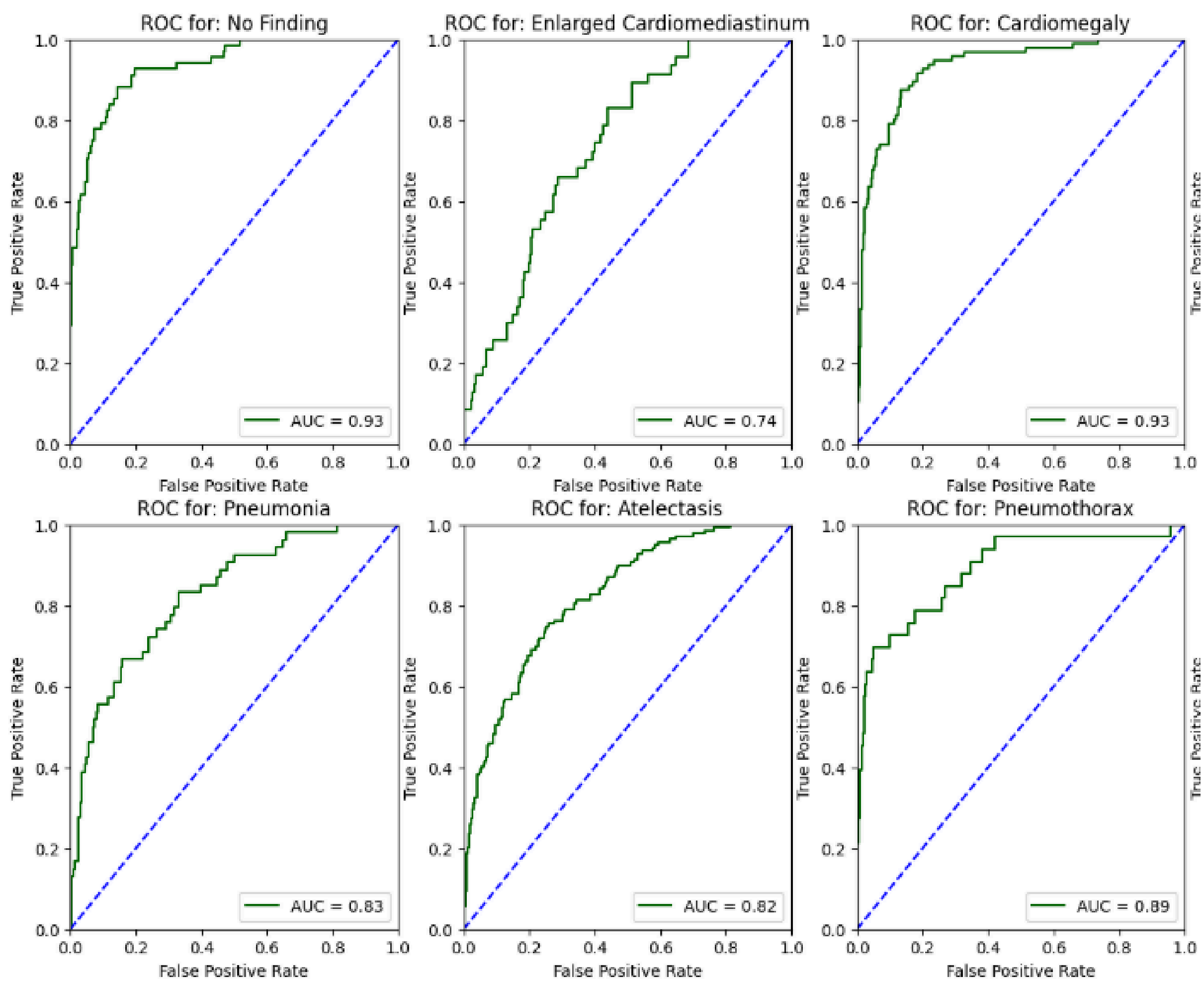
6.0 Results and Discussion

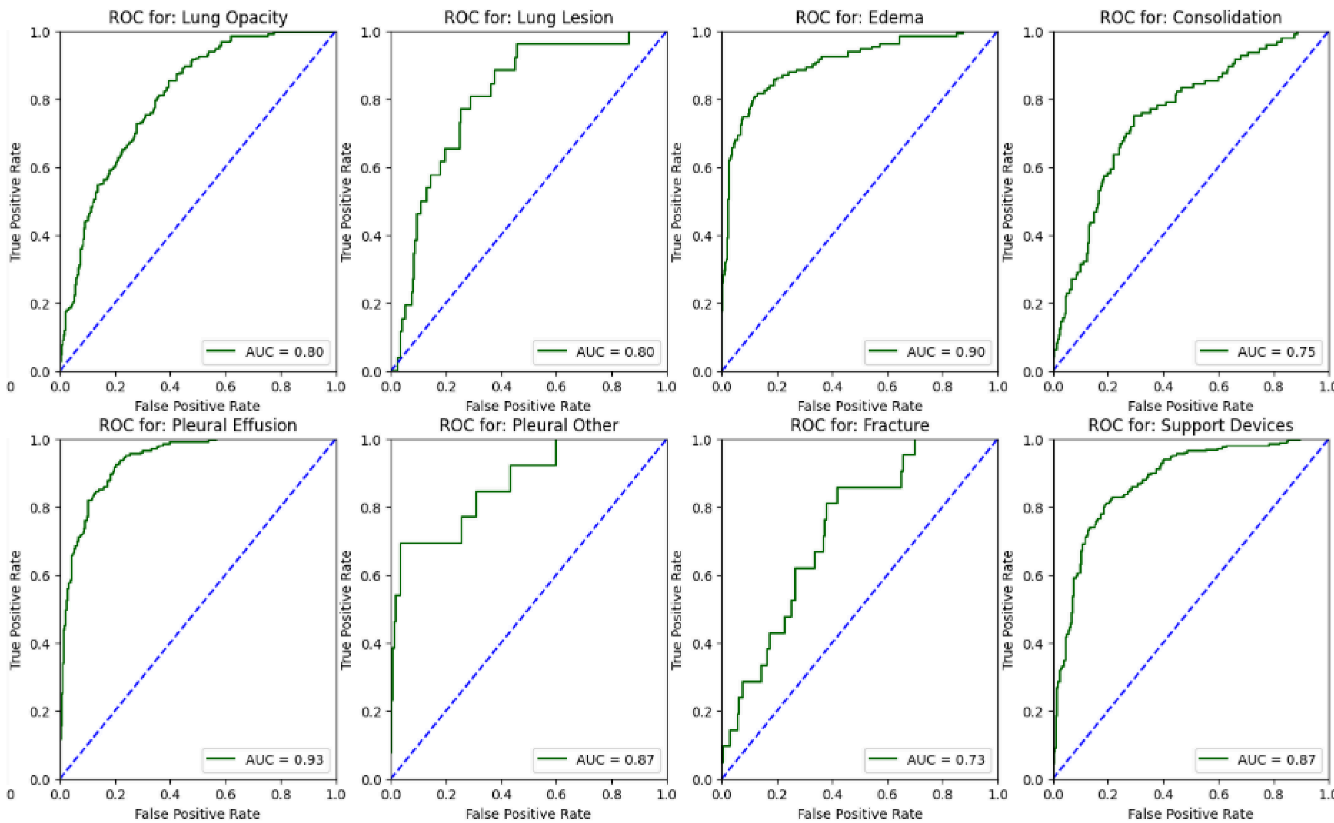
In this section, we will evaluate the performance of our model by comparing the AUC (Area Under the Curve) scores of our model across various conditions to those reported in the original paper. This comparison will help us understand how well our model performs under similar conditions and provide insights into its effectiveness.

Condition	**Our AUC Scores**	**Original AUC Scores**	**Difference**
Consolidation	0.75	0.90	- 0.15
Edema	0.90	0.92	- 0.02
Pleural Effusion	0.93	0.97	- 0.04
Cardiomegaly	0.93	0.90	+ 0.03
Atelectasis	0.82	0.85	- 0.03
Average AUC	0.86	0.91	- 0.05

As we can see even with a scaled down model (20,000 training images) we are able to perform relatively well, although not exactly as well as the model in the original paper (trained over a much larger dataset)

We will focus on evaluating the AUROC (Area Under the Receiver Operating Characteristic) for each class individually, as this metric is commonly used to assess the classification performance of multi-class models, particularly in imbalanced datasets. By comparing the AUC scores from both the original paper and our model, we aim to assess whether our model demonstrates comparable or improved performance in terms of classification accuracy and generalization across different classes.





AUC SCORES FOR EACH CLASS

We observe that the AUC score for 11 classes is above 0.8, and only 3 classes with a score in the range 0.7 - 0.8.

This shows that our model performs quite well across all classes, and perhaps certain classes can be better classified through some other methods (more data, class imbalances, etc.)

7.0 Conclusion

In this project, we developed and evaluated a model aimed at achieving robust classification performance, with a focus on comparing its effectiveness to the benchmarks established in the original paper. Through a comprehensive analysis, including the comparison of AUROC scores, we found that our model performed quite well across all classes, demonstrating strong generalization and accurate predictions.

While our model's AUROC scores were slightly below those reported in the original paper, the differences were marginal and indicate that our approach is highly competitive. These results highlight the potential of our chosen methods and architecture in addressing classification challenges effectively, even when evaluated against high-performing benchmarks.

Overall, this project provides valuable insights into the performance and applicability of advanced classification techniques, reaffirming their role in tackling complex datasets.

8.0 Future Work

While the model demonstrated strong performance and achieved competitive AUROC scores, there is room for further improvement. In future iterations of this work, we aim to explore the following directions to enhance model performance and broaden its applicability:

1. **Model Architecture Enhancements**: Experimenting with more advanced architectures, could improve classification accuracy and model efficiency.
2. **Data Augmentation**: Incorporating additional augmentation techniques could improve the robustness of the model by enabling it to learn from a more diverse dataset.
3. **Transfer Learning and Pretraining**: Utilizing larger, pre-trained models on similar datasets may help leverage prior knowledge and improve performance on our dataset.
4. **Addressing Class Imbalance**: Implementing techniques like focal loss or oversampling underrepresented classes could mitigate potential bias and ensure consistent performance across all classes.
5. **Ensemble Learning**: Combining predictions from multiple models could improve accuracy and reduce the risk of overfitting to specific patterns in the data.

By testing these enhancements, we may achieve better performance in this domain.

9.0 References

[1] CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison, Irvin et al., 2019 [arXiv:1901.07031]

[2] Densely Connected Convolutional Networks, Huang et al., 2018 [arXiv:1608.06993v5]

[3] Communication-Efficient Learning of Deep Networks from Decentralized Data, McMahan et al., 2017 [arXiv:1602.05629v3]