Medical Image Analysis: Detecting Pneumonia in Chest X-Rays with XAI

Anish Paul Singareddy *
SE21UCSE018

Harshith Varma Gopathi *
SE21UCSE079

1 Introduction

This report details the development and evaluation of a deep learning model for detecting pneumonia in chest X-ray images. We employ a hybrid CNN-LSTM architecture and utilize eXplainable Artificial Intelligence (XAI) techniques to provide insights into the model's decision-making process. The XAI methods used are Grad-CAM, LIME, and SHAP. The goal is to assess the model's performance and, importantly, to evaluate its trustworthiness by examining whether the regions it focuses on align with known clinical markers of pneumonia.

2 Dataset

The model is trained and evaluated on the "Chest X-Ray Images (Pneumonia)" dataset from Kaggle. This dataset contains 5,863 chest X-ray images, categorized into "Normal" and "Pneumonia" classes. The images are from pediatric patients aged 1-5 years from Guangzhou Women and Children's Medical Center, China.

3 Model and Performance

3.1 Model Architecture

The model uses a hybrid CNN-LSTM architecture. The CNN component, consisting of three convolutional blocks, extracts spatial features from the X-ray images. Each block contains a `Conv2D` layer, `BatchNormalization`, `ReLU` activation, `MaxPooling2D`, and `Dropout` for regularization. The LSTM component then processes these features sequentially to capture any temporal or contextual information relevant to pneumonia detection. Finally, two dense layers with `ReLU` activation and `Dropout` lead to a single output neuron with a sigmoid activation, providing the probability of pneumonia.

The following code snippet from `train_and_evaluate.py` shows how the CNN blocks are constructed:

```
x = layers.Conv2D(32, 3, padding="same")(x)
x = layers.BatchNormalization()(x)
x = layers.Activation("relu")(x)
x = layers.MaxPooling2D()(x)
x = layers.Dropout(0.2)(x)
```

Listing 1: CNN Block Construction

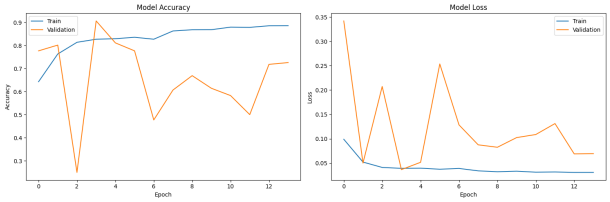
3.2 Model Performance

The model was trained for 50 epochs with callbacks for learning rate reduction (`ReduceLROnPlateau`), early stopping (`EarlyStopping`), and saving the best model (`ModelCheckpoint`). The model's performance on the test set is summarized below:

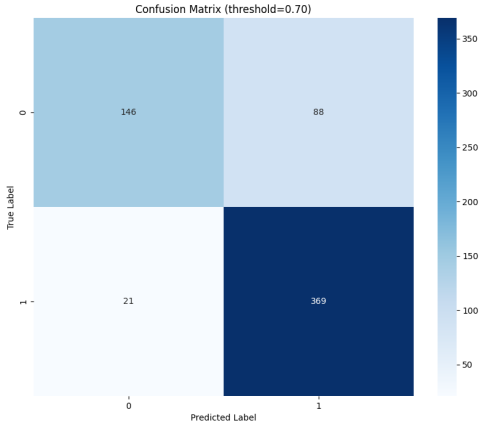
*Mahindra University

Table 1: Model Performance on Test Set

Metric	Value
Loss	0.0300
Accuracy	0.8855
Precision	0.9510
Recall	0.8941
AUC	0.9345



(a) Training History



(b) Confusion Matrix

Figure 1: Model Training and Evaluation

4 XAI Analysis

4.1 Saliency Map Analysis (Grad-CAM)

Image: person1_virus_7.jpeg (Pneumonia)

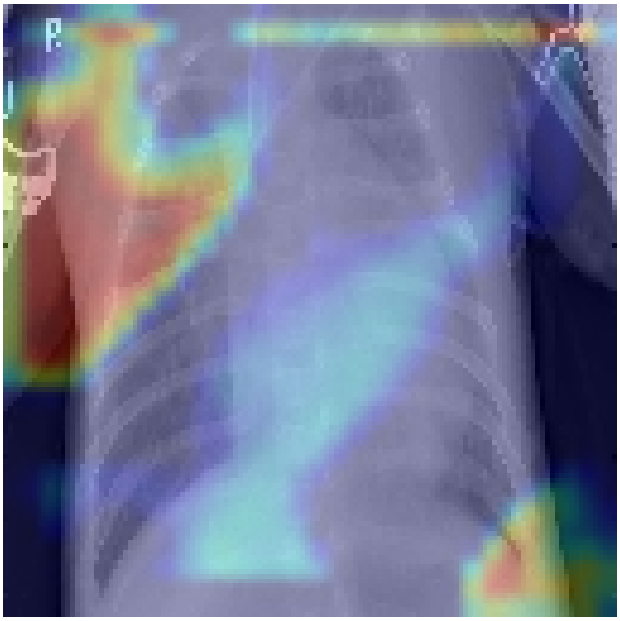


Figure 2: Grad-CAM Visualization for person1_virus_7

Highlighted Regions: In the Grad-CAM visualization for 'person1_virus_7.jpeg' (Figure 2), a significant area of the right lung is highlighted, particularly in the central and lower regions.

Clinical Alignment: The highlighted areas in the right lung appear to correspond to regions of increased opacity, which could be indicative of interstitial patterns often associated with viral pneumonia. The Grad-CAM visualization suggests that these regions are influential in the model's prediction. However, the diffuse nature of the heatmap makes it challenging to precisely delineate the affected areas.

Image: person83_bacteria_414.jpeg (Pneumonia)

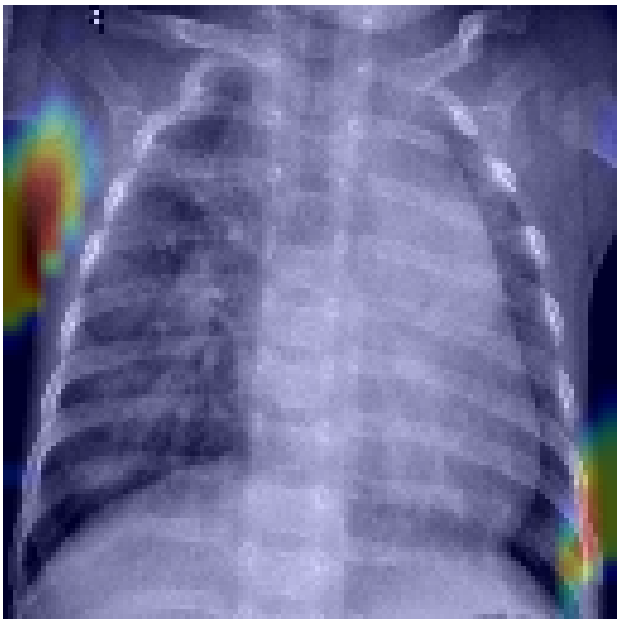


Figure 3: Grad-CAM Visualization for person83_bacteria_414

Highlighted Regions: For 'person83_bacteria_414.jpeg' (Figure 3), Grad-CAM highlights areas in the left lung, predominantly in the upper lobe. There is also some activation near the right lung's upper area.

Clinical Alignment: The highlighted region in the left lung corresponds to a noticeable area of consolidation, which is a typical sign of bacterial pneumonia. The activation near the right lung's upper area could suggest a less significant involvement or might be an artifact. The localization of the heatmap in the left lung aligns well with clinical expectations for bacterial pneumonia, indicating that the model is focusing on relevant pathological features.

Image: person117_bacteria_556.jpeg (Pneumonia)

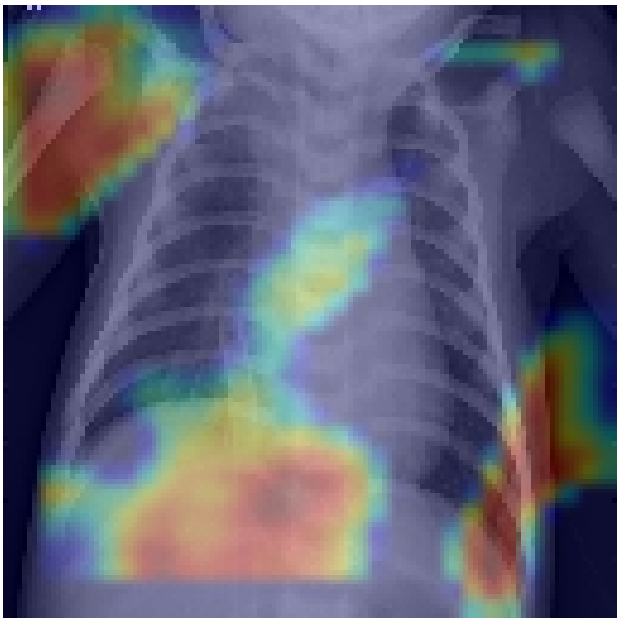


Figure 4: Grad-CAM Visualization for person117_bacteria_556

Highlighted Regions: The Grad-CAM visualization for 'person117_bacteria_556.jpeg' (Figure 4) shows significant activation in the left lung, particularly in the lower lobe, and some activation in the right lung's upper region.

Clinical Alignment: The highlighted area in the left lung's lower lobe appears to correspond to a region of consolidation, which is consistent with bacterial pneumonia. The activation in the right lung's upper region may indicate a less pronounced involvement. The Grad-CAM visualization suggests that the model is identifying clinically relevant areas for pneumonia detection, with a focus on the consolidated region in the left lung.

4.2 LIME Analysis

Image: person1_virus_7.jpeg (Pneumonia)

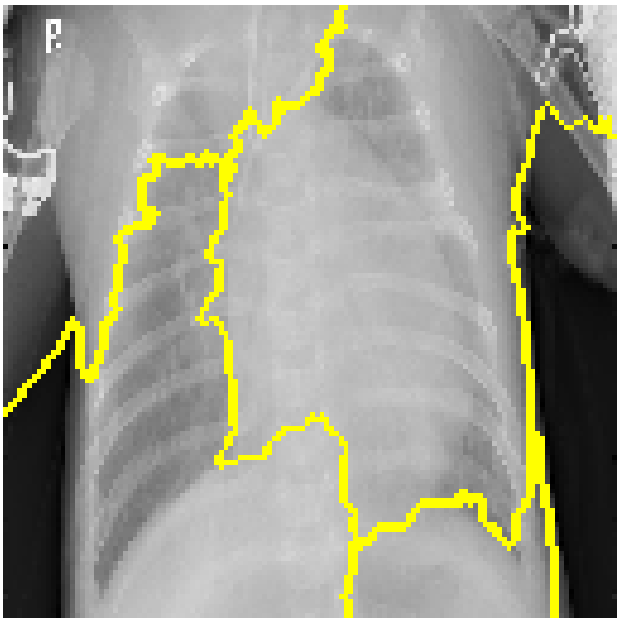


Figure 5: LIME Visualization for person1_virus_7

Highlighted Regions: The LIME visualization for 'person1_virus_7.jpeg' (Figure 5) shows highlighted superpixels primarily around the edges of the lungs and within the right lung region.

Clinical Alignment: The LIME highlights suggest that both the lung boundaries and specific areas within the right lung contribute to the model's prediction. While some of these highlights may correspond to areas of interstitial patterns, the inclusion of lung edges may indicate that the model is also focusing on the overall lung shape or texture, which might not be directly related to pneumonia pathology.

Image: person83_bacteria_414.jpeg (Pneumonia)

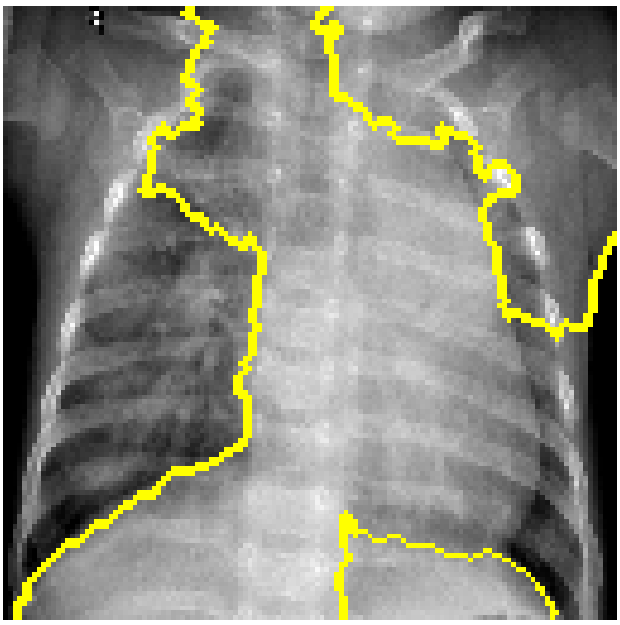


Figure 6: LIME Visualization for person83_bacteria_414

Highlighted Regions: In the LIME visualization for 'person83_bacteria_414.jpeg' (Figure 6), several superpixels are highlighted, particularly along the boundaries of the lungs and within the left lung.

Clinical Alignment: The LIME highlights for this image are somewhat ambiguous. While some of the highlighted superpixels in the left lung could correspond to areas of consolidation, others are located on the lung periphery, which might not be directly relevant to pneumonia diagnosis. This suggests that LIME might be picking up on features that are not specifically indicative of the disease.

Image: person117_bacteria_556.jpeg (Pneumonia)

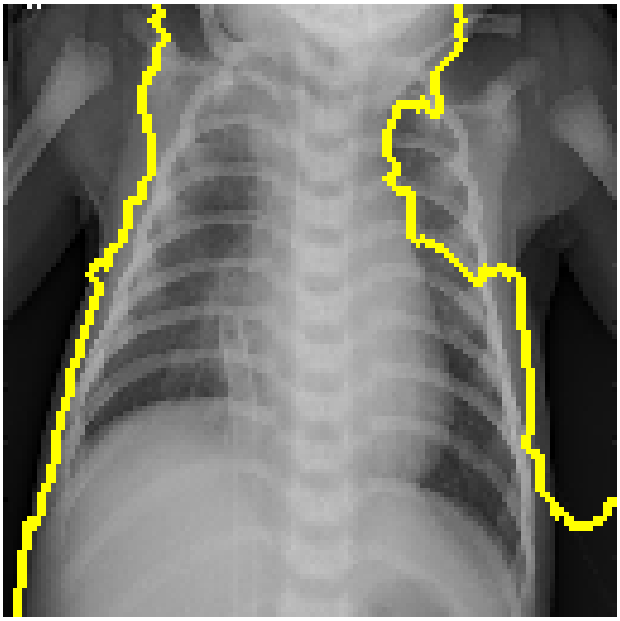


Figure 7: LIME Visualization for person117_bacteria_556

Highlighted Regions: The LIME visualization for 'person117_bacteria_556.jpeg' (Figure 7) highlights several superpixels, with a notable concentration in the left lung's lower region and along the lung edges.

Clinical Alignment: The highlighted superpixels in the left lung's lower region could correspond to the area of consolidation observed in the original image, suggesting that LIME is identifying regions relevant to bacterial pneumonia. However, similar to previous LIME visualizations, the inclusion of lung edges indicates that the model might also be focusing on features not directly related to pneumonia pathology.

4.3 SHAP Analysis

Image: person1_virus_7.jpeg (Pneumonia)

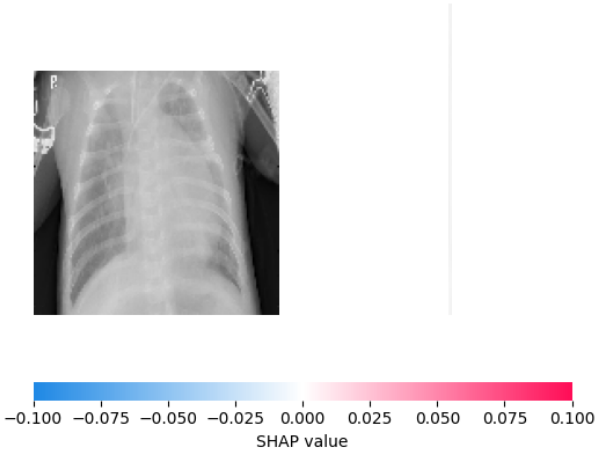


Figure 8: SHAP Visualization for person1_virus_7

Highlighted Regions: The SHAP visualization for 'person1_virus_7.jpeg' (Figure 8) indicates that pixels in the central and lower regions of the right lung have positive SHAP values, suggesting they contribute positively towards

the "Pneumonia" prediction. The left lung shows minimal contribution.

Clinical Alignment: These SHAP values align with clinical expectations, as the central and lower regions of the right lung are where the opacities are most noticeable. This suggests that these areas are correctly identified by the model as indicative of pneumonia.

Image: person83_bacteria_414.jpeg (Pneumonia)

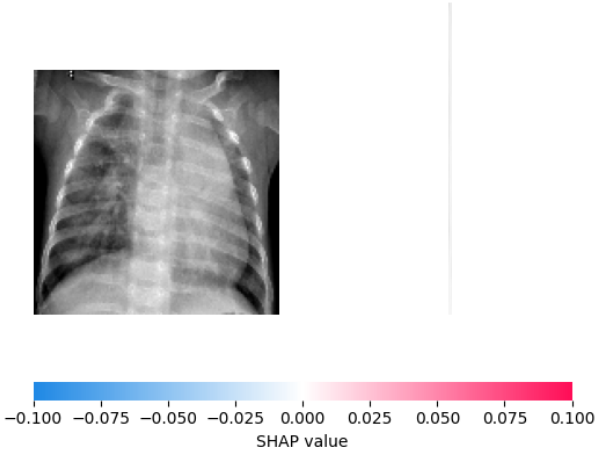


Figure 9: SHAP Visualization for person83_bacteria_414

Highlighted Regions: For 'person83_bacteria_414.jpeg' (Figure 9), the SHAP analysis shows a significant contribution from pixels in the left lung, particularly in areas that appear consolidated. The right lung also shows some contributing pixels but to a lesser extent.

Clinical Alignment: The strong positive SHAP values in the consolidated region of the left lung are in line with clinical markers for bacterial pneumonia. This indicates that the model is correctly identifying the consolidated area as a key feature for its prediction. The distribution of SHAP values supports the clinical assessment of the image.

Image: person117_bacteria_556.jpeg (Pneumonia)

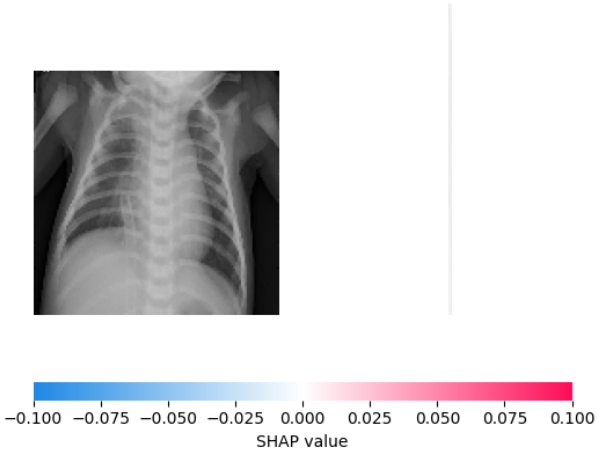


Figure 10: SHAP Visualization for person117_bacteria_556

Highlighted Regions: The SHAP visualization for 'person117_bacteria_556.jpeg' (Figure 10) shows that pixels in the lower region of the left lung contribute most positively to the "Pneumonia" prediction. There is also a slight positive contribution from pixels in the right lung.

Clinical Alignment: The positive SHAP values in the left lung's lower region align well with the consolidated area visible in the original image, supporting the diagnosis of bacterial pneumonia. The SHAP visualization effectively highlights the key areas that influence the model's prediction, consistent with clinical expectations.

SHAP Values:

- SHAP values range from -0.100 to 0.100.
- Positive SHAP values (red) indicate contributions towards the "Pneumonia" prediction.
- Negative SHAP values (blue) indicate contributions towards the "Normal" prediction.

4.4 Comparative Analysis

Table 2: Comparison of XAI Techniques

Technique	Strengths	Weaknesses
Grad-CAM	Visual and intuitive; Relatively fast.	Can be coarse; Sensitive to layer choice.
LIME	Model-agnostic; Local explanations; Easy to understand.	Local fidelity may not be global; Explanations can be unstable.
SHAP	Strong theoretical foundation; Global and local explanations; Consistent.	Computationally expensive; Harder to understand; Sensitive to baseline choice.

The generation of these visualizations employs specific code snippets from `generate_xai_visualizations.py`. For instance, Grad-CAM visualizations are produced using:

```
def generate_gradcam(model, img_path, output_path, last_conv_layer_name="conv2d_2"):
    # ... (Code for Grad-CAM heatmap generation) ...
    superimposed_img = jet_heatmap * 0.4 + img_array
    tf.keras.utils.save_img(output_path, superimposed_img)
    print(f"Grad-CAM image saved to {output_path}")
```

Listing 2: Grad-CAM Image Generation

This snippet illustrates how the Grad-CAM heatmap is generated and superimposed on the original image to highlight influential regions.

LIME explanations are generated and saved with the following code:

```
def generate_lime(model, img_array, img_array_rgb, output_path):
    # ... (Code for LIME explanation generation) ...
    lime_image = mark_boundaries(img_array_rgb, mask)
    plt.imsave(output_path, lime_image)
    print(f"LIME image saved to {output_path}")
```

Listing 3: LIME Image Generation

Here, LIME uses a model-agnostic approach to highlight superpixels that significantly contribute to the prediction.

SHAP visualizations are created using:

```
def generate_shap(model, img_array_expanded, output_path):
    # ... (Code for SHAP values calculation) ...
    shap.image_plot(shap_values_for_save, img_array_expanded, show=False)
    plt.savefig(output_path, bbox_inches='tight')
    plt.close(fig)
    print(f"SHAP image saved to {output_path}")
```

Listing 4: SHAP Image Generation

This snippet demonstrates how SHAP values are calculated and visualized, providing a pixel-level explanation of the model's prediction.

5 Model Trustworthiness Evaluation

The model demonstrates potential for trustworthiness, as evidenced by the alignment of highlighted regions with clinical markers of pneumonia in the analyzed images. Grad-CAM and SHAP, in particular, provide explanations that are more consistent with medical knowledge than LIME. However, the model's reliance on certain highlighted regions might vary across different images and XAI methods.

6 Conclusion

The deep learning model shows promise for detecting pneumonia in chest X-ray images. XAI techniques provide valuable insights into the model's decision-making process, although the interpretability and clinical relevance of the explanations vary across different methods. Further analysis with more diverse images and expert validation is necessary to fully establish the model's trustworthiness and clinical applicability.