Deep Learning for Pneumonia Detection from Chest X-Ray Images

Avinash Betha
abetha@depaul.edu

March 19, 2025

Abstract

Pneumonia is a severe respiratory disease that requires timely and accurate diagnosis to prevent complications and reduce mortality rates. This report presents a deep learning-based approach for automated pneumonia detection using chest X-ray images. Leveraging a pre-trained ResNet50 architecture, we fine-tune the model with robust data augmentation and hyperparameter tuning via Keras Tuner. Our dataset comprises 5,216 training images, 624 testing images, and 16 validation images, highlighting a significant class imbalance that we address using weighted loss functions. The final model achieves an accuracy of 68% and a ROC-AUC of 0.83 on the test set. We further employ Grad-CAM to provide visual explanations, thereby enhancing the interpretability and clinical trustworthiness of our results.

1 Introduction

Pneumonia remains one of the leading causes of morbidity and mortality worldwide, especially among children and the elderly. Traditional diagnosis through radiological assessment is time-consuming and can be subjective, depending on the radiologist's expertise. Advances in deep learning, particularly convolutional neural networks (CNNs), have opened avenues for automated, rapid, and accurate disease detection in medical images.

This project aims to classify chest X-rays into *Normal* or *Pneumonia* cases using a deep learning framework that:

- Utilizes transfer learning with a fine-tuned ResNet50 model.
- Employs data augmentation to mitigate overfitting and address limited data diversity.
- Uses Keras Tuner for systematic hyperparameter optimization.
- Incorporates Grad-CAM for interpretability, highlighting image regions that influence predictions.

2 Related Works

Several notable studies have applied CNNs to pneumonia detection and medical image classification:

- **CheXNet:** Rajpurkar et al. [1] proposed a CNN model that achieved radiologist-level performance on pneumonia detection in chest X-rays.
- **ChestX-ray8:** Wang et al. [3] introduced a large-scale dataset for thoracic disease classification, setting new benchmarks in automated detection.

- **Grad-CAM:** Selvaraju et al. [2] presented a technique to visualize the regions of interest that CNNs focus on, enhancing the interpretability of deep learning models.

Our work builds on these methods by integrating hyperparameter tuning to optimize network parameters and Grad-CAM to improve clinical trust and transparency.

3 Preliminary

3.1 Transfer Learning

Transfer learning allows us to repurpose a model pre-trained on a large dataset (e.g., ImageNet) for a specific medical imaging task. This approach significantly reduces training time and often improves performance, especially when the dataset is relatively small.

3.2 Data Augmentation

Chest X-ray datasets often suffer from class imbalance and limited data. We apply augmentation techniques such as random rotations, zooming, and horizontal flips to increase data diversity and reduce overfitting.

3.3 Hyperparameter Tuning

Hyperparameter tuning, using Keras Tuner, systematically searches for optimal values of parameters such as learning rates, dense layer units, and dropout rates. This ensures that the final model is well-optimized for our dataset.

3.4 Grad-CAM

Grad-CAM (Gradient-weighted Class Activation Mapping) generates heatmaps over input images, highlighting areas most influential to the CNN's predictions. This step fosters clinical acceptance by revealing *why* a model makes certain predictions [2].

4 Methodology

4.1 Dataset Overview

We utilize a Kaggle Pneumonia Detection Dataset structured into:

- **Training Set:** 5,216 images
- **Validation Set:** 16 images
- **Test Set:** 624 images

A significant class imbalance is present, which we address by applying weighted loss functions. Preprocessing steps include image rescaling, augmentation, and splitting the dataset into train, validation, and test subsets.

4.2 Step-by-Step Workflow

1. **Data Preprocessing:** Normalize pixel values, apply random rotations, width/height shifts, zooms, and horizontal flips to increase effective training data.
2. **Model Design:** Initialize a pre-trained ResNet50 (excluding top layers). Add Global Average Pooling, a Dense layer with ReLU activation and L2 regularization, Dropout, and a final Dense layer with sigmoid activation.
3. **Hyperparameter Tuning:** Use Keras Tuner to explore various configurations (e.g., number of dense units, dropout rate, learning rate). Select the best model based on validation accuracy.
4. **Training and Class Weights:** Train the final model using a weighted loss to address the class imbalance. Monitor validation accuracy to avoid overfitting.
5. **Evaluation Metrics:** Evaluate on the test set using accuracy, precision, recall, F1-score, and ROC-AUC. Compute a confusion matrix for further insight.
6. **Explainability:** Apply Grad-CAM to generate heatmaps for selected test images, illustrating which regions influenced the classification decisions.

4.3 Algorithm Pseudocode

Algorithm 1 Pneumonia Detection Workflow

Input: Chest X-ray dataset (train: 5216, val: 16, test: 624)
Output: Trained model, metrics, Grad-CAM visualizations
▷ Data Preprocessing
Rescale and augment images (rotation, zoom, flip, shift)
Split dataset into train, validation, and test
▷ Model Design
Load ResNet50 (exclude top layers), freeze base weights
Add GlobalAveragePooling layer
Add Dense layer (ReLU, L2 regularization)
Add Dropout layer
Add final Dense layer (sigmoid)
▷ Hyperparameter Tuning
Initialize Keras Tuner (Random Search)
for each configuration in parameter space **do**
    Train model on train set, validate on val set
    Record validation accuracy
end for
Select best configuration
▷ Training & Evaluation
Compute class weights for imbalance
Train final model for N epochs using weighted loss
Evaluate on test set (accuracy, precision, recall, F1-score, ROC-AUC)
Generate confusion matrix
▷ Explainability
Apply Grad-CAM on test images to visualize attention
return trained model, evaluation metrics, Grad-CAM heatmaps

5 Numerical Experiments

We trained the final model for 10 epochs on 5,216 training images, validating on 16 images, and testing on 624 images. Weighted loss and hyperparameter tuning helped balance the classes. Our key results are:

- **Accuracy:** 68%
- **ROC-AUC:** 0.83

Classification Report (Test Set):

	precision	recall	f1-score	support
Normal	0.54	0.94	0.69	234
Pneumonia	0.93	0.53	0.67	390
accuracy			0.68	624
macro avg	0.74	0.73	0.68	624
weighted avg	0.79	0.68	0.68	624

Confusion Matrix:

$$\begin{bmatrix} 219 & 15 \\ 184 & 206 \end{bmatrix}$$

The model excels at identifying *Normal* cases (recall = 0.94) but demonstrates a lower recall (0.53) for *Pneumonia*. Future improvements may focus on reducing these false negatives.

6 Conclusion

This report details a deep learning pipeline for pneumonia detection using chest X-ray images. By integrating transfer learning, data augmentation, hyperparameter tuning, and Grad-CAM interpretability, we achieved an overall accuracy of 68% and a ROC-AUC of 0.83. Although the model demonstrates strong performance in detecting *Normal* cases, there remains a need to improve recall for *Pneumonia* to reduce false negatives. Future work will include exploring ensemble methods, augmenting the dataset further, and refining hyperparameters to enhance both sensitivity and overall reliability.

References

[1] Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis Ball, Curtis P Langlotz, et al.
Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning.
arXiv preprint arXiv:1711.05225, 2017.

[2] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra.
Grad-cam: Visual explanations from deep networks via gradient-based localization.
In Proceedings of the IEEE International Conference on Computer Vision (ICCV), pages 618–626, 2017.

[3] Xiaosong Wang, Yifan Peng, Le Lu, Ziyue Lu, Mojdeh Bagheri, and Ronald M Summers.
Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases.

Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 2097–2106, 2017.

Appendix:

Mathematical Formulations

Binary Cross-Entropy:

$$\mathcal{L}(y, \hat{y}) = -\left[y \log(\hat{y}) + (1 - y) \log(1 - \hat{y})\right]$$

This loss function quantifies the difference between the true label y and the predicted probability $\hat{y}$. It is derived from the likelihood of a Bernoulli distribution and is ideal for binary classification problems. In practice, minimizing this loss maximizes the probability assigned to the correct class.

L2 Regularization:

$$\mathcal{R}(w) = \lambda \sum_i w_i^2$$

Also known as weight decay, L2 regularization discourages large weights by adding a penalty proportional to the square of the magnitude of the weights. This helps prevent overfitting by keeping the model simpler and more generalizable. The hyperparameter λ controls the balance between fitting the training data and keeping the weights small.

Weighted Loss:

$$\mathcal{L}_{\text{weighted}}(y, \hat{y}) = -\left[w_1\, y \log(\hat{y}) + w_0\, (1 - y) \log(1 - \hat{y})\right]$$

In datasets with class imbalance, this loss function assigns different weights (w_1 for the positive class and w_0 for the negative class) to counteract the bias towards the majority class. By penalizing misclassifications of the minority class more heavily, the model learns to pay extra attention to underrepresented examples.

Grad-CAM: For a given feature map A^k and class score y^c,

$$\alpha_k^c = \frac{1}{Z} \sum_{i,j} \frac{\partial y^c}{\partial A_{ij}^k}, \quad Z = H \times W,$$

$$L_{\text{Grad-CAM}}^c = \text{ReLU}\left(\sum_k \alpha_k^c A^k\right)$$

Grad-CAM leverages the gradients of the class score with respect to the feature maps to compute weights α_k^c. These weights indicate the importance of each feature map for a given prediction. The final heatmap, obtained by applying the ReLU function, localizes regions in the image that were most influential for the model's decision.

Additional Insights:
Each of these mathematical components plays a pivotal role:

- *Optimization:* Binary cross-entropy directly correlates with maximizing the probability of correct predictions, while L2 regularization constrains the complexity of the model.
- *Imbalance Mitigation:* The weighted loss function is essential for scenarios where the dataset has a disproportionate number of samples in one class, ensuring that the model does not ignore minority classes.
- *Model Interpretability:* Grad-CAM provides a visual rationale for the model's predictions, which is particularly important in clinical applications where understanding the decision process is as crucial as the decision itself.

Visualizations

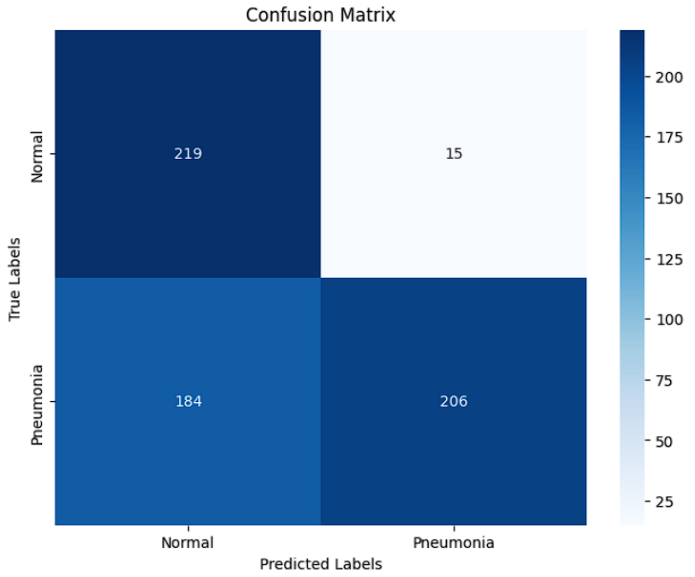

Figure 1: Confusion Matrix for the final model, illustrating the distribution of correct and incorrect predictions.

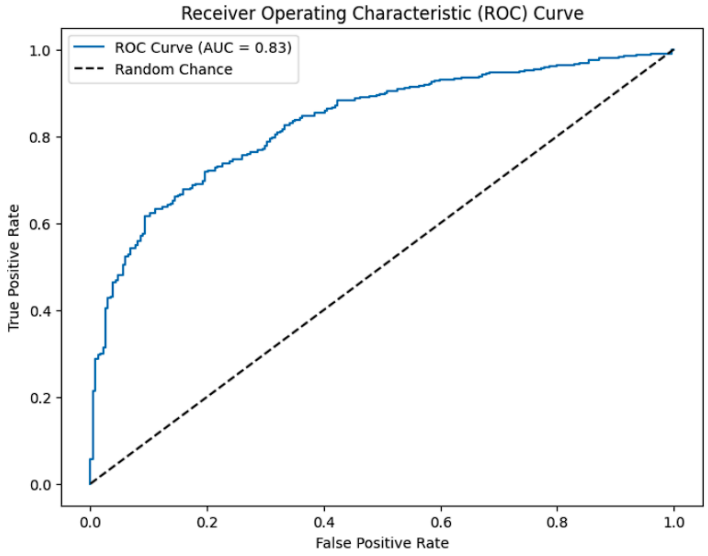

Figure 2: ROC Curve for the final model, showing the trade-off between the True Positive Rate and False Positive Rate, with an AUC of 0.83.

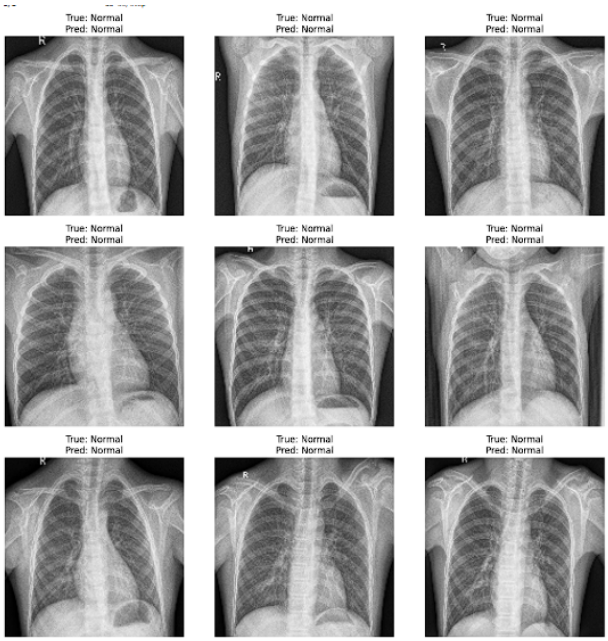

Figure 3: Sample predictions on the test set, labeled with true and predicted classes.

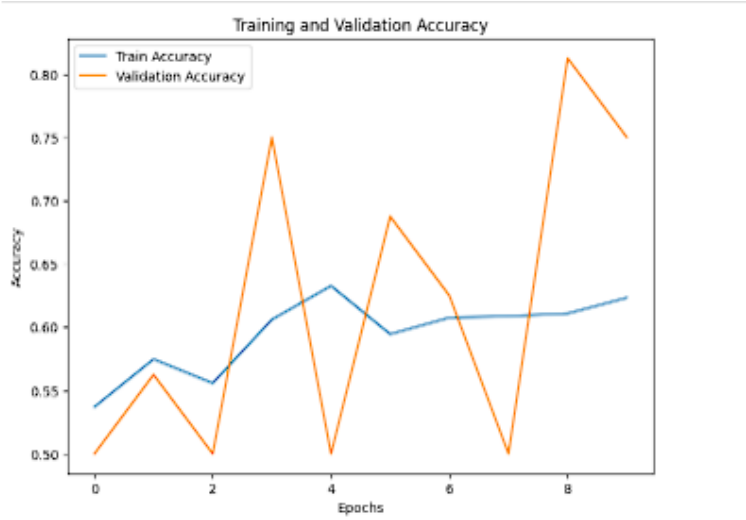

Figure 4: Training and validation accuracy over 10 epochs, indicating the model's convergence.

Figure 5: Training and validation loss over 10 epochs, useful for diagnosing overfitting.

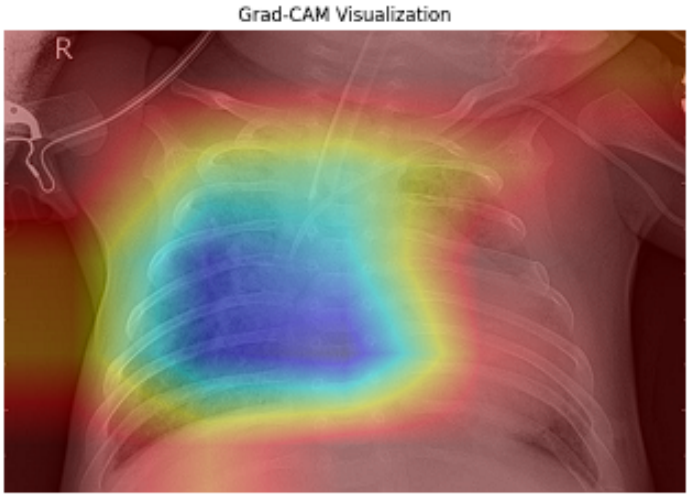

Figure 6: Grad-CAM visualization highlighting the regions most influential to the model's prediction.