Interpretable Deep Learning for Automated Lung Disease Detection

Jishnu Moorthy I Khushi I Shayan Hasan I Shivakumar Vinod

Table of Contents

Introduction

Background

Lung diseases, including pneumonia, tuberculosis, and interstitial lung disease (ILD), pose a severe global health challenge, accounting for over **3 million deaths annually** (WHO).

Chest X-rays are a critical and cost-effective diagnostic tool for identifying lung diseases. Despite their widespread use, the process of interpreting X-rays relies heavily on radiologist expertise, making it prone to human error due to **fatigue, workload, and variability in training**.

Challenges in AI-based Radiology

Deep learning and artificial intelligence have shown great promise in medical imaging, but their **"black box"** nature limits adoption.

Clinicians often distrust these models due to lack of **interpretability**, leaving them uncertain about the reasoning behind AI predictions.

Other challenges include the **imbalance in medical datasets**, **poor representation of rare diseases**, and **ethical concerns** such as algorithmic bias and data privacy.

Project Goals

This project addresses the gaps in AI-based radiology by:

1. **Developing Accurate Models**
   - Creating convolutional neural network (CNN)-based systems capable of identifying common lung diseases like pneumonia, pleural effusion, and cardiomegaly with high accuracy.
2. **Enhancing Interpretability**
   - Implementing Grad-CAM, a visual explanation tool, to highlight the key regions in chest X-rays that influenced the AI model's predictions.
3. **Addressing Dataset Challenges**
   - Using innovative techniques to handle class imbalance and improve the robustness of predictions.
4. **Promoting Ethical AI Design**
   - Focusing on transparency, fairness, and accountability to build clinician trust and ensure safe implementation in clinical environments.

Phase I Overview

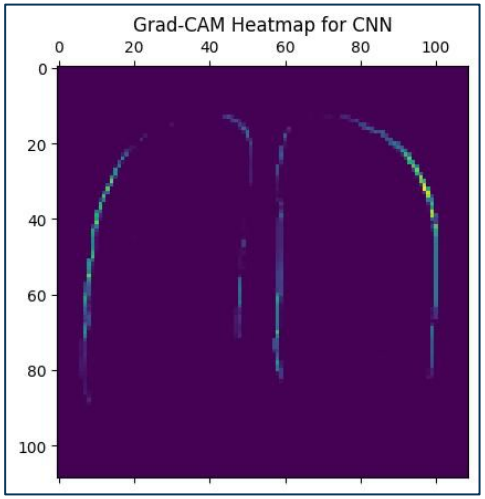
Phase 1 Achievements

Exploratory Data Analysis (EDA) and Data Insights:

- Analyzed the CXLSeg dataset (243,324 chest X-ray images).
- Identified Class imbalance and rare disease underrepresentation

Baseline Models:

- Developed Custom CNN as benchmark
- Established performance benchmarks for future improvements



Literature Review

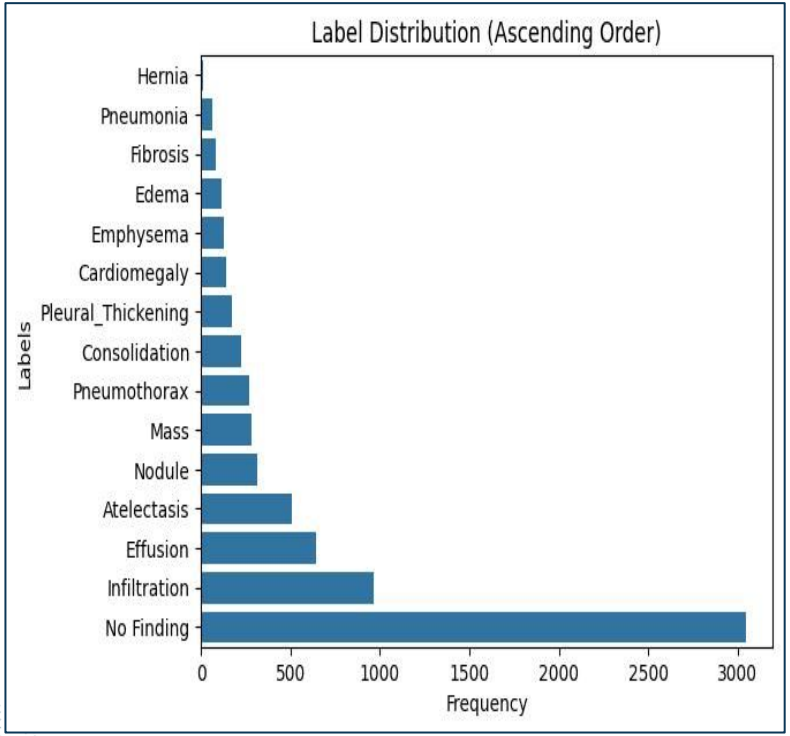
AI in Medical Imaging

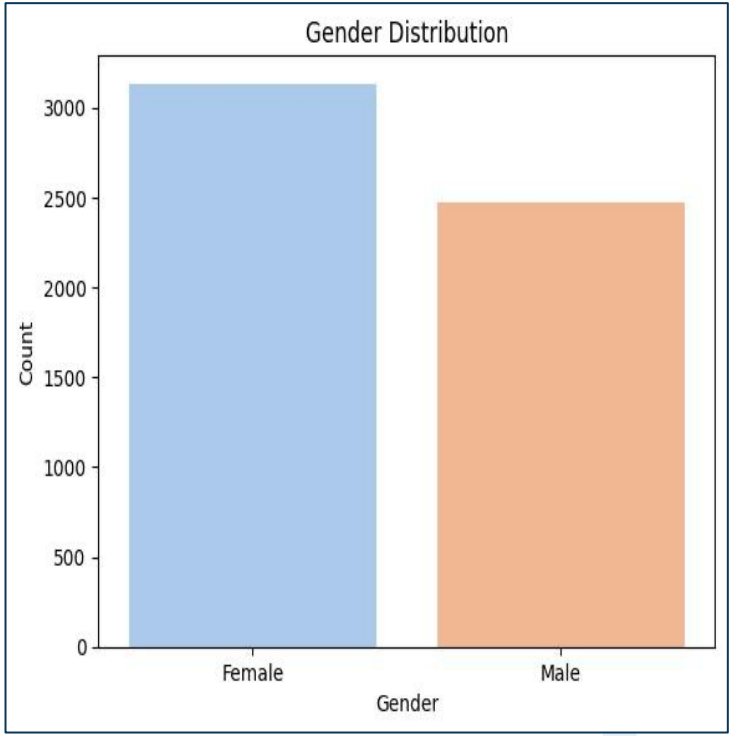
- Deep learning techniques, particularly CNNs, have significantly enhanced diagnostic accuracy in radiology. Studies show CNNs achieve radiologist-level performance, with metrics like AUC often exceeding 0.9.
- Publicly available datasets like the NIH Chest X-ray dataset have been instrumental in advancing research by providing large-scale annotated medical images.

Challenges Identified

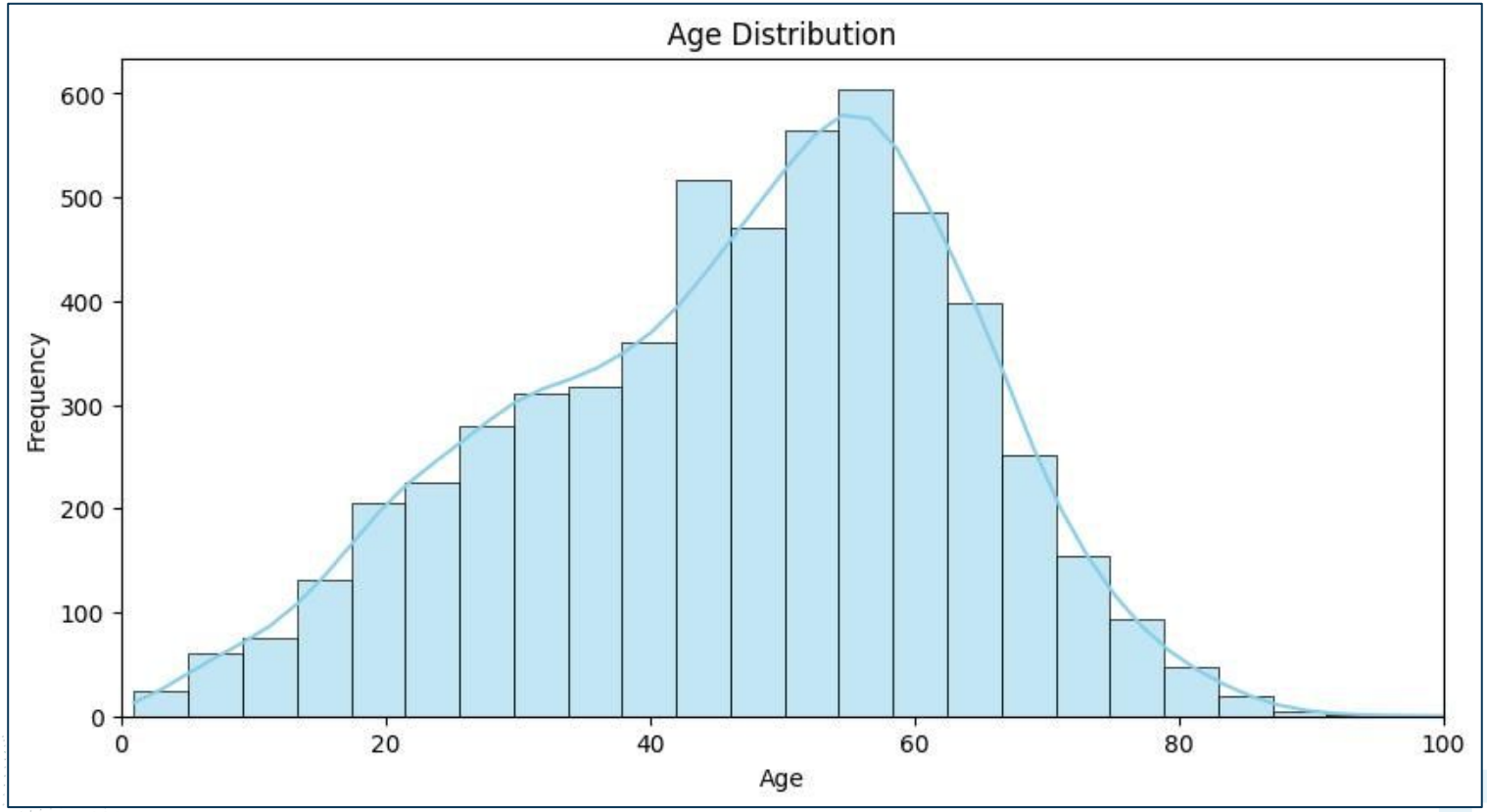
- **Class Imbalance**: Medical datasets often overrepresent common conditions while underrepresenting rare but clinically critical diseases, skewing AI predictions.
- **Interpretability**: Tools like Grad-CAM and LIME improve transparency, but their effectiveness in aligning AI reasoning with clinical understanding remains limited.
- **Ethical Concerns**: Algorithm bias and privacy issues persist, emphasizing the need for equitable and transparent AI systems.

Exploring the Dataset

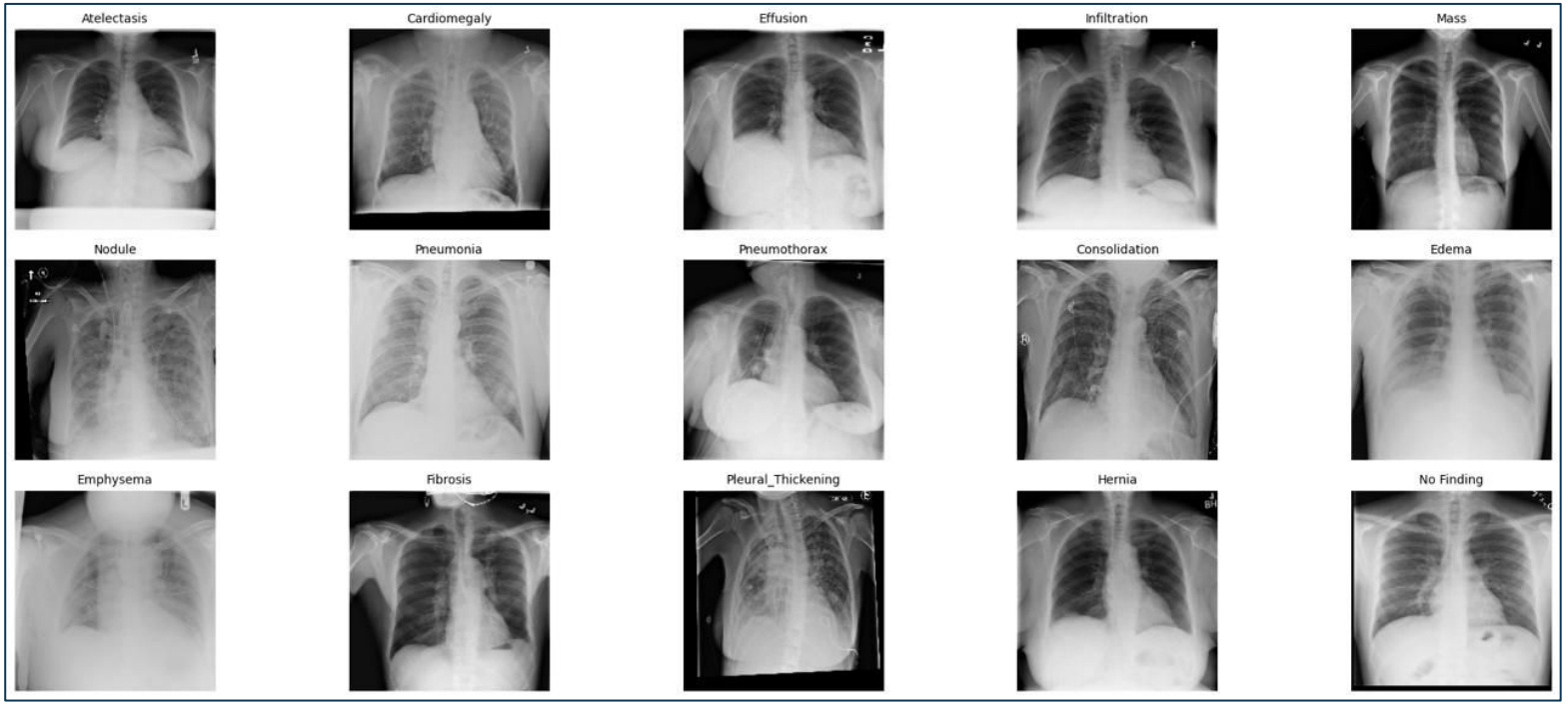




Exploring the Dataset



Exploring the Dataset



Methodology

Data Overview and Preparation

- The NIH Chest X-ray dataset was selected for its comprehensive coverage of 14 disease categories and inclusion of unsegmented chest X-rays.
- Preprocessing involved:
    - Normalizing pixel values to a standard range for model compatibility.
    - Resizing images to 224×224 pixels to match model input requirements.
    - Augmenting data with random transformations to address overfitting.
    - Including weighted classes to address class imbalance.

Model Architectures

- **Custom CNN**: A lightweight model built from scratch to serve as a baseline.
- **Pretrained ResNet50**: A sophisticated model leveraging transfer learning from ImageNet, fine-tuned for chest X-ray classification.

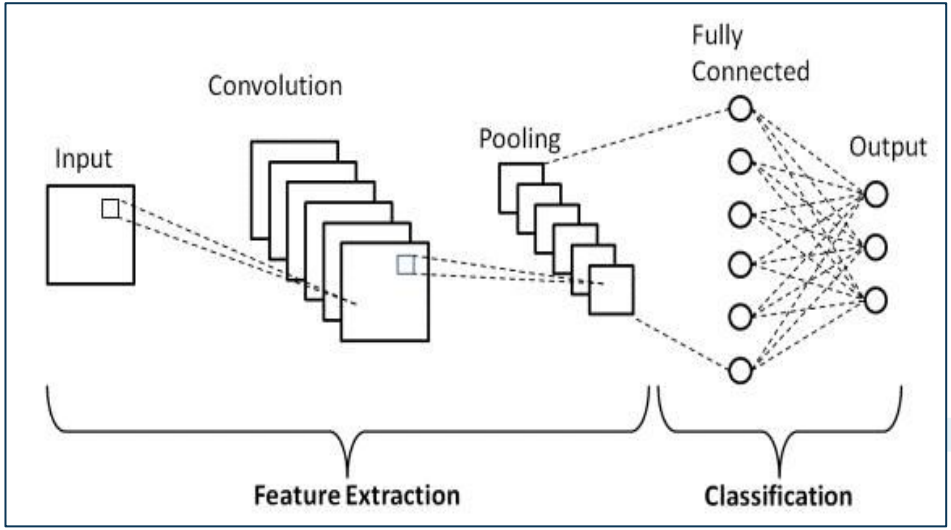
Model Architecture : Custom CNN

Convolutional Layers: Apply filters to detect patterns like edges and textures, reducing spatial dimensions while retaining important features.

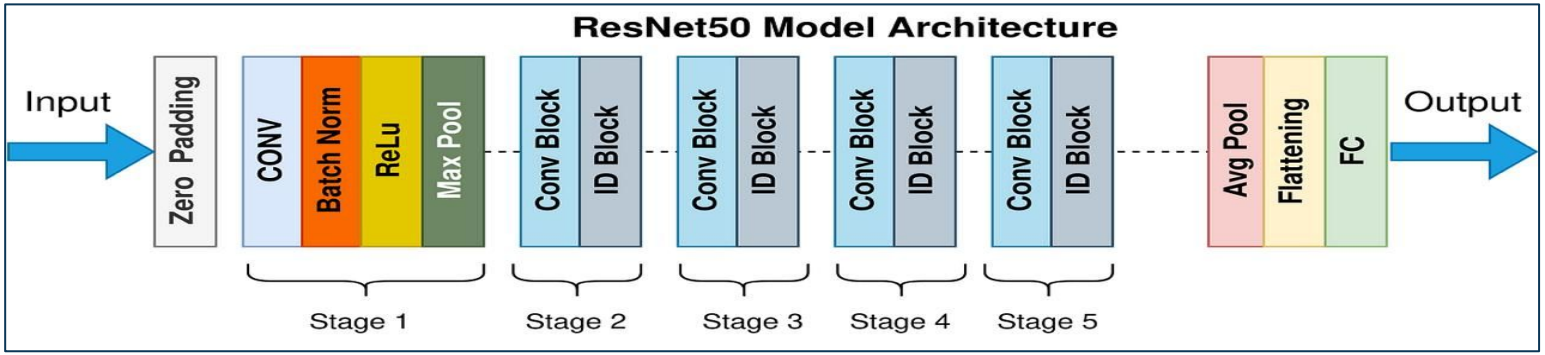
Activation Functions: Non-linear activation functions (e.g., ReLU) are used after convolutional layers to enable the network to learn complex patterns.

Pooling Layers: Downsample the output to reduce computational complexity and make the model invariant to small translations and distortions.

Fully Connected Layers: Combine features extracted by convolutions to make final predictions, using softmax or sigmoid for classification.



Model Architecture : ResNet 50



Residual Connections: Skip connections allow gradients to flow more easily, mitigating the vanishing gradient problem and helping deep networks learn efficiently.

Residual Blocks: Consist of multiple convolutional layers with skip connections that add the input to the output, improving learning.

Deep Architecture: ResNet models are very deep (e.g., ResNet-50, ResNet-101), enabling the network to learn complex features without performance degradation.

Efficient Training: Residual connections allow deeper models to be trained effectively, overcoming issues with vanishing gradients and improving accuracy.

Results

ResNet50 Performance:

- Accuracy: **7.66%**
- AUC-ROC: **0.46**
- Grad-CAM visualizations lacked clinical focus, limiting interpretability.

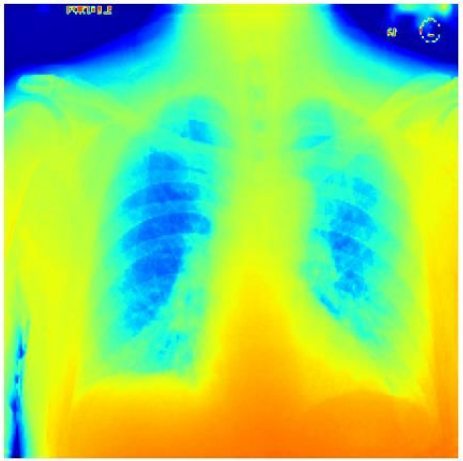
Custom CNN Performance:

- Accuracy: **11.59%**
- AUC-ROC: **0.53**
- Grad-CAM visualizations showed slightly better alignment but were still not clinically impactful.

Reasons for Suboptimal Performance

1. **Dataset Limitations**:
   - Only a **5% sample** of the NIH dataset was used, reducing diversity and disease representation.
   - Severe **class imbalance** made it challenging to train models, particularly for rare diseases.
2. **Computational Constraints**:
   - Limited resources restricted the use of advanced architectures and extended training epochs.
3. **Transfer Learning Challenges**:
   - ResNet50 struggled to generalize effectively due to the **small dataset size** and insufficient augmentation.

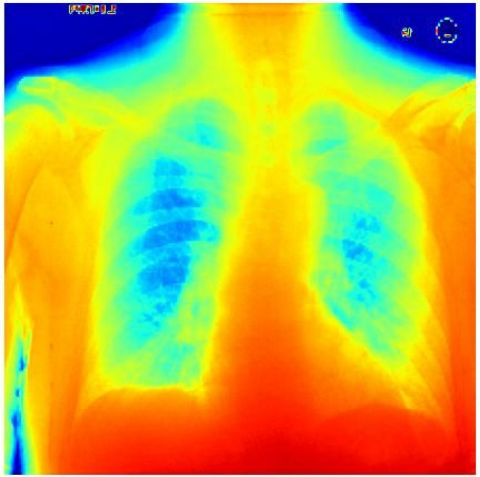

Original Image


Grad-CAM Heatmap
Predicted Class: Effusion (0.12)

CNN

ResNet


Original Image


Grad-CAM Heatmap
Predicted Class: Effusion (0.26)

Key Findings

Model Performance

- Our custom CNN outperformed the ResNet50 across all metrics, demonstrating its ability to generalize better and identify rare diseases with higher accuracy.

Interpretability

- Grad-CAM proved essential for building clinician trust by offering visual insights into the AI's decision-making process.
- It is also crucial in helping us fine-tuning future models, understanding which sections to augment, crop or pre-process the image for better accuracy.

Dataset Challenges

- Weighted loss functions partially addressed class imbalance, but certain rare conditions remain difficult to classify.

Ethical Considerations

- **Dataset Imbalance**: Overrepresentation of certain classes skews predictions, potentially leading to biased outcomes and unequal care.
- **Demographic Diversity**: Limited dataset diversity affects the model's ability to generalize across different patient populations.
- **Transparency**: The use of AI in healthcare requires trust from clinicians. Without clear explanations for AI predictions, clinicians may hesitate to adopt these tools.

Limitations

1. **Limited Representation of Rare Conditions**: Rare diseases like fibrosis and hernia were underrepresented in the dataset, reducing the model's ability to generalize and affecting prediction accuracy for less common conditions.

2. **Computational Constraints**: Limited resources restricted the use of advanced architectures and larger datasets, impacting model sophistication and overall performance.

3. **Generalizability**: Training on a specific dataset may limit the model's accuracy in diverse real-world settings. Broader validation on varied datasets is needed for better applicability and reliability.

Future Work

Dataset Enhancements

- Incorporate more diverse datasets, such as MIMIC-CXR, to improve generalizability across healthcare settings.

Model Advancements

- Investigate attention-based models and ensemble learning for improved performance and interpretability.

Ethical Improvements

- Conduct fairness audits and enhance compliance with evolving regulatory standards.

Real-World Integration

- Collaborate with clinicians to refine deployment strategies and integrate AI into existing workflows.

Conclusion

- This project successfully developed interpretable deep learning models capable of classifying chest X-ray abnormalities, addressing critical challenges in healthcare diagnostics.
- By emphasizing transparency and ethical AI design, the research bridges the gap between artificial intelligence and clinical practice, fostering trust and usability among clinicians.
- The foundation of this work allows for scalability, where larger datasets and computationally advanced resources can be integrated to achieve more robust and clinically impactful results in the future.
- These contributions pave the way for the development of safer, reliable, and widely adoptable AI tools, transforming diagnostic workflows and improving healthcare outcomes globally.

Q&A

Thank you! Any questions or feedback?