

Weakly Supervised Pneumonia Localization from Chest X-Rays Using Deep Neural Network and Grad-CAM Explanations

Kiran Shahi ^{1,*} and Anup Bagale  ²

¹MBS Survey Software LTD., Steyning, United Kingdom

²Frontline Hospital, Kathmandu, Nepal

*Corresponding author(s). E-mail(s): ks@mbs-software.co.uk;
Contributing authors: bagaleanup1@gmail.com;

August 2025

Abstract

Chest X-ray imaging is commonly used to diagnose pneumonia, but accurately localizing the pneumonia affected regions typically requires detailed pixel-level annotations, which are costly and time consuming to obtain. To address this limitation, this study proposes a weakly supervised deep learning framework for pneumonia classification and localization using Gradient-weighted Class Activation Mapping (Grad-CAM). Instead of relying on costly pixel-level annotations, the proposed method utilizes image-level labels to generate clinically meaningful heatmaps that highlight pneumonia affected regions. Furthermore, we evaluate seven pre-trained deep learning models including a Vision Transformer under identical training conditions, using focal loss and patient-wise splits to prevent data leakage. Experimental results suggest that all models achieved high classification accuracy (96–98%), with ResNet-18 and EfficientNet-B0 showing the best overall performance and MobileNet-V3 providing an efficient lightweight alternative. Grad-CAM heatmap visualizations in this study confirm that the proposed methods focus on clinically relevant lung regions, supporting the use of explainable AI for radiological diagnostics. Overall, this work highlights the potential of weakly supervised, explainable models that enhance the transparency and clinical trust in AI-assisted pneumonia screening.

Keywords Chest X-ray, Explainable AI, Grad-CAM, Pneumonia Detection, Pneumonia Localization, Weak Supervision

1 Introduction

Pneumonia is still a leading cause of morbidity and mortality worldwide, especially among children and elderly individuals. Although chest X-ray imaging is the most common diagnostic tool [1], interpreting chest X-rays can be a challenging task due to easily missed, subtle, and ambiguous lesions. Variability in radiologists' interpretations and the frequent oversight of small abnormalities can make consistent diagnosis difficult [2]. These limitations highlight the need for reliable solutions. As artificial intelligence systems continue to mature, deep learning-based methods offer strong potential for the accurate and efficient detection of pneumonia.

Prior research has demonstrated the state-of-the-art capabilities of Convolutional Neural Networks (CNNs) and Vision Transformers (ViT) in various medical image analysis tasks, including pneumonia detection from chest X-rays[3]. However, most solutions operate in a "black-box" approach, offering limited insight into influential regions in the X-ray image that drive their decisions. Since radiologists require transparent, localized explanations to verify model outputs, this lack of interpretability restricts clinical adoption. Moreover, pixel-level annotations, such as segmentation masks or bounding boxes, are necessary for fully supervised localization techniques; however, they are costly and challenging to acquire at scale [4].

Weakly supervised learning (WSL) techniques offer a practical approach for spatial localization using only image level labels, thereby avoiding the burdens of manual pixel wise annotations. Among various WSL approaches, Gradient-weighted Class Activation Mapping (Grad-CAM) has been widely adopted for visual explanations in radiology, highlighting the most influential regions that contribute to model predictions. Therefore, interpretable heatmaps generated using Grad-CAM enhance interpretability and provide clinicians with intuitive visual cues linking predictions to underlying radiographic features.

This study presents a unified framework for pneumonia classification and Grad-CAM based weakly supervised localization of pneumonia using chest X-rays. We benchmark seven different pre-trained model, such as ResNet-18[5], ResNet50[5], DenseNet121[6], EfficientNet-B0[7], MobileNetV2[8], MobileNetV3[9], and the transformer-based ViT-B16[10], under identical training conditions. This article highlights the potential of explainable and weakly supervised AI methods to narrow the gap between automated image interpretation and practical clinical decision-making.

The main contributions of this paper are as follows:

- We evaluate a Chest X-Rays dataset [3] with strict patient level split to prevent the data leakage.
- We benchmark six pretrained CNN architectures and a Vision Transformer backbones under identical training and evaluation settings.
- We integrate Grad-CAM and token activation visualization to produce

radiologically meaningful heatmaps aligned with lung regions, offering interpretable AI insights for clinicians.

- We identify MobileNet-V3 as an optimal trade off between accuracy and computational cost, supporting real-time, edge and mobile health application.

The remainder of the paper is organised as follows. Section 2 reviews previous studies on pneumonia detection, weakly supervised learning and the use of explainability in pneumonia localisation. Section 3 describes the methods and neural architectures employed in the experiments. Section 4 presents the experimental setup, datasets, and analysis of the results. Finally, Section 5 provides the conclusion and future work.

2 Related Work

In 2018, Kermany et al. [3] introduced a large chest X-ray dataset dedicated to pneumonia detection, which opened new opportunities for researchers in medical image analysis. Early research on pneumonia detection primarily relied on supervised learning methods and focused mainly on pneumonia classification. For instance, Tilve et al. [11] benchmarked pneumonia detection using both traditional machine learning techniques, such as k-nearest neighbors (KNN), and modern convolutional neural network (CNN) approaches, demonstrating the superior performance of CNN-based supervised methods. Similarly, Erdem and Aydin [12] further proposed a novel CNN framework with separable blocks and transfer learning for efficient pneumonia detection. Similarly, Zavaleta et al. [13] demonstrated that lightweight architectures such as MobileNetV2 achieve a favorable balance between predictive accuracy and computational efficiency. Although these supervised models achieved strong classification performance, they relied heavily on large, manually labeled datasets, making them costly to train and prone to overfitting and poor generalization.

Weakly supervised learning (WSL) has emerged as a promising solution to reduce dependence on expensive, pixel-level annotations required for fully supervised models [14]. WSL utilizes incomplete or inexact supervision, such as image-level labels or free-text radiology reports, to enable large-scale model training [15, 16]. Tam et al. [15] introduced a multimodal framework combining object detection with natural language processing (NLP) for semantically grounded localization. Subsequent work extended these ideas using transformer and generative architectures. Saber et al. [17] proposed a multi-scale transformer with lung segmentation and attention mechanisms, while Keshavamurthy et al. [18] developed a GAN-based WSL model for fine-grained pneumonia localization without bounding-box labels. Other CNN-based WSL methods [19, 20, 21] demonstrated accurate localization using only image-level supervision.

While weakly supervised methods are continuously advancing in pneumonia detection, most studies still rely on complex architectures or domain-specific an-

notations, which limit reproducibility and clinical deployment. Moreover, few works systematically compare CNN and transformer backbones under identical training and evaluation settings. In addition, due to the black-box nature of AI models, many prior studies remain limited to non-interpretable approaches, creating hesitation toward clinical adoption. This study addresses these gaps by introducing a unified benchmarking framework for weakly supervised pneumonia localization using Grad-CAM across seven pretrained models, emphasizing interpretability, computational efficiency, and clinical relevance.

3 Methods

As illustrated in FIGURE 2, the proposed framework follows a standard deep learning pipeline consisting of dataset preprocessing, feature extraction using pretrained model, training and evaluation of model performance. Furthermore, we compute the class activation maps to create heatmaps that localize the pneumonia affected regions.

3.1 Dataset

We used the publicly available Chest X-ray dataset [3]. In this dataset, 1583 X-ray images are in normal class and 4273 are in pneumonia class including both train and test set. In train set 1349 images are in normal class and 3884 images are in pneumonia class. Similarly, in test set 234 images are in normal class and 390 images are in pneumonia class. However, during dataset inspection, we observed that some patients ids were on both training and test sets, which could cause data leakage. To address this issue, we merged the original splits and re-partitioned the dataset at the patient level into training (70%), validation (15%), and test (15%) sets. Each image was resized to 224 x 224 pixels to match the input requirements of ImageNet-pretrained backbones. Since the original images were grayscale, we duplicated the channel three times to create a pseudo-RGB input to match the input shape for pretrained backbones. Further, to enhance generalization, we applied data augmentation including random rotation, horizontal flipping, brightness/contrast adjustment, and Gaussian noise. Dataset splitting was performed at the patient level to prevent data leakage, with 70% of patients for training, 15% for validation, and 15% for testing.

Table 1: Dataset distribution after splitting.

Subset	NORMAL	PNEUMONIA
Train (70%)	1,114	2,951
Validation (15%)	232	653
Test (15%)	237	669

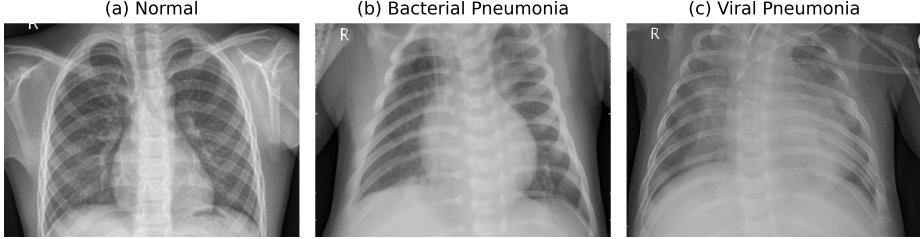


Figure 1: Sample chest X-ray images from the dataset: (a) Normal, (b) Pneumonia (Bacterial), and (c) Pneumonia (Viral).

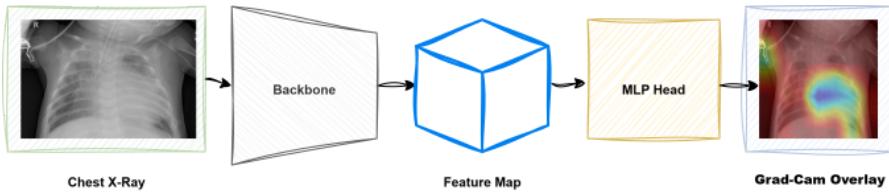


Figure 2: Model Architecture

3.2 Model Architectures

In this study, we evaluated seven different widely used ImageNet pretrained models to explore different trade-offs between accuracy, efficiency and representational power. These include residual networks, densely connected networks, parameter-efficient scaling methods, mobile optimized networks, and transformer based models.

- **ResNet18 and ResNet50 [5]:** Residual networks (ResNet) were introduced by Kaiming He et al. in 2015 to address the vanishing gradient problem by introducing skip connections that enable more stable gradients to flow across layers. ResNet-18, with its 18 layers serves as a lightweight baseline, whereas ResNet-50 with its deeper 50 layers architecture, captures more complex hierarchical features.
- **DenseNet121 [6]:** In 2016, Gao Huang et al. introduced DenseNet, which improves feature reuse and gradient propagation by connecting each layer to all subsequent layers. This design leads to compact models with fewer parameters while retaining strong representational capacity.
- **EfficientNet-B0 [7]:** EfficientNet introduces a compound scaling method that uniformly scales depth, width and resolution using a fixed coefficient. This results in highly parameter-efficient models that achieve strong accuracy with fewer computational resources.
- **MobileNetV2 [8] and MobileNetV3 [9]:** MobileNetV2 and MobileNetV3 are lightweight architectures designed for efficient deployment on mobile

and edge devices. MobileNetV2 employs inverted residual blocks with linear bottlenecks to reduce computational cost, while MobileNetV3 further integrates squeeze-and-excitation modules and neural architecture search to improve the latency–accuracy trade-off.

- **ViT-B16 [10]:** Transformer architectures have become the de facto standard for natural language processing tasks. Building on this success, Alexey Dosovitskiy et al. extended the transformer framework to vision by proposing the Vision Transformer (ViT), which replaces convolutional operations with self-attention and processes images as sequences of non-overlapping patches. In this study, we include the ViT-B/16 model to compare transformer-based architectures with traditional CNNs. ViT-B/16 splits each image into 16×16 pixel patches and processes the resulting sequence using a transformer encoder.

All of the above mentioned pretrained models were integrated with a custom classification head consisting of fully connected layers, batch normalization, ReLU activation and dropout layers ensuring fair comparison.

3.3 Training Procedure

All models were initialized with ImageNet-pretrained weights to leverage transfer learning. Training was conducted under identical protocols to ensure a fair comparison between backbones.

- **Input preprocessing:** Each image was resized to 224×224 pixels and normalized with ImageNet mean and standard deviation. Since the dataset is grayscale, the channel was duplicated three times to create a pseudo-RGB input to match the input shape of pretrained models.
- **Data augmentation:** To improve generalization and mitigate overfitting, $\pm 15^\circ$ rotations, $\pm 5\%$ affine transformation, 5% brightness contrast adjustment, CLAHE, gamma correction, Gaussian noise, motion blur, median blur and coarse dropout were applied.
- **Loss functions:** We evaluated three options Cross-Entropy Loss, Weighted Cross-Entropy Loss, and Focal Loss [22]. Focal Loss was ultimately chosen as it provided improved handling of the severe class imbalance (404 normal vs. 3692 pneumonia images).

The Focal Loss is an extension of the standard binary cross-entropy to better handle class imbalance by reducing the relative loss for well-classified class. It introduces a focusing parameter γ that down-weights easy samples, allowing the model to concentrate more on hard or misclassified cases.

Mathematically the binary focal loss function is defined as:

$$L(y, \hat{p}) = -\alpha y(1 - \hat{p})^\gamma \log(\hat{p}) - (1 - \alpha)(1 - y)\hat{p}^\gamma \log(1 - \hat{p})$$

where $y \in \{0, 1\}$ is the ground-truth label and $\hat{p} \in [0, 1]$ is the predicted probability for the positive class. The parameter γ controls how strongly easy examples are down weighted higher values increase the focus on hard samples while α balances the importance between positive and negative classes. When $\gamma = 0$, the Focal Loss simplifies to the standard weighted binary cross-entropy loss.

- **Class imbalance strategies:** In addition to Focal Loss, we applied random over-sampling of minority class during training. This ensured that each mini-batch was more balanced and prevented the model from being biased toward the pneumonia class.
- **Optimizer and hyperparameters:** All models were trained using the Adam optimizer with a learning rate of 1×10^{-4} and a weight decay of 1×10^{-4} . Training was performed with a batch size of 32 for up to 10 epochs, with early stopping applied to prevent overfitting.
- **Model checkpoint and early stopping:** For each training loop, the best model checkpoint was selected according to validation accuracy and ROC-AUC score. Early stopping was employed to mitigate overfitting when no improvement in validation performance was observed for three consecutive epochs.
- **Evaluation:** After each training loop, each model was evaluated on an independent test set of chest X-ray images. To ensure fairness and reproducibility, we assessed our methods using standard evaluation metrics, including accuracy, ROC-AUC, PR-AUC, and the best F1-score. Each evaluation metric is explained in the following section with its mathematical formulation.

3.4 Performance Evaluation Metrics

To systematically assess model performance, we employed a set of evaluation metrics designed to measure both classification accuracy and clinical relevance in class-imbalanced conditions. To formulate evaluation metrics mathematically, let us assume TP, TN, FP, and FN represents true positives, true negatives, false positives, and false negatives, respectively.

Accuracy:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

Accuracy measures the proportion of correct predictions both true pneumonia cases (TP) and true normal cases (TN) out of all predictions. However, accuracy can be misleading in imbalanced datasets because it may overestimate performance by favoring the majority class. Therefore, we evaluate our models using additional class-imbalance-aware metrics.

Precision (Positive Predictive Value):

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

Precision quantifies how many of model’s positive predictions were actually true positive.

Recall (Sensitivity / True Positive Rate):

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

Similarly, recall measures how many of the actual positive cases (pneumonia) the model correctly identifies. A high recall indicates that the model misses very few pneumonia cases. This is clinically important because false negatives failing to detect pneumonia can lead to potentially serious consequences.

Specificity (True Negative Rate):

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}$$

Specificity measures how well the model identifies normal cases. A high specificity indicates that few normal X-rays are incorrectly predicted as pneumonia.

F1-score:

$$F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

The F1-score is the harmonic mean of precision and recall. In this study, we report the best F1-score obtained across all classification thresholds.

ROC-AUC (Receiver Operating Characteristic – Area Under the Curve): The ROC-AUC represents the model’s overall ability to distinguish between pneumonia and normal cases across all classification thresholds. A higher ROC-AUC indicates stronger discriminative performance, independent of the decision threshold.

PR-AUC (Precision–Recall – Area Under the Curve): The PR-AUC summarizes the trade-off between precision and recall across all thresholds. It is particularly informative in imbalanced datasets because it emphasizes the model’s ability to correctly detect the minority class.

3.5 Pneumonia Localization

To highlight the most influential regions in predictions, we employ Gradient weighted Class Activation Mapping (Grad-CAM) [23] serving as a weakly supervised localization mechanism and enhancing clinical interpretability. For CNN based architectures, Grad-CAM is computed using the feature maps and gradients of the last convolution layer, which provide a direct spatial correspondence with the input image. Whereas, ViT operate on patch embeddings instead of convolution features, therefore we extend Grad-CAM formulation by capturing activations and gradients from the final MLP block of the last transformer encoder layer.

3.5.1 Grad-CAM for CNN Architectures

In convolutional architectures such as ResNet, DenseNet, and MobileNet, Grad-CAM is applied to the *last convolutional layer*, which retains the highest-level semantic and spatial information. Let $A \in \mathbb{R}^{C \times H \times W}$ denote the activation maps of this layer, and $\nabla Y_c \in \mathbb{R}^{C \times H \times W}$ represent the gradients of the predicted class score Y_c with respect to these activations. The channel-wise importance weights are obtained by global average pooling of the gradients:

$$\alpha_k = \frac{1}{HW} \sum_{i=1}^H \sum_{j=1}^W \nabla Y_c[k, i, j], \quad (1)$$

and the class-discriminative heatmap is computed as:

$$\text{CAM}(i, j) = \text{ReLU} \left(\sum_{k=1}^C \alpha_k A_k(i, j) \right). \quad (2)$$

The resulting activation map is upsampled to match the original image resolution and combined with the input image to generate a heatmap overlay that highlights the regions most responsible for the model’s decision.

3.5.2 Grad-CAM for Vision Transformers

Vision Transformers (ViTs) replace convolutional filters with tokenized patch embeddings, requiring a modification of the Grad-CAM formulation. We capture activations from the *final Linear layer of the last MLP block* within the last transformer encoder, which preserves spatially meaningful representations for all image patches. Let $A \in \mathbb{R}^{N \times C}$ be the activations of N patch tokens (excluding the class token) and $\nabla Y_c \in \mathbb{R}^{N \times C}$ the corresponding gradients of the predicted class. The importance weights are computed as:

$$\alpha_k = \frac{1}{N} \sum_{i=1}^N \nabla Y_c[i, k], \quad (3)$$

and the patch-level class activation map is obtained as:

$$\text{CAM}(i) = \text{ReLU} \left(\sum_{k=1}^C \alpha_k A[i, k] \right). \quad (4)$$

The one-dimensional patch map is reshaped into a 2D grid ($H_p \times W_p$) based on the number of patches and subsequently upsampled to the input image resolution. The resulting heatmap is overlaid on the original image to visualize the spatial contribution of each patch to the prediction.

This unified Grad-CAM framework provides consistent visual interpretability across both CNN and transformer-based backbones, enabling qualitative comparison of their attention on diagnostically relevant regions.

3.6 Quantitative Localization Evaluation

Since pixel-level annotations are not available in chest X-ray dataset, we adopt a lightweight quantitative metric to evaluate the anatomical consistency of Grad-CAM explanations. Specifically, we compute a Lung Attention Ratio (LAR), defined as the proportion of Grad-CAM activation energy that falls within a coarse lung region of interest (ROI).

The lung ROI is defined using a fixed thoracic anatomical prior that excludes image borders and sub-diaphragmatic regions. This ROI does not represent precise lung segmentation and is used solely for evaluation purposes. For each input image, Grad-CAM heatmaps are normalized and only the top 20% of activation values are retained to suppress background noise. LAR is then computed as the ratio of activation within the lung ROI and the total activation across the image, as shown in Eq (5).

Quantitative evaluation is performed on a fixed representative subset of test images from each class, and the same subset is used across all evaluated architectures.

$$\text{LAR} = \frac{\sum_{(x,y) \in \Omega_{\text{lung}}} A(x,y)}{\sum_{(x,y)} A(x,y)} \quad (5)$$

where $A(x,y)$ denotes the Grad-CAM activation at spatial location (x,y) , Ω_{lung} represents the coarse lung region of interest, and Ω denotes the full image domain.

4 Experiments

The details of the experiments, including the datasets, loss functions, model training, results and analysis are described as follows:

4.1 Experimental Setup

All models were trained under identical conditions to ensure fair comparison. Training and evaluation were performed using PyTorch 2.8.0+cu126 on an NVIDIA T4 GPU with 15 GB VRAM. The batch size was 32, learning rate 1×10^{-4} , and weight decay 1×10^{-4} . Each model was trained for 10 epochs with early stopping based on validation ROC-AUC. The best checkpoint per backbone was saved and later evaluated on the independent test split. Evaluation metrics include Accuracy, ROC-AUC, PR-AUC and Best F1.

4.2 Result

Overall, all evaluated architectures achieved strong discriminative performance on the pneumonia classification task, with test accuracies ranging between 96–98%. Among them, ResNet-18 and EfficientNet-B0 achieved the highest test accuracy of 98% with an F1-score of 0.987, while maintaining ROC-AUC and PR-AUC values above 0.997. Despite its smaller size, MobileNet-V3 Large

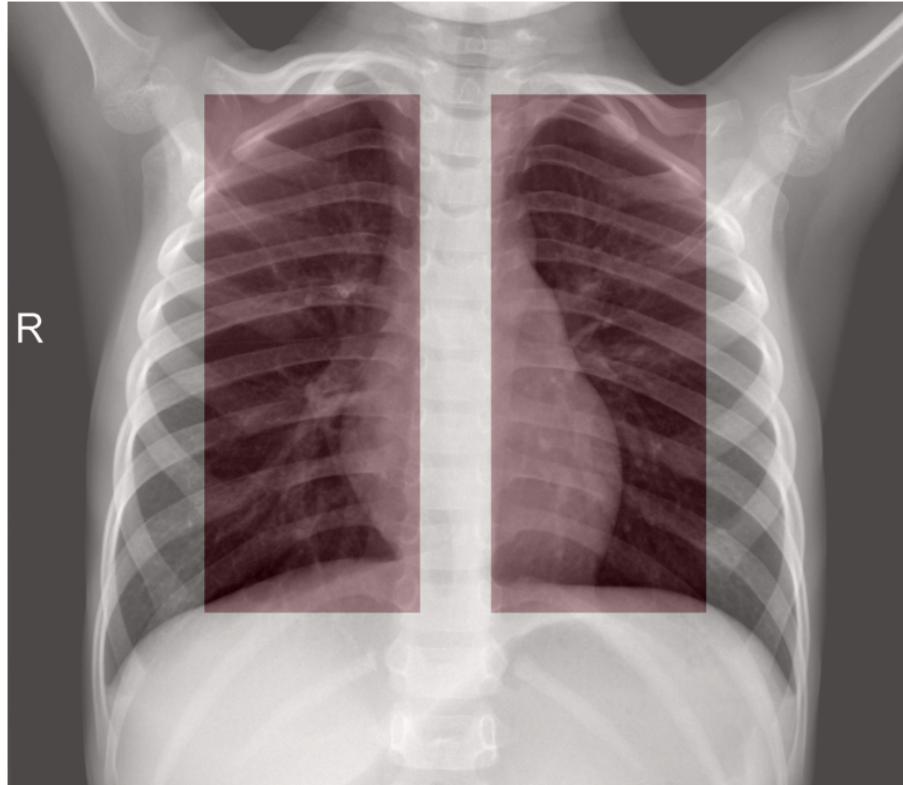


Figure 3: Illustration of the coarse lung region of interest (ROI) used for quantitative localization evaluation.

Table 2: Performance comparison among the evaluated architectures on the Chest X-Rays dataset [3].

Model	Val Acc	Test Acc	ROC-AUC	PR-AUC	F1	Params (M)
ResNet-18	97.5%	98%	0.9971	0.9990	0.987	11.5
ResNet-50	96.8%	96%	0.9952	0.9983	0.981	24.6
DenseNet-121	97.9%	97%	0.9955	0.9984	0.984	7.5
EfficientNet-B0	96.9%	98%	0.9971	0.9989	0.987	4.7
MobileNet-V2	95.6%	97%	0.9946	0.9980	0.982	2.9
MobileNet-V3	96.2%	97%	0.9971	0.9990	0.987	4.9
ViT	96.2%	97%	0.9971	0.9990	0.987	86.2

Table 3: Per class quantitative evaluation of performance (Precision, Recall, Specificity) for all evaluated architectures on the Chest X-Ray test set [3].

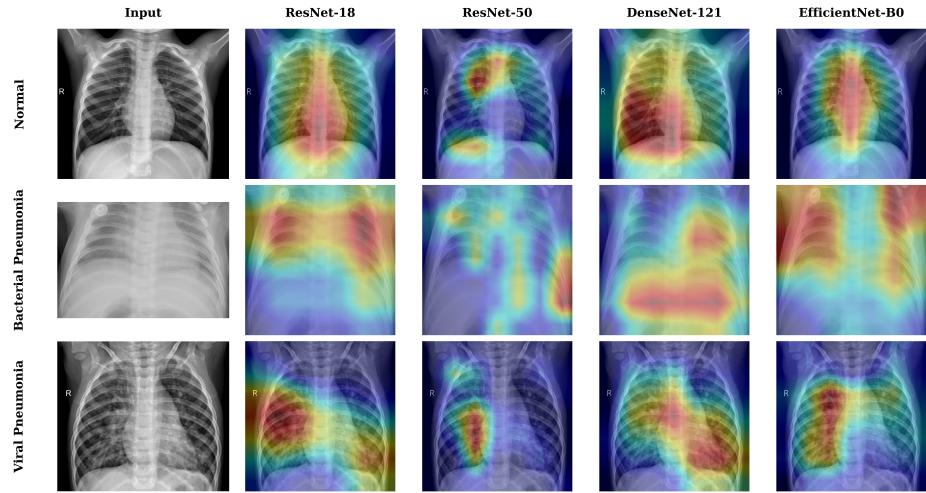
Model	Class	Precision	Recall	Specificity
ResNet-18	Normal	0.97	0.96	0.994
	Pneumonia	0.99	0.99	0.958
ResNet-50	Normal	0.93	0.94	0.993
	Pneumonia	0.98	0.97	0.966
DenseNet-121	Normal	0.97	0.93	0.928
	Pneumonia	0.99	0.96	0.991
EfficientNet-B0	Normal	0.95	0.97	0.966
	Pneumonia	0.99	0.98	0.983
MobileNet-V2	Normal	0.94	0.95	0.993
	Pneumonia	0.98	0.98	0.966
MobileNet-V3	Normal	0.93	0.97	0.970
	Pneumonia	0.99	0.97	0.975
ViT	Normal	0.95	0.95	0.953
	Pneumonia	0.98	0.98	0.981

delivered comparable accuracy of 97%, demonstrating its suitability for mobile and embedded clinical applications. In contrast, deeper backbones such as ResNet-50 and DenseNet-121 exhibited marginally lower generalization performance, suggesting mild overfitting. These results indicate that compact architectures, when combined with focal loss and patient-wise splitting, can achieve high diagnostic accuracy while remaining computationally efficient.

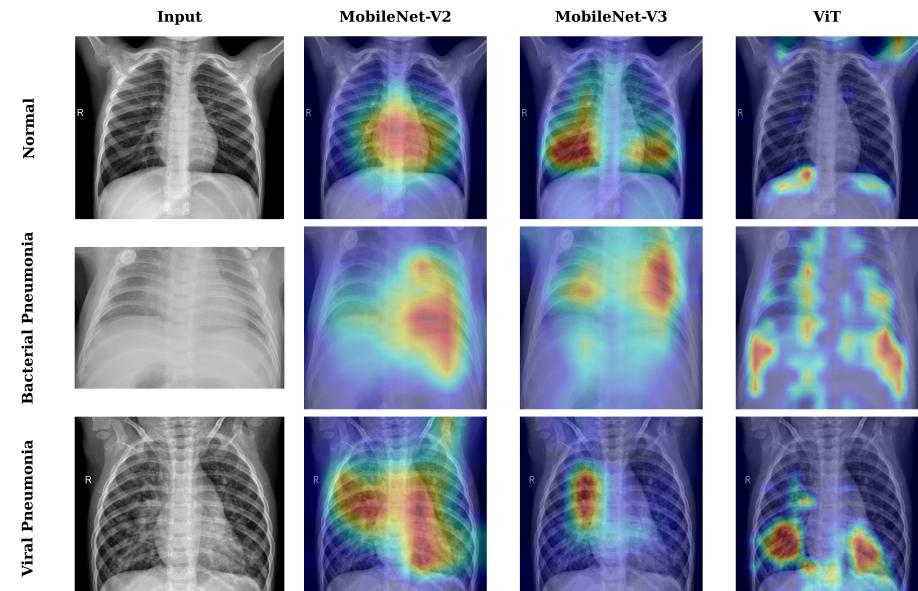
FIGURE 4 illustrates Grad-CAM overlays for representative Normal, Bacterial, and Viral Pneumonia samples across the evaluated architectures. For Normal chest X-rays, activation responses are generally weak and spatially diffuse, often extending beyond lung boundaries, indicating low diagnostic confidence in the absence of pathology. An exception is DenseNet-121, which exhibits spurious activation along the left lung region, suggesting mild sensitivity to background intensity variations or residual noise.

Table 4: Quantitative Grad-CAM localization using Lung Attention Ratio (LAR) on a representative subset of the test set.

Model	Normal	Bacterial Pneumonia	Viral Pneumonia
ResNet-18	0.547 ± 0.157	0.242 ± 0.137	0.375 ± 0.158
ResNet-50	0.593 ± 0.119	0.320 ± 0.086	0.272 ± 0.190
DenseNet-121	0.448 ± 0.241	0.417 ± 0.128	0.429 ± 0.143
EfficientNet-B0	0.589 ± 0.163	0.343 ± 0.163	0.225 ± 0.214
MobileNet-V2	0.612 ± 0.128	0.410 ± 0.201	0.409 ± 0.180
MobileNet-V3	0.553 ± 0.184	$\mathbf{0.693 \pm 0.072}$	$\mathbf{0.584 \pm 0.280}$
ViT-B/16	$\mathbf{0.172 \pm 0.038}$	0.605 ± 0.133	0.381 ± 0.096



(a) ResNet / DenseNet / EfficientNet.



(b) MobileNet and ViT.

Figure 4: Grad-CAM overlays for a normal chest X-ray and two pneumonia cases (bacterial, viral) across seven backbones. Bright regions indicate strong model attention toward pneumonia related features.

In contrast, pneumonia cases produce focused and high-intensity activations within pulmonary regions corresponding to radiographic opacities, particularly in the middle and lower lung zones. Among the CNN backbones, MobileNet-V3 produces the most compact and noise-free localization across classes, while ResNet-18 and DenseNet-121 also demonstrate well-defined activations for pneumonia cases. Although EfficientNet-B0 achieves high classification accuracy (98%), its Grad-CAM visualizations are comparatively diffuse and occasionally midline-biased. Similarly, ResNet-50 displays intermittent off-target hotspots.

Quantitative localization results are summarized in TABLE 4. MobileNet-V3 achieves stable lung-focused attention for pneumonia cases, with a Lung Attention Ratio (LAR) of 0.693 with deviation of 0.072 for Bacterial Pneumonia, indicating low variance and consistent localization behavior. In contrast, ViT-B/16 exhibits clearer discrimination between Normal and Bacterial Pneumonia samples, with a substantially lower LAR for Normal images 0.172 with deviation of 0.038 and higher LAR for Bacterial Pneumonia 0.605. However, its separation for Viral Pneumonia is less pronounced 0.381 with deviation of ± 0.096 , reflecting broader and more diffuse attention patterns associated with global self-attention.

Overall, the combined qualitative and quantitative analyses demonstrate that the proposed models predominantly attend to clinically meaningful lung regions. In particular, MobileNet-V3 achieves a favorable balance between localization stability, interpretability, and computational efficiency, reinforcing its potential for trustworthy and deployable AI-assisted pneumonia screening.

5 Conclusion and Future Works

This study benchmarked multiple CNN backbones and Vision Transformer for weakly supervised pneumonia localization using only image level supervision. All models achieved high discriminative performance, with test accuracies ranging from 96% to 98%. ResNet-18 and EfficientNet-B0 consistently outperformed deeper networks, demonstrating that compact architectures can generalize well when trained with class-balanced sampling and focal loss. Grad-CAM heatmaps confirmed that attention focused on radiologically relevant opacities, validating interpretability and trustworthiness. The results further show that lightweight models, such as MobileNet-V3, can deliver near state-of-the-art (SOTA) accuracy with low computational cost, facilitating edge device or mobile health deployments.

Although the proposed framework demonstrates the effectiveness of Grad-CAM based weakly supervised localization for pneumonia detection, several opportunities for extension remain open for future investigations, as outlined below.

- This research is currently limited to a single dataset. Further work should involve evaluation on larger and more diverse datasets, such as RSNA Pneumonia and NIH ChestX-ray14, to enhance robustness and generalization.

- Further extensions may explore multi-label thoracic disease localization, radiologist reader studies, and mobile deployment optimizations to strengthen 11 the framework’s clinical relevance and translational impact.

Overall, this study highlights that explainable and weakly supervised deep-learning methods can bridge the gap between black-box image classification and clinically interpretable decision support for pneumonia detection.

Author Contributions

K.S.: conceptualization, methodology, software; K.S. and A.B.: data curation, writing—original draft preparation; K.S. and A.B.: visualization, investigation; K.S.: supervision; K.S. and A.B.: software, validation; writing—reviewing and editing. All authors have read and agreed to the published version of the manuscript.

Funding

This research received no external funding

Institutional Review Board Statement

Not applicable

Informed Consent Statement

Not applicable

Data Availability Statement

The dataset used in this study is publicly available (see Reference [3] for the Kermany dataset). The source code supporting the findings of this study is available at: <https://github.com/kiranshahi/pneumonia-analysis>.

Acknowledgments

The author acknowledges the use of the Chest X-ray dataset by Kermany et al. [3].

Conflict of Interest Disclosure

The author declares no conflict of interest.

References

- [1] Angela Bartolf and Catherine Cosgrove. Pneumonia. *Medicine*, 44(6):373–377, 2016. Respiratory Disorders (Part 3 of 3).
- [2] Alan Ropp, Stephen Waite, Deborah Reede, and Jay Patel. Did i miss that: Subtle and commonly missed findings on chest radiographs. *Current Problems in Diagnostic Radiology*, 44(3):277–289, 2015.

- [3] Daniel S. Kermany, Michael Goldbaum, Wenjia Cai, Carolina C.S. Valentim, Huiying Liang, Sally L. Baxter, Alex McKeown, Ge Yang, Xiaokang Wu, Fangbing Yan, Justin Dong, Made K. Prasadha, Jacqueline Pei, Magdalene Y.L. Ting, Jie Zhu, Christina Li, Sierra Hewett, Jason Dong, Ian Ziyar, Alexander Shi, Runze Zhang, Lianghong Zheng, Rui Hou, William Shi, Xin Fu, Yaou Duan, Viet A.N. Huu, Cindy Wen, Edward D. Zhang, Charlotte L. Zhang, Oulan Li, Xiaobo Wang, Michael A. Singer, Xiaodong Sun, Jie Xu, Ali Tafreshi, M. Anthony Lewis, Huimin Xia, and Kang Zhang. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell*, 172(5):1122–1131.e9, 2018.
- [4] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen A.W.M. van der Laak, Bram van Ginneken, and Clara I. Sánchez. A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42:60–88, 2017.
- [5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.
- [6] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.
- [7] Mingxing Tan and Quoc Le. EfficientNet: Rethinking model scaling for convolutional neural networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 6105–6114. PMLR, 09–15 Jun 2019.
- [8] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.
- [9] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, Quoc V. Le, and Hartwig Adam. Searching for mobilenetv3. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.
- [10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

- [11] Ashitosh Tilve, Shrameet Nayak, Saurabh Vernekar, Dhanashri Turi, Pratiksha R. Shetgaonkar, and Shailendra Aswale. Pneumonia detection using deep learning approaches. In *2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE)*, pages 1–8, 2020.
- [12] Ebru Erdem and Tolga Aydin. Detection of pneumonia with a novel cnn-based approach. *Sakarya University Journal of Computer and Information Sciences*, 4(1):26–34, 2021.
- [13] Renzo Zavaleta, Eduardo Bautista, Luis Peña, Claudio Bances, and Luis Salazar. Pneumonia detection system using convolutional neural networks. *Journal of Artificial Intelligence and Autonomous Intelligence*, 02(03):264–292, Jul 2025.
- [14] Anne L Martel, Purang Abolmaesumi, Danail Stoyanov, Diana Mateus, Maria A Zuluaga, S Kevin Zhou, Daniel Racoceanu, and Leo Joskowicz. *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part I*, volume 12261. Springer Nature, 2020.
- [15] Leo K Tam, Xiaosong Wang, Evrim Turkbey, Kevin Lu, Yuhong Wen, and Daguang Xu. Weakly supervised one-stage vision and language disease detection using large scale pneumonia and pneumothorax studies. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 45–55. Springer, 2020.
- [16] Ahmad T Al-Taani and Ishraq T Al-Dagamseh. Automatic detection of pneumonia using concatenated convolutional neural network. *Preprint*, 2022.
- [17] Alireza Saber, Pouria Parhami, Alimohammad Siahkarzadeh, Mansoor Fateh, and Amirreza Fateh. Efficient and accurate pneumonia detection using a novel multi-scale transformer approach. *arXiv preprint arXiv:2408.04290*, 2024.
- [18] Krishna Nand Keshavamurthy, Carsten Eickhoff, and Krishna Juluru. Weakly supervised pneumonia localization in chest x-rays using generative adversarial networks. *Medical physics*, 48(11):7154–7171, 2021.
- [19] Kairou Guo, Jiangbo Cheng, Kaiyuan Li, Lanhui Wang, Yadong Lv, and Desen Cao. Diagnosis and detection of pneumonia using weak-label based on x-ray images: a multi-center study. *BMC Medical Imaging*, 23(1):209, 2023.
- [20] David Odaibo, Zheng Zhang, Frank Skidmore, and Murat Tanik. Detection of visual signals for pneumonia in chest radiographs using weak supervision. In *2019 SoutheastCon*, pages 1–5. IEEE, 2019.

- [21] Philip Müller, Felix Meissen, Georgios Kaassis, and Daniel Rueckert. Weakly supervised object detection in chest x-rays with differentiable roi proposal networks and soft roi pooling. *IEEE Transactions on Medical Imaging*, 2024.
- [22] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(2):318–327, 2020.
- [23] 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 *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 618–626, Los Alamitos, CA, USA, October 2017. IEEE Computer Society.