

A Comparative Study of ResNet50, VGG16, VGG19, And InceptionV3 Transfer-Learning Architectures for Pneumonia Detection and Classification from Chest X-Ray Images

K Ruchitha¹, Jhansi rani²

^{1,2}Department of Computer Science and Systems Engineering, Andhra University College of Engineering, Visakhapatnam, India

Abstract—Pneumonia remains an important respiratory disease for which rapid and reliable image-based screening can assist clinical decision-making. This study presents a controlled comparative evaluation of four convolutional neural network (CNN) architectures—ResNet50, VGG16, VGG19, and InceptionV3—for binary classification of chest X-ray images into pneumonia-positive and normal categories. All networks are initialized with ImageNet-pre-trained weights and adapted using a common classification head containing convolution, max pooling, dropout, batch normalization, a dense layer, and a sigmoid output. To reduce the effect of limited training data and class imbalance, training-only augmentation using horizontal flipping, rotation, and zoom is employed. A two-stage transfer-learning strategy is used: the convolutional base is first frozen and the newly added classification head is trained, followed by selective unfreezing and fine-tuning of deeper layers at a lower learning rate. The reported test results show accuracies of 85.7%, 86.2%, 88.9%, and 91.3% for VGG16, VGG19, InceptionV3, and ResNet50, respectively. ResNet50 also reports the highest F1-score (91%) and AUC (0.96) in the supplied experimental results. The study further examines preprocessing, augmentation, generalization, computational considerations, and practical deployment issues. The results indicate that architectural design and staged fine-tuning are important factors when transfer learning is applied to relatively small and imbalanced chest X-ray datasets.

Index Terms—Deep Learning, Transfer Learning, Pneumonia Detection, Chest X-Ray, ResNet50, VGG16, VGG19, InceptionV3, CNN, Medical Image Classification.

I. INTRODUCTION

Pneumonia is an infection that affects the lungs and can produce abnormalities visible in chest

radiographs. Chest X-ray examination is widely used because it is comparatively accessible and provides useful information about pulmonary patterns. However, interpretation can be affected by image quality, overlapping anatomical structures, disease presentation, and differences in reader experience. These factors motivate the development of computer-aided systems that can provide an additional automated screening signal.

Deep learning has become a major approach for medical image classification because convolutional neural networks can learn hierarchical visual features directly from images. Training a deep CNN completely from scratch, however, normally requires large labelled datasets and substantial computational resources. Transfer learning addresses this limitation by starting from a model trained on a large image corpus such as ImageNet and adapting its learned representations to a target task. The supplied study applies this principle to four established CNN families and keeps the downstream training pipeline as consistent as possible.

The central research question is how different backbone architectures behave when the same pneumonia-classification pipeline is applied to a relatively small, imbalanced chest X-ray dataset. The comparison focuses on VGG16, VGG19, InceptionV3, and ResNet50. The study also examines the contribution of augmentation and partial fine-tuning, because a frozen ImageNet representation may not fully capture the domain-specific texture and spatial patterns present in chest radiographs.

II. OBJECTIVES OF THE STUDY

- To develop a transfer-learning pipeline for binary classification of chest X-ray images as normal or pneumonia-positive.
- To compare ResNet50, VGG16, VGG19, and InceptionV3 under a common preprocessing, augmentation, classification-head, and fine-tuning framework.
- To investigate whether staged fine-tuning improves the performance obtained from a frozen pre-trained backbone.
- To evaluate the model's using accuracy, precision, recall, F1-score, and AUC rather than relying only on accuracy.
- To discuss generalization, class imbalance, overfitting, computational cost, and practical considerations for computer-aided screening.

III. CONTRIBUTIONS

The main contribution of this work is a controlled architecture comparison in which the classification head and major training procedures are held constant. This design reduces the possibility that performance differences are caused simply by changing multiple experimental factors at the same time. A second contribution is the explicit examination of staged fine-tuning. The paper also expands the discussion to include generalization, deployment constraints, and limitations that should be considered before applying a model to clinical data.

IV. RELATED WORK

Deep CNNs have been increasingly studied for chest radiograph analysis because their learned feature hierarchies can represent edges, textures, shapes, and higher-level image patterns. Earlier work demonstrated that features learned from natural-image datasets can be transferred to medical images. Subsequent research introduced large chest X-ray datasets and multi-label frameworks for identifying thoracic abnormalities. These developments established transfer learning as a practical strategy when medical datasets are smaller than general-purpose image datasets. CheXNet, for example, investigated a 121-layer densely connected architecture for pneumonia-related

chest X-ray analysis. Other studies explored transfer learning with VGG, ResNet, Inception, and related CNN families.

Comparative studies have also examined ensembles, where predictions from multiple backbones are combined. Although ensembles can improve robustness, they generally increase inference cost and make deployment more complex.

The present work differs in emphasis by using an identical classification head, augmentation strategy, and staged fine-tuning concept across four backbones. The intention is not to claim that one architecture is universally superior, but to document how the supplied experimental pipeline behaves for the selected dataset and metrics.

V. PROBLEM STATEMENT

A chest X-ray contains complex anatomical and pathological information. Pneumonia may appear through diffuse or localized opacities, and the visual appearance can vary between patients. A model therefore needs to learn patterns that are sufficiently discriminative while avoiding dependence on irrelevant image characteristics. The problem is further complicated by class imbalance, limited labelled examples, and differences between training and real-world clinical images.

A useful automated system should therefore provide more than a high training accuracy. It should generalize to unseen images, maintain acceptable sensitivity to pneumonia-positive cases, limit unnecessary false positives, and remain computationally practical. These requirements motivate the combined use of transfer learning, augmentation, regularization, and multi-metric evaluation.

VI. DATASET DESCRIPTION

The supplied paper uses the publicly available chest X-ray pneumonia dataset containing 5,863 anterior-posterior radiographs organized into training, validation, and test partitions. The two target classes are pneumonia-positive and normal. The dataset is imbalanced, with pneumonia-positive images substantially more numerous than normal images. The original study addresses this imbalance through targeted training augmentation.

Dataset characteristic	Description
Image modality	Chest X-ray radiographs
Image view	Anterior-posterior images
Total images	5,863
Target classes	Normal; Pneumonia
Data partitions	Training; Validation; Test
Class distribution	Pneumonia-positive class is larger
Learning approach	ImageNet transfer learning

The separation of training, validation, and test data is important for an unbiased evaluation. Augmentation is restricted to the training partition in the supplied framework. Validation and test images are resized and rescaled without random augmentation so that reported performance reflects the model's behavior on fixed unseen samples.

VII. PREPROCESSING AND DATA AUGMENTATION

- Images are resized to 224×224 pixels for VGG16, VGG19, and ResNet50, while InceptionV3 uses 299×299 pixels.
- Pixel intensity values are rescaled to the range $[-1, 1]$ to match the expected input distribution used by the transfer-learning pipeline.
- Random horizontal flipping is used to increase appearance variability during training.
- Random rotation of approximately $\pm 20^\circ$ is used to expose the model to small orientation changes.
- Random zoom of approximately $\pm 20\%$ is used to reduce sensitivity to scale and framing differences.
- Augmentation is applied only to training images; validation and test data are not augmented.

Augmentation does not create genuinely new patient information, but it can expose the network to plausible variations of existing training examples. This is especially useful when the available dataset is limited. The transformations must nevertheless remain clinically reasonable; excessively strong transformations can alter disease-relevant patterns.

VIII. TRANSFER-LEARNING FRAMEWORK

Each selected backbone is initialized with ImageNet-pre-trained weights. The original classification layer is

removed because the original network was designed for a different multi-class image-recognition problem. A common task-specific head is then attached. Using a shared head is useful in a comparative experiment because the major variable becomes the feature-extraction backbone.

Component	Role in the proposed pipeline
Pre-trained CNN backbone	Extracts hierarchical visual features from chest X-ray images
2D convolution layer	Provides an additional task-specific feature transformation
Max pooling	Reduces spatial dimensions and retains strong responses
Dropout (0.2)	Regularizes the classifier and reduces co-adaptation
Batch normalization	Stabilizes intermediate activations during training
Flatten	Converts feature maps into a vector representation
Dense layer (128, ReLU)	Learns higher-level task-specific combinations
Dropout (0.2)	Adds further regularization before classification
Sigmoid output	Produces a probability for binary classification

IX. ARCHITECTURE DESCRIPTION

A. RESNET50

ResNet50 is a deep residual network built from bottleneck blocks. Each block uses a sequence of 1×1 , 3×3 , and 1×1 convolutions together with an identity shortcut. The shortcut allows information and gradients to travel through the network without depending entirely on every intermediate transformation. This residual formulation helps make deep networks easier to optimize. In the supplied comparison, ResNet50 contains approximately 25.6 million parameters and is used as an ImageNet feature extractor followed by the shared pneumonia-classification head.

B. VGG16

VGG16 uses a relatively uniform sequence of small 3×3 convolutional filters followed by pooling operations and fully connected layers. The architecture contains 13 convolutional layers and 3 fully connected

layers, giving 16 weight layers. It's simple and regular structure makes it easy to understand, but its large parameter count can increase memory use and overfitting risk when the target dataset is comparatively small.

C. VGG19

VGG19 extends the VGG design with additional convolutional depth. It contains 19 weight layers and retains the same general pattern of small convolutions and pooling.

The additional depth can provide richer feature representations, but the model remains computationally heavier than architectures that use residual connections or factorized convolutions. In this study, VGG19 is evaluated with exactly the same task-specific head and staged training concept as the other models.

D. INCEPTIONV3

InceptionV3 uses factorized and multi-branch convolutional operations to capture information at different spatial scales while controlling computational cost. Larger convolutions can be factorized into smaller or asymmetric operations, reducing the cost of direct large-kernel processing. The supplied paper reports approximately 23.9 million parameters across the architecture and uses a 299×299 input size.

X. CLASSIFICATION PIPELINE

For an input image, the preprocessing stage first converts the image to the required size and intensity range. The selected backbone then generates feature maps.

The shared classification head transforms these maps into a compact representation and finally produces a sigmoid probability. A threshold can be applied to convert the probability into the two output classes: normal or pneumonia.

In a deployment-oriented system, the probability should be treated as a model output rather than as an independent medical diagnosis.

Clinical use would require external validation, appropriate calibration, assessment across different hospitals and devices, and oversight by qualified healthcare professionals.

XI. TRAINING AND FINE-TUNING METHODOLOGY

The training procedure is divided into two phases. In Phase 1, the pre-trained convolutional base is frozen and only the newly added classification head is trained. This allows the new layers to learn the binary pneumonia task without immediately changing all pre-trained features. The supplied paper uses 20 epochs for this initial stage.

In Phase 2, selected deeper convolutional blocks are unfrozen and trained jointly with the classification head. The learning rate is reduced so that the pre-trained representation is adapted gradually rather than overwritten too aggressively. The supplied framework uses a further 10 epochs for this phase. Early stopping and ReduceLROnPlateau are monitored using validation loss to reduce unnecessary training and to adapt the learning rate when progress slows.

Training element	Configuration in supplied paper
Loss function	Binary cross-entropy
Optimizer	RMSProp
Initial learning rate	0.0001
Phase 1	Frozen convolutional base; 20 epochs
Phase 2	Selective deeper-layer unfreezing; 10 epochs
Learning rate in Phase 2	Reduced relative to initial phase
Callbacks	Early stopping; ReduceLROnPlateau
Monitoring	Validation loss

XII. EVALUATION METRICS

Accuracy measures the proportion of correctly classified images among all evaluated images. It is easy to interpret but can be misleading when classes are imbalanced. Precision measures how many images predicted as pneumonia are actually pneumonia-positive.

Recall, also called sensitivity, measures how many pneumonia-positive cases are detected by the model. F1-score is the harmonic mean of precision and recall and provides a single measure that balances the two. AUC summarizes the model's ability to separate the two classes across classification thresholds.

Metric	Interpretation
Accuracy	$(TP + TN) / (TP + TN + FP + FN)$
Precision	$TP / (TP + FP)$
Recall / Sensitivity	$TP / (TP + FN)$
F1-score	$2 \times \text{Precision} \times \text{Recall} / (\text{Precision} + \text{Recall})$
AUC	Area under the ROC curve across thresholds

For pneumonia screening, recall deserves particular attention because a false negative means that an image containing pneumonia is classified as normal. However, recall should not be interpreted in isolation. A model can increase recall by producing more false positives, so precision, specificity, F1-score, ROC-AUC, and confusion-matrix analysis should also be considered.

XIII. OVERFITTING AND GENERALIZATION

Generalization refers to how well a trained model performs on previously unseen images rather than only on the images used during optimization. In medical imaging, generalization is particularly important because images can differ by hospital, X-ray machine, acquisition protocol, patient population, image quality, and disease prevalence.

The study addresses generalization through several mechanisms. First, transfer learning provides a useful starting representation. Second, augmentation exposes the model to controlled input variation. Third, dropout and batch normalization provide regularization within the classification head. Fourth, the test partition is kept separate from model fitting. Finally, staged fine-tuning allows domain-specific adaptation while using a smaller learning rate during deeper-layer updates.

A high-test score on one public dataset does not automatically establish clinical generalization. Stronger evidence would require external datasets, cross-institutional validation, patient-level separation, and evaluation on images collected under different acquisition conditions.

XIV. EXPERIMENTAL RESULTS

Table I reproduces the comparative performance reported in the supplied paper. All values are percentages except AUC. The table should be

interpreted as results for the specified dataset and experimental framework, not as universal performance values for the architectures.

Model	Accuracy (%)	Precision (%)	Recall (%)	F1 (%)	AUC
VGG16	85.7	86	84	85	0.91
VGG19	86.2	87	85	86	0.92
InceptionV3	88.9	89	87	88	0.94
ResNet50	91.3	91	90	91	0.96

XV. EFFECT OF STAGED FINE-TUNING

The supplied study also reports an ablation-style comparison for ResNet50. This comparison shows how the performance changes as additional training components are introduced.

Configuration	Accuracy (%)	Precision (%)	Recall (%)	F1 (%)
Frozen base only	83.4	84	82	83
Frozen base + augmentation	86.8	87	85	86
Partial fine-tuning (proposed)	91.3	91	90	91

The reported sequence indicates that augmentation provides an improvement over the frozen-base configuration and that partial fine-tuning produces a further increase. A reasonable interpretation is that the pre-trained representation contains useful generic visual information, while deeper adaptation allows the model to better represent chest-radiograph-specific patterns. This interpretation is consistent with the purpose of transfer learning: retain useful learned features while adapting them to the target domain.

XVI. COMPARATIVE DISCUSSION

Within the supplied experiment, ResNet50 reports 91.3% accuracy, 91% precision, 90% recall, 91% F1-score, and 0.96 AUC. InceptionV3 reports 88.9% accuracy and 0.94 AUC, while VGG16 and VGG19 report 85.7% and 86.2% accuracy, respectively. The numerical differences describe the observed performance on the stated test partition.

The architecture-level explanation proposed in the paper is that residual connections in ResNet50 and factorized multi-scale processing in InceptionV3 can provide efficient feature learning for the target

problem. The VGG models use a simpler sequential design but have substantially larger parameter counts, especially in their fully connected portions. With a limited dataset, a large parameter count can increase the opportunity for overfitting unless sufficient regularization and data are available.

The results should also be considered together with computational requirements. ResNet50 and InceptionV3 are structurally different from VGG16 and VGG19, so accuracy alone does not capture all deployment trade-offs. Inference time, memory usage, model size, energy consumption, calibration, and ease of maintenance are also relevant when selecting a model for a real application.

XVII. CONFUSION MATRIX INTERPRETATION

A confusion matrix divides predictions into true positives, true negatives, false positives, and false negatives. For this task, a true positive is a pneumonia-positive X-ray correctly identified as pneumonia, while a true negative is a normal X-ray correctly identified as normal. A false negative is a pneumonia-positive X-ray incorrectly labelled normal, and a false positive is a normal X-ray incorrectly labelled pneumonia.

The confusion matrix is useful because two models with similar accuracy can have different error patterns. For a medical screening application, the relative frequency of false negatives and false positives should be examined together with the intended clinical workflow.

XVIII. ADDITIONAL ANALYSIS: CLASS IMBALANCE

The dataset used in the study contains more pneumonia-positive images than normal images. If a classifier is trained without considering this imbalance, the loss function can be influenced more strongly by the larger class. The supplied framework uses targeted augmentation of the minority class during training. More generally, class weighting, balanced sampling, focal loss, or carefully designed resampling can also be investigated, provided that the method is applied only to the training data.

Accuracy should be interpreted carefully under imbalance. For example, a model could obtain a

seemingly high accuracy by performing well on the larger class while failing to detect enough samples from the smaller class. This is why the present paper reports precision, recall, F1-score, and AUC in addition to accuracy.

XIX. ADDITIONAL ANALYSIS: DATA LEAKAGE

Data leakage occurs when information from the validation or test distribution indirectly influences training. In medical imaging, leakage can happen if multiple images from the same patient are split across training and testing, or if preprocessing statistics are calculated using the complete dataset before partitioning.

The safest evaluation design uses patient-level separation whenever patient identifiers are available and keeps all test information isolated until final evaluation.

The supplied paper describes separate training, validation, and test partitions and specifies that augmentation is restricted to training. For a future experimental version, the study can explicitly document whether the dataset partitions are patient-independent, because this detail affects how strongly the reported test results can be generalized.

XX. ADDITIONAL ANALYSIS: EXPLAINABILITY

A medical image classifier should ideally provide more than a class label. Explainability methods such as Grad-CAM can be used to visualize image regions that contribute strongly to a prediction. For chest X-rays, such visualizations can help researchers inspect whether the network is responding to clinically meaningful lung regions or to irrelevant artifacts such as markers, borders, or acquisition-related patterns.

Heatmaps should not be treated as proof that the model has discovered the correct pathology. They are diagnostic tools for model inspection. A robust explainability study should compare visualizations across correct predictions, false positives, and false negatives and, where possible, involve clinical experts in interpretation.

XXI. ROBUSTNESS AND GENERALIZATION TESTING

An extended evaluation can test the trained models against controlled changes such as mild contrast variation, noise, compression, cropping, and image-resolution differences.

These tests help determine whether the classifier is overly sensitive to image-processing conditions. External validation on an independent dataset is stronger than synthetic perturbation because it tests naturally occurring differences in acquisition and patient populations.

Generalization test	Purpose
Independent dataset	Measures cross-dataset transfer
Cross-device evaluation	Tests sensitivity to acquisition hardware
Image-quality variation	Tests robustness to noise and resolution
Patient-level split	Reduces dependence between training and testing
Threshold analysis	Examines precision/recall trade-offs
Calibration analysis	Checks whether predicted probabilities are meaningful

XXII. COMPUTATIONAL AND DEPLOYMENT CONSIDERATIONS

A practical screening system must consider model size, inference latency, available memory, and hardware constraints in addition to predictive performance. A model that is accurate but too large or slow for the target environment may require compression, quantization, pruning, or a lighter backbone. Conversely, reducing the model too aggressively can change its sensitivity and should therefore be validated after optimization.

XXIII. PROPOSED SYSTEM WORKFLOW

The complete workflow can be represented as a sequence of data preparation, model adaptation, training, evaluation, and prediction stages. The same basic workflow is used for all four architectures so that the comparison remains controlled.

- Step 1 – Dataset collection: obtain the chest X-ray images and the normal/pneumonia labels.
- Step 2 – Dataset partitioning: maintain separate training, validation, and test partitions.
- Step 3 – Preprocessing: resize images according to the selected backbone and rescale pixel values.
- Step 4 – Training augmentation: apply horizontal flipping, rotation, and zoom only to training images.
- Step 5 – Backbone initialization: load ImageNet-pre-trained VGG16, VGG19, InceptionV3, or ResNet50.
- Step 6 – Classification head: attach the common convolution, pooling, dropout, normalization, dense, and sigmoid layers.
- Step 7 – Phase 1 training: freeze the backbone and train the task-specific head.
- Step 8 – Phase 2 fine-tuning: selectively unfreeze deeper layers and continue training at a lower learning rate.
- Step 9 – Evaluation: calculate accuracy, precision, recall, F1-score, AUC, and confusion matrix.
- Step 10 – Model comparison: examine predictive performance together with computational and generalization considerations.
- Step 11 – Prediction: pass a new X-ray through the selected model and report the predicted class and probability.

XXIV. IMPLEMENTATION ENVIRONMENT

Category	Recommended / stated implementation
Programming language	Python
Deep learning framework	TensorFlow / Keras
Array and data processing	NumPy and related Python tools
Image processing	OpenCV / Keras image utilities
Training hardware	GPU-enabled environment
Model input	224 × 224 for VGG/ResNet; 299 × 299 for InceptionV3
Loss	Binary cross-entropy
Optimizer	RMSProp
Evaluation	Accuracy, precision, recall, F1, AUC, confusion matrix

XXV. PRACTICAL APPLICATION SCENARIO

A possible application is a computer-aided screening interface in which a chest X-ray is uploaded to a trained model. The system performs the same preprocessing used during training, sends the image through the selected backbone, and returns a normal or pneumonia-positive classification with a probability value. An optional explanation map can be generated for model inspection.

The appropriate role of such a system is decision support rather than replacement of radiological or clinical judgement. A real clinical deployment would require prospective validation, regulatory review where applicable, security and privacy controls, monitoring for distribution shift, and a clear process for handling uncertain or low-quality images.

XXVI. LIMITATIONS OF THE PRESENT STUDY

- The comparison is based on one public dataset and therefore may not represent the diversity of images encountered in different hospitals.
- The dataset is class-imbalanced, so performance should not be interpreted from accuracy alone.
- The reported results come from the supplied experimental framework and should be reproduced before being used as final claims for a new implementation.
- External validation and patient-level independence should be explicitly documented for stronger evidence of generalization.
- Chest X-ray findings can overlap with other diseases and technical artifacts; a binary classifier may not represent the full clinical differential diagnosis.
- Model probabilities may require calibration before they are interpreted as reliable confidence estimates.

XXVII. ETHICAL AND SAFETY CONSIDERATIONS

Medical AI systems can affect decisions about patient care, so privacy, fairness, transparency, and human oversight are important. Training data should be handled according to applicable privacy requirements. Evaluation should include relevant patient groups and acquisition settings where possible. Developers should document intended use, known limitations, and

situations in which the model should not be relied upon.

XXVIII. FUTURE ENHANCEMENTS

- Evaluate additional CNN backbones such as DenseNet121 and EfficientNet under the same controlled experimental protocol.
- Investigate Vision Transformer and hybrid CNN-transformer approaches for comparison with convolutional backbones.
- Use multi-institutional and geographically diverse datasets to test cross-domain generalization.
- Perform patient-level splitting and external validation to reduce the possibility of optimistic estimates.
- Investigate ensemble methods while reporting the additional memory and inference cost.
- Add Grad-CAM or related explanation methods for systematic model inspection.
- Evaluate calibration and clinically relevant operating thresholds rather than relying on a single default threshold.
- Explore model compression, quantization, and lightweight architectures for resource-constrained deployment.
- Study robustness to image-quality changes and acquisition differences.
- Develop a clear human-in-the-loop workflow in which uncertain predictions are referred for expert review.

XXIX. DISCUSSION AND KEY FINDINGS

The comparative experiment demonstrates that the choice of backbone can materially affect pneumonia classification performance when the rest of the training pipeline is controlled. In the supplied results, ResNet50 records 91.3% accuracy and 0.96 AUC, while InceptionV3 records 88.9% accuracy and 0.94 AUC. VGG16 and VGG19 record 85.7% and 86.2% accuracy, respectively. The staged ResNet50 experiment further reports a progression from 83.4% accuracy with a frozen base, to 86.8% after augmentation, and to 91.3% after partial fine-tuning. These observations support the practical importance of transfer-learning strategy rather than simply selecting a deep network. The results suggest that useful generic

ImageNet features can be retained initially, while domain-specific adaptation of deeper layers can improve performance. At the same time, the numerical results should be interpreted within the boundaries of the dataset and experimental protocol.

For a research paper, the most important next step is therefore not only to increase the numerical score but to strengthen evidence for generalization. Independent datasets, patient-level separation, calibration, robustness testing, explainability, and transparent reporting can provide a more complete evaluation of a pneumonia screening model.

XXX. CONCLUSION

This paper presented a comparative transfer-learning study of ResNet50, VGG16, VGG19, and InceptionV3 for binary pneumonia detection from chest X-ray images. A common classification head, training augmentation strategy, and staged fine-tuning concept were used to make the architectural comparison more consistent. The supplied experimental results report 91.3% accuracy for ResNet50, 88.9% for InceptionV3, 86.2% for VGG19, and 85.7% for VGG16, with ResNet50 also reporting the highest F1-score and AUC in the presented table.

The study highlights the usefulness of transfer learning, augmentation, regularization, and partial fine-tuning when working with a limited medical image dataset. It also emphasizes that accuracy alone is insufficient for assessing a medical screening model. Future work should focus on external validation, generalization across institutions, explainability, calibration, robustness, and efficient deployment.

The proposed system should be considered a computer-aided screening and research framework rather than an autonomous diagnostic replacement. Its practical value depends on rigorous validation, appropriate clinical oversight, and responsible handling of patient data.

REFERENCES

[1] World Health Organization, "Pneumonia," 2021. [Online]. Available: <https://www.who.int/health-topics/pneumonia>.

[2] Kaggle, "Chest X-Ray Images (Pneumonia) dataset," publicly available chest radiograph dataset used in the supplied study.

[3] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2016, pp. 770–778.

[4] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv:1409.1556, 2014.

[5] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the Inception architecture for computer vision," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2016, pp. 2818–2826.

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

[7] V. Chouhan et al., "A novel transfer learning-based approach for pneumonia detection in chest X-ray images," Appl. Sci., vol. 10, no. 2, p. 559, 2020.

[8] T. Rahman et al., "Transfer learning with deep convolutional neural network (CNN) for pneumonia detection using chest X-ray," Appl. Sci., vol. 10, no. 9, p. 3233, 2020.

[9] M. Mujahid et al., "Pneumonia classification from X-ray images with Inception-V3 and convolutional neural network," Diagnostics, vol. 12, no. 5, p. 1280, 2022.

[10] M. K. Dalei and S. Mahapatra, "Pneumonia detection and classification from chest X-ray images using modified DenseNet121 deep learning approach," in Proc. ICAIHC, 2025.

[11] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2017.

[12] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in Proc. Int. Conf. Learning Representations (ICLR), 2015.

[13] I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning. Cambridge, MA, USA: MIT Press, 2016.