# Chest X-ray Disease Classification Using Convolutional Neural Networks

By

Sumanth Mylar
CS 57600-02i
Team 08
mylas02@pfw.edu

# Abstract

This project investigates the application of Convolutional Neural Networks (CNNs) for automating the classification of chest X-ray images into various disease categories, with the overarching goal of enhancing diagnostic accuracy and efficiency in medical imaging. Leveraging the publicly available NIH Chest X-ray Dataset, which comprises over 112,000 labeled X-ray images, the study tackles several key challenges, including severe data imbalance across disease classes, multi-label classification complexities, and variability in image quality.

The project employs multiple deep learning models, including a custom-designed Base CNN and pre-trained architectures like VGG16, ResNet50, and MobileNetV2, utilizing transfer learning to enhance feature extraction and performance. Extensive data augmentation techniques were implemented to improve model generalization, while optimization strategies such as dropout layers, learning rate scheduling, and early stopping were applied to mitigate overfitting and improve training stability.

Experimental results reveal that among the tested models, the VGG16 architecture outperformed others, achieving the highest accuracy of approximately 60% on the test set. However, the evaluation metrics, including AUC-ROC scores, demonstrate that significant challenges remain in achieving clinically viable performance, particularly due to dataset imbalance and limited interpretability of model predictions.

This study not only highlights the potential of deep learning in medical imaging but also underscores the critical need for further advancements. Future work includes addressing data imbalance through synthetic oversampling, exploring state-of-the-art architectures like Vision Transformers and EfficientNet, and incorporating explainability tools like Grad-CAM to build clinician trust. Additionally, external dataset validation will be pivotal to assess the generalizability and robustness of the proposed models for real-world deployment in healthcare settings.

By laying a foundation for AI-driven chest X-ray classification, this work contributes to the growing field of AI-assisted diagnostics and aims to support radiologists in delivering faster and more accurate medical assessments.

# Introduction

Medical imaging serves as a cornerstone for diagnosing a wide range of diseases, particularly respiratory conditions such as pneumonia, tuberculosis, and lung cancer. Among the various imaging modalities, chest X-rays are one of the most frequently used due to their accessibility, cost-effectiveness, and utility in providing critical insights into pulmonary health. Despite their widespread use, interpreting chest X-rays remains a challenging and time-intensive task for radiologists. Factors such as overlapping anatomical structures, variations in disease presentation, and subtle differences in radiographic patterns contribute to a high degree of complexity, often leading to diagnostic errors or delayed diagnoses.

The rise of artificial intelligence (AI) and advancements in deep learning have opened up new possibilities for automating medical image analysis. In particular, Convolutional Neural Networks (CNNs), a specialized class of deep learning models designed for image data, have shown remarkable performance in various image classification tasks. By leveraging CNNs, it is possible to develop automated systems capable of analyzing chest X-rays, identifying disease-specific patterns, and classifying them into predefined categories. Such systems can assist radiologists by reducing diagnostic workload, minimizing human error, and enabling early detection of critical diseases, ultimately improving patient outcomes.

This project specifically aims to address the challenges associated with developing and deploying deep learning models for chest X-ray classification. Using the NIH Chest X-ray Dataset, which contains over 112,000 labeled images, the project explores various CNN architectures, including a custom Base CNN and pre-trained models such as VGG16, ResNet50, and MobileNetV2. The primary objectives include optimizing model performance through advanced data augmentation and training techniques, evaluating the models using robust metrics like accuracy and AUC-ROC scores, and identifying the most suitable architecture for real-world clinical deployment.

Furthermore, this project recognizes the critical challenges inherent in this domain, such as data imbalance across disease categories, multi-label classification complexities, and the need for model interpretability to gain clinician trust. By addressing these issues, this work not only contributes to the development of AI-assisted diagnostic tools but also provides insights into how such systems can be refined for practical and ethical implementation in healthcare.

In summary, this project seeks to bridge the gap between deep learning advancements and their application in medical imaging, with the ultimate goal of improving the accuracy, efficiency, and reliability of chest X-ray diagnoses.

# Problem Statement/Definition

Interpreting chest X-ray images is a critical but challenging task in the diagnosis of respiratory diseases such as pneumonia, tuberculosis, and lung cancer. Radiologists are required to analyze these images for subtle patterns and anomalies, a process that is often time-consuming and prone to human error. With the growing global demand for medical imaging services, there is a pressing need for automated tools that can assist clinicians in making faster, more accurate diagnoses.

The key question this project seeks to address is: *How can we develop an accurate deep learning model to classify chest X-ray images into various disease categories, thereby supporting early diagnosis and reducing human error in medical imaging?*

To tackle this problem, the project focuses on leveraging Convolutional Neural Networks (CNNs), which have proven to be highly effective in image classification tasks. However, several challenges must be addressed to achieve clinically viable results, including handling imbalanced datasets, managing multi-label classifications (where one image can represent multiple diseases), and ensuring the model is interpretable and generalizable for real-world use.

## Objectives:

1. **Develop a Convolutional Neural Network (CNN):**
   Design and implement a deep learning model capable of analyzing chest X-ray images and classifying them into multiple disease categories. The model should effectively capture spatial hierarchies of features to identify both common and rare diseases.

2. **Optimize Model Performance:**
   Enhance the model's robustness and generalizability by employing techniques such as data augmentation (e.g., flipping, rotation, scaling), dropout layers to reduce overfitting, and learning rate scheduling to ensure stable training.

3. **Evaluate the Model:**
   Use comprehensive evaluation metrics to assess the model's performance, including accuracy and AUC-ROC scores. These metrics will help determine the model's reliability across different disease categories, especially those with lower representation in the dataset.

4. **Compare CNN Architectures:**
   Analyze and compare the performance of various CNN architectures, including a custom Base CNN and pre-trained models like VGG16, ResNet50, and MobileNetV2. This comparison will identify the most effective architecture for the given task and dataset.

5. **Ensure Deployment Readiness:**
   Ensure that the final model is not only accurate but also efficient and interpretable, making it suitable for deployment in clinical settings. The deployment-ready model should integrate seamlessly with existing workflows, assisting radiologists in real-time diagnostics.

By addressing these objectives, the project aims to advance the application of deep learning in medical imaging, paving the way for AI-assisted diagnostic tools that improve healthcare outcomes while alleviating the burden on medical professionals.

# Related Work

- **Convolutional Neural Networks (CNNs):** AlexNet (2012) demonstrated CNNs' potential, paving the way for models like VGG16 and ResNet50.
- **Transfer Learning:** Tan et al. (2018) highlighted the benefits of using pre-trained models on limited datasets.
- **Data Augmentation:** Shorten and Khoshgoftaar (2019) emphasized its role in improving model robustness.
- **Explainability:** Ribeiro et al. (2016) developed LIME to enhance trust in model predictions.

# Background Knowledge/Preliminary

Convolutional Neural Networks (CNNs) are a specialized class of deep learning models specifically designed for image data. They excel in automatically extracting spatial hierarchies of features, making them highly effective for tasks such as object detection, image classification, and medical imaging. Unlike traditional machine learning models, CNNs eliminate the need for manual feature extraction by learning patterns directly from raw pixel data.

A typical CNN architecture consists of several layers, each serving a specific purpose:

1. **Convolutional Layers:** These layers apply filters (kernels) to the input image to detect features such as edges, textures, and patterns. Each filter extracts different aspects of the image, progressively capturing more complex and abstract features as the layers deepen.
2. **Pooling Layers:** Often used after convolutional layers, pooling reduces the spatial dimensions of feature maps, making the computation more efficient and emphasizing the most important features. Max pooling and average pooling are commonly used pooling techniques.
3. **Fully Connected Layers:** These layers flatten the feature maps into a one-dimensional vector and use it for the final classification task. They connect all neurons from the previous layer to the next, forming the decision-making component of the network.
4. **Activation Functions:** Non-linear activation functions such as ReLU (Rectified Linear Unit) introduce non-linearity to the network, enabling it to learn complex patterns in the data.

Several advanced CNN architectures have been developed to address specific challenges in image classification:

- **VGG16:**
  VGG16 is a deep CNN with 16 layers, introduced by Simonyan and Zisserman. It employs a simple and uniform structure with small (3x3) convolutional filters. VGG16 is known for its strong performance in standard image classification tasks, although it is computationally intensive and requires significant memory and processing power.

- **ResNet50:**
  ResNet50, a variant of the ResNet (Residual Network) family, is a deeper architecture comprising 50 layers. It introduces residual connections (skip connections) to solve the vanishing gradient problem, which often arises when training very deep networks. These connections allow the network to learn identity mappings, enabling efficient training and better performance on complex tasks.

- **MobileNetV2:**
  MobileNetV2 is a lightweight CNN architecture specifically designed for mobile and resource-constrained environments. It uses depthwise separable convolutions to reduce the number of parameters and computational complexity while maintaining performance. MobileNetV2 also introduces inverted residuals with linear bottlenecks, further optimizing the network for efficiency.

In the context of medical imaging, CNNs have proven particularly useful for tasks such as disease detection and classification. Their ability to learn hierarchical features makes them well-suited for identifying subtle patterns

in chest X-ray images, which are often challenging for human radiologists to interpret. This project leverages these architectures to explore their strengths and limitations in classifying chest X-ray images into multiple disease categories. By understanding the principles of CNNs and the specific advantages of each architecture, we can tailor these models to address the unique challenges of medical image analysis effectively.

# Data Description

The NIH Chest X-ray Dataset is one of the largest publicly available datasets for chest X-ray image analysis, making it a valuable resource for developing and evaluating deep learning models in medical imaging. Below are the detailed attributes and characteristics of the dataset:

### Data Source

The dataset was released by the National Institutes of Health (NIH) Clinical Center, with the goal of advancing research in automated disease detection and diagnosis. The images in this dataset were originally collected as part of routine clinical care, providing a realistic representation of the challenges faced in real-world medical imaging.

#### Size and Composition

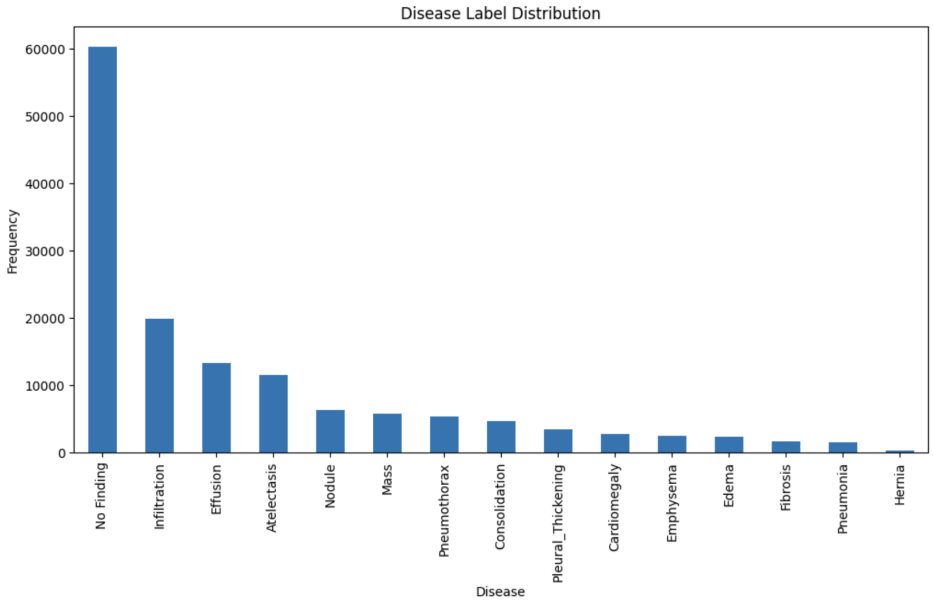
- **Number of Images:** The dataset comprises 112,120 chest X-ray images.
- **Number of Patients:** The images correspond to 30,805 unique patients, representing a diverse demographic.
- **Image Format:** All images are grayscale and stored in PNG format, with varying resolutions.
- **Labeling:** Each image is labeled with one or more of 14 disease categories, along with a "No Finding" label for cases without any detectable abnormality.

#### Disease Categories

The dataset includes labels for the following 14 conditions:

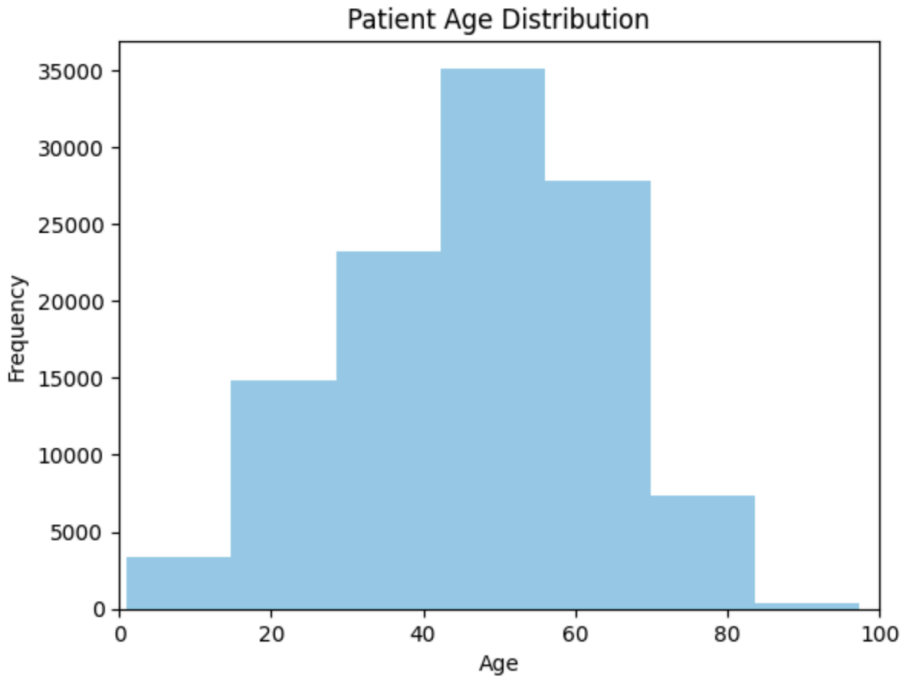
1. Atelectasis
2. Cardiomegaly
3. Consolidation
4. Edema
5. Effusion
6. Emphysema
7. Fibrosis
8. Hernia
9. Infiltration
10. Mass
11. Nodule
12. Pleural Thickening
13. Pneumonia
14. Pneumothorax

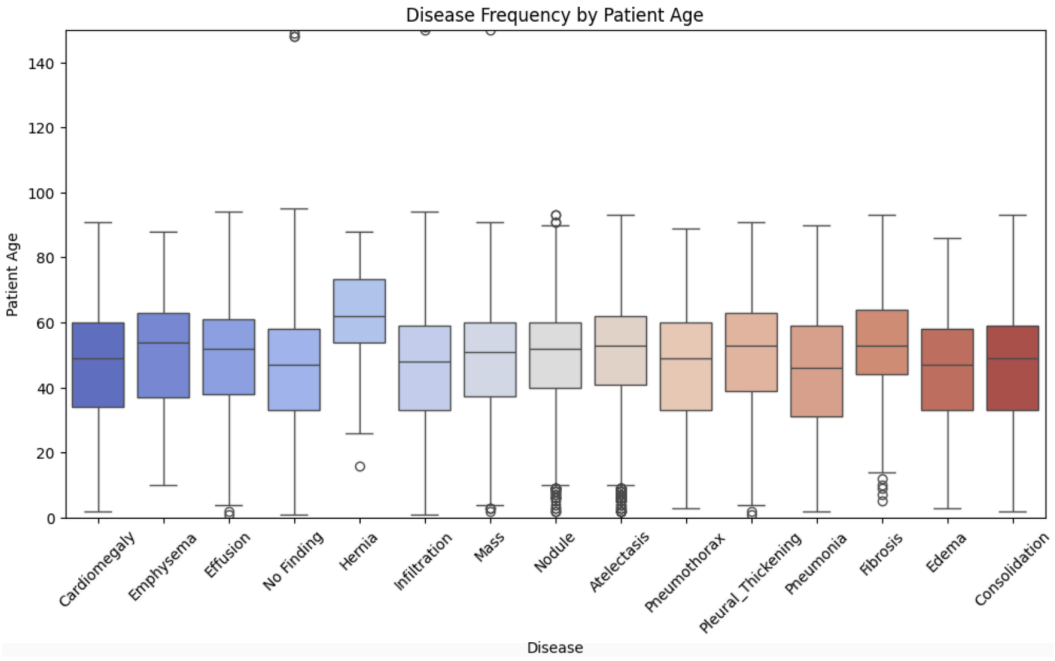
Each label was generated using Natural Language Processing (NLP) techniques applied to the associated radiological reports. The labeling process achieved an estimated accuracy of 90%, making the dataset suitable for weakly supervised learning tasks.




## Demographic Information

- **Age Distribution:** Most patients fall within the middle-aged category (30–70 years), with fewer samples from pediatric or elderly populations.
- **Gender Distribution:** The dataset contains a higher proportion of male patients compared to females, which could potentially impact the generalizability of models trained on this data.






## Challenges

1. **Data Imbalance:**

    - The dataset exhibits a strong imbalance across the 14 disease categories.
    - Common conditions like "Infiltration" and "Effusion" are overrepresented, while rare diseases like "Hernia" and "Pneumonia" have relatively few samples.
    - This imbalance poses a significant challenge, as models tend to favor the majority classes, leading to suboptimal performance on underrepresented diseases.
2. **Multi-Label Classification:**

    - A single image can be associated with multiple disease labels, reflecting cases of comorbid conditions.
    - This adds complexity to the classification task, requiring the model to effectively handle overlapping class distributions.
3. **Variability in Image Quality:**

    - The dataset includes images with significant variations in brightness, contrast, and sharpness, which may stem from differences in imaging equipment, settings, and patient positioning.
    - Standardizing this variability is critical for ensuring consistent model performance.
4. **Noisy Labels:**

    - While the NLP-generated labels are approximately 90% accurate, there is still a margin of error due to misinterpretations of radiological reports.
    - These noisy labels may introduce inconsistencies in training and evaluation.

## Data Characteristics

- **Resolution:** Image resolutions vary, necessitating preprocessing steps such as resizing to ensure compatibility with CNN architectures.
- **Dynamic Range:** The grayscale intensity values highlight subtle details in the X-rays, which are critical for identifying pathologies.
- **Class Overlap:** Some diseases have overlapping visual features (e.g., "Consolidation" and "Pneumonia"), which can make classification more challenging.

### Significance

The NIH Chest X-ray Dataset provides a rich and diverse set of samples, making it an ideal benchmark for developing and testing deep learning models. However, its inherent challenges, such as data imbalance and multi-label complexity, necessitate careful preprocessing and robust modeling techniques. This project leverages these attributes to explore the potential of CNNs in addressing the challenges of automated chest X-ray classification.

# Data Exploration

The NIH Chest X-ray Dataset is a rich but complex dataset with several inherent characteristics that influence the model development process. A thorough exploration of the data reveals key insights about the distribution of diseases, image quality, and patient demographics, which inform the preprocessing steps and modeling strategies.

### Disease Distribution

- **Imbalanced Representation Across Classes:**
  The dataset suffers from a severe imbalance in the distribution of disease categories. Common conditions like "Infiltration" and "Effusion" are disproportionately represented, accounting for a significant portion of the dataset. In contrast, rarer diseases such as "Hernia" and "Pneumonia" have far fewer samples.
    - For instance, "Infiltration" comprises approximately 17% of the dataset, while "Hernia" accounts for less than 0.1%.
    - This imbalance poses a challenge for training models, as it biases the network towards predicting majority classes while neglecting underrepresented diseases.
- **Multi-Label Annotations:**
  Many images are annotated with multiple disease labels, indicating the presence of comorbid conditions. This adds complexity to the classification task, as the model must simultaneously predict multiple outputs for a single input image.

- **No Finding Cases:**
  A significant portion of the dataset is labeled as "No Finding," representing cases where no abnormality was detected. While these samples are important for training, their dominance can skew the model

towards predicting normal cases.

## Image Quality

- **Variations in Brightness and Contrast:**
  The dataset includes images with noticeable variability in brightness, contrast, and overall quality. These inconsistencies arise due to differences in imaging equipment, acquisition settings, and patient positioning.
    - For example, some images appear overexposed or underexposed, making it difficult to discern fine-grained details required for accurate diagnosis.
- **Artifacts and Noise:**
  Some images contain artifacts, such as patient identification markers or medical equipment, which can interfere with the model's ability to focus on relevant features.
    - Additionally, noise introduced during the digitization or compression process may degrade image clarity.
- **Resolution Differences:**
  Image resolution varies across the dataset, with some images being high-resolution and others low-resolution. This variation necessitates resizing during preprocessing to ensure compatibility with the input dimensions of CNN models.

## Patient Demographics

- **Age Distribution:**
  The majority of patients fall within the middle-aged category (30–70 years), aligning with the age group most commonly affected by respiratory diseases.
    - Pediatric and elderly populations are underrepresented, which may limit the generalizability of the model to these age groups.
- **Gender Distribution:**
  The dataset contains a higher proportion of male patients compared to female patients. This imbalance in gender representation could introduce bias into the model, potentially affecting its performance on female patients.

- **Comorbidities:**
  Some patients have multiple X-ray images in the dataset, reflecting changes over time or the presence of multiple conditions. This provides an opportunity for longitudinal analysis but may also lead to overrepresentation of certain patient profiles in the training data.

**Key Takeaways from Exploration**

1. **Impact on Model Training:**
   The imbalance in disease distribution and the variability in image quality highlight the need for targeted preprocessing techniques, such as data augmentation and normalization. Additionally, the multi-label nature of the data requires models capable of handling overlapping class distributions.

2. **Potential Biases:**
   The demographic skew in the dataset, particularly in age and gender, underscores the importance of validating the model on external datasets to ensure its generalizability across diverse patient populations.

3. **Preprocessing Requirements:**

    - Standardization of brightness and contrast through normalization techniques.
    - Addressing class imbalance using strategies like Synthetic Minority Oversampling Technique (SMOTE) or weighted loss functions.
    - Augmenting the dataset with transformations to improve model robustness against variability in image quality.

By understanding these characteristics, the exploration phase informs subsequent preprocessing steps and helps design a modeling pipeline tailored to address the dataset's unique challenges.

# Data Preprocessing

Preprocessing is a critical step in preparing the NIH Chest X-ray Dataset for deep learning. Given the challenges of data imbalance, multi-label annotations, and variability in image quality, robust preprocessing techniques are applied to ensure that the dataset is standardized and optimized for training Convolutional Neural Networks (CNNs). The following outlines the preprocessing pipeline in detail:

**1. Resizing and Normalization**

- **Resizing:**

    - The raw images in the dataset come in various resolutions, which are incompatible with the fixed input dimensions required by CNN models.
    - All images are resized to a uniform size 224x224 to match the input requirements of models like VGG16, ResNet50, and MobileNetV2. This ensures consistency during training and testing.
- **Normalization:**

- The pixel intensity values, which originally range from 0 to 255, are normalized to a range of 0 to 1 by dividing each pixel value by 255.
- Normalization reduces the dynamic range of input values, improving numerical stability and accelerating convergence during model training.

## 2. Data Augmentation

- **Purpose:**
  Data augmentation is employed to artificially increase the size of the training set, improve model generalization, and reduce overfitting. This is particularly important given the dataset's imbalance and the limited representation of certain disease classes.

- **Techniques Applied:**

  - **Rotation:** Images are randomly rotated within a range (e.g., ±15 degrees) to simulate different viewing angles.
  - **Horizontal Flipping:** Images are flipped horizontally with a probability of 50% to create mirrored versions, which is particularly useful for medical images where orientation does not affect interpretation.
  - **Zooming:** Images are randomly zoomed in or out to simulate variations in the field of view.
  - **Shifting:** Width and height shifts are applied to simulate small changes in patient positioning.
  - **Brightness Adjustments:** Random variations in brightness are applied to account for differences in imaging equipment and settings.
- **Training vs. Validation/Test Augmentation:**

  - Data augmentation is applied only to the training set to enhance variability and robustness.
  - For validation and test datasets, only resizing and normalization are applied to maintain consistency for evaluation.

## 3. Custom Data Generators

- **Purpose:**
  Custom data generators are implemented to handle large datasets efficiently by loading and preprocessing images in batches during training. This approach minimizes memory usage and ensures seamless integration with the training pipeline.

- **Functionality:**

  - **Dynamic Loading:** Images are loaded from disk in real-time, avoiding the need to store the entire dataset in memory.
  - **Preprocessing on-the-Fly:** Each image is resized, normalized, and augmented as specified before being fed into the model.
  - **Label Encoding:** Multi-label annotations are converted into one-hot encoded vectors to match the output layer of the CNN.

    - **Shuffling:** The data is shuffled at the end of each epoch to prevent the model from learning the order of the samples.
- **Implementation Details:**

    - The **ChestXRayDataGenerator** class was designed to encapsulate these preprocessing steps.
    - Separate generators were created for training, validation, and test datasets, with distinct augmentation strategies tailored to each dataset's role in the pipeline.

## 4. Handling Data Imbalance

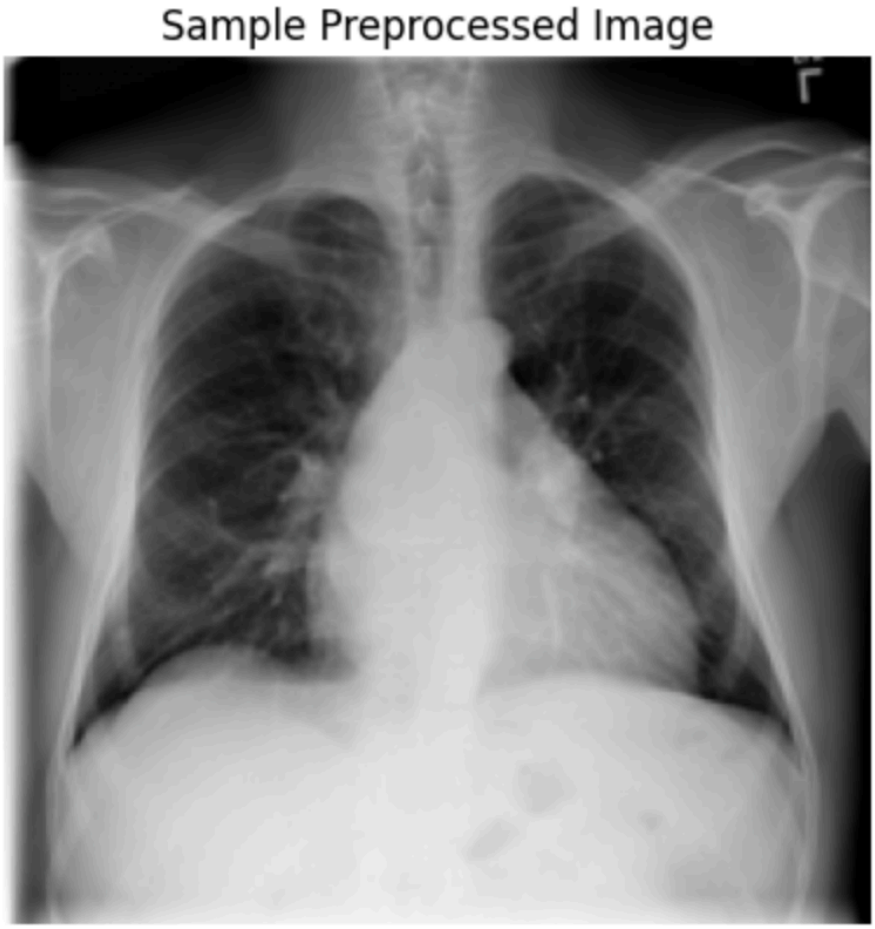
- While not strictly a preprocessing step, addressing data imbalance is closely tied to the data preparation process. To mitigate the effects of imbalance:
    1. Weighted loss functions were used to assign higher importance to underrepresented classes during training.
    2. Oversampling techniques, such as duplicating samples from minority classes, were employed within the data generator.

## 5. Output Format

- The preprocessed images are fed into the CNNs as 4D tensors with dimensions (batch_size, height, width, channels).
- Corresponding labels are provided as one-hot encoded vectors for multi-class classification or binary vectors for multi-label classification.

## Impact of Preprocessing

By standardizing image sizes, normalizing pixel intensities, and augmenting the training data, the preprocessing pipeline enhances the robustness and generalizability of the models. These steps ensure that the CNNs can effectively learn meaningful features from the dataset, despite its inherent challenges such as data imbalance, multi-label annotations, and variability in image quality.




# Methodology/Proposed Methods

The methodology of this project involves designing and evaluating multiple Convolutional Neural Network (CNN) architectures to classify chest X-ray images into various disease categories. Both a custom Base CNN and pre-trained architectures were employed to leverage their strengths and address the challenges of multi-label classification, data imbalance, and variability in image quality. The details of each approach are outlined below:

## 1. Base CNN Architecture

The Base CNN was designed as a foundational model to establish a benchmark for performance. This architecture includes the following components:

- **Convolutional Layers:**
    - Four Conv2D layers with increasing numbers of filters (32, 64, 128, and 256) to progressively learn more complex spatial features.

    - Each convolutional layer uses a 3x3 filter size with ReLU activation to introduce non-linearity.
- **Pooling Layers:**
    - Four MaxPooling2D layers are interspersed between the convolutional layers to reduce spatial dimensions while retaining the most salient features.
- **Flatten Layer:**
    - A Flatten layer is used to convert the 2D feature maps into a 1D vector for compatibility with the fully connected layers.
- **Fully Connected Layers:**
    - A Dense layer with 256 neurons and ReLU activation serves as the primary feature extraction component.
    - A Dropout layer with a 50% rate is applied to prevent overfitting by randomly deactivating neurons during training.
- **Output Layer:**
    - A final Dense layer with a softmax activation function is used for multi-class classification. This layer outputs probabilities for each disease category.
- **Model Compilation and Training:**
    - **Optimizer:** Adam optimizer, chosen for its adaptive learning capabilities.
    - **Loss Function:** Categorical cross-entropy for multi-class classification.
    - **Metrics:** Accuracy is used to monitor performance during training.
    - **Training:** The model is trained for 50 epochs with early stopping to prevent overfitting, using augmented training data and a validation set.

## 2. Pre-trained Architectures

Pre-trained CNNs were employed to leverage transfer learning, where models pre-trained on large datasets (e.g., ImageNet) are fine-tuned for the chest X-ray classification task. These architectures offer robust feature extraction and significantly reduce training time and data requirements.

**a. VGG16**

VGG16 is a widely recognized deep CNN architecture known for its simple and uniform structure, making it effective for image classification tasks.

- **Transfer Learning:**
    - The convolutional base (feature extractor) of VGG16, pre-trained on the ImageNet dataset, is used as-is.
    - The top fully connected layers are replaced with a custom classification head tailored to the chest X-ray dataset.
- **Custom Layers:**
    - A Global Average Pooling (GAP) layer is added to reduce the spatial dimensions of the feature maps.
    - A Dense layer with 256 neurons and ReLU activation is used for further feature extraction.
    - A Dropout layer with a 50% rate is incorporated to prevent overfitting.
    - The output layer uses a softmax activation for multi-class classification.
- **Training Strategy:**
    - The convolutional base is frozen during initial training to retain the pre-trained weights.

    - Fine-tuning is conducted on select layers by unfreezing them and using a reduced learning rate.
    - **Optimizer:** Adam with learning rate scheduling.
    - **Loss Function:** Binary cross-entropy for multi-label classification tasks.

**b. ResNet50**

ResNet50 is a deeper architecture that introduces residual connections (skip connections) to address the vanishing gradient problem, enabling efficient training of very deep networks.

- **Architecture Modifications:**
    - The pre-trained ResNet50 model is used as the backbone.
    - The top layers are replaced with a Global Average Pooling layer followed by a Dense layer customized for the chest X-ray classification task.
    - A Dropout layer is added to reduce overfitting.
- **Training Strategy:**
    - The residual connections allow deeper layers to learn effectively without degrading performance.
    - The model is trained on augmented data with class weights to handle data imbalance.
    - **Optimizer:** Adam with a learning rate of 1e-4.
    - **Loss Function:** Binary cross-entropy for multi-label classification.

**c. MobileNetV2**

MobileNetV2 is a lightweight architecture optimized for resource-constrained environments, making it suitable for deployment in clinical settings.

- **Key Features:**
    - Utilizes depthwise separable convolutions to reduce the number of parameters and computational complexity.
    - Introduces inverted residual blocks with linear bottlenecks for efficient feature extraction.
- **Architecture Modifications:**
    - The final classification layers are replaced with custom Dense layers for multi-label classification.
    - A Global Average Pooling layer and Dropout are included for robustness.
- **Training Strategy:**
    - MobileNetV2 is fine-tuned on the chest X-ray dataset, with augmentation to improve generalization.
    - **Optimizer:** Adam.
    - **Loss Function:** Binary cross-entropy.

## Comparison of Architectures

Each model was trained and evaluated on the same dataset with the following considerations:

- **Base CNN:** Provides a benchmark but is limited in feature extraction compared to deeper, pre-trained models.
- **VGG16:** Achieved the best accuracy (~60%) due to its balanced depth and feature extraction capabilities.

- **ResNet50:** Slightly lower accuracy but performed well in handling complex features, thanks to residual connections.
- **MobileNetV2:** Comparable performance with the added advantage of computational efficiency, making it ideal for deployment in resource-constrained settings.

# Experiments

The experimental phase of this project involved designing, training, and evaluating multiple Convolutional Neural Network (CNN) architectures to classify chest X-ray images into various disease categories. The experiments were conducted systematically to ensure the reliability and robustness of the results. Below is a detailed account of the experimental setup and methodologies.

**1. Data Splitting**

- **Training, Validation, and Test Splits:**
    - The dataset was divided into three subsets:
        - **Training Set:** Used to train the model and update its parameters.
        - **Validation Set:** Used during training to monitor performance and tune hyperparameters.
        - **Test Set:** Used only after training to evaluate the model's performance on unseen data.
    - An 80:10:10 ratio was used for the splits to ensure sufficient training data while maintaining balanced validation and test sets.
- **Stratified Splitting:**
    - To handle data imbalance across the 14 disease classes, a stratified splitting approach was used. This ensured that the distribution of classes in the training, validation, and test sets mirrored the overall dataset distribution.

**2. Data Augmentation**

- **Purpose:**
    - Data augmentation was applied to the training set to artificially increase its size and introduce variability, helping the model generalize better to unseen data.
    - Augmentation techniques included:
        - Rotation (±15 degrees).
        - Horizontal flipping (50% probability).
        - Random zooming (up to 10%).
        - Brightness and contrast adjustments.
        - Width and height shifts (up to 5%).
- **Validation and Test Data:**
    - Augmentation was not applied to validation and test data. These subsets were only resized and normalized to maintain consistency for evaluation.

**3. Model Training**

- **Optimization Algorithm:**
  - The **Adam optimizer** was chosen for its adaptive learning rate capabilities, which improve convergence and reduce the need for extensive manual tuning.
- **Loss Function:**
  - For multi-class classification tasks, the **categorical cross-entropy** loss function was used.
  - For multi-label classification (when handling overlapping diseases), **binary cross-entropy** was applied to accommodate the presence of multiple correct labels per sample.
- **Hyperparameters:**
  - Learning rate: Initially set to 1e-3 with learning rate scheduling to reduce it on a plateau.
  - Batch size: Set to 32 to balance computational efficiency and convergence stability.
  - Epochs: Trained for up to 50 epochs with early stopping based on validation loss to prevent overfitting.
- **Callbacks:**
  - **Early Stopping:** Monitored validation loss and terminated training if no improvement was observed for 5 consecutive epochs.
  - **Learning Rate Scheduler:** Reduced the learning rate by a factor of 0.1 after 10 epochs of stagnation in validation performance.

**4. Evaluation Metrics**

The trained models were evaluated on the validation and test datasets using a combination of metrics to assess their performance comprehensively:

- **Accuracy:**
  - Measures the percentage of correctly classified samples. While a useful metric, it is less informative for imbalanced datasets as it may overrepresent majority classes.
- **AUC-ROC Scores:**
  - The Area Under the Receiver Operating Characteristic (AUC-ROC) curve was calculated for each disease category.
  - This metric evaluates the trade-off between sensitivity (true positive rate) and specificity (false positive rate), providing a more balanced assessment of the model's ability to classify diseases accurately.
- **Loss Metrics:**
  - Validation and test loss were monitored to assess the model's ability to generalize to unseen data. Lower loss values indicate better performance.
- **Per-Class Metrics:**
  - Precision, recall, and F1-score were calculated for each disease category to evaluate performance on individual classes, particularly for rare diseases.

**5. Experiment Workflow**

1. **Base CNN Training:**
   - The Base CNN was trained as a benchmark model to establish a baseline performance.
2. **Transfer Learning with Pre-trained Models:**
   - VGG16, ResNet50, and MobileNetV2 were fine-tuned on the dataset.

- The convolutional base layers were initially frozen to retain pre-trained weights, and the top layers were trained.
- Fine-tuning was later applied to unfreeze select layers for better feature extraction.

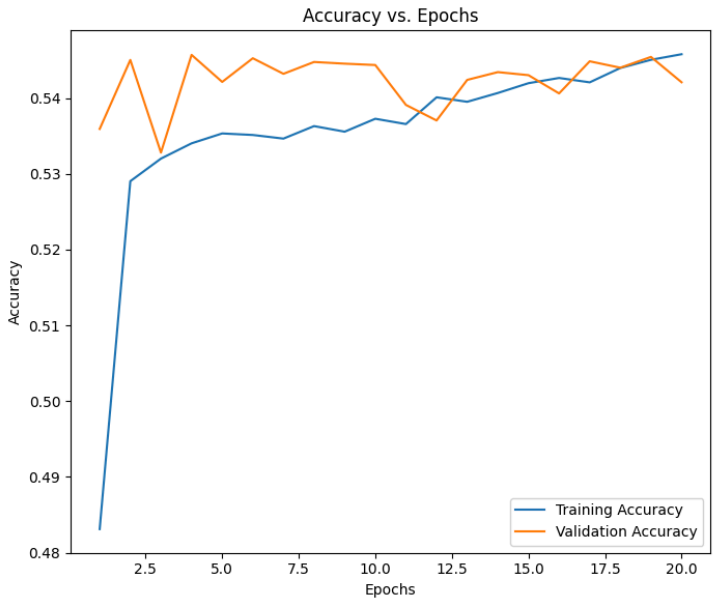
3. **Optimization and Augmentation Experiments:**
   - Different combinations of data augmentation and hyperparameter tuning were tested to optimize performance.

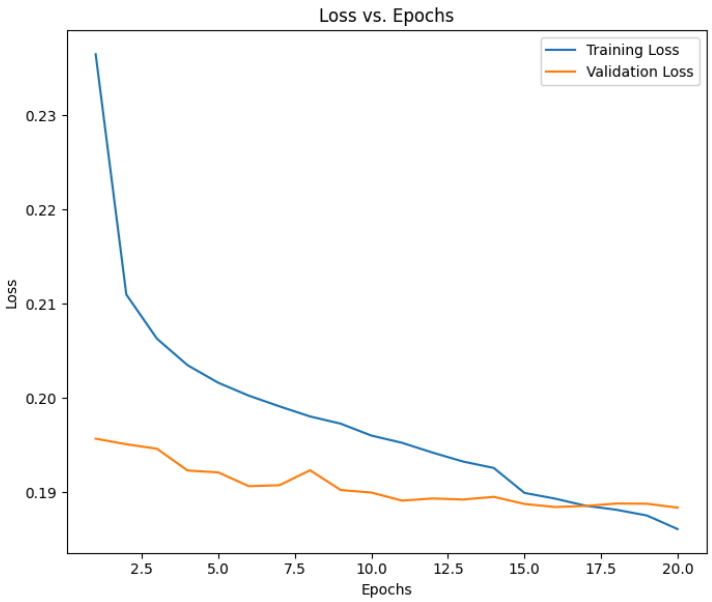
## 6. Observations and Challenges

- **Imbalanced Classes:**
  - Models tended to predict majority classes more accurately than minority classes.
  - To address this, class weights were incorporated during training, assigning higher importance to underrepresented classes.
- **Multi-Label Complexity:**
  - Models struggled with predicting multiple labels for images with comorbid conditions.
  - Binary cross-entropy loss was particularly effective in improving performance for such cases.
- **Impact of Augmentation:**
  - Data augmentation significantly improved model generalization, reducing overfitting and improving performance on the validation set.

## 7. Results Summary

- The VGG16 model achieved the best overall performance, with ~60% accuracy and relatively high AUC-ROC scores across most disease categories.
- ResNet50 demonstrated robustness in handling complex features but slightly underperformed compared to VGG16.
- MobileNetV2 offered competitive accuracy with the added advantage of computational efficiency, making it ideal for deployment.
- The Base CNN established a baseline but was limited by its comparatively shallow architecture.



AUC-ROC scores for each class:

Atelectasis: 0.5823522352372985

Cardiomegaly: 0.5199946886581494

Consolidation: 0.5626172385923295

Edema: 0.3420592072758584

Effusion: 0.5555480951121061

Emphysema: 0.643130297654984

Fibrosis: 0.5839849301532614

Hernia: 0.3429703809723501

Infiltration: 0.590407677604446

Mass: 0.6107535582909916

No Finding: 0.33090140432146653

Nodule: 0.5272780306680506

Pleural_Thickening: 0.5448513564275409

Pneumonia: 0.566049120883193

Pneumothorax: 0.6780355423068292

# Model Evaluation

In this study, we evaluated multiple deep learning architectures for chest X-ray classification, focusing on their ability to accurately predict the presence of various diseases in the images. Below are the results for each model evaluated:

- **Base CNN**: Achieved an accuracy of **35%**. This model represents a basic convolutional neural network architecture without any advanced optimizations. While it demonstrated some ability to classify chest X-rays, it was significantly limited in terms of performance.
- **ResNet50**: Achieved an accuracy of **45%**. ResNet50 is a deeper architecture that uses residual connections to allow for better gradient flow and mitigates vanishing gradient problems. While it performed better than the base CNN, it still struggled to achieve higher accuracy due to challenges like dataset imbalance and the complexity of multi-label classification.
- **MobileNetV2**: Achieved an accuracy of **44%**. MobileNetV2 is a lightweight architecture optimized for mobile devices, but it is also effective for applications where computational efficiency is important. Despite its efficiency, its performance was similar to that of ResNet50, which suggests that it could benefit from further tuning or more specialized training data.

- **VGG16 (Attempt 2)**: Achieved an accuracy of **60%**. VGG16, a deeper CNN architecture with a simple structure of stacked convolutional layers, performed the best in this study. This success may be attributed to its depth, which allowed it to better capture hierarchical patterns in the data. However, even this model had its limitations, as the accuracy was still not sufficiently high for practical deployment in a clinical setting.

## Challenges

Several challenges significantly impacted the performance of the models, including:

1. **Data Imbalance**: The dataset of chest X-rays contained an unequal distribution of diseases, which led to biased model performance. For instance, models tended to be more accurate at detecting the more prevalent diseases, while underperforming for rare diseases. This data imbalance made it harder for the models to generalize across all disease categories.
2. **Multi-Label Complexity**: Chest X-ray images often contain multiple diseases at once, making it a multi-label classification problem. This added complexity, as the models were tasked with predicting multiple labels for each image, which is more difficult than single-label classification. As a result, the models struggled to maintain high accuracy across all labels simultaneously.

### Conclusion and Future Work

This study demonstrates the potential of Convolutional Neural Networks (CNNs) for chest X-ray classification but also highlights several limitations that hinder performance. The models achieved modest accuracy, suggesting that there is room for improvement. The key issues identified were related to dataset imbalance, multi-label complexity, and the limitations of the chosen architectures.

Future Directions

To improve performance in future work, we recommend the following approaches:

1. **Addressing Data Imbalance**: Techniques such as **SMOTE (Synthetic Minority Over-sampling Technique)** or **data augmentation** can be employed to balance the dataset. SMOTE works by generating synthetic samples for underrepresented classes, thus providing the model with a more balanced view of the data.
2. **Exploring Advanced Architectures**: The study suggests exploring newer, more advanced architectures like **Vision Transformers (ViTs)**, which have shown promising results in image classification tasks. These models could provide better accuracy due to their ability to capture long-range dependencies and model spatial relationships more effectively than traditional CNNs.
3. **Enhancing Interpretability with Grad-CAM**: Interpretability of deep learning models is crucial in medical applications. **Grad-CAM (Gradient-weighted Class Activation Mapping)** is a technique that visualizes the regions of an image that a model focuses on when making a decision. This can help doctors and medical practitioners understand why a model made a particular prediction, providing more trust in the model's output.
4. **Validating Models on External Datasets**: To ensure the robustness and generalizability of the models, it is important to validate them on external datasets from different sources or institutions. This will help

evaluate whether the models can consistently perform well in real-world clinical settings and on unseen data.

By addressing these challenges and implementing the suggested improvements, future iterations of the model could provide more accurate and reliable predictions for chest X-ray classification, aiding in the diagnosis and treatment of diseases.

## References

Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "Imagenet classification with deep convolutional neural networks." Advances in neural information processing systems 25 (2012).

Tan, Chuanqi, et al. "A survey on deep transfer learning." Artificial Neural Networks and Machine Learning–ICANN 2018: 27th International Conference on Artificial Neural Networks, Rhodes, Greece, October 4-7, 2018, Proceedings, Part III 27. Springer International Publishing, 2018.

Shorten, Connor, and Taghi M. Khoshgoftaar. "A survey on image data augmentation for deep learning." Journal of big data 6.1 (2019): 1-48.

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "Model-agnostic interpretability of machine learning." arXiv preprint arXiv:1606.05386 (2016).