2023/09/08

# Automated Chest X-Ray Report Generator Using Multi-Model Deep Learning Approach

**1st Arief Purnama Muharram**; 2nd Hollyana Puteri Haryono;
3rd Abassi Haji Juma; 4th Ira Puspasari; 5th Nugraha Priya Utama



# Presentation Outline



Introduction

Methodology





Result & Discussion

Conclusion





# Introduction

# Background

- Deep learning for image processing has rapidly advanced in recent years [1] [2].
  - Demonstrating promising results across a variety of domains, including **radiology** [2].
- There is a **demand** for automated systems that can **accurately** and **quickly** process the **growing volume of medical images**.
- Chest X-ray (CXR) imaging, in particular, is one of the most often utilized radiologic diagnostic tools, although **its interpretation can be difficult even for the experienced radiologists** [3].
- Radiologists can benefit from automated CXR report generation since it allows them to focus on more difficult cases while reducing the possibility of errors.



1. Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," Nature, vol. 521, pp. 436-444, 2015.
2. J. H. Thrall, X. Li, Q. Li, C. Cruz, S. Do, K. Dreyer, and J. Brink, "Artificial Intelligence and Machine Learning in Radiology: Opportunities, Challenges, Pitfalls, and Criteria for Success," J. Am. Coll. Radiol., vol. 15, no. 3, pp. 504-508, Mar. 2018.
3. C.S.Lee,P.G.Nagy,S.J.Weaver,andD.E.Newman-Toker,"Cognitive and system factors contributing to diagnostic errors in radiology," AJR. American journal of roentgenology, vol. 201, no. 3, pp. 611-617, 2013.

# Motivation

- There was growing research on deep learning in radiology images, especially in disease and abnormality detection.
- However, most of the studies have primarily focused on detecting or classifying a **single abnormality** in the image.
- Meanwhile, when dealing with an image, a radiologist must find **pathologic or abnormal features** that could appear in the image.
- Therefore, we proposed a multi-model deep learning-based automated CXR report generator system to help radiologists' work better and faster.



# Literature Review

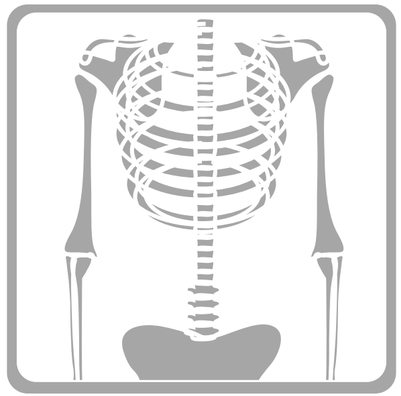
| Author | Year | Aim | Algorithm Used | Performance |
|---|---|---|---|---|
| Rajkomar et al. [4] | 2017 | Developing a deep convolutional neural network for classifying the view orientations of chest radiographs. | GoogLeNet | Accuracy of 100% and 38 image classifications per second. |
| Lakhani et al. [6] | 2017 | Developing CNNs for automated classification of pulmonary tuberculosis using chest radiograph | AlexNet, GoogLeNet | The best performance is achieved by the ensemble of AlexNet and GoogLeNet, with an Area Under the Curve (AUC) of 0.99. |
| Cicero et al. [5] | 2017 | Developing and test a deep convolutional neural network for frontal chest radiograph abnormality classification and computer-aided detection | GoogLeNet | The sensitivity, specificity, and AUC for pleural effusion were 0.91, 0.91, and 0.962, respectively; for pulmonary edema, they were 0.82, 0.82, and 0.868; for consolidation, 0.74, 0.75, and 0.850; for cardiomegaly, 0.81, 0.80, and 0.875; and for pneumothorax, 0.78, 0.78, and 0.861. |
| Saha et al. [7] | 2020 | Developing multi model based ensemble to detect COVID-19 from CXR Images | Multiple CNN models (CNNs, MobileNet, InceptionV3, DenseNet201, DenseNet121 and Xception) ensembled using feature concatenation and decision fusion | Accuracy of 96% using decision fusion scheme for 3 class classification and 89.21% for 4 class classification. Accuracy of 95.84% using feature concatenation scheme for 3 class classification and 89.26% for 4 class classification |
| Deb et al. [8] | 2022 | Developing a multi model ensemble based DCNN structure | VGGNet, GoogLeNet, DenseNet, and NASNet | An accuracy of 88.98% was achieved for the three-class classification, and a 98.58% accuracy was attained for the binary class classification using the publicly available dataset. |
| Tang et al. [9] | 2020 | Developing deep CNN model to classify normal vs abnormal CXRs | AlexNet, VGG, GoogLeNet, ResNet, and DenseNet | ROC score for ResNet18, ResNet50, GoogLeNet, and DenseNet are by around 98% using both NIH CXR and WCMC pediatric dataset. |
| Bhatt et al. [10] | 2021 | Developing classification model for pulmonary consolidations detection in CXR images using transfer learning and progressive resizing techniques | VGG-19 (baseline for binary classification) and EfficientNet-B3 (baseline for multi-class classification) | Achieved 100% accuracy |
| Amjoud et al. [11] | 2021 | Developing automated report generation for CXR using Transformer-based model | Transfer learning and transformer-based approach | Achieved BLUE-1, BLUE-2, and ROUGE metrics, where the scores were 0.479, 0.359, and 0.380, respectively. |
| Ghadekar et al. [12] | 2021 | Developing report generation using VGG for chest diseases diagnosis and encoder-decoder model for report generation. | VGG16 and encoder-decoder model | Accuracy of 88% for classification images and 96% for report generation. |

# Objective

**Create a system capable of detecting multiple abnormalities in radiology images.**

We **hypothesized** that by using a **multi-model approach**, with each model responsible for identifying a specific abnormality, we **could detect multiple abnormalities** within a single image.

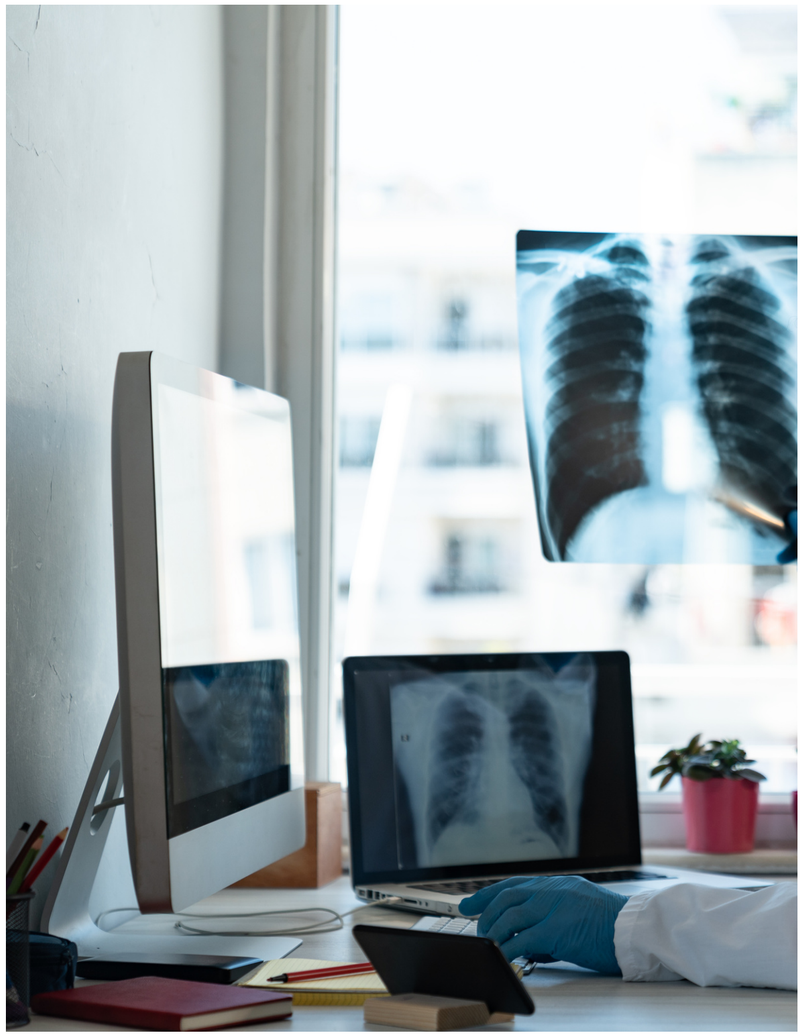
# Problem Limitations



**Adult Frontal CXR Dataset**

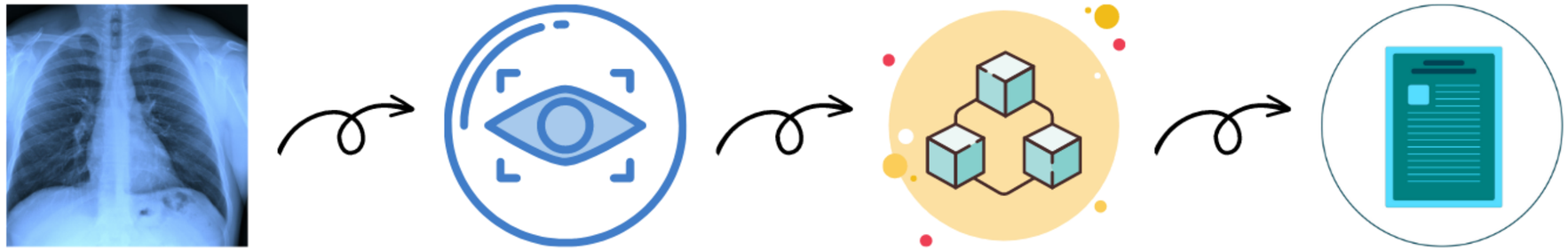



Abnormalities to detect:
**Cardiomegaly, Lung Effusion, and Consolidation**



# Methodology

# System Design



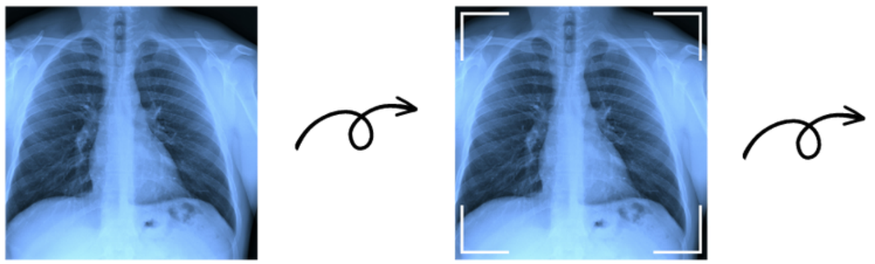
Chest
X-Ray Image

Image
Pre-processing

CXR
Deep Learning
Models

Report
Generation

# Step 1: Image Pre-processing



CXR Image

Square Crop
and
Image Resizing

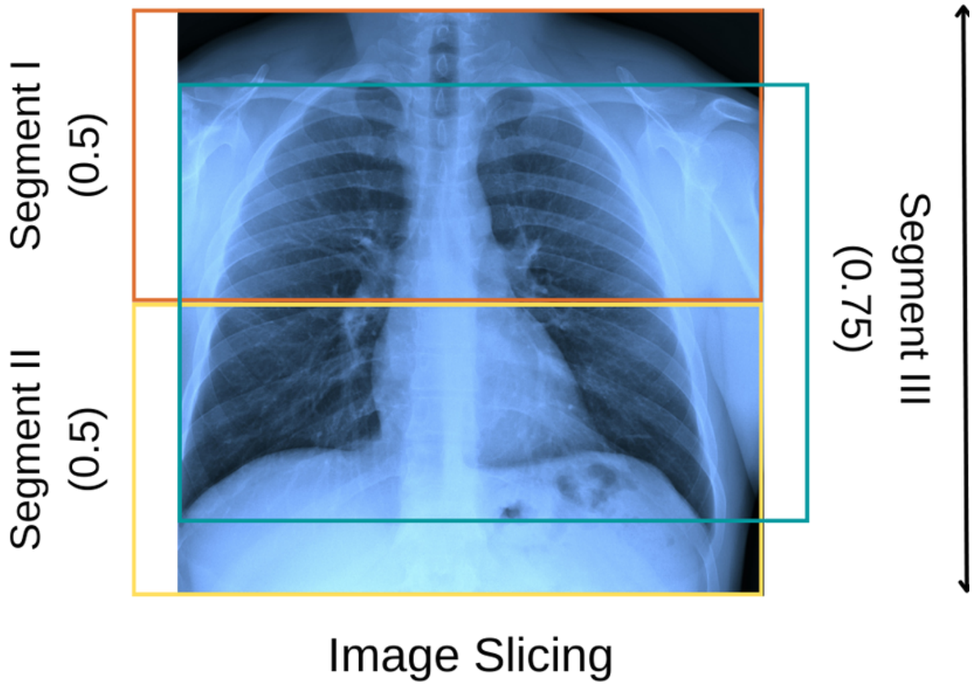
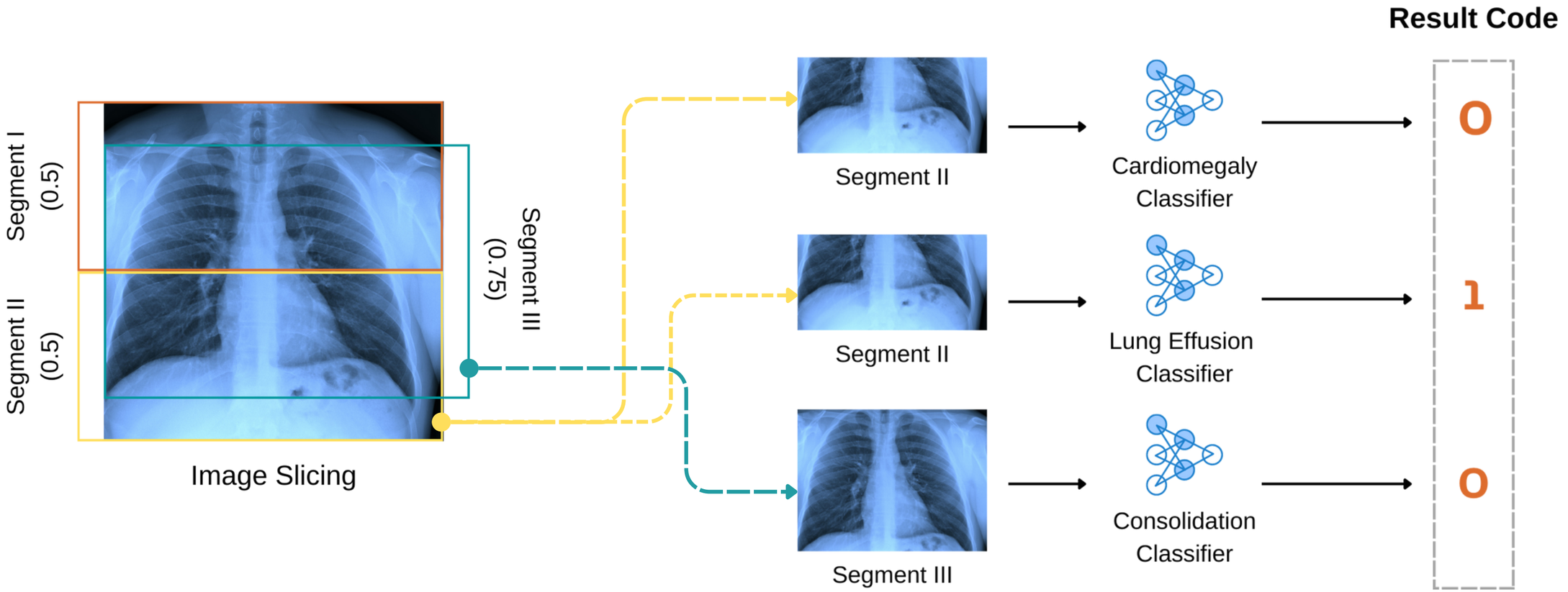



Image Slicing

Resize to 128x128 pixels

# Step 2: Deep Learning Models



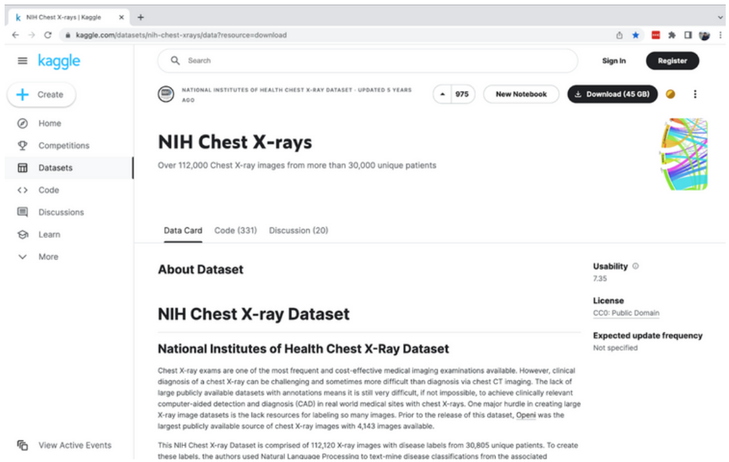
# Step 3: Report Generation

| Result Code | Master Text |
| --- | --- |
| | Pada foto radiologi dada |
| 0 | (0) Bentuk jantung baik, tidak ditemukan kardiomegali. / ~~(1) Terdapat kardiomegali, CTR > 50%.~~ |
| 1 | ~~(0) Tidak ditemukan tanda-tanda efusi.~~ / (1) Ditemukan tanda-tanda efusi. |
| 0 | (0) Tidak ditemukan tanda-tanda konsolidasi. / ~~(1) Ditemukan tanda-tanda konsolidasi.~~ |

# Step 3: Report Generation

Pada foto radiologi dada bentuk jantung baik, tidak ditemukan kardiomegali. Ditemukan tanda-tanda efusi. Tidak ditemukan tanda-tanda konsolidasi.

# Dataset Preparation




NIH Chest X-Rays Dataset [1]

Model Development Dataset

Splitting

Train Dataset (70%)

Test Dataset (30%)

System Evaluation Dataset



Manually Reannotated

Exclude images that meet the following criteria:

- Low image quality
- Incorrect orientation
- Misdiagnosed

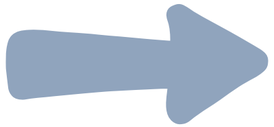


1.X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, "NIH Chest X-rays [Data set]," [Online]. Available: https://nihcc.app.box.com/v/ChestXray-NIHCC

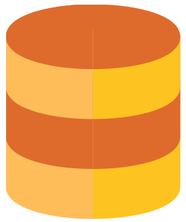
# Dataset Preparation



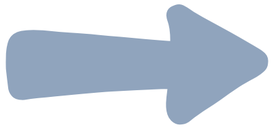
**Model Development Dataset**



| Abnormalities | Labels | | Total (N) |
|---|---|---|---|
| | False (n) | True (n) | |
| Cardiomegaly | 615 | 793 | 1408 |
| Lung Effusion | 658 | 896 | 1554 |
| Consolidation | 454 | 652 | 1106 |



**System Evaluation Dataset**



**200 Data**



1.X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, "NIH Chest X-rays [Data set]," [Online]. Available: https://nihcc.app.box.com/v/ChestXray-NIHCC

# Design of Experiment

**‘One factor at a time’**
Hyperparameter Tuning strategy

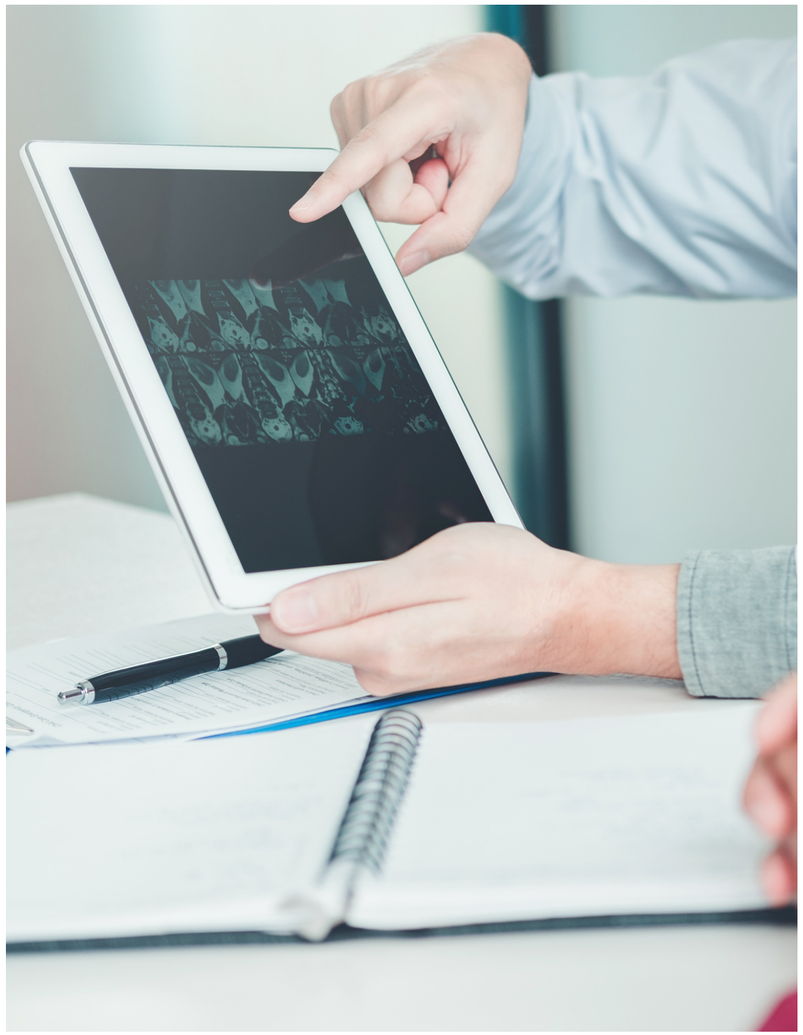
| Hyperparameter | Value |
|---|---|
| Pretrained model | ResNet50*, ResNet18, GoogLeNet |
| Learning rate | 1e-2, 1e-3*, 1e-4, 1e-5 |
| Optimizer | Adam*, SGD |

default parameter used

**Evaluation metric:**
Accuracy

$$accuracy = \frac{TP + TN}{TP + FP + FN + TN}$$

1. K.He,X.Zhang,S.Ren,andJ.Sun,"DeepResidualLearningforImage Recognition," in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770-778.
2. C. Szegedy et al., "Going deeper with convolutions," in 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1-9.



# Result & Discussion

# Model Development

HYPERPARAMETER TUNING RESULT FOR CARDIOMEGALY

| Step | Hyperparameter | Value | Accuracy | |
|---|---|---|---|---|
| | | | Training | Testing |
| 1 | Learning Rate | 1e-2 | 75.4% | 80.5% |
| | | **1e-3** (default) | **87.1%** | **87%** |
| | | 1e-4 | 82.6% | 81.5% |
| | | 1e-5 | 78.1% | 79.8% |
| 2 | Optimizer | **Adam** (default) | **87.1%** | **87%** |
| | | SGD | 72.4% | 80.5% |
| 3 | Pre-trained Model | ResNet50 (default) | 87.1% | 87% |
| | | ResNet18 | 85% | 90.3% |
| | | **GoogLeNet** | **90.1%** | **90.1%** |

**Selected Hyperparameter:**

1e-3, Adam, GoogLeNet

HYPERPARAMETER TUNING RESULT FOR LUNG EFFUSION

| Step | Hyperparameter | Value | Accuracy | |
|---|---|---|---|---|
| | | | Training | Testing |
| 1 | Learning Rate | 1e-2 | 77.5% | 78.4% |
| | | **1e-3** (default) | **90.2%** | **91.8%** |
| | | 1e-4 | 87% | 87% |
| | | 1e-5 | 82% | 84.6% |
| 2 | Optimizer | **Adam** (default) | **90.2%** | **91.8%** |
| | | SGD | 76.4% | 79.4% |
| 3 | Pre-trained Model | **ResNet50** (default) | **90.2%** | **91.8%** |
| | | ResNet18 | 91.5% | 87.4% |
| | | GoogLeNet | 87.8% | 89.4% |

**Selected Hyperparameter:**

1e-3, Adam, ResNet50

HYPERPARAMETER TUNING RESULT FOR CONSOLIDATION

| Step | Hyperparameter | Value | Accuracy | |
|---|---|---|---|---|
| | | | Training | Testing |
| 1 | Learning Rate | 1e-2 | 67.5% | 68.9% |
| | | 1e-3 (default) | 79.1% | 78% |
| | | **1e-4** | **79.1%** | **81%** |
| | | 1e-5 | 69.6% | 74.7% |
| 2 | Optimizer | **Adam** (default) | **79.1%** | **78%** |
| | | SGD | 66% | 65.6% |
| 3 | Pre-trained Model | ResNet50 (default) | 79.1% | 78% |
| | | **ResNet18** | **82.7%** | **78.4%** |
| | | GoogLeNet | 78.5% | 73.3% |

**Selected Hyperparameter:**

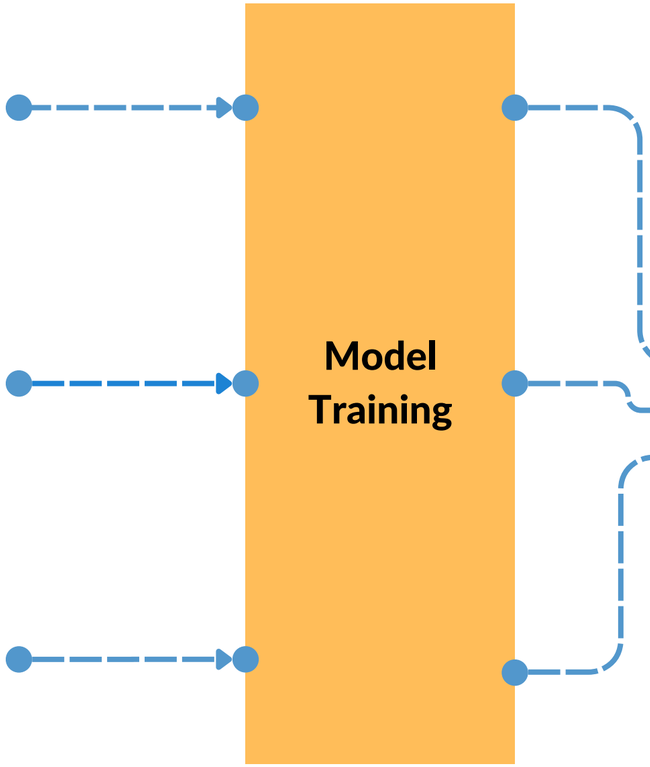
1e-4, Adam, ResNet18

# Model Development

**Cardiomegaly Model Hyperparameter:**
1e-3, Adam, GoogLeNet

**Lung Effusion Model Hyperparameter:**
1e-3, Adam, ResNet50

**Consolidation Model Hyperparameter:**
1e-4, Adam, ResNet18

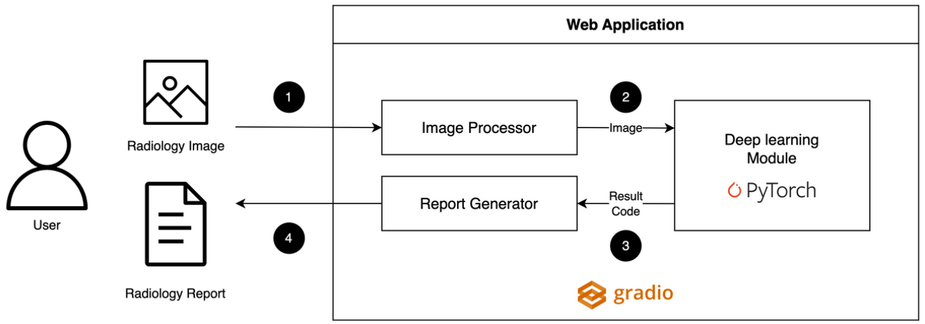


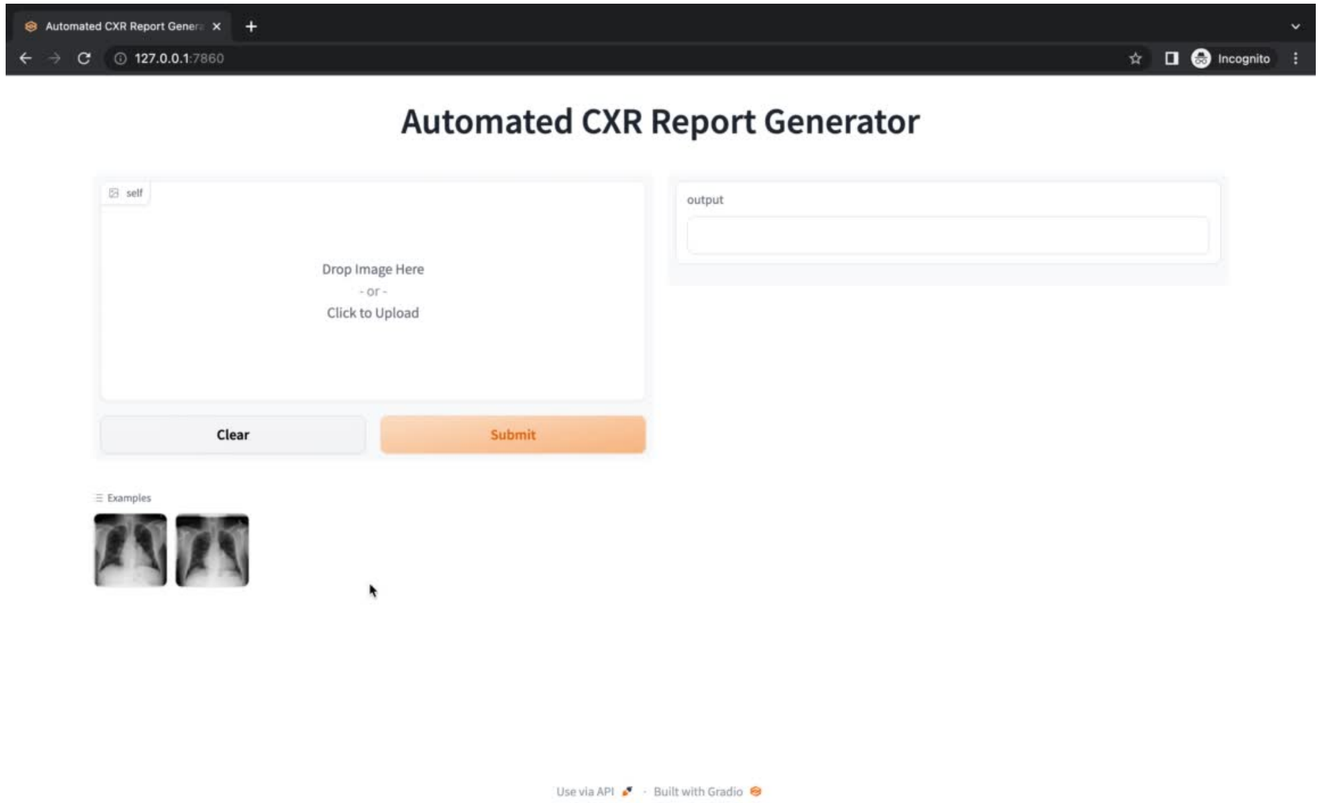

RESULT OF EACH FINAL MODEL

| Abnormality | Accuracy | |
|---|---|---|
| | Training | Testing |
| Cardiomegaly | 89.2% | 88.7% |
| Lung Effusion | 88.6% | 86.1% |
| Consolidation | 75.9% | 81% |

# System Evaluation

- The system was deployed as a monolithic web-based application using Gradio.
- The system architecture is as follows:



# System Evaluation

Results with **system evaluation dataset**:

| Abnormality | Accuracy |
|---|---|
| Cardiomegaly | 52% |
| Lung Effusion | 80% |
| Consolidation | 40% |
| **Cardiomegaly + Lung Effusion + Consolidation** | **20%** |

Comparison with the other studies:

- Amjoud et al., using transfer learning and transformer- based approach, achieved BLUE-1, BLUE-2, and ROUGE metrics, where the scores were 0.479, 0.359, and 0.380, respectively.
- Ghadekar et al., using VGG16 and encoder-decoder model, achieved accuracy of 88% for classification images and 96% for report generation.
- **Our proposed model, using multi-model deep learning approach, ahieved accuracy of 20%.**

Possible explanations:

- Medical images are **highly heterogeneous**
  - Increase the risk of **overfitting** and lead to a **loss of generalizability** for the models.
- **The errors of each model** contribute to the poor performance of the system.
  - The error rates of the cardiomegaly, lung effusion, and consolidation models were **48%**, **20%**, and **60%**, respectively.
  - The combination of these errors could make it difficult for the system to make accurate predictions.

# System Evaluation

**Mathematical explanation:**

- Let A, B, and C denote the cardiomegaly, lung effusion, and consolidation models, respectively.
- Pcorrect(A), Pcorrect(B), and Pcorrect(C) represent the probabilities of accurately predicting the cardiomegaly, lung effusion, and consolidation models, respectively.

Results with **system evaluation dataset**:

| Abnormality | Accuracy |
|---|---|
| Cardiomegaly | 52% |
| Lung Effusion | 80% |
| Consolidation | 40% |
| **Cardiomegaly + Lung Effusion + Consolidation** | **20%** |

$$P_{error}(A) = 1 - P_{correct}(A) = 1 - 0.52 = 0.48 \quad (1)$$

$$P_{error}(B) = 1 - P_{correct}(B) = 1 - 0.8 = 0.2 \quad (2)$$

$$P_{error}(C) = 1 - P_{correct}(C) = 1 - 0.4 = 0.6 \quad (3)$$

$$P_{correct}(A \cap B \cap C) = P_{correct}(A) \times P_{correct}(B) \times P_{correct}(C) = 0.52 \times 0.8 \times 0.4 = 0.1664 \quad (4)$$

$$\begin{aligned} P_{error}(A \cup B \cup C) &= P_{error}(A) + P_{error}(B) + P_{error}(C) \\ &\quad - P_{error}(A \cap B) - P_{error}(A \cap C) \\ &\quad - P_{error}(B \cap C) + P_{error}(A \cap B \cap C) \\ &= 0.48 + 0.2 + 0.6 - (0.48 \times 0.2) \\ &\quad - (0.48 \times 0.6) - (0.2 \times 0.6) \\ &\quad + (0.48 \times 0.2 \times 0.6) \\ &= 0.8336 \end{aligned} \quad (5)$$



# Conclusion

# Conclusion

- We developed an automated CXR report generator based on multi-model deep learning for detecting and classifying abnormalities in CXR images.
- Although each individual model yields good accuracy (> 80%), the system's accuracy, when tested using the system evaluation dataset, only reached **20%**.
  - This finding indicates that the multi-model approach is **not promising**, as its performance is **significantly dependent on and affected by the individual models** that constitute it.
  - The poor performance in one of the constituent models could significantly affect the overall system performance.
- Further investigations are needed to develop a robust and responsible method that can detect multiple abnormalities present in the CXR.

# Thank You

