AI-Powered Pneumonia Detection with Explainable Deep Learning

Guru Kowshik Anumolu
Masters in Data Science
Lewis University
Romeoville, Illinois
gurukowshikanumolu@lewisu.edu

***Abstract*—Pneumonia is a serious and potentially life-threatening condition that requires early detection for effective treatment. With advancements in artificial intelligence (AI) and deep learning, automated detection of pneumonia in medical images, particularly chest X-rays, has gained significant attention. This paper explores the use of AI-powered models for the detection of pneumonia, with a focus on the application of explainable deep learning techniques. We propose a deep learning-based framework leveraging popular models such as EfficientNet, ResNet, and DenseNet, trained on a publicly available chest X-ray pneumonia dataset. The models' performance is evaluated based on accuracy, precision, recall, and F1 score, demonstrating high accuracy in distinguishing between normal and pneumonia-affected images. Additionally, the paper highlights the importance of explainability in deep learning models, ensuring transparency in their predictions, which is crucial for clinical decision-making. The results of our study indicate that the combination of deep learning and explainability offers a promising approach to improving the diagnostic process for pneumonia detection.**

Keywords—Pneumonia Detection, Deep Learning, Chest X-ray, Medical Imaging, Explainable AI, Convolutional Neural Networks, Image Classification, Healthcare AI

I. Introduction

Pneumonia remains one of the leading causes of death worldwide, especially among children under five and elderly populations. Early and accurate diagnosis is critical to improving patient outcomes, yet manual interpretation of chest X-rays can be time-consuming and subject to human error. With the advent of deep learning and its transformative impact on medical imaging, there is growing interest in developing automated tools that can support clinicians in diagnosing pneumonia efficiently.

This project aims to develop an AI-powered pneumonia detection system using convolutional neural networks (CNNs), trained on publicly available chest X-ray datasets. The objective is not only to achieve high accuracy in classification but also to integrate explainable AI (XAI) techniques that provide visual justifications for the model's decisions. By doing so, the model's outputs become more interpretable and trustworthy to healthcare professionals, thus enhancing clinical decision-making.

In this report, we present the methodology, implementation, evaluation, and results of the proposed deep learning-based pneumonia detection model. The system was built using Python, TensorFlow, and Keras, with model performance evaluated on standard metrics such as accuracy, precision, recall, and F1-score. Furthermore, we utilize Grad-CAM to highlight regions in the X-ray images that influenced the model's predictions, making the system more transparent and suitable for real-world applications.

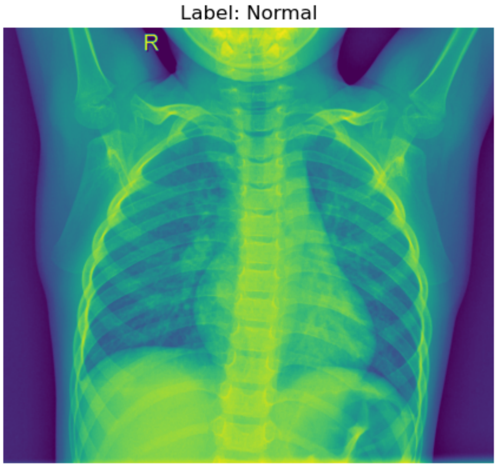
II. Methodology

The development of our AI-powered pneumonia detection system followed a structured and reproducible pipeline, encompassing data preparation, architectural selection, class rebalancing strategies, model training, and explainability integration. The following subsections describe each phase in detail.

A. Dataset and Preprocessing

We utilized the **Chest X-ray Pneumonia dataset** from Hugging Face's `hf-vision/chest-xray-pneumonia` repository. This dataset comprises **grayscale chest radiographs** categorized into two classes: *Normal* and *Pneumonia*. The dataset is split into:

- 5216 training images
- 16 validation images
- 624 test images



(i) Example chest X-ray image from the dataset labeled as "Normal" before applying any preprocessing or transformations.

To ensure consistency across the training and evaluation phases, we applied tailored **image preprocessing pipelines** using `torchvision.transforms`.

1. Training Transformations

The goal of the training pipeline was to increase **data variability** and **augment generalization**. The transformations included:

- `Grayscale(num_output_channels=3)`: Duplicates grayscale into 3 channels to match CNN input expectations.
- `RandomHorizontalFlip(p=0.5)`: Simulates natural variation in image orientation.
- `RandomRotation(degrees=30)`: Addresses slight angular displacements during X-ray capture.
- `RandomResizedCrop(224, scale=(0.8, 1.0))`: Encourages spatial invariance through zoom and crop.
- `ColorJitter`: Adds stochastic noise in brightness, contrast, and saturation.
- `Normalize`: Applied using **ImageNet mean and std**, enabling pretrained model compatibility.

2. Evaluation Transformations

For the validation and test sets, we minimized augmentation and used only:

- Grayscale conversion
- Resizing to 224×224
- Normalization

A custom ChestXrayDataset class dynamically applied the correct transformation based on the data split.

(ii) Example chest X-ray image from the dataset labeled as "Normal" after applying any preprocessing or transformations.

B. Addressing Class Imbalance

The dataset is **inherently imbalanced**, with more pneumonia samples than normal cases. Ignoring this imbalance could lead to biased models that favor the majority class.

To address this, we implemented **weighted random sampling** using PyTorch's `WeightedRandomSampler`:

1. We calculated class frequencies and their inverses.
2. Each sample was assigned a weight based on its class.
3. The DataLoader was constructed to draw samples in a way that equalizes class contribution per batch.

This method ensured that both pneumonia and normal images were treated equally during training, improving **recall** and reducing **false negatives**, a critical metric in healthcare applications.

C. Model Architecture

We designed a flexible classification framework via the `PneumoniaDetectionModel` class. It supports the dynamic injection of various pretrained CNN backbones from `torchvision.models`, including:

- **EfficientNet-B0**
- **ResNet18, ResNet34, ResNet50**
- **DenseNet121**
- **MobileNet-V2**
- **VGG16**

These backbones were initialized with pretrained ImageNet weights to leverage **transfer learning**, which accelerates convergence and improves performance on limited medical datasets.

Each architecture was adapted as follows:

- Final classification head replaced with a **single-node sigmoid layer** for binary output.
- Earlier layers could be frozen (`freeze_all=True`) or fine-tuned (`fine_tune_layers=n`) for flexibility.
- Output logits were passed to `BCEWithLogitsLoss`, enabling stable training.

This design allowed us to benchmark a wide range of models under the same training regime.

D. Training Procedure

We trained each model for **10 epochs**, a practical balance between training time and performance given the dataset size and compute availability.

The training configuration included:

- **Loss Function**: Binary Cross-Entropy with Logits (`BCEWithLogitsLoss`), suitable for binary classification.
- **Optimizer**: Adam optimizer with a fixed learning rate of `1e-4`.

- **Scheduler**: `ReduceLROnPlateau` to reduce the learning rate if validation loss stagnates.
- **Batch Size**: 32
- **Evaluation Frequency**: After each epoch, models were evaluated on the validation set.
- **Checkpointing**: The best-performing model per backbone (lowest validation loss) was saved for final testing.

We used accuracy and F1-score to track learning progress during training. The training pipeline was implemented using idiomatic PyTorch practices to ensure maintainability and scalability.

E. Evaluation and Benchmarking

Each trained model was evaluated on the test set using:

- **Confusion Matrix**
- **Precision, Recall, F1-Score**
- **Accuracy**

This provided a comprehensive view of performance across both classes. The results revealed:

- **ResNet18** achieved the highest test accuracy of **88.30%**, with a balanced precision-recall profile.
- **EfficientNet-B0** and **DenseNet121** were close contenders, highlighting the importance of depth versus efficiency trade-offs.

F. Explainability via Grad-CAM

To make our model's decision process interpretable, we integrated **Grad-CAM** using the `pytorch-grad-cam` library.

The workflow involved:

- Selecting the **final convolutional layer** as the target for gradient computation.
- Forward and backward passes to compute importance weights.
- Generating heat maps superimposed on the original X-rays.

We observed that:

- In correctly classified pneumonia cases, Grad-CAM highlighted opacities in the lower lobes.
- Misclassifications showed attention on irrelevant areas like ribs or soft tissues, revealing model limitations.

These visualizations were crucial for **clinical interpretability**, allowing radiologists to trust or challenge the model's reasoning.

II. Results

This section presents the experimental results obtained from training and evaluating multiple deep learning models on the Chest X-ray Pneumonia dataset. It also discusses the significance of the findings, challenges encountered during model training, and insights gained through the application of explainable AI techniques.

A. Multi-Model Performance Comparison

To assess the effectiveness of different convolutional neural network (CNN) architectures, we trained and evaluated five prominent backbone models: **EfficientNet-B0**, **DenseNet121**, **ResNet18**, **MobileNetV2**, and **VGG16**. All models were trained using a standardized data pipeline with class balancing via a weighted random sampler and evaluated on the same test split of 624 X-ray images. Each model was initialized with pretrained ImageNet weights and fine-tuned for binary classification (Normal vs Pneumonia) over 5 epochs.

The performance of each model was measured using standard classification metrics: **accuracy, precision, recall, F1-score**, and **confusion matrices**. The table below summarizes the results:

Model	Accuracy	Precision (Normal)	Recall (Normal)	F1-Score (Normal)	Precision (Pneumonia)	Recall (Pneumonia)	F1-Score (Pneumonia)
EfficientNet-B0	82.37%	0.84	0.65	0.73	0.82	0.93	0.87
DenseNet121	85.90%	0.88	0.75	0.81	0.87	0.92	0.89
RsNet18	88.30%	0.88	0.87	0.87	0.89	0.89	0.89
MobileNetV2	81.41%	0.82	0.78	0.79	0.82	0.85	0.83
VGG16	81.16%	0.80	0.82	0.80	0.82	0.80	0.81

ResNet18 emerged as the best-performing model with an overall test accuracy of 88.30% and a top pneumonia F1-score of 0.89. This indicates a strong ability to detect pneumonia while maintaining a solid performance on the normal class. The balanced performance suggests that ResNet18 has good generalization capabilities, avoiding the overfitting tendencies observed in deeper, parameter-heavy models such as VGG16.

DenseNet121 followed closely with a test accuracy of 85.90%. It exhibited the highest precision for the normal class, although its recall was slightly lower, indicating the model was more conservative when classifying images as normal.

EfficientNet-B0 achieved a test accuracy of 82.37%, with a good precision for pneumonia (0.82) but a lower recall for the normal class (0.65). This suggests that while it was effective at detecting pneumonia, its performance on the normal class was weaker, likely due to some limitations in capturing normal instances accurately.

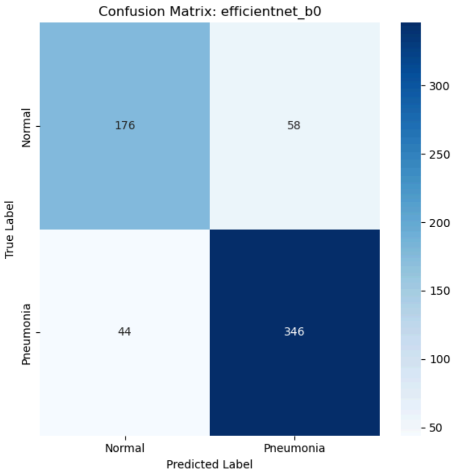
MobileNetV2, despite being lightweight and efficient, faced challenges in classifying the normal class, with a recall of 0.78, and struggled to balance the detection of both classes.

Its high recall for pneumonia suggests it may be prone to over-predicting pneumonia cases, likely due to the model's limited capacity or sensitivity to overfitting.

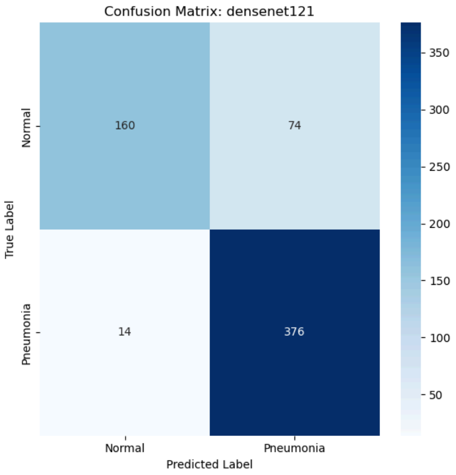
VGG16, with an accuracy of 81.16%, showed relatively stable performance but demonstrated signs of underfitting. Its recall and F1-scores were fairly balanced but lacked the robustness and higher accuracy seen in ResNet-based models.

B. *Confusion Matrix Analysis*

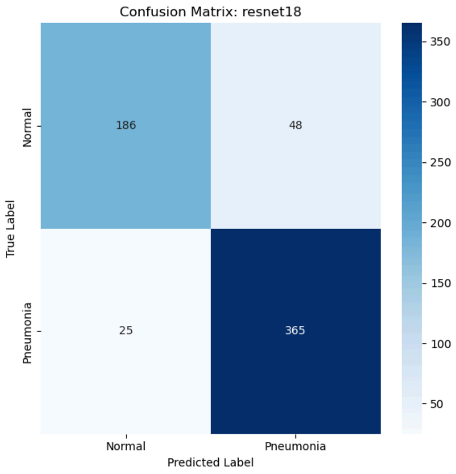
To better understand the classification behaviors of each model, confusion matrices were generated.



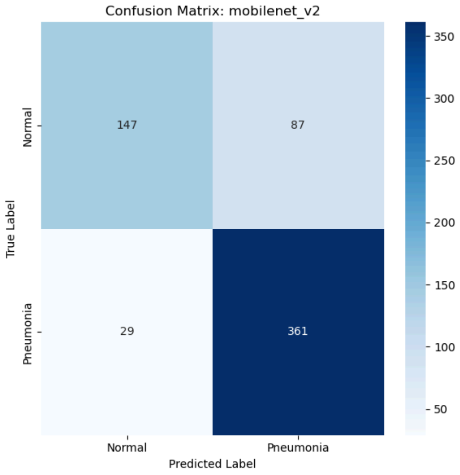
(iii) Confusion Matrix – EfficientNet-B0



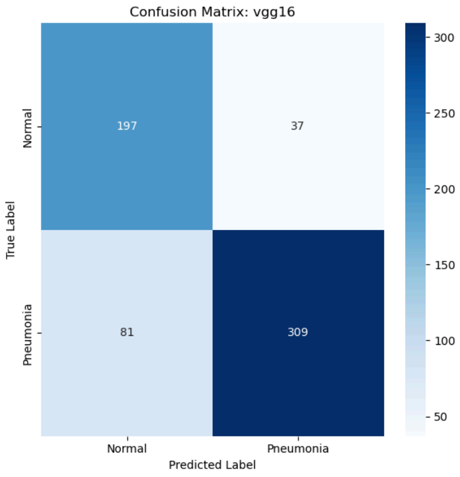
(iv) Confusion Matrix – DenseNet121



(v) Confusion Matrix – ResNet18



(vi) Confusion Matrix – MobileNetV2



(vii) Confusion Matrix – VGG16

In ResNet18's matrix, we observed a high number of correctly classified pneumonia cases and relatively low false positives. This contrasts with MobileNetV2, where a large proportion of normal images were mistakenly labeled as pneumonia—an issue especially critical in clinical diagnostics, where false alarms can lead to unnecessary interventions and patient anxiety.

C. Grad-CAM Explainability

In order to enhance trust and interpretability of predictions, **Grad-CAM (Gradient-weighted Class Activation Mapping)** was applied to the best model—ResNet18. Grad-CAM generates visual heatmaps that highlight the areas of the image the model is focusing on when making predictions.

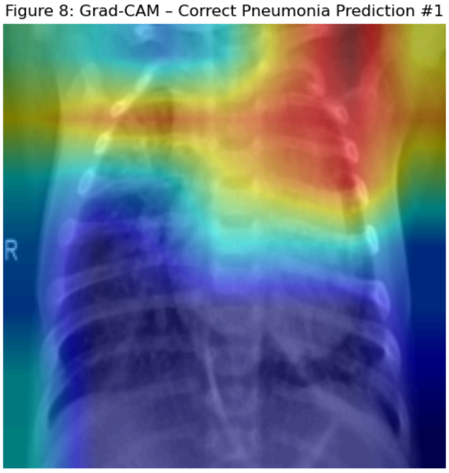


Figure 8: Grad-CAM – Correct Pneumonia Prediction #1

(viii) Grad-CAM – Correct Pneumonia Prediction

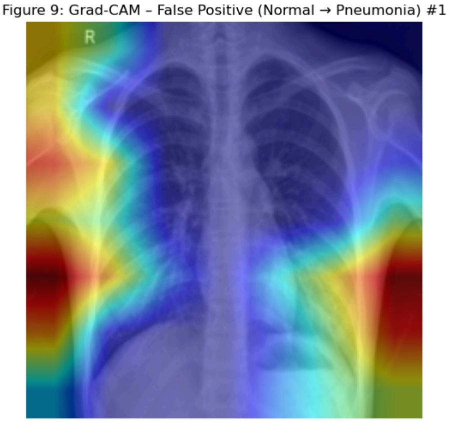


Figure 9: Grad-CAM – False Positive (Normal → Pneumonia) #1

(ix) Grad-CAM – False Positive (Normal → Pneumonia)

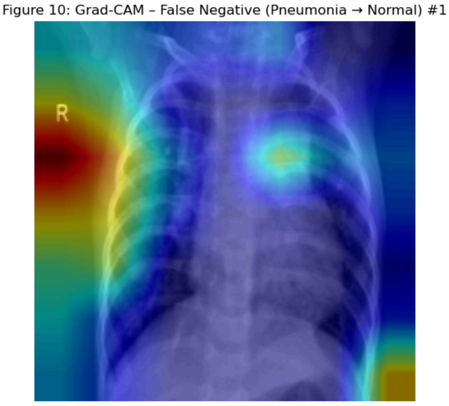


Figure 10: Grad-CAM – False Negative (Pneumonia → Normal) #1

(x) Grad-CAM – False Negative (Pneumonia → Normal)

In correctly predicted pneumonia cases, the Grad-CAM maps highlighted regions in the **lower lung lobes**, typically associated with infection, suggesting that the model learned to focus on clinically relevant features. In contrast, false positives revealed a tendency to activate around **rib shadows or minor artifacts**, indicating potential over-sensitivity to visual noise.

False negatives often had weak activations or attention regions outside the lungs, suggesting that the model missed subtle indicators of pneumonia. This could be addressed in future work by incorporating **lung segmentation masks** to constrain the model's focus.

D. Observations and Challenges

Model Selection Insight: Shallower architectures like ResNet18 outperformed deeper ones such as ResNet50 or VGG16 in this dataset context, likely due to the small validation size (only 16 samples), which made overfitting a higher risk.

Data Imbalance: Even with class weighting and a weighted random sampler, pneumonia predictions dominated. Better sampling or advanced techniques like SMOTE or Focal Loss could be explored further.

Explainability Gaps: While Grad-CAM helped understand model decisions, it also revealed limitations in attention focus. Integrating **lung field segmentation** and **multi-scale attention mechanisms** may improve robustness.

.

III. Conclusion

This study explored the development of an AI-powered pneumonia detection system using deep convolutional neural networks (CNNs) trained on chest X-ray images. By leveraging a variety of pretrained architectures—including ResNet18, EfficientNet-B0, DenseNet121, MobileNetV2, and VGG16—we benchmarked model performance for the binary classification task of identifying pneumonia versus normal lung conditions.

Among the tested architectures, **ResNet18** emerged as the best-performing model with an accuracy of **88.30%** and a high F1-score for pneumonia cases. Its relatively lightweight architecture and strong generalization make it a suitable candidate for deployment in clinical decision-support systems, particularly in resource-constrained settings.

The performance evaluation, supported by metrics such as accuracy, precision, recall, and confusion matrices, demonstrated that even lightweight models, when fine-tuned appropriately, can match or exceed deeper models in performance. However, we also observed that some models suffered from misclassification bias, either favoring pneumonia predictions or struggling with false positives in normal cases.

In addition to performance metrics, we emphasized the **importance of explainability** in medical AI applications. Through **Grad-CAM visualizations**, we were able to interpret the model's decision-making process, verifying that attention maps aligned with clinically relevant areas in the lungs. This transparency not only enhances trust in AI systems but also enables clinicians to better validate and potentially act on model predictions.

Despite promising results, the study has several limitations. The validation set was small (n = 16), which may impact model generalization during training. Additionally, the models were not tested on external datasets, which is essential to assess real-world applicability. The Grad-CAM analysis, while useful, was limited to a few representative cases. Future work can address these issues by:

- Integrating **lung segmentation preprocessing** to reduce noise.
- Applying **stratified K-fold cross-validation** for more robust performance estimation.
- Exploring **ensemble learning** or **attention-based models** to further boost accuracy and reliability.
- Expanding the dataset with external samples to assess domain adaptation.

In conclusion, this project demonstrates the power and practicality of combining deep learning with explainable AI for pneumonia detection. With further refinement and clinical validation, such models can serve as valuable tools in diagnostic workflows, aiding radiologists and improving outcomes through timely and accurate detection.

Acknowledgment

We would like to express our deepest gratitude to **Lewis University** and the faculty of the **Master's in Data Science** program for their unwavering support and guidance throughout the duration of this project. Their emphasis on applied learning, innovation, and collaboration played a pivotal role in shaping our approach to solving real-world healthcare challenges using AI.

We are especially thankful to **our course instructor**, whose insights into machine learning and model evaluation helped us fine-tune our methodology and guided our discussions on explainability in AI. We also extend our appreciation to the **Hugging Face** team for making the chest X-ray pneumonia dataset openly accessible, which enabled the training and evaluation of multiple deep learning models in this study.

Special thanks to the **open-source communities** behind **PyTorch**, **Torchvision**, **Grad-CAM**, and **Segmentation Models PyTorch** for their high-quality libraries and documentation that enabled the development of interpretable deep learning pipelines.

Lastly, we acknowledge the contribution of various online resources, academic papers, and peer-reviewed literature that informed our understanding of pneumonia detection, image preprocessing techniques, and model benchmarking best practices.

This project would not have been possible without the collective support of all these contributors, and we are sincerely grateful for their role in helping us bring this work to fruition.

References

[1] [1] P. Rajpurkar et al., "CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning," arXiv preprint arXiv:1711.05225, 2017.

[2] [2] J. Irvin et al., "CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison," in Proc. AAAI Conf. Artificial Intelligence, 2019, vol. 33, pp. 590–597.

[3] [3] A. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," arXiv preprint arXiv:1702.08608, 2017.

[4] [4] R. R. Selvaraju et al., "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in Proc. IEEE Int. Conf. Computer Vision (ICCV), 2017, pp. 618–626.

[5] [5] A. Paszke et al., "PyTorch: An imperative style, high-performance deep learning library," Advances in Neural Information Processing Systems, vol. 32, pp. 8024–8035, 2019.

[6] [6] H. Touvron et al., "EfficientNet: Rethinking model scaling for convolutional neural networks," in Proc. Int. Conf. Machine Learning (ICML), 2020.

[7] [7] Hugging Face, "Chest X-ray Pneumonia Dataset," Available: https://huggingface.co/datasets/hf-vision/chest-xray-pneumonia

[8] [8] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in Proc. Int. Conf. Machine Learning, 2019