Comparative Evaluation of CNN Architectures for Pneumonia Detection from Chest X ray Images

Albahlool Abood
Faculty of Information Technology
University of Gharyan, Libya
Albahlool.abood@gu.edu.ly

Khairia Mabrok
Faculty of Information Technology
University of Gharyan, Libya
Khiriakhaledalshebani@gmail.com

Rodaina Ahmad
Faculty of Information Technology
University of Gharyan, Libya

Aimen Ahmad
Faculty of Information Technology
University of Gharian, Libya
aimen.ahmad@gu.edu.ly

Abstract*— Pneumonia remains a major global health burden, where timely recognition on chest X ray images is clinically important yet often challenged by subtle radiographic signs and variability in interpretation. This paper presents a controlled comparative evaluation of four convolutional neural network architectures, MobileNet, ResNet50, VGG, and InceptionV3, for binary classification of chest X ray images into diseased and normal cases. Experiments were conducted using a publicly available Kaggle dataset of 4,479 images under a unified preprocessing and evaluation protocol. Performance was assessed on a held out test set of 300 images, including 200 diseased and 100 normal cases, using accuracy and macro averaged precision, recall, and F1 score, supported by confusion matrix analysis. The results show that MobileNet achieved the highest test accuracy at 95.0 percent, while ResNet50 and VGG achieved 94.7 percent, and InceptionV3 achieved 92.0 percent. Confusion matrix inspection indicates that MobileNet produced the fewest false negatives for diseased cases in this setting, which is important for screening oriented use. Inference time measurements using batch size 1 at 180 × 180 input on CPU further highlight the efficiency advantage of lightweight architectures. Overall, these findings provide a reproducible benchmark to support architecture selection for computer assisted pneumonia screening and clinical triage .

***Keywords*—Chest X ray, Pneumonia detection, Convolutional neural networks, Deep learning, Medical imaging, Inference time**

I. Introduction

Pneumonia remains a major cause of morbidity and mortality worldwide, with elevated risk among children under five years of age, older adults, and immunocompromised patients. The World Health Organization identifies pneumonia as a significant public health concern, particularly in low and middle income countries where access to diagnostic resources and specialist expertise may be limited [1], [2]. Early and accurate diagnosis is therefore essential to reduce complications, support timely treatment decisions, and improve survival outcomes.

Chest X ray imaging is widely used for suspected pneumonia because it is affordable, accessible, and noninvasive. In routine clinical practice, clinicians and radiologists examine radiographic patterns such as opacities and consolidations to determine whether infection is present. However, reliable interpretation can be challenging when findings are subtle, anatomical overlap obscures relevant regions, image quality varies across acquisition conditions, and decisions depend on reader experience [3]. In high volume environments, time pressure and interobserver variability may further increase the risk of delayed detection or misclassification.

Recent advances in artificial intelligence, particularly deep learning, have improved automated analysis of medical images. Convolutional neural networks can learn hierarchical feature representations directly from pixel level data and have demonstrated strong performance in radiological tasks without reliance on handcrafted feature extraction [4], [5]. Prior studies also report that CNN based systems can achieve performance comparable to expert radiologists for pneumonia detection, which supports the feasibility of computer assisted screening and triage [3], [6]. Fig. 1 summarizes the workflow adopted in this study, from preprocessing to model inference and output interpretation.

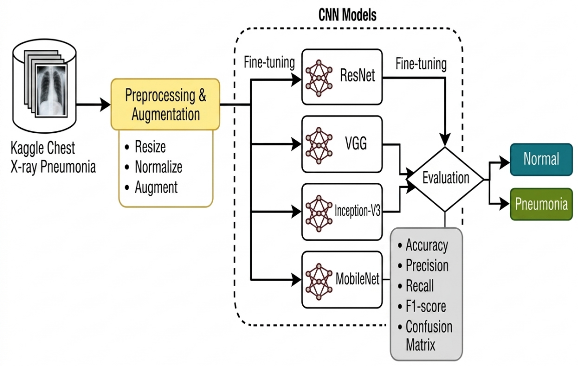


Fig. 1. Standard CNN workflow for pneumonia detection from chest X ray images.

DOI:10.5281/zenodo.20088614

Despite this progress, several limitations remain in the literature and restrict fair benchmarking and reproducibility. First, many studies evaluate a single architecture, which limits direct architectural comparison under consistent conditions. Second, reported results are often not comparable across studies due to differences in preprocessing pipelines, dataset splitting strategies, augmentation policies, training configurations, and evaluation metrics [7], [8]. Third, fewer studies jointly consider diagnostic performance and practical deployment factors, including computational efficiency and inference latency, which are important for real time screening in resource constrained settings.

To address these limitations, this paper presents a controlled comparative evaluation of four widely used CNN architectures, ResNet, VGG, InceptionV3, and MobileNet, for binary pneumonia detection from chest X ray images. All models are trained and evaluated under a unified experimental protocol using identical preprocessing procedures, training configurations, and evaluation criteria. This design enables transparent comparison and clearer identification of architectural strengths and limitations.

The contributions of this study are summarized as follows:

1. A unified comparative evaluation of four widely used CNN architectures for binary pneumonia detection under identical experimental conditions.

2. A consistent experimental framework that supports transparent and reproducible architectural comparison using the same preprocessing, training, and evaluation settings.

3. A deployment oriented perspective through analysis of model accuracy and inference efficiency to support informed architecture selection for AI assisted pneumonia screening.

The remainder of this paper is organized as follows. Section II reviews related studies and clarifies the research gap. Section III describes the dataset, preprocessing procedures, model architectures, and experimental setup. Section IV presents the experimental results. Section V discusses the findings, including clinical implications and study limitations. Section VI concludes the paper and outlines directions for future research.

II. Related Work

Deep learning, particularly Convolutional Neural Networks, has substantially advanced automated chest X ray interpretation by enabling end to end feature learning directly from image data. Early progress in this area was accelerated by the availability of large scale radiographic datasets and the success of transfer learning, where models pretrained on natural image benchmarks are adapted to medical imaging tasks. Nevertheless, reported performance in pneumonia detection remains highly dependent on dataset characteristics, label quality, preprocessing pipelines, and evaluation protocols, which makes direct comparison across studies inherently difficult. Recent survey studies have also highlighted the rapid growth of CNN based pneumonia detection research and emphasized the continuing challenges related to dataset quality, preprocessing variability, and inconsistent evaluation protocols [9].

One of the most influential studies in this field is CheXNet, proposed by Rajpurkar et al. [3], which employed a 121 layer DenseNet model on the ChestX ray14 dataset and reported strong performance for pneumonia detection. This work demonstrated the potential of deep learning for large scale chest radiograph analysis and stimulated extensive follow up research. At the same time, it also exposed a major challenge in medical imaging datasets, namely the reliance on labels automatically extracted from radiology reports, which may introduce noise and uncertainty into model development and evaluation. Such limitations are important because weakly labeled data can inflate performance estimates and reduce clinical interpretability.

Subsequent studies increasingly investigated transfer learning with established architectures such as VGG, ResNet, DenseNet, and Inception variants. Kermany et al. [6] showed that pretrained deep learning models can achieve strong performance in pediatric pneumonia classification from chest X ray images, highlighting the practical value of transfer learning when annotated medical datasets are limited. Similarly, Rahman et al. [10] demonstrated that transfer learning with deep convolutional neural networks can also provide strong performance for pneumonia detection from chest X ray images, further reinforcing the effectiveness of pretrained models in this domain. Stephen et al. [11] reported high classification accuracy using a CNN based approach on pediatric chest radiographs, further supporting the effectiveness of deep feature learning for pneumonia recognition. These studies collectively confirmed that CNN based methods can outperform conventional machine learning pipelines that rely on handcrafted features. However, many of them were evaluated on a single dataset and provided limited insight into cross dataset generalization or robustness under varying clinical conditions.

More recent work has shifted toward broader comparative evaluation and deployment oriented design. Sanida and Dasygenis [12] explored a lightweight CNN for chest X ray based lung disease identification on heterogeneous embedded systems, emphasizing the importance of reducing model complexity for practical implementation in resource limited settings. In addition to lightweight model design, application oriented studies have also explored practical deployment pathways, including web based clinical support solutions for pneumonia detection, as reported in [13]. Hashmi et al. [14] proposed a hybrid framework that combines deep learned features with handcrafted descriptors and classical classification strategies, illustrating that feature fusion may improve performance under certain experimental conditions. In parallel, multi disease classification frameworks have gained growing attention because they better reflect real

DOI:10.5281/zenodo.20088614

diagnostic scenarios in which pneumonia must be distinguished from several thoracic abnormalities rather than only from normal cases [15]. Explainability has also become an increasingly important direction, as demonstrated by Wang et al. [16], who incorporated Integrated Gradients to improve transparency and support clinician trust in CNN based pneumonia diagnosis. Comparative pediatric studies such as Choudhury [17] further showed that fine tuned pretrained architectures often outperform custom built CNNs, which confirms the continued relevance of transfer learning in this domain.

Recent studies published in 2024 to 2026 have extended this line of research in several important directions. Chandranegara et al. [18] proposed a ResNet RS based framework for pneumonia identification and incorporated Contrast Limited Adaptive Histogram Equalization together with data augmentation to improve image quality and expand the effective training distribution. Their results indicated that the combined use of preprocessing and augmentation can noticeably enhance model performance. Al Foysal and Sultana [19] presented a comparative CNN based study for automated pneumonia diagnosis using chest X ray images and applied augmentation techniques such as rotation, scaling, and flipping to improve generalization. Their findings support the value of comparative deep learning analysis, although the conclusions remain dependent on the specific dataset and experimental setting used. More recently, Saranyaraj et al. [20] introduced PneuNet, a lightweight architecture that combines depthwise separable convolutions, squeeze and excitation blocks, atrous spatial pyramid pooling, and a learnable pooling mechanism for multiscale feature fusion. This design reflects a clear shift toward computationally efficient models that still aim to preserve discriminative diagnostic capability.

Despite these advances, several important gaps remain in the literature. First, many studies still evaluate only a single architecture, making it difficult to determine whether observed gains arise from model design or from favorable experimental settings. Second, even when multiple models are compared, the studies often rely on heterogeneous preprocessing methods, augmentation strategies, image resolutions, dataset partitions, and evaluation metrics, which limits reproducibility and weakens the fairness of cross model comparison. Third, several studies continue to emphasize overall accuracy while offering limited analysis of clinically important behavior such as sensitivity, specificity, and class wise error distribution. Finally, although lightweight and deployment aware architectures are gaining attention, consistent comparative evaluation under unified conditions is still limited. To address these limitations, the present study adopts a controlled and unified evaluation framework for four widely used CNN architectures, namely ResNet, VGG, InceptionV3, and MobileNet. All models are trained and evaluated under identical preprocessing, training, and assessment conditions. This standardized design improves reproducibility, supports transparent architectural comparison, and enables clearer conclusions regarding the relative strengths and limitations of each model for AI assisted pneumonia screening from chest X ray images.

Table 1. Comparative summary of representative CNN based studies for pneumonia detection from chest X ray images

Ref	Study focus / approach	Main contribution	Main limitation	Relevance to the present study
[6]	Deep learning for pediatric pneumonia classification	Demonstrated that pretrained deep learning models can achieve strong performance in pediatric pneumonia related image classification	The study is not designed as a controlled comparison among several standard CNN architectures under identical settings	Shows the practical value of transfer learning in medical image analysis
[11]	CNN based pneumonia classification in healthcare	Reported strong classification performance on pediatric chest X ray images and supported the effectiveness of CNN based feature learning	Relies on a single dataset and offers limited evidence on generalization across settings	Supports the use of deep CNNs for pneumonia screening, but leaves fair cross model comparison unresolved
[12]	Lightweight CNN for chest X ray based lung disease identification on embedded systems	Emphasized deployment oriented design and computational practicality in resource constrained environments	Focuses on lightweight implementation, but does not provide a unified comparison against multiple standard architectures in the same protocol	Motivates inclusion of MobileNet type models when deployment efficiency matters
[14]	Hybrid deep learning and handcrafted feature fusion for pneumonia detection	Explored combining deep features with handcrafted descriptors and classical classifiers to improve performance	Hybrid designs may improve accuracy, but they complicate direct architectural comparison with standard end to end CNNs	Highlights an alternative design direction beyond standard pretrained CNN baselines
[16]	CNN based pneumonia diagnosis with Integrated Gradients	Added explainability to pneumonia classification, improving model transparency and interpretability	Explainability is addressed, but not within a unified multi architecture comparison framework	Important for clinical trust, though the current study focuses primarily on comparative performance
[17]	Comparative study of transfer learning and custom CNNs for pediatric pneumonia	Showed that fine tuned pretrained architectures often outperform custom CNN models	The comparison remains study specific and may still depend on its own dataset and protocol choices	Directly supports the importance of comparing pretrained architectures in a fair setting
[18]	ResNet RS with CLAHE preprocessing and data augmentation	Showed that contrast enhancement and augmentation can strengthen pneumonia identification performance	Improvement may depend heavily on the preprocessing pipeline and testing scenario used	Demonstrates the impact of preprocessing choices on reported performance

Ref	Study focus / approach	Main contribution	Main limitation	Relevance to the present study
[19]	Comparative CNN analysis for AI driven pneumonia diagnosis	Compared CNN based diagnosis on chest X ray images and used augmentation to improve generalization	Conclusions remain tied to the selected dataset and training conditions	Reinforces the need for careful, fair comparison across CNN models
[20]	PneuNet, lightweight CNN with multiscale feature fusion	Proposed an efficiency oriented architecture using depthwise separable convolutions, squeeze and excitation, ASPP, and learnable pooling	Represents a specialized architecture rather than a direct standardized baseline comparison with common pretrained models	Highlights the trend toward lightweight yet discriminative pneumonia detection models
This study	Unified comparison of ResNet, VGG, InceptionV3, and MobileNet	Evaluates four widely used CNN architectures under identical preprocessing, training, and evaluation settings	Limited to the selected dataset and binary classification setting	Provides a controlled baseline for fair architectural comparison

III. Methodology

This section describes the experimental methodology adopted for pneumonia detection from chest X ray images. It presents the dataset, preprocessing procedures, selected convolutional neural network architectures, model training strategy, and evaluation metrics. A unified experimental framework was used across all models to ensure a fair, consistent, and reliable comparison.

A. Dataset Description

The dataset used in this study was obtained from a publicly available Kaggle repository [21]. It contains chest X ray images labeled as either normal or pneumonia, where the pneumonia category includes both bacterial and viral infection cases. The dataset was organized into separate subsets for model development and performance assessment in order to support unbiased evaluation under consistent experimental conditions.

B. Data Preprocessing

Several preprocessing steps were applied to the chest X ray images before training. First, all images were resized to 180 × 180 pixels in order to ensure a uniform input format across the evaluated models. Next, pixel values were normalized to the range from 0 to 1 to improve numerical stability during training and to facilitate faster convergence of the optimization process [22].

Data augmentation was not applied in the reported experiments. The preprocessing pipeline focused on resizing and intensity normalization to ensure consistent input formatting across models. This choice allows the reported results to reflect architectural behavior under a controlled setting[23].

C. Convolutional Neural Network Architectures

This study evaluates four widely used convolutional neural network architectures for binary pneumonia classification, namely MobileNet, ResNet50, VGG, and InceptionV3. CNN based models are well suited to medical image analysis because they can automatically learn hierarchical spatial features from image data without relying on handcrafted feature engineering. This capability is enabled by the hierarchical interaction of convolution, activation, and pooling operations, which progressively transform low level visual patterns into more abstract and discriminative representations [24].

The selected models represent distinct architectural design principles. ResNet is based on residual learning, which improves gradient flow and enables effective training of deeper networks [25]. VGG employs a simple and uniform stack of convolutional layers that facilitates hierarchical feature extraction. InceptionV3 uses multi scale convolutional operations to capture patterns at different spatial resolutions. MobileNet is a lightweight architecture designed for efficient computation in resource constrained settings.

In this implementation, InceptionV3 was initialized with ImageNet pretrained weights within a transfer learning setting [26]. The remaining architectures were trained from scratch under the same optimization and evaluation protocol.

D. Model Training Configuration

All models were trained under identical experimental conditions to ensure a fair comparison. The Adam optimizer was used for parameter optimization, and binary cross entropy was adopted as the loss function for the two class classification task [27]. In addition, early stopping, batch normalization, and regularization techniques were applied to improve generalization and reduce overfitting. The same preprocessing pipeline, training protocol, and evaluation strategy were maintained across all architectures so that observed performance differences could be attributed more reliably to model design rather than to variations in the experimental setup.

E. Evaluation Metrics

Model performance was evaluated using accuracy, precision, recall, F1 score, and confusion matrix analysis [28]. Accuracy provides a general indication of overall classification performance, whereas precision and recall offer deeper insight into false positive and false negative behavior. In medical diagnosis, recall is especially important because false negative predictions may delay treatment and increase clinical risk. The F1 score provides a balanced measure of precision and recall, while confusion matrices allow detailed examination of class wise prediction behavior and diagnostic reliability.

IV. Results

This section presents the quantitative results of the evaluated convolutional neural network models for pneumonia detection from chest X ray images. Model performance was assessed using accuracy, precision, recall, F1 score, and confusion matrix analysis in order to examine both overall classification performance and class wise prediction behavior.

A. Experimental Setting and Evaluation Criteria

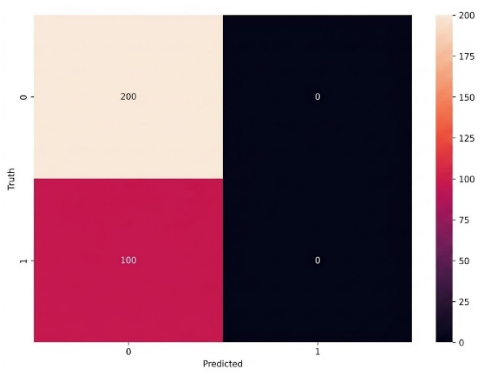
All evaluated models were trained and tested under identical experimental conditions to ensure a fair comparison. Prior to training, the input images were resized to 180 × 180 pixels and normalized. Model performance was then assessed using accuracy, precision, recall, F1 score, and confusion matrices. These metrics are particularly important in medical image classification, where false negative predictions may have serious clinical consequences.

B. Performance of Individual CNN Architectures

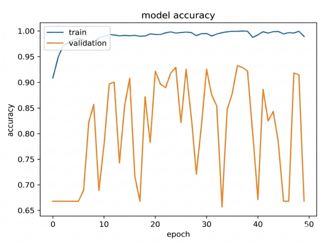
1) MobileNet

The MobileNet based model achieved strong classification performance under the adopted experimental protocol, with a test accuracy of 95.0 percent. Confusion matrix analysis indicates reliable sensitivity to pneumonia cases, where 198 out of 200 pneumonia images were correctly classified, and only 2 pneumonia cases were missed. For the normal class, 87 out of 100 images were correctly classified, while 13 normal images were incorrectly predicted as pneumonia. These results suggest that, within the adopted input resolution and training configuration, the lightweight MobileNet architecture was able to learn discriminative radiographic patterns for pneumonia detection while maintaining efficient inference behavior. The confusion matrix, accuracy curve, and loss curve are presented in Fig. 2.

(a)



(b) (c)



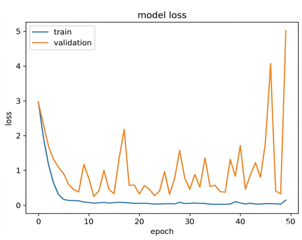
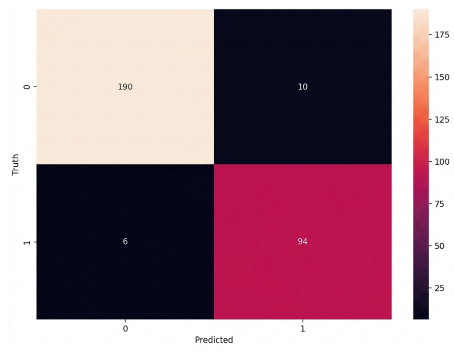


Fig. 2. MobileNet-Based Model performance (a) confusion matrix, (b) accuracy, and (c) loss.

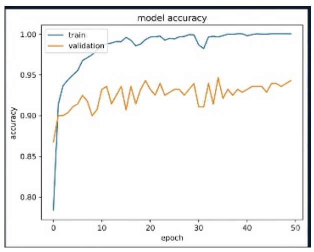
2) ResNet

The ResNet50 based model achieved strong classification performance on the test set, with an accuracy of 94.7 percent. Confusion matrix analysis indicates high sensitivity to pneumonia cases, where 196 out of 200 pneumonia images were correctly classified and 4 cases were missed. For the normal class, 88 out of 100 images were correctly classified, while 12 normal images were incorrectly predicted as pneumonia. These results suggest that residual learning supported effective feature representation for pneumonia related patterns while maintaining stable class separation under the adopted experimental protocol. Model performance is presented in Fig. 3.

(a)



(b) (c)



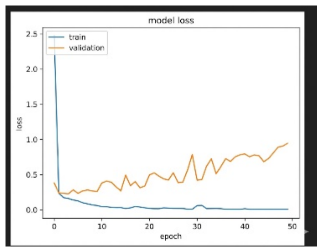


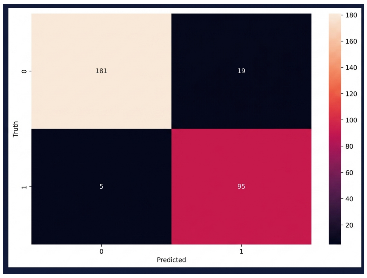
Fig. 3. ResNet-Based Model performance (a) confusion matrix, (b) accuracy, and (c) loss.

3) VGG

The VGG based model achieved strong classification performance on the test set, with an accuracy of 94.7 percent. Confusion matrix analysis indicates that 190 out of 200 pneumonia images were correctly classified, while 10 pneumonia cases were missed. For the normal class, 94 out of 100 images were correctly classified, and 6 normal images were incorrectly predicted as pneumonia. These results confirm that VGG provides reliable discrimination between pneumonia and normal cases under the adopted protocol, with a comparatively low false positive rate for normal images. However, the number of missed pneumonia cases was higher than that observed for MobileNet and ResNet50 in this evaluation, which suggests a different balance between sensitivity to pneumonia and specificity for normal images. Model performance is presented in Fig. 4.

DOI:10.5281/zenodo.20088614

(a)



(b) (c)

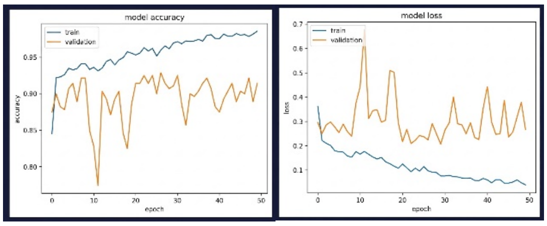
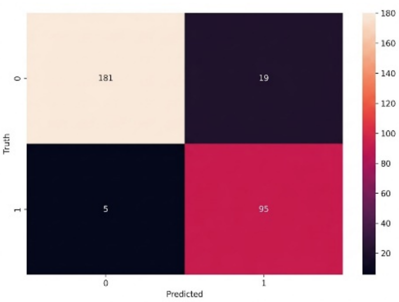


Fig. 4. VGG model performance: (a) confusion matrix, (b) accuracy, and (c) loss

4) InceptionV3

The InceptionV3 based model achieved an accuracy of 92.0 percent on the test set. Confusion matrix analysis shows that 181 out of 200 pneumonia images were correctly classified, while 19 pneumonia cases were missed. For the normal class, 95 out of 100 images were correctly classified, and 5 normal images were incorrectly predicted as pneumonia. Although the model maintained a relatively low false positive count for normal cases, it produced the highest number of missed pneumonia cases among the evaluated architectures in this setting. This indicates that InceptionV3 offered weaker sensitivity to pneumonia related patterns under the adopted training configuration when compared with MobileNet, ResNet50, and VGG. Model performance is presented in Fig. 5.

(a)



(b) (c)

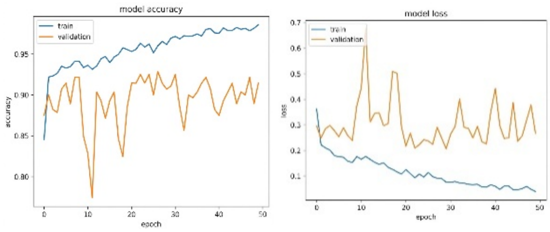


Fig. 5. InceptionV3 model performance: (a) confusion matrix, (b) accuracy, and (c) loss

C. Comparative Analysis

Table 2 summarizes the overall classification performance of the evaluated CNN architectures on the held out test set of 300 images, including 200 pneumonia cases and 100 normal cases. The results show that MobileNet achieved the highest accuracy at 95.0 percent, while ResNet50 and VGG achieved 94.7 percent, and InceptionV3 achieved 92.0 percent. Although the top three models have similar accuracy, their confusion matrices reveal different error patterns. MobileNet produced the fewest false negatives for pneumonia cases, which is important for screening oriented use, but it exhibited lower recall for normal cases compared with VGG and InceptionV3. InceptionV3 achieved the lowest accuracy and generated the highest number of missed pneumonia cases in this test setting. Fig. 6 provides a visual comparison of accuracy values across models.
To complement performance metrics, inference latency was measured to provide a deployment oriented view. Table 3 reports inference time per image using batch size 1 at an input resolution of 180 × 180 on CPU. MobileNet achieved the lowest latency, followed by InceptionV3 and ResNet50, while VGG incurred the highest computational cost. These results highlight the trade off between diagnostic performance and computational efficiency that should be considered when selecting models for practical deployment.

Table 2. Overall performance metrics of CNN models for pneumonia detection .

Model	Accuracy (%)	Precision (Macro) (%)	Recall (Macro) (%)	F1-score (Macro) (%)	Recall Pneumonia (%)	Recall Normal (%)
MobileNet	95.0	95.8	93.0	94.2	99.0	87.0
ResNet50	94.7	94.9	93.0	93.9	98.0	88.0
VGG	94.7	93.7	94.5	94.1	95.0	94.0
InceptionV3	92.0	90.3	92.8	91.3	90.5	95.0

Note: Macro averaged metrics are computed across the two classes. The test set contains 200 pneumonia cases and 100 normal cases.

Table 3. Inference time comparison across CNN models (batch size = 1)

Model	Input size	Inference time(ms)	Std (ms)	Hardware
MobileNet	180 × 180 × 3	34.60	9.74	CPU
InceptionV3	180 × 180 × 3	97.46	24.63	CPU
ResNet50	180 × 180 × 3	154.39	39.62	CPU
VGG	180 × 180 × 3	421.80	100.67	CPU

Note: Inference time is reported as mean and standard deviation per image after warm up, using batch size 1 on the same CPU environment and input resolution.

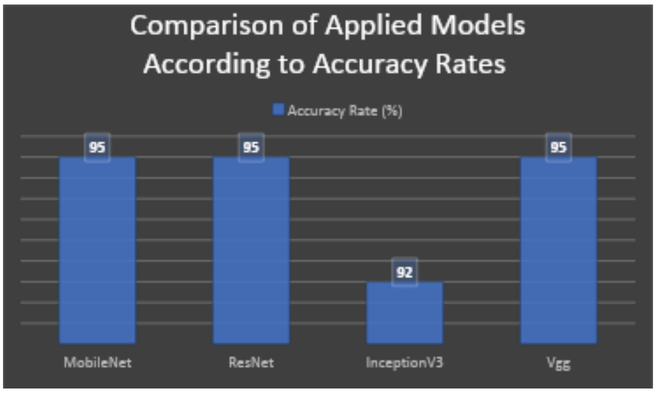


Fig. 6. Accuracy comparison of the evaluated CNN models.

To further inspect error distribution, Fig. 7 presents confusion matrices for the best performing and weakest performing models, highlighting missed pneumonia cases and false alarms on normal images .

(a) (b)

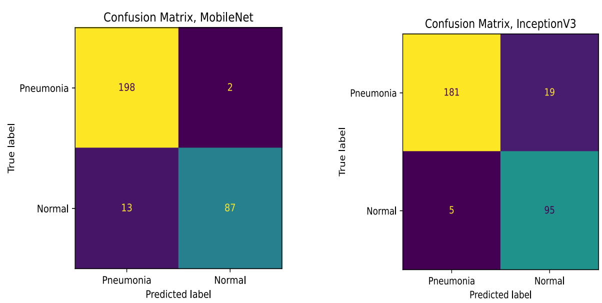


Fig. 7. Confusion matrices for the best performing and weakest performing models on the test set: (a) MobileNet, (b) InceptionV3.

V. Discussion

This study compared four convolutional neural network architectures for pneumonia detection from chest X ray images under a unified experimental protocol and a fixed test set of 300 images, including 200 pneumonia cases and 100 normal cases. The corrected results indicate that architectural ranking depends not only on overall accuracy but also on error distribution across classes, particularly false negatives in pneumonia cases, which are clinically more critical than false positives. In this setting, MobileNet achieved the highest accuracy and produced the fewest missed pneumonia cases, while InceptionV3 showed weaker performance and a higher number of missed pneumonia cases. ResNet50 and VGG achieved comparable overall accuracy and demonstrated strong diagnostic behavior, with differences primarily reflected in the balance between false positives and false negatives.

MobileNet achieved the strongest screening oriented behavior in this evaluation, as indicated by its confusion matrix and its low false negative count for pneumonia cases, with 2 missed cases out of 200. This suggests that, under the adopted input resolution and training configuration, the lightweight architecture learned discriminative cues for pneumonia while preserving an efficient design. From a deployment perspective, this advantage is supported by the inference time results, where MobileNet exhibited the lowest latency among the evaluated models using batch size 1 on CPU. These results support the suitability of MobileNet for time sensitive screening and triage in resource constrained environments, provided that external validation confirms stability across different imaging conditions and patient populations.

ResNet50 and VGG also demonstrated strong classification performance, achieving accuracy close to MobileNet, with reliable separation between pneumonia and normal cases. ResNet50 is commonly associated with stable optimization due to residual learning, which supports effective representation learning in deeper networks. VGG also produced competitive performance, and its error profile suggests a different trade off between sensitivity to pneumonia and specificity for normal cases. These findings indicate that deeper architectures remain strong choices when diagnostic reliability is the primary requirement, although their higher computational cost should be considered in deployment planning.

InceptionV3 achieved lower accuracy in the corrected evaluation and produced the highest number of missed pneumonia cases. While multi scale feature extraction can be beneficial in chest radiography, the present results suggest that this advantage did not translate into superior diagnostic reliability under the adopted protocol. This outcome highlights the importance of controlled benchmarking, since architectural strengths reported in other contexts may not generalize across datasets, preprocessing choices, and training settings.

A. Clinical Implications

From a clinical perspective, sensitivity to pneumonia cases is a key requirement for screening and triage, since false negative predictions may delay treatment and increase risk of

DOI:10.5281/zenodo.20088614

complications. The corrected confusion matrix analysis indicates that MobileNet achieved the lowest false negative count for pneumonia in this setting, which strengthens its potential value for screening oriented workflows. However, these systems should be used as decision support tools rather than replacements for clinical expertise. Real clinical decisions depend on patient history, laboratory findings, and physician judgment, and AI models are best positioned to assist prioritization and second look assessment.

B. Training and Validation Behavior

Training and validation curves indicate stable convergence across models under the adopted settings, as shown in Figs. 2 to 5. However, the corrected results emphasize that generalization should be assessed using class specific error patterns rather than accuracy alone. Confusion matrix analysis clarifies whether performance gains reflect genuine sensitivity to pneumonia patterns or reflect bias toward the majority class. This supports reporting class specific metrics alongside macro averaged measures in pneumonia detection studies.

C. Interpretation of Classification Performance

Reporting multiple metrics is essential for clinically meaningful evaluation. While accuracy provides an overall summary, precision, recall, F1 score, and confusion matrix inspection indicate how errors are distributed across classes. In this study, the corrected results show that models with similar accuracy may differ substantially in their pneumonia false negative behavior. Therefore, model selection should align with the intended clinical objective, with emphasis on minimizing false negatives for screening use cases.

D. Comparative Interpretation of the Evaluated Models

Overall, the corrected comparison supports two practical conclusions. First, lightweight architectures can provide strong diagnostic performance and favorable latency under certain training settings, as demonstrated by MobileNet in this evaluation. Second, deeper architectures remain competitive and may offer robust representation capacity, but their computational cost may be higher. These findings reinforce the value of unified benchmarking, since fair architectural assessment requires identical preprocessing, training, and evaluation conditions.

VI. Conclusion and Future Work

This study presented a controlled comparative evaluation of four convolutional neural network architectures, namely MobileNet, ResNet50, VGG, and InceptionV3, for binary pneumonia detection from chest X ray images. All models were trained and evaluated under a unified experimental protocol, enabling a fair comparison of architectural behavior and diagnostic performance. Evaluation was conducted on a held out test set of 300 images, including 200 pneumonia cases and 100 normal cases, using accuracy, macro averaged metrics, and confusion matrix analysis.

The results show that MobileNet achieved the highest test accuracy at 95.0 percent, while ResNet50 and VGG achieved 94.7 percent, and InceptionV3 achieved 92.0 percent. Although the top three models exhibited similar accuracy, confusion matrix analysis revealed meaningful differences in error distribution. MobileNet produced the fewest missed pneumonia cases, with 2 missed cases out of 200, which is important for screening oriented use where reducing false negatives is a priority. ResNet50 and VGG also delivered strong diagnostic performance, with 4 and 10 missed pneumonia cases, respectively, and differences mainly reflected in their balance between missed pneumonia cases and false alarms on normal images. InceptionV3 produced the weakest performance and the highest number of missed pneumonia cases in this evaluation, with 19 missed cases out of 200.

In addition to diagnostic quality, deployment feasibility requires attention to computational efficiency. Inference time measurements using batch size 1 at an input resolution of 180 × 180 on CPU show that MobileNet provides a clear efficiency advantage, with an average latency of 34.60 ms per image, while VGG incurs substantially higher computational cost at 421.80 ms per image. ResNet50 and InceptionV3 fall between these extremes at 154.39 ms and 97.46 ms per image, respectively. These findings highlight that architecture selection should consider both clinical risk, particularly missed pneumonia cases, and practical constraints such as inference latency in resource constrained environments.

This study has limitations that should be considered. The evaluation relied on a single public dataset and a specific input resolution, and performance may vary across institutions due to differences in imaging protocols, devices, and patient populations. In addition, the current task is binary classification and does not represent the full range of thoracic conditions encountered in clinical practice.

Future work will extend the evaluation in several directions. First, multi class classification can be explored to distinguish pneumonia subtypes and other thoracic abnormalities. Second, explainable AI methods such as Grad CAM and saliency based visualization can be integrated to improve transparency and support clinical interpretability. Third, external validation on larger and more diverse multi institutional datasets is essential to assess generalizability and reduce dataset specific bias. Finally, deployment oriented testing on representative hardware platforms and evaluation under realistic clinical workflow constraints can further clarify real world latency and reliability.

References

[1] World Health Organization, “Pneumonia,” WHO Fact Sheet, 2023. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/pneumonia. **Accessed on: Feb. 10, 2026.**

[2] R. Rudan, C. Boschi-Pinto, Z. Biloglav, K. Mulholland, and H.

Campbell, “Epidemiology and etiology of childhood pneumonia,” Bulletin of the World Health Organization, vol. 86, no. 5, pp. 408–416, 2008.

[3] A. Rajpurkar et al., “CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning,” arXiv preprint arXiv:1711.05225, 2017.

[4] G. Litjens et al., “A survey on deep learning in medical image analysis,” Medical Image Analysis, vol. 42, pp. 60–88, 2017.

[5] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” Nature, vol. 521, pp. 436–444, 2015.

[6] D. S. Kermany, M. Goldbaum, W. Cai, et al., “Identifying medical diagnoses and treatable diseases by image-based deep learning,” Cell, vol. 172, no. 5, pp. 1122–1131.e9, 2018.

[7] C. Gu and M. Lee, “Deep transfer learning using real-world image features for medical image classification, with a case study on pneumonia X-ray images,” Bioengineering, vol. 11, no. 4, Art. no. 406, 2024, doi: 10.3390/bioengineering11040406.

[8] I. D. Apostolopoulos and T. A. Mpesiana, “COVID-19: Automatic detection from X-ray images utilizing transfer learning with convolutional neural networks,” Physical and Engineering Sciences in Medicine, vol. 43, pp. 635–640, 2020.

[9] R. Siddiqi and S. Javaid, “Deep learning for pneumonia detection in chest X-ray images: A comprehensive survey,” Journal of Imaging, vol. 10, no. 8, Art. no. 176, 2024.

[10] T. Rahman, M. E. H. Chowdhury, A. Khandakar, K. R. Islam, K. F. Islam, Z. B. Mahbub, M. A. Kadir, and S. Kashem, “Transfer learning with deep convolutional neural network (CNN) for pneumonia detection using chest X-ray,” arXiv preprint arXiv:2004.06578, 2020.

[11] O. Stephen, M. Sain, U. J. Maduh, and D.-U. Jeong, “An efficient deep learning approach to pneumonia classification in healthcare,” Journal of Healthcare Engineering, vol. 2019, Art. no. 4180949, 2019.

[12] T. Sanida and M. Dasygenis, “A novel lightweight CNN for chest X-ray-based lung disease identification on heterogeneous embedded system,” Applied Intelligence, vol. 54, pp. 4756–4780, 2024.

[13] A. Y. Nageye et al., “Enhancing deep learning for pneumonia detection: Developing web based solution for Dr. Sumait Hospital in Mogadishu Somalia,” Discover Applied Sciences, vol. 7, Art. no. 309, 2025, doi: 10.1007/s42452-025-06735-6.

[14] M. F. Hashmi et al., “Hybrid deep learning and handcrafted feature fusion for pneumonia detection in chest X-rays,” Journal of Electrical Systems and Information Technology, vol. 12, Art. no. 94, 2025.

[15] P. Sahoo et al., “A multistage framework for respiratory disease detection and assessing severity in chest X-ray images,” Scientific Reports, vol. 14, Art. no. 12380, 2024, doi: 10.1038/s41598-024-60861-6.

[16] J. Rabbah, M. Ridouani, and L. Hassouni, “Improving pneumonia diagnosis with high-accuracy CNN-based chest X-ray image classification and integrated gradient,” Biomedical Signal Processing and Control, vol. 101, Art. no. 107239, 2025, doi: 10.1016/j.bspc.2024.107239.

[17] A. R. Choudhury, “Pediatric pneumonia detection from chest X-rays: A comparative study of transfer learning and custom CNNs,” arXiv preprint arXiv:2601.00837 [cs.CV], 2025, doi: 10.48550/arXiv.2601.00837.

[18] D. R. Chandranegara, V. D. Vitanti, W. Suharso, H. Wibowo, and S. Arifianto, “Analysis of pneumonia on chest X-ray images using convolutional neural network model iResNet-RS,” JOIV: International Journal on Informatics Visualization, vol. 8, no. 1, pp. 183–189, Mar. 2024, doi: 10.62527/joiv.8.1.1728.

[19] A. Al Foysal and S. Sultana, “AI-driven pneumonia diagnosis using deep learning: A comparative analysis of CNN models on chest X-ray images,” Open Access Library Journal, vol. 12, Art. no. e12899, 2025, doi: 10.4236/oalib.1112899.

[20] D. Saranyaraj, V. Shrinaath, A. Nayak, and R. Vishal, “PneuNet: A lightweight convolutional neural network with multiscale feature fusion for automated pneumonia detection from chest X-rays,” Frontiers in Medicine, vol. 12, Art. no. 1713587, 2026, doi: 10.3389/fmed.2025.1713587.

[21] Kaggle, “Chest X-Ray Images (Pneumonia) Dataset.” [Online]. Available: https://www.kaggle.com/datasets/paultimothymooney/chest-xray-pneumonia. Accessed on: Des. 22, 2025.

[22] M. Elhaddad and S. Hamam, “AI-driven clinical decision support systems: An ongoing pursuit of potential,” Cureus, vol. 16, no. 4, e57728, 2024, doi: 10.7759/cureus.57728.

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

[24] L. Perez and J. Wang, “The effectiveness of data augmentation in image classification using deep learning,” arXiv preprint arXiv:1712.04621, 2017.

[25] S. Albawi, T. A. Mohammed, and S. Al-Zawi, “Understanding of a convolutional neural network,” in 2017 International Conference on Engineering and Technology (ICET), Antalya, Turkey, Aug. 21–23, 2017, pp. 1–6, doi: 10.1109/ICEngTechnol.2017.8308186.

[26] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet classification with deep convolutional neural networks,” in Advances in Neural Information Processing Systems, 2012, pp. 1097–1105.

[27] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778.

[28] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” in International Conference on Learning Representations (ICLR), 2015.

[29] L. Sokolova and G. Lapalme, “A systematic analysis of performance measures for classification tasks,” Information Processing & Management, vol. 45, no. 4, pp. 427–437, 2009.

[30] W. Akbar, A. Soomro, A. Hussain, T. Hussain, F. Ali, M. Inam Ul Haq, R. W. Attar, A. Alhomoud, A. A. Alzubi, and R. Alsagri, “Pneumonia detection: A comprehensive study of diverse neural network architectures using chest X-rays,” International Journal of Applied Mathematics and Computer Science, vol. 34, no. 4, pp. 679–699, 2024, doi: 10.61822/amcs-2024-0045.

[31] C.-T. Yen and C.-Y. Tsao, “Lightweight convolutional neural network for chest X-ray images classification,” Scientific Reports, vol. 14, Art. no. 29759, 2024, doi: 10.1038/s41598-024-80826-z.

[32] J. Hofmeister, N. Garin, X. Montet, M. Scheffler, A. Platon, P.-A. Poletti, J. Stirnemann, M.-P. Debray, Y.-E. Claessens, X. Duval, and V. Prendki, “Validating the accuracy of deep learning for the diagnosis of pneumonia on chest X-ray against a robust multimodal reference diagnosis: A post hoc analysis of two prospective studies,” European Radiology Experimental, vol. 8, Art. no. 20, 2024, doi: 10.1186/s41747-023-00416-y.