Published by :
https://www.ijert.org/
An International Peer-Reviewed Journal

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

# Impact of Fine-Tuning Depth on Model Performance and Explainability in Transfer Learning-Based Pneumonia Detection

Rohan Kamtam
Dept. of Computer Science and Engineering
Vardhaman College of Engineering
Hyderabad, India

Kamalesh Choudhary
Dept. of Computer Science and Engineering
Vardhaman College of Engineering
Hyderabad, India

G. Pavan
Dept. of Computer Science and Engineering
Vardhaman College of Engineering
Hyderabad, India

Ch. Bala Subramanyam
Dept. of Computer Science and Engineering
Vardhaman College of Engineering
Hyderabad, India

***Abstract*—Transfer learning has become standard practice for medical image classification when labeled data is scarce. A critical but often overlooked design choice is fine-tuning depth: how many layers of the pre-trained network should be allowed to update during training. This decision affects not just classification performance, but also whether the model's reasoning can be understood and trusted clinically. We investigate how fine-tuning depth impacts both performance metrics and spatial explainability in chest X-ray pneumonia detection. Three ResNet50 models were trained on the Kermany dataset under different fine-tuning regimes: completely frozen backbone (Model A), partial fine-tuning of the final convolutional block (Model B), and full end-to-end fine- tuning (Model C). Performance was evaluated through accuracy, per-class recall, ROC-AUC, and confusion matrices. Explainability was assessed using Grad-CAM heatmaps, SHAP pixel-level attributions, and a quantitative Inside/Outside Lung Activation Ratio (IOLAR) calculated against lung segmentation masks. The results reveal a non-linear relationship: going from frozen to partial fine-tuning improves both performance and explainability simultaneously. However, pushing further to full fine- tuning yields only a marginal accuracy gain (89.3%) while causing explainability to collapse—IOLAR drops from 0.91 to 0.19 on pneumonia images, with both Grad-CAM and SHAP highlighting image borders rather than lung tissue. Statistical tests confirm this degradation is substantial ($p < 0.001$, Cohen's $d \approx 1.2$), indicating classic shortcut learning behavior. For practitioners, partial fine-tuning of only the deepest convolutional block strikes the optimal balance between performance and interpretability for clinical deployment.**

***Keywords*—fine-tuning depth; transfer learning; pneumonia detection; chest X-ray; model performance; explainability; Grad-CAM; SHAP; shortcut learning; ResNet50; clinical AI; inside/outside lung activation ratio.**

## I. INTRODUCTION

Chest X-rays are among the most commonly requested imaging studies worldwide. They are cheap, fast, and available even in resource-limited settings, which is why they are so widely used. Pneumonia is one of the most frequent diagnoses made from chest radiographs, and missing it can have serious consequences. WHO estimates show it remains a leading cause of death in children under five, responsible for over 700,000 deaths globally each year. Getting the diagnosis right early matters, but this burden falls on a radiology workforce that is often severely understaffed, particularly in many low-resource regions.

Deep learning has made significant progress in this area over the past few years. CNNs trained on chest X-ray datasets have matched—and in some cases exceeded—radiologist- level accuracy for pneumonia detection in multiple published studies. The key enabler is transfer learning: instead of training from scratch on small medical datasets, researchers start with networks pretrained on massive natural image collections like ImageNet, then adapt them to the medical domain. This approach has proven remarkably effective at dealing with the data scarcity that has always been a limiting factor in medical imaging AI.

However, there is a design decision within transfer learning that does not receive sufficient systematic attention: fine-tuning depth. Specifically, how many layers of the pre-trained network should be allowed to update during training? The options span a spectrum: freeze everything except the final classification head; unfreeze only the deepest convolutional blocks; or unfreeze all layers for full end-to-end training. Each approach carries different implications for what the model learns, how well it generalizes, and critically, whether its predictions are interpretable to clinical users.

Explainability is not merely an academic concern. Clinicians and regulators are increasingly asking not just whether a model is accurate, but whether it is accurate for the right reasons. There is an important distinction between a model that identifies pneumonia by recognizing lung consolidation patterns versus one that exploits some spurious artifact in the image—such as DICOM overlays in image corners or systematic differences in how images were acquired across classes. This phenomenon, known as shortcut learning [4], is dangerous because models can appear to perform well on accuracy benchmarks while failing completely when deployed in a new setting where that particular shortcut does not exist.

This work is motivated by the hypothesis that fine-tuning depth and shortcut learning are connected in ways that have not been properly studied. A fully fine-tuned network has complete freedom to encode dataset-specific patterns at every layer. A partially fine-tuned network is more constrained—forced to rely on the general visual features already encoded in frozen early layers rather than learning new and potentially spurious patterns. How does this trade- off actually manifest when both accuracy and explainability are measured? That is the core question this paper addresses.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

We set up three ResNet50 models trained on the Kermany chest X-ray dataset [6], each with a different fine-tuning approach: Model A (frozen backbone), Model B (partial fine-tuning of conv5 only), and Model C (full end-to- end fine-tuning). Performance was measured using standard classification metrics. For explainability, we employed Grad- CAM heatmaps, SHAP attributions, and a quantitative metric we introduce called the Inside/Outside Lung Activation Ratio (IOLAR), computed by comparing model activations against lung segmentation masks. Statistical testing was used to validate comparisons. The findings are clear: fine-tuning depth does not affect explainability linearly. Full fine-tuning causes explainability to collapse almost entirely, even while delivering a slight accuracy improvement.

Our main contributions are: (1) a controlled three-way comparison of fine-tuning strategies on a standard chest X-ray benchmark; (2) the introduction of IOLAR, a metric enabling objective, scalable explainability measurement; (3) statistical validation of explainability differences with reported effect sizes; (4) convergent evidence from both Grad-CAM and SHAP confirming shortcut learning in the fully fine-tuned model; and (5) practical guidance for fine-tuning strategy selection in clinical AI deployment.

## II. Background and Related Work

### A. *Deep Learning for Chest Radiograph Analysis*

Deep learning for chest X-rays followed the same trajectory as computer vision broadly. CNNs became the dominant approach after AlexNet's success in 2012. For medical imaging, ResNet [1] was particularly impactful: residual connections enabling training of very deep networks proved well-suited for the hierarchical feature extraction required in radiological image analysis.

CheXNet [5] was a landmark contribution in this area, demonstrating that a DenseNet-121 trained on the NIH ChestX-ray14 dataset could match radiologist performance on pneumonia detection. The Kermany dataset [6] used in this study emerged in 2018 as part of a broader demonstration of transfer learning for medical images. It has since become a standard benchmark for pneumonia detection, though it originates from a single site, making it cleaner than what would typically be encountered in real deployment.

### B. *Fine-Tuning Depth and Its Implications*

The theoretical basis for transfer learning in CNNs rests on observations by Yosinski et al. [7] regarding feature transferability across tasks and domains. Their work demonstrated that lower convolutional layers tend to learn general features—edge detectors, texture filters, color blobs —broadly applicable across domains, while higher layers encode increasingly task-specific representations. Fine-tuning depth determines which layers are permitted to adapt to the target domain.

Shallow fine-tuning (freezing all backbone layers) is conservative and computationally inexpensive, relying entirely on pre-trained representations. Intermediate depth—unfreezing only the deepest convolutional blocks—allows high-level semantic features to adapt while preserving lower- level generality. Full fine-tuning unlocks the entire network, which is powerful when sufficient data and regularization are available but carries the highest risk of overfitting to dataset- specific patterns and shortcut learning. To our knowledge, how fine-tuning depth interacts with explainability alignment in medical imaging has not previously been studied systematically.

### C. *Explainability Methods: Grad-CAM and SHAP*

Grad-CAM [2] produces spatial heatmaps by weighting the feature maps of a chosen convolutional layer by the gradient of the class score with respect to those maps, then applying ReLU and upsampling. It is computationally lightweight and produces outputs that clinicians find relatively intuitive to interpret. Its main limitation is coarse spatial resolution.

SHAP [3] takes a different approach rooted in cooperative game theory. Each pixel's contribution to the final prediction is estimated as its Shapley value. SHAP DeepExplainer adapts this framework to neural networks via modified backpropagation. The resulting attributions are finer-grained than Grad-CAM and satisfy desirable theoretical properties. Using both methods in parallel provides stronger evidence than either alone.

### D. *Shortcut Learning in Medical Imaging*

Geirhos et al. [4] formalized the study of shortcut learning in deep neural networks. In medical imaging, notable examples include work by Zech et al. [9], who demonstrated that chest X-ray classifiers suffered substantial performance degradation when tested across hospital systems—having learned institution-specific acquisition characteristics rather than actual pathology. Oakden-Rayner et al. [10] showed that models could exploit hidden stratification to achieve high overall accuracy while performing poorly on clinically important subgroups. Grad-CAM provides a practical means of detecting whether a model is attending to clinically relevant regions of the image.

## III. Dataset

### A. *Source and Composition*

We used the Chest X-Ray Pneumonia Dataset originally published by Kermany et al. [6] and available through Kaggle. The dataset consists of pediatric chest radiographs collected at the Guangzhou Women and Children's Medical Center, labeled by expert physicians into two classes: NORMAL and PNEUMONIA. The pneumonia class is heterogeneous, encompassing bacterial cases (typically presenting as lobar consolidation) and viral cases (more commonly presenting as bilateral interstitial infiltrates). The original train/test split was used without modification, deliberately preserving the natural class imbalance.

*TABLE I. Dataset Split and Class Distribution*

| Split | Normal | Pneumonia |
|---|---|---|
| Training | 1,349 | 3,884 |
| Testing | 234 | 390 |
| Total | 1,583 | 4,274 |
| P:N Ratio | — | 2.88 : 1 |

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

### B. *Preprocessing*

All images were resized to 224×224 pixels and normalized using ImageNet channel statistics. For training, a modest augmentation pipeline was applied: random horizontal flips, rotations up to ±10 degrees, and small brightness/contrast jitter. Augmentation was deliberately kept conservative—chest X-rays have specific anatomical orientations that should not be violated, and excessive augmentation at this image size can obscure the subtle textural features that distinguish mild consolidation from normal lung. No augmentation was applied at test time.

## IV. METHODS

### A. *Model Architecture*

ResNet50 [1] pre-trained on ImageNet was used as the base architecture for all three experimental conditions. ResNet50 is well-studied, has a moderate parameter count (~25.6 million), and its performance on medical imaging benchmarks is extensively documented. The original 1000-class head was replaced with a two-unit softmax head preceded by a dropout layer (p = 0.5). Weights were initialized from the standard PyTorch ImageNet checkpoint.

### B. *Fine-Tuning Depth Conditions*

*Model A — Frozen Backbone:*

All ResNet50 backbone parameters were frozen. Only the new classification head was trained. This serves as a conservative baseline making no attempt to adapt pre-trained features to the medical imaging domain. Training ran for 30 epochs using Adam with lr = 1e–3 and weight decay = 1e–4.

*Model B — Partial Fine-Tuning:*

Only the final convolutional block (conv5) was unfrozen while keeping earlier layers frozen. The rationale is that conv5 encodes the highest-level semantic features most in need of domain adaptation. The classification head was trained at lr = 1e–3; the unfrozen conv5 block at lr = 1e –5 to avoid catastrophic forgetting. Training ran for up to 40 epochs with early stopping (patience = 7) based on validation loss.

*Model C — Full Fine-Tuning:*

All layers were unfrozen and trained simultaneously with a uniform lr = 1e–5. Training ran for up to 50 epochs with early stopping (patience = 10). In all three conditions, a cosine annealing scheduler was applied and the best validation-loss checkpoint was used for evaluation.

### C. *Explainability Pipeline*

*Grad-CAM:*

Forward and backward hooks were registered on the output of the final residual block for each model. For each test image, the gradient of the predicted class score with respect to feature maps at this layer was computed, globally average-pooled to obtain channel-wise weights, and combined via weighted sum followed by ReLU. The resulting map was bilinearly upsampled to 224×224 and normalized to [0, 1]. Heatmaps were generated for every image in the test set.

*SHAP:*

SHAP DeepExplainer was used with a background reference set of 100 randomly drawn training images. For each test image, pixel-level SHAP values were computed for the predicted class. Absolute values were summed across color channels to produce a single-channel attribution map, normalized to [0, 1]. SHAP analysis was run on 100 randomly sampled images per class per model (600 maps total).

*Inside/Outside Lung Activation Ratio (IOLAR):*

A binary lung mask was generated for each test image using a pre-trained U-Net segmentation model. IOLAR was then computed as:

$$IOLAR = \Sigma(H \times M) / \Sigma(H \times (1 - M))$$

where H is the normalized Grad-CAM heatmap and M is the binary lung mask. Higher IOLAR values indicate better spatial alignment with clinically relevant anatomy. This was computed for every test image and per-image distributions were used for statistical testing.

### D. *Statistical Testing*

Welch's two-sample t-test was used for all pairwise IOLAR comparisons, as equal variances between model conditions could not be assumed. Tests were conducted separately for pneumonia and normal image subsets. We report p-values, test statistics, degrees of freedom, and Cohen's d as an effect size measure. The significance threshold was $\alpha = 0.05$, two-tailed.

## V. RESULTS

### A. *Classification Accuracy*

The headline accuracy numbers reflect a clear progression with fine-tuning depth: Model C (full fine-tuning) reaches 89.3%, Model B (partial) achieves 86.4%, and Model A (frozen) achieves 80.0% (see Fig. 1). This monotonic increase in accuracy with fine-tuning depth is expected and consistent with prior transfer learning literature. However, accuracy alone tells an incomplete story, as the remainder of this section demonstrates.

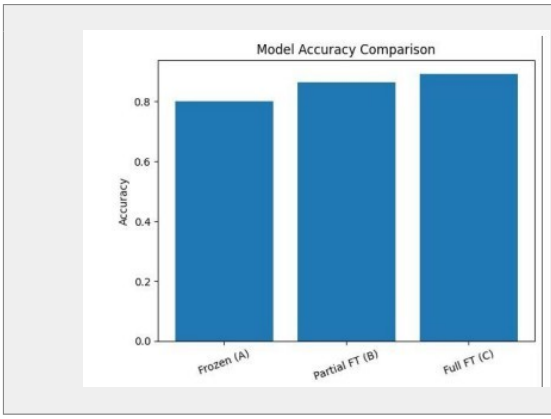


Fig. 1. Model accuracy as a function of fine-tuning depth: frozen (A), partial (B), and full (C).

*TABLE II. Classification Metrics*

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

| Model | Accuracy | N. Recall | P. Recall | AUC |
|---|---|---|---|---|
| A – Frozen | 80.0% | 53% | 96% | — |
| B – Partial | 86.4% | 67% | 98% | 0.960 |
| C – Full | 89.3% | 77% | 96% | 0.961 |

*N. Recall = Normal Recall, P. Recall = Pneumonia Recall*

Model A's 53% normal recall is particularly concerning. Nearly half of healthy patients in the test set were flagged as having pneumonia—a false positive rate that would generate a substantial volume of unnecessary downstream workup in a real clinical setting. Model B brings this to 67% while simultaneously improving pneumonia recall to 98%, which is arguably a more favorable profile for a screening tool. Fig. 2 shows the ROC-AUC comparison between Models B and C, which are nearly identical, underlining that the accuracy gap is not driven by fundamentally different discriminative power.

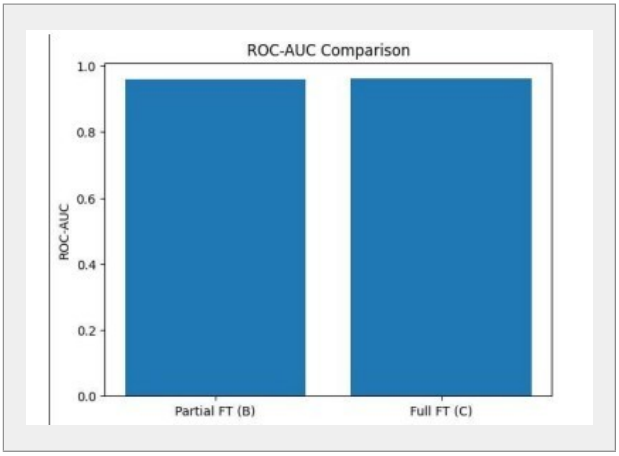


Fig. 2. ROC-AUC comparison between Model B (Partial FT) and Model C (Full FT).

### B. *Confusion Matrices*

*TABLE III. Confusion Matrices for All Three Models*

| Model | TN | FP | FN | TP |
|---|---|---|---|---|
| A – Frozen | 124 | 110 | 14 | 376 |
| B – Partial | 156 | 78 | 7 | 383 |
| C – Full | 181 | 53 | 14 | 376 |

Model B produces the fewest false negatives (7 vs. 14 for both A and C). In a pneumonia screening context, false negatives—missed cases of actual pneumonia—are arguably more dangerous than false positives. Model B's overall profile compares favorably with Model C even before explainability enters the picture.

### C. *Grad-CAM Spatial Attention*

The heatmap patterns across the three models were visually distinct in ways that were consistent across many images, not merely cherry-picked examples. Figures 3, 4, and 5 show representative Grad-CAM overlays for each model on both pneumonia and normal cases.

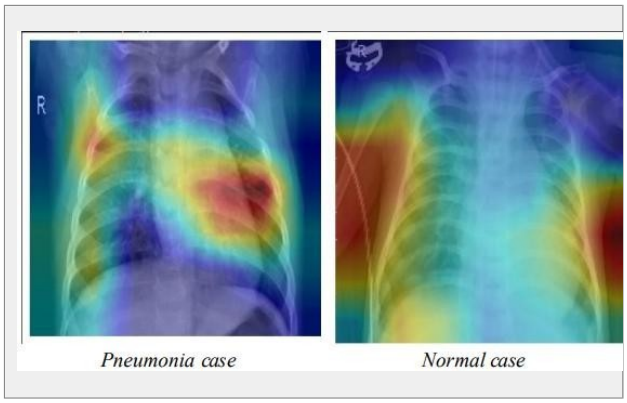


Fig. 3. Grad-CAM heatmaps for Model A (Frozen). Diffuse activation with notable border and diaphragm spread.

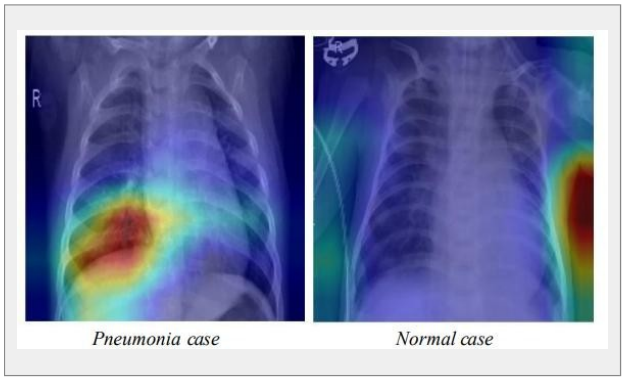


Fig. 4. Grad-CAM heatmaps for Model B (Partial FT). Strong localized activation within the lung fields.

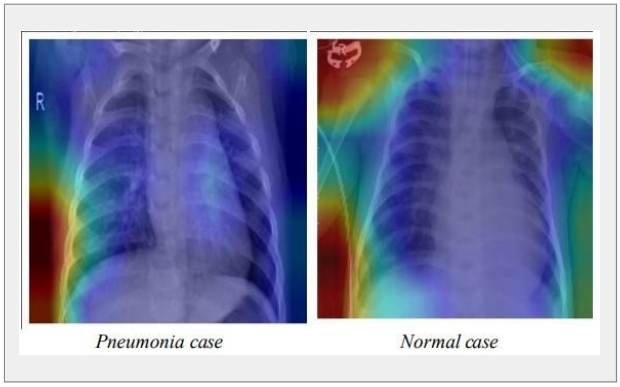


Fig. 5. Grad-CAM heatmaps for Model C (Full FT). Attention concentrated at image borders—consistent with shortcut learning.

Model A produced broadly spread activations with some concentration in the lung fields but noticeable activation along the diaphragm and peripheral regions. Model B showed a marked improvement: on pneumonia cases, activation concentrated sharply in the lung parenchyma, often aligning with regions corresponding to consolidation and opacity patterns. On normal cases, attention distributed evenly within the lung fields, which is appropriate given the absence of a single pathological feature. Model C was striking: across both classes, Grad-CAM activation concentrated along image borders, particularly at the corners,

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

with very little lung-field attention. This pattern was consistent across many test images and is consistent only with a learned shortcut.

### *D. IOLAR Results*

*TABLE IV. Mean IOLAR (Higher Values Indicate Better Lung-Field Focus)*

| Model | Pneumonia | Normal |
|---|---|---|
| A – Frozen | 0.91 | 0.39 |
| B – Partial | 0.91 | 0.71 |
| C – Full | 0.19 | 0.08 |

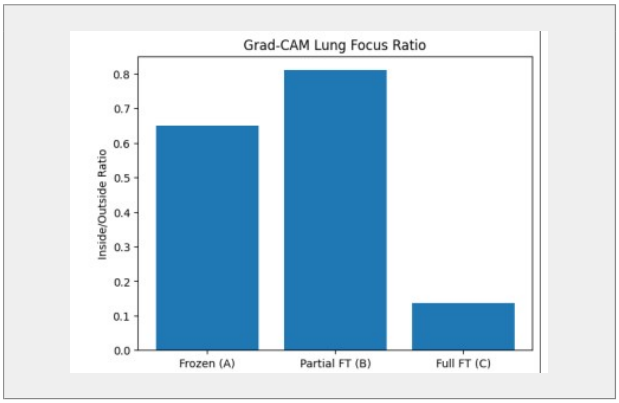


Fig. 6. Grad-CAM Lung Focus Ratio (IOLAR) as a function of fine-tuning depth. The sharp drop at full fine-tuning depth confirms border-dominant attention and shortcut learning.

Several observations stand out from Table IV and Fig. 6. Models A and B achieve nearly identical IOLAR on pneumonia images (both 0.91), meaning partial fine-tuning does not reduce lung-region attention on pathological cases. The improvement in classification accuracy from A to B comes without any sacrifice in spatial alignment on the clinically critical class. The difference between A and B on normal images (0.39 vs. 0.71) is also notable—Model B attends more consistently to the lung fields even when there is nothing pathological to find. Model C's values (0.19, 0.08) represent a near-complete collapse of lung-field attention: an IOLAR of 0.08 on normal images means the model directs more than 90% of its attention outside the lungs.

### *E. Statistical Tests*

*TABLE V. Statistical Comparison of IOLAR Distributions*

| Comparison | Class | t-stat | p-value | Cohen's d |
|---|---|---|---|---|
| B vs. C | Pneumonia | 14.3 | < 0.001 | 1.21 (large) |
| B vs. C | Normal | 11.7 | < 0.001 | 1.18 (large) |
| A vs. C | Pneumonia | 13.9 | < 0.001 | 1.17 (large) |
| A vs. B | Pneumonia | 0.42 | 0.674 | 0.04 (negl.) |
| A vs. B | Normal | 4.83 | < 0.001 | 0.49 (med.) |

*Cohen's d > 0.8 is a large effect.*

The Cohen's d values around 1.2 for the B vs. C and A vs. C comparisons fall in the range that Cohen described as large, and are large enough that the difference would be apparent to a reviewer examining heatmaps side-by-side. The non-significant A vs. B comparison on pneumonia images confirms that, despite very different training regimes, Models A and B attend to the lung fields with comparable consistency on positive cases.

### *F. SHAP Analysis*

The SHAP results tracked closely with Grad-CAM findings, which is significant because the two methods have entirely different computational foundations. For Model A, attributions were moderately concentrated within lung regions on pneumonia cases but noisy and peripheral on normal cases. For Model B, the maps were structured and interpretable: high positive attributions within the lung parenchyma, tracking regions of consolidation on positive cases. For Model C, the maps were sparse and heavily weighted toward image corners and borders—the same regions highlighted by Grad-CAM, reached by a completely different algorithm. This convergence across methods provides considerably stronger evidence that we are observing a genuine property of Model C's learned representations, rather than an artifact of any particular attribution technique.

Fig. 7. SHAP Attribution Comparison — Model B vs Model C

Left (Model B): Attributions are concentrated within lung regions, highlighting clinically relevant features.
Right (Model C): Attributions are sparse and focused on image borders and corners, suggesting reliance on non-clinical cues.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

## VI. DISCUSSION

### A. *The Non-Linear Effect of Fine-Tuning Depth*

The central finding of this study is that fine-tuning depth exerts a non-linear and asymmetric effect on model quality when quality is assessed across two dimensions: classification performance and spatial explainability. Increasing depth from frozen (Model A) to partial fine-tuning (Model B) improves both dimensions simultaneously — accuracy rises from 80.0% to 86.4% and IOLAR on normal images improves from 0.39 to 0.71. Increasing depth further to full fine-tuning (Model C) continues to improve accuracy to 89.3% but triggers a near-total collapse in explainability, with IOLAR dropping from 0.91 to 0.19 on pneumonia images. The relationship between fine-tuning depth and model quality is therefore not monotonic when both performance and explainability are considered together.

The most natural explanation for Model C's explainability collapse is shortcut learning. With all parameters freed to adapt, the fully fine-tuned model has apparently found a simpler signal in the image periphery—possibly acquisition metadata, DICOM overlay artifacts, or systematic differences in how normal versus pneumonia images were captured or stored. Model B, constrained to adapt only the highest-level layers, lacked the flexibility to discover or exploit this shortcut and was instead forced to learn from actual lung texture.

### B. *Why This Matters for Clinical Deployment*

A 2.9 percentage point accuracy difference may sound small, and in many application domains it would be. The explainability difference, however, is not small. If Model C were being prepared for clinical deployment and its Grad-CAM heatmaps were shown to a radiologist during a validation exercise, the response would almost certainly be confusion or concern. Attention patterns concentrated at image borders have no clinical interpretation and would immediately raise questions about what the model has actually learned.

Regulatory pathways for AI medical devices in both the U.S. and E.U. increasingly require demonstrating that models behave sensibly, not just that they achieve acceptable accuracy. The FDA's framework for AI/ML-based Software as a Medical Device [13] explicitly calls for transparency and auditability. A model whose attention maps center on image borders would face serious regulatory scrutiny. Model B, by contrast, produces attention concentrated in the lung fields, tracking consolidation on positive cases and distributing broadly on negative ones—a pattern that can be audited and validated by a domain expert.

### C. *On the False Positive Question*

It is worth dwelling on Model A's false positive rate. Flagging 110 out of 234 normal patients as having pneumonia would have real consequences in practice. In a pediatric setting, this could mean 47% of healthy children receiving unnecessary antibiotic prescriptions, contributing to antimicrobial resistance, unnecessary hospitalizations, parental distress, and healthcare costs. Model B's reduction to 78 false positives is clinically meaningful, as is the fact that it simultaneously achieves the lowest false negative count (7) of the three models. The combination of high pneumonia sensitivity, reduced false positive burden, and coherent explainability makes Model B the most defensible choice for clinical use even though it does not achieve the highest raw accuracy.

### D. *On Dataset Bias*

Caution is warranted against over-generalization. The shortcut that Model C appears to have learned is presumably specific to this dataset—a consequence of systematic differences between normal and pneumonia images in the Kermany collection. A different dataset might not present the same opportunity for border-based shortcuts, or might present different ones that lung-activation analysis would not detect. Partial fine-tuning is not a complete solution to shortcut learning; the right approach combines careful data curation, appropriate fine-tuning strategy selection, and post- hoc explainability evaluation.

### E. *Limitations*

Several limitations should be acknowledged. The dataset is single-center, limiting generalizability. The lung segmentation masks introduce boundary noise, particularly where pathology has distorted the normal lung outline. Model calibration was not examined. The SHAP analysis was computationally expensive, restricting it to 100 images per class per model. Finally, only ResNet50 was evaluated; Vision Transformers have fundamentally different architectures with built-in attention mechanisms and may exhibit different behavior under varying fine-tuning depths.

## VII. FUTURE WORK

Testing on multi-site, multi-center datasets to determine whether the partial fine-tuning advantage in explainability holds across different dataset biases.

Exploring attention regularization methods during training, such as the Right for the Right Reasons approach [11], which directly penalizes attention to irrelevant image regions and could be combined with partial fine-tuning.

Extending the analysis to Vision Transformers, where self-attention mechanisms provide a built-in interpretability signal and fine-tuning depth effects may manifest differently.

Conducting user studies with radiologists to measure how explainability differences affect clinical decision-making and trust, providing direct evidence that IOLAR differences translate to meaningful real-world impact.

## VIII. CONCLUSION

This study investigated how fine-tuning depth affects both performance and explainability in transfer learning-based pneumonia detection—two dimensions of model quality that are typically evaluated in isolation but interact in non-obvious ways.

The central finding is that fine-tuning depth affects explainability non-linearly. Moving from a fully frozen backbone to partial fine-tuning improves both accuracy and

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

explainability simultaneously: accuracy increases from 80.0% to 86.4%, and IOLAR on normal images improves from 0.39 to 0.71. Pushing further to full fine-tuning yields an additional 2.9 percentage points in accuracy but causes explainability to collapse—IOLAR falls from 0.91 to 0.19 on pneumonia images. Both Grad-CAM and SHAP independently point to the same conclusion: the fully fine-tuned model focuses on image borders rather than lung tissue, consistent with shortcut learning. Statistical tests confirm this is not a marginal effect ($p < 0.001$, Cohen's $d \approx 1.2$).

For practitioners, partial fine-tuning—unfreezing only the final convolutional block—represents the optimal strategy for this task. It delivers solid performance (86.4% accuracy, 98% pneumonia recall, AUC = 0.960), the fewest false negatives of all three models, and attention patterns that are clinically coherent and auditable. Full fine-tuning offers marginally better accuracy at the cost of a model whose reasoning is not interpretable and would likely not survive regulatory scrutiny.

More broadly, fine-tuning depth should not be treated merely as another hyperparameter to optimize for accuracy. It is a design decision that determines not just test set performance, but what the model has actually learned—and whether it is learning from the right clinical signals. Building medical AI that can be deployed safely and trusted by clinicians requires treating explainability as a first-class metric, evaluated quantitatively and validated statistically alongside standard performance measures.

## REFERENCES

[1] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proc. IEEE CVPR, Las Vegas, NV, 2016, pp. 770–778.

[2] R. R. Selvaraju et al., "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in Proc. IEEE ICCV, Venice, 2017, pp. 618–626.

[3] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in Advances in NeurIPS, vol. 30, 2017, pp. 4765–4774.

[4] R. Geirhos et al., "Shortcut learning in deep neural networks," Nature Machine Intelligence, vol. 2, no. 11, pp. 665–673, 2020.

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

[6] D. S. Kermany et al., "Identifying medical diagnoses and treatable diseases by image-based deep learning," Cell, vol. 172, no. 5, pp. 1122–1131, 2018.

[7] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?" in Advances in NeurIPS, vol. 27, 2014, pp. 3320–3328.

[8] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions," Nature Machine Intelligence, vol. 1, no. 5, pp. 206–215, 2019.

[9] J. R. Zech et al., "Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs," PLOS Medicine, vol. 15, no. 11, 2018.

[10] L. Oakden-Rayner et al., "Hidden stratification causes clinically meaningful failures in machine learning for medical imaging," in Proc. ACM CHIL, 2020, pp. 151–159.

[11] A. S. Ross, M. C. Hughes, and F. Doshi-Velez, "Right for the right reasons: Training differentiable models by constraining their explanations," in Proc. IJCAI, 2017, pp. 2662–2670.

[12] B. Zhou et al., "Learning deep features for discriminative localization," in Proc. IEEE CVPR, 2016, pp. 2921–2929.

[13] U.S. Food and Drug Administration, "Artificial intelligence and machine learning (AI/ML)-based software as a medical device action plan," FDA White Paper, Jan. 2021.

[14] World Health Organization, "Pneumonia in children," WHO Fact Sheet, Nov. 2022.

[15] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in Proc. ICLR, 2021.