Automated Detection and Localization of Abnormalities in Chest X-rays using Weakly-Supervised Learning

Song Yiran (A0294421H), Wu Yuhang (A0294425Y), and Lu Jinzhou (A0329981M)
Department of Biomedical Engineering
National University of Singapore

Abstract

Medical AI holds immense promise, yet its widespread clinical adoption is hampered by two persistent challenges: the prohibitive cost of pixel-level annotations and the 'black box' nature of models. We propose a novel, AI-powered diagnostic framework to overcome these critical barriers, combining deep learning with explainable AI(XAI) to achieve both high accuracy and interpretability. The framework's performance was evaluated on the MIMIC-CXR dataset through a comparative analysis of the fine-tuned DenseNet-121 and ResNet-50 architectures. Furthermore, the integration of Grad-CAM heatmaps provides compelling visual evidence of our model's interpretability, provides an insight into the model's decision-making process.

1 Introduction

In the rapidly evolving landscape of modern medicine, chest X-rays remain an essential diagnostic cornerstone for multiple cardiopulmonary conditions. However, traditional interpretation is time-consuming and relies heavily on the physician's experience. Benefiting from advances in deep learning, the integration of AI into medical image analysis presents a promising future. Yet, despite the advantages, there are two non-negligible challenges: Challenge 1: High Cost of Data Annotation: Pixel-level annotation for medical images is expensive and time-consuming, a major bottleneck for AI applications. Challenge 2: "Black Box" nature of model: The opaque decision-making process of deep learning models leads to a lack of trust from clinicians. To address these critical challenges, we aim to develop an AI-assisted diagnostic framework with both high accuracy and high interpretability through two main functions: classification and localization. Our research uses the extensive MIMIC-CXR dataset, employing automated weak label extraction from radiology reports using the CheXpert NLP tool to circumvent challenge 1. Concurrently, we integrate advanced Explainability Analysis (XAI) techniques, transforming opaque model decisions into verifiable insights to overcome Challenge 2. In addition, we compare the performance of two widely adopted Convolutional Neural Networks in medical imaging, DenseNet-121 and ResNet-50, in this project. Through this comparative analysis, we aim to not only enhance diagnostic efficiency and accuracy but, crucially, to provide clinicians with transparent and reliable visual evidence, thereby fostering profound confidence in AI-driven diagnostic assistance.

2 Related work

Our project pipeline is built upon several state-of-the-art studies in the field of multi-label categorization of medical images like chest X-rays (CXR). The two primary inspirations are: Tanno and Barrett (2025)[1]and Thapa and Kaur (2025)[2]. Tanno and Barrett[1] pioneered a Vision-Language

Model (VLM) framework designed for the automated generation of structured radiology reports. Our project took inspiration from their NLP methodology and structural report processing for the efficient extraction of image-level 'weak labels', thus circumventing the prohibitive cost of pixel-level manual annotation, addressing 'Challenge 1' outlined in our introduction. Thapa and Kaur (2025)[2] proposed a deep-learning model (DenseNet121) for multi-label image classification of chest X-rays with three distinct techniques: Grad-CAM, LIME, and DeepLIFT, to offer comprehensive discernment into the model's decision-making process. This study serves as a direct and invaluable reference for our explainable deep-learning-based multi-label image classification approach for chest X-rays. Collectively, these two studies provide a robust theoretical and practical foundation for our project, guiding our choices in automated weak labeling and the integration of explainability.

3 Methodology

Our project flow is segmented into four main stages: Dataset Acquisition and Weak Labeling, Deep Learning Model Training, Explainability Analysis, along with Evaluation and Validation (Figure 1).

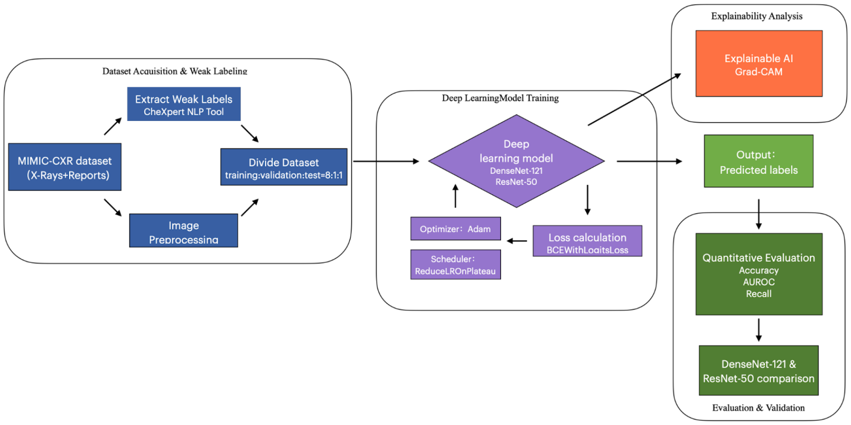


Figure 1: Deep Learning Model Training Workflow for Medical Image Analysis

3.1 Dataset Acquisition and Weak Labeling

Our project used the publicly available MIMIC-CXR dataset[3], a comprehensive collection of chest X-ray images and their corresponding free-text radiology reports from 1,000 patients, comprising 5,534 DICOM images. We employed the CheXpert NLP tool[4] to automatically extract image-level 'weak labels' from the free-text radiology reports. This powerful tool provides annotations for 14 distinct thoracic pathologies, allowing us to generate a rich, multi-label dataset without any manual pixel-level annotation effort.

Before feeding into the deep learning models, all images were resized to 224x224 pixels and normalized using ImageNet's mean and standard deviation values. The dataset was divided the dataset into training (80%: 4,594 images), validation (10%: 499 images), and test (10%: 441 images) sets.

Our data pipeline included a notable technical innovation: a custom PyTorch Dataset class that directly loads DICOM images. This method bypassed the slow pre-conversion process, significantly speeding up training setup and ensuring the preservation of original image integrity.

3.2 Deep Learning Model Training

We fine-tuned two state-of-the-art Convolutional Neural Networks (CNNs) widely recognized for their efficacy in medical imaging: DenseNet-121[5] (approximately 7M parameters) and ResNet-50[6] (approximately 23.5M parameters) for our multi-label classification task. A comprehensive comparison between DenseNet-121 and ResNet-50 can be found in section 4.2.

DenseNet-121 was selected as our primary architecture due to its unique dense connectivity pattern, which promotes feature reuse and significantly enhances parameter efficiency while achieving high accuracy. This efficiency is crucial for potential deployment in resource-constrained clinical environments and helps mitigate overfitting risks. The architecture of DenseNet-121 is composed of 121 layers and approximately 7 million parameters, making it significantly more parameter-efficient than many other deep learning models. It is structured around four DenseBlocks, with 6, 12, 24, and 16 layers respectively, interspersed by three Transition Layers(Figure 2, right).

In the meantime, ResNet-50 is a foundational architecture known for its deep residual learning, serving as a robust comparative benchmark to assess the performance advantages of DenseNet-121 within our framework. The ResNet-50 architecture is shown in Figure 2(left).

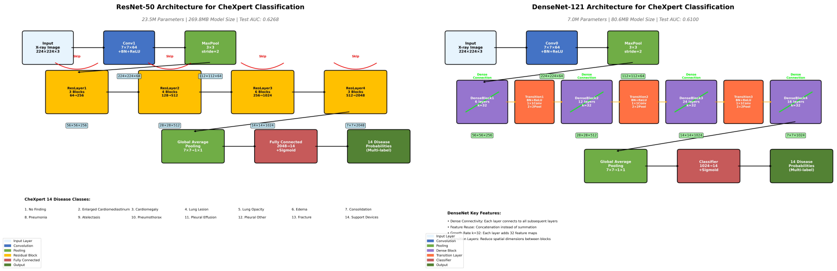

ResNet-50 Architecture for CheXpert Classification
23.5M Parameters | 269.8MB Model Size | Test AUC: 0.6268
Input X-ray Image 224×224×3
Conv1 7×7×64 +BN+ReLU
MaxPool 3×3 stride=2
Skip
ResLayer1 3 Blocks 64→256
ResLayer2 4 Blocks 128→512
ResLayer3 6 Blocks 256→1024
ResLayer4 3 Blocks 512→2048
224×224×64
112×112×64
56×56×256
28×28×512
14×14×1024
7×7×2048
Global Average Pooling 7×7→1×1
Fully Connected 2048→14 +Sigmoid
14 Disease Probabilities (Multi-label)
CheXpert 14 Disease Classes:
1. No Finding
2. Enlarged Cardiomediastinum
3. Cardiomegaly
4. Lung Lesion
5. Lung Opacity
6. Edema
7. Consolidation
8. Pneumonia
9. Atelectasis
10. Pneumothorax
11. Pleural Effusion
12. Pleural Other
13. Fracture
14. Support Devices
Input Layer
Convolution
Pooling
Residual Block
Fully Connected
Output
DenseNet-121 Architecture for CheXpert Classification
7.0M Parameters | 80.6MB Model Size | Test AUC: 0.6100
Input X-ray Image 224×224×3
Conv0 7×7×64 +BN+ReLU
MaxPool 3×3 stride=2
Dense Connection
DenseBlock1 6 layers k=32
Transition1 BN+ReLU 1×1Conv 2×2Pool
DenseBlock2 12 layers k=32
Transition2 BN+ReLU 1×1Conv 2×2Pool
DenseBlock3 24 layers k=32
Transition3 BN+ReLU 1×1Conv 2×2Pool
DenseBlock4 16 layers k=32
224×224×64
112×112×64
56×56×256
28×28×512
14×14×1024
7×7×1024
Global Average Pooling 7×7→1×1
Classifier 1024→14 +Sigmoid
14 Disease Probabilities (Multi-label)
DenseNet Key Features:
• Dense Connectivity: Each layer connects to all subsequent layers
• Feature Reuse: Concatenation instead of summation
• Growth Rate k=32: Each layer adds 32 feature maps
• Transition Layers: Reduce spatial dimensions between blocks
Input Layer
Convolution
Pooling
Dense Block
Transition Layer
Classifier
Output

Figure 2: DenseNet-121(right) and ResNet-50(left) architecture

Both models were initialized with weights pre-trained on the ImageNet dataset, followed by transfer learning to adapt their powerful feature extraction capabilities to our specific chest X-ray analysis task.

The training objective for our multi-label classification task was defined by the Binary Cross-Entropy Loss with Logits (BCEWithLogitsLoss). This loss function is particularly well-suited for scenarios where multiple independent labels can be present in a single image, providing stable gradients for learning. Uncertain labels (-1) were treated as positive (1), following the CheXpert convention. To iteratively minimize this loss and guide model convergence, we employed the Adam (Adaptive Moment Estimation) optimizer. This was paired with a ReduceLROnPlateau scheduler to dynamically reduce the learning rate when validation loss plateaued, further optimizing the training process.

3.3 Explainability Analysis

To directly confront challenge 2: the 'black box' nature, we integrated advanced Explainable AI (XAI) techniques, which are crucial for providing clinicians with insights into the model's decision-making process and enhancing trust. Specifically, we focused on Grad-CAM (Gradient-weighted Class Activation Mapping)[7] as our primary method for visual interpretability. Grad-CAM facilitates local interpretability by generating visual activation heatmaps. These heatmaps highlight the decisive regions of interest (ROIs) in the chest X-rays that most strongly influence the model's classification decisions, thereby offering a crucial visual understanding of where the model is 'looking' and aiding in lesion localization.

3.4 Evaluation and Validation

The model checkpoint with the highest validation Area Under the Receiver Operating Characteristic Curve(AUROC) was saved for final evaluation, model performance was rigorously evaluated on an independent test dataset (441 images), processed following the same pipeline as the training data to ensure consistency and prevent bias. Key quantitative evaluation metrics included the AUROC, which serves as our primary metric for assessing the model's discriminative ability across different pathologies. Additionally, F1-Score, Accuracy, Precision, and Recall were calculated at a 0.5 probability threshold to provide a comprehensive understanding of classification performance. Confusion matrices were also generated to visualize per-class classification performance and identify potential biases.

4 Experimental Results and Analysis

4.1 Experimental Setup

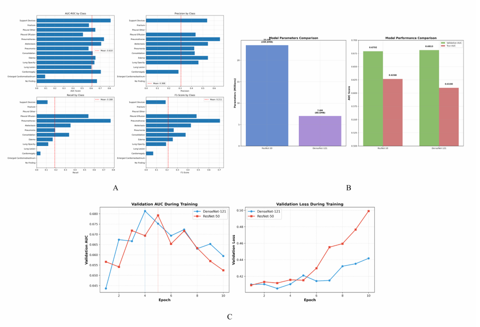
All models were trained on an NVIDIA GeForce RTX 4060 GPU, using the PyTorch 2.0+ framework for deep learning operations. We utilized torchvision for accessing pre-trained models, scikit-learn for comprehensive metric calculation, and pydicom for efficient DICOM image processing, ensuring a streamlined and optimized workflow.

Our training configuration employed an initial learning rate of 1×10^{-4}, a batch size of 16, and the Adam optimizer for efficient convergence. Models were trained for 10 epochs using Binary Cross-Entropy Loss with Logits (BCEWithLogitsLoss). Furthermore, a ReduceLROnPlateau scheduler (with patience=2 and factor=0.5) was implemented to dynamically reduce the learning rate when validation loss plateaued, strategically optimizing the training process and preventing overfitting.

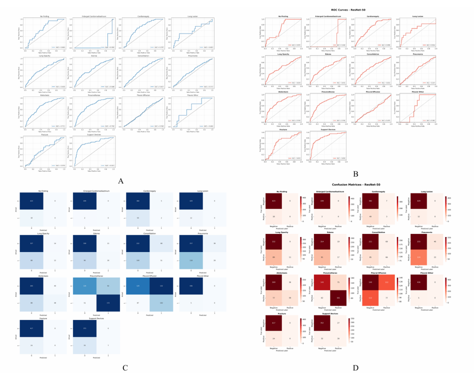
4.2 Results

Our experiments on the MIMIC-CXR dataset showed: a mean AUROC of 0.610, a mean F1-Score of 0.211, a mean Recall of 0.188, and an Accuracy of 0.806 on the test set were achieved in the DenseNet-121 model(shown in Figure 3. a. A). The mean AUROC was achieved with a model size of 80.6MB and training convergence in 4 epochs, while ResNet-50 achieved a Mean AUROC of 0.627, a F1-score of 0.237, and an Accuracy of 0.810 on the test set.

Both architectures demonstrated comparable performance, with ResNet-50 achieving slightly higher AUROC (0.627 vs 0.610) while DenseNet-121 excelled in parameter efficiency (7M vs 23.5M parameters). (comparison shown in Figure 3. a. B, C, and Figure 3. b)



(a) The Performance of DenseNet-121 and ResNet-50 models. A. The DenseNet-121 AUROC, Precision, Recall, and F1-Score results by class B. DenseNet-121 and ResNet-50 comparision C. DenseNet-121 and ResNet-50 training comparison



(b) ROC Curves of DenseNet-121(A) and ResNet-50(B), Confusion matrices of DenseNet-121(C) and ResNet-50(D)

Figure 3: Comparative Analysis of DenseNet-121 and ResNet-50: Performance Metrics and ROC Evaluation

To provide a granular understanding of our final model's (DenseNet-121) diagnostic capabilities, we also report the AUROC for each of the 14 individual thoracic pathologies, presented in Table 1. The model's performance varies discriminatively across different conditions, with particularly strong performance on common and well-defined abnormalities. The overall performance is further summarized by a Macro-average AUROC of 0.610 and a Weighted accuracy of 0.806.

Grad-CAM heatmaps provided compelling visual evidence of our model's interpretability. The heatmaps consistently demonstrated that the model correctly focused on anatomically relevant regions for its diagnostic decisions. For instance, in cases of Cardiomegaly, the heatmaps prominently highlighted the cardiac silhouette; for Pneumothorax, attention was drawn to the pleural space(see Figure 4).

Table 1: AUROC Performance Comparison of DenseNet-121 and ResNet-50 Models

Pathology	AUROC	Pathology	AUROC
Support Devices	0.815 / 0.826	Lung Opacity	0.636 / 0.639
Pneumothorax	0.742 / 0.746	Consolidation	0.617 / 0.631
Atelectasis	0.711 / 0.749	Lung Lesion	0.603 / 0.654
Cardiomegaly	0.707 / 0.688	Fracture	0.576 / 0.650
No Finding	0.665 / 0.695	Pneumonia	0.572 / 0.554
Edema	0.648 / 0.641	Pleural Effusion	0.510 / 0.475
Pleural Other	0.646 / 0.581	Enlarged Cardiomediastinum	0.092 / 0.246
Mean AUROC	**0.610 / 0.627**	**Weighted Accuracy**	**0.806 / 0.810**

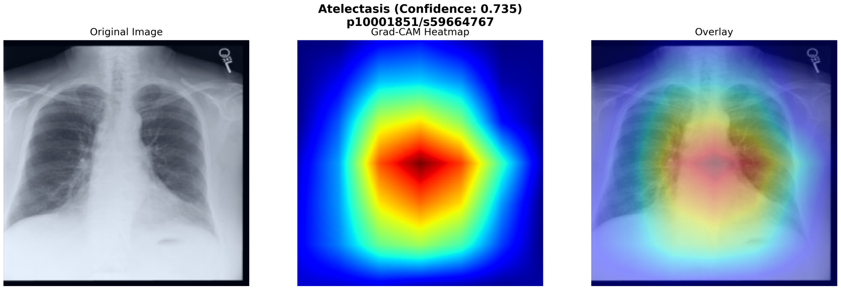


Figure 4: Grad-CAM visualization showing model attention on the lung apex for Atelectasis detection

5 Conclusion and Future Direction

5.1 Discussion

The comparative results on our test set illustrate the different performance trade-offs of DenseNet-121 and ResNet-50. This analysis highlights the trade-off between parameter efficiency (7M for DenseNet-121 vs. 23.5M for ResNet-50) and the specific diagnostic strengths demonstrated by each architecture.

A deeper analysis of the confusion matrices (Figure 3(b)) reveals complex performance trade-offs across the 14 individual pathologies, indicating distinct diagnostic preferences between the two architectures. For instance, ResNet-50 exhibited higher Recall (lower false negative rate) for specific categories such as Atelectasis and Consolidation. Conversely, DenseNet-121 showed stronger detection capabilities for other key pathologies, including Pneumothorax and Pleural Effusion.

The variability observed in AUROC scores across different CheXpert classes (e.g., extremely low scores for Enlarged Cardiomediastinum or Pleural Other) is primarily due to intrinsic dataset limitations. Specifically, severe class imbalance (rare pathologies providing few positive samples) leads to unstable model learning, while the NLP-based labeling introduces noise and uncertainty, particularly for ambiguous findings like Pneumonia or Lung lesions. The relatively high AUROCs in frequent classes (e.g., Support Devices) further suggest partial overfitting to dominant categories. Overall, these results highlight the critical need for class-balanced loss functions and label uncertainty modeling in future work, rather than indicating model instability.

Regarding the Grad-CAM result, the consistent alignment of highlighted regions with clinically relevant anatomical structures provides compelling evidence of the model's decision-making process. Even though the visualization results demonstrate broad patterns rather than precise and localized identification, this outcome precisely highlights the value of interpretability, as it reveals the model's true behavior, including both its strengths and limitations. This level of transparency is crucial for the clinical adoption of medical artificial intelligence.

5.2 Future Work

Several promising avenues for future research can emerge to further enhance the model's capabilities and clinical utility: First of all, advanced models can be explored, such as Vision-Language Models (VLMs), inspired by works like Tanno and Barrett (2025)[1]. These models can not only classify and localize but also generate radiology reports directly from images, thereby providing a more comprehensive AI assistant.

Future efforts will also be focused on refining the localization capabilities, which can also be combined with the possibility of translating these visual insights into radiological findings or full image-to-text reports, to provide a more transparent decision-making process and enhance clinicians' trust. In addition, to qualitatively validate the Grad-CAM results, experts' confirmation is expected to be integrated. Moreover, to translate this project from lab to clinic, future work will involve conducting longitudinal studies and initiating pilot deployments in real-world clinical settings, allowing for continuous feedback and iterative improvement.

5.3 Limitations

Despite the significant advancements achieved, our current framework has several limitations. This project is designed to automatically extract the weak labels from reports by the CheXpert NLP tool, however, this is more of a trade-off solution. While it is highly resource-efficient compared to traditional manual annotation, these labels do not offer the same exact precision, which might lower the model's final performance. It can also lead to mistakes if the original radiology reports are vague or misinterpreted. Our model's ability to generalize depends heavily on the MIMIC-CXR dataset's diversity. Even though the dataset is large, it might not fully cover all patient types, imaging methods, or disease variations seen in different hospitals. Therefore, the confirmation of external and diverse datasets is essential.

Although Grad-CAM[7] is great for visualizing the model's focus, it only gives us broad localization (coarse heatmaps) instead of precise, pixel-level boundaries. This can be an issue for finding small or subtle abnormalities. Also, like many XAI tools, Grad-CAM's explanations are based on the model's learned patterns, it's hard to truly measure how accurate these explanations are. Last but not least, as discussed in section 5.2, this project lacks validation from clinicians, which is essential to fully evaluate the framework's impact on diagnostic efficiency, accuracy, and workflow in a live clinical environment.

References

[1] Tanno, R., Barrett, D.G.T., Sellergren, A., et al. (2025) Collaboration between clinicians and vision-language models in radiology report generation. *Nature Medicine* **31**(2):599–608.

[2] Thapa, M. & Kaur, R. (2025) An explainable deep-learning based multi-label image classification for chest X-rays. *Procedia Computer Science* **239**:281–288.

[3] Johnson, A.E., et al. (2019) MIMIC-CXR: A large publicly available database of labeled chest radiographs. *Scientific Data* **6**(1):317.

[4] Irvin, J., et al. (2019) CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *AAAI Conference on Artificial Intelligence*.

[5] Huang, G., et al. (2017) Densely connected convolutional networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2261–2269.

[6] He, K., et al. (2016) Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778.

[7] Selvaraju, R.R., et al. (2017) Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *IEEE International Conference on Computer Vision (ICCV)*, pp. 618–626.

[8] Rohmah, L.N. & Bustamam, A. (2020) Improved classification of coronavirus disease (COVID-19) based on combination of texture features using CT scan and X-ray images. In *2020 3rd International Conference on Information and Communications Technology (ICOIACT)*, pp. 105–109.

[9] Khader, F., Kather, J.N., et al. (2023) Medical transformer for multimodal survival prediction in intensive care: integration of imaging and non-imaging data. *Scientific Reports* **13**(1):10666.