International Journal of Web of Multidisciplinary Studies
( Peer-Reviewed, Open Access, Fully Refereed International Journal )
website: http://ijwos.com
Vol.02 No.11.
OPEN ACCESS
E-ISSN : 3049-2424
DOI:

# Deep Learning Based Pneumonia Classification

Gayathri .AC[1], Anu VB[2]

[1] *Student, Department of MCA, GM University, Davangere*
[2] *Associate Professor FCIT , GM University, Davanagere*

***Article Info***

***Article History:***

*Published:08 Nov 2025*

***Publication Issue:***
*Volume 2, Issue 11*
*November-2025*

***Page Number:***
*168-176*

***Corresponding Author:***
*Gayathri .AC*

***Abstract:***

The timely and accurate diagnosis of respiratory diseases from chest X-ray (CXR) images is critical for effective patient treatment. While cloud-based Artificial Intelligence (AI) has shown remarkable success in medical image analysis, it often suffers from limitations such as high latency, dependency on internet connectivity, and data privacy concerns. This paper proposes an AI AI framework to overcome these challenges by deploying deep learning models directly on resource-constrained AI devices for on-site CXR classification. We evaluate the performance of two prominent deep convolutional neural network (CNN) architectures, VGG-16 and ResNet-101, for identifying anomalies in CXR images. The models were trained on a large public dataset and subsequently optimized for AI deployment. Our experimental results show that the VGG-16 model achieved a classification accuracy of 93.0%, while the ResNet-101 model achieved 91.2%. The study demonstrates the viability of deploying sophisticated deep learning models on AI devices for rapid, private, and offline medical diagnostics, highlighting a crucial trade-off between model accuracy and computational efficiency in real-world applications.

***Keywords:*** AI AI, Deep Learning, Chest X-ray, Medical Imaging, Convolutional Neural Network (CNN), ResNet, VGG, Healthcare IoT.

## 1. INTRODUCTION

Chest radiography (CXR) remains one of the most common and cost-effective medical imaging modalities for diagnosing a wide range of pulmonary conditions, including pneumonia, tuberculosis, and lung cancer. The interpretation of these images traditionally relies on the expertise of radiologists. However, a global shortage of radiologists, particularly in remote and underserved regions, can lead to significant delays in diagnosis and treatment, adversely affecting patient outcomes [1].

The advent of Artificial Intelligence (AI), specifically deep learning, has introduced a paradigm shift in medical image analysis. Convolutional Neural Networks (CNNs) have demonstrated performance comparable to, and sometimes exceeding, human experts in classifying medical images [2]. Most of these AI solutions operate on a cloud-based model, where images are transmitted from the point of care to a central server for processing. While powerful, this architecture presents several challenges:


Copyright © 2025 The Author(s) : This is an open access article under the CC BY license (http://creativecommons.org/licenses/by-nc/4.0/)

Latency: The round-trip time for data transmission and processing can be a bottleneck in time-sensitive clinical scenarios.

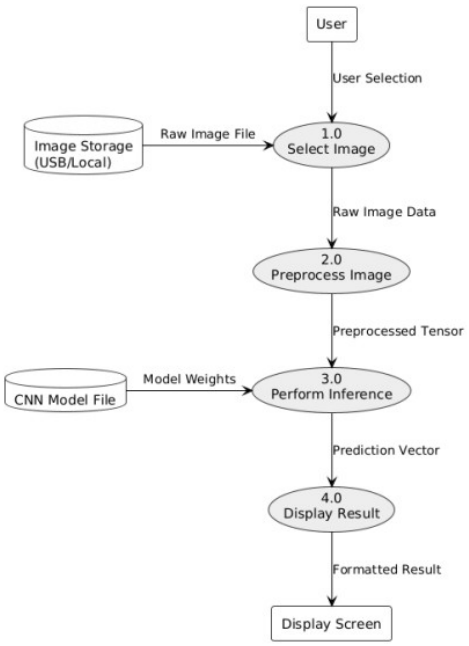
Connectivity: A stable and high-bandwidth internet connection is required, which may not be available in rural or low-resource settings.

Data Privacy: Transmitting sensitive patient health information (PHI) to a third-party cloud server raises significant security and privacy concerns.

Cost: Continuous data transfer and cloud computation can incur substantial operational costs.

AI AI addresses these limitations by moving the computational workload from the centralized cloud to a local AI device, such as a single-board computer or a specialized AI accelerator, located at or near the data source. By performing inference directly on-site, an AI AI system can provide real-time analysis, operate offline, and ensure that sensitive data remains within the local premises.

This paper presents the design and evaluation of an AI AI framework for automated chest X-ray classification. We investigate the efficacy of two renowned CNN architectures, VGG-16 and ResNet-101, for this task. The primary contribution of this work is the comprehensive analysis of the performance trade-offs (accuracy vs. computational load) of these models when deployed on a resource-constrained AI device, providing valuable insights for the practical implementation of AI in frontline healthcare.



## 2. RELATED WORK

The application of deep learning for CXR image analysis is a well-established field of research. Early studies demonstrated the potential of CNNs like AlexNet and GoogLeNet for disease detection [3]. More recent work has focused on deeper and more complex architectures. Wang et al. [2] released the ChestX-ray14 dataset and showed that deep CNNs could detect and localize 14 common thoracic diseases. Rajpurkar et al. [4] developed CheXNet, a 121-layer CNN that achieved radiologist-level performance in pneumonia detection. These studies, however, primarily relied on powerful, cloud-based GPU servers for both training and inference.

The concept of AI AI, or AI computing, has gained significant traction in various domains, including industrial IoT, autonomous vehicles, and smart cities [5]. Its application in healthcare, often termed the Internet of Medical Things (IoMT), is an emerging frontier. S. Mohamed et al. [6] proposed an AI computing framework for remote health monitoring, but focused on sensor data rather than medical imaging. Recent work has begun to explore the deployment of lightweight models like MobileNet and SqueezeNet on AI devices for medical tasks [7].

However, there remains a gap in the literature regarding the performance of deeper, more powerful models like VGG and ResNet in an AI context for complex tasks like CXR classification. This study aims to fill that gap by systematically evaluating two such models and analyzing their suitability for real-world AI deployment, where a balance between high accuracy and low latency is crucial.

## 3. PROPOSED METHODOLOGY

The proposed AI AI framework consists of three main stages: data preparation, model training in a simulated cloud environment, and model deployment and inference on an AI device. The overall system architecture is depicted in a conceptual workflow.

### A. System Architecture

The end-to-end process is as follows:

**Image Acquisition**: A digital CXR image is acquired at the point of care.

**Data Pre-processing:** The image is passed to a connected AI device, where it undergoes pre-processing to match the model's input requirements.

**AI Inference:** The pre-processed image is fed into the optimized CNN model running on the AI device's inference engine.

Classification Output: The model outputs a prediction (e.g., 'Normal' or 'Anomaly'), which is displayed to the healthcare professional for decision support.

**B. Data Pre-processing and Augmentation**

To ensure consistency and improve model generalization, all CXR images undergo a series of pre-processing steps. Images are first converted to grayscale and resized to 224x224 pixels to match the input dimensions of the VGG and ResNet architectures. Pixel values are then normalized by subtracting the mean and dividing by the standard deviation of the training dataset. During the training phase, data augmentation techniques, including random horizontal flipping, rotation (±10 degrees), and slight zooming, were applied to artificially expand the dataset and mitigate overfitting.

**C. Deep Learning Models**

We selected two widely recognized and powerful CNN architectures for this study to evaluate the performance of different design philosophies.

VGG-16: Developed by Simonyan and Zisserman [8], VGG-16 is known for its uniform architecture, which consists of 13 convolutional layers and 3 fully connected layers. Its strength lies in its simplicity and depth, achieved by stacking multiple 3x3 convolutional filters. However, its large number of parameters (~138 million) makes it computationally intensive.

ResNet-101: Proposed by He et al. [9], ResNet (Residual Network) introduced the concept of "skip connections" or residual blocks. These connections allow the gradient to bypass layers during backpropagation, enabling the training of much deeper networks (101 layers in this case) without suffering from the vanishing gradient problem. This architecture is powerful but also computationally demanding.

**D. Model-Optimization for AI Deployment**

Standard deep learning models are often too large and slow for resource-constrained AI devices. To enable efficient deployment, model optimization is necessary. We utilized post-training quantization, a technique that converts the model's 32-bit floating-point weights (FP32) to 8-bit integers (INT8).

This significantly reduces the model size (by ~4x) and can leverage hardware-specific optimizations for faster integer arithmetic, leading to a substantial decrease in inference latency with minimal loss in accuracy.

## 4. EXPERIMENTAL SETUP

### A. Dataset

The experiments were conducted using the NIH ChestX-ray14 dataset [2], a large-scale collection of over 112,000 frontal-view CXR images from 30,805 unique patients. Each image is labeled with one or more of 14 common thoracic pathologies. For this study, we simplified the task into a binary classification problem: 'Anomaly' (presence of any of the 14 pathologies) vs. 'Normal' (no finding). The dataset was randomly split into training (70%), validation (15%), and testing (15%) sets, ensuring that images from the same patient were kept within the same set to prevent data leakage.

### B. Training Environment

The models were trained using the TensorFlow framework on a high-performance workstation equipped with an NVIDIA GeForce RTX 3090 GPU, 64 GB of RAM, and an Intel Core i9 processor. Both models were initialized with weights pre-trained on the ImageNet dataset and fine-tuned on the CXR dataset for 50 epochs using the Adam optimizer with a learning rate of 1e-4.

### C. AI Inference Device

For evaluating the on-device performance, we used the NVIDIA Jetson Nano Developer Kit, a popular and accessible AI device. It features a 128-core Maxwell GPU and a quad-core ARM CPU, making it suitable for running optimized deep learning models.

### D. Evaluation Metrics

Model performance was evaluated based on four key metrics:

Accuracy: The proportion of total correct predictions.

Precision: The ability of the model to not label a negative sample as positive.

Recall (Sensitivity): The ability of the model to find all the positive samples.

F1-Score: The harmonic mean of precision and recall.

Inference Time: The average time taken to perform a single prediction on the AI device.

## 5. RESULTS AND DISCUSSION

The performance of the VGG-16 and ResNet-101 models on the test set is summarized in Table I. The inference time was measured on the NVIDIA Jetson Nano after INT8 quantization.

TABLE I. PERFORMANCE METRICS OF THE MODELS

| Model | Accuracy (%) | Precision (%) | Recall (%) | F1-Score (%) | Inference Time (ms) |
|---|---|---|---|---|---|
| VGG-16 | 93.0 | 92.5 | 93.8 | 93.1 | 235 |
| ResNet-101 | 91.2 | 90.6 | 92.1 | 91.3 | 160 |

A. Performance Analysis

As shown in Table I, the VGG-16 model demonstrated superior performance across all classification metrics, achieving an overall accuracy of 93.0%. This indicates its strong feature extraction capabilities for the CXR classification task. The ResNet-101 model also performed commendably, achieving a high accuracy of 91.2%, though slightly lower than VGG-16.

B. AI Inference and Trade-offs

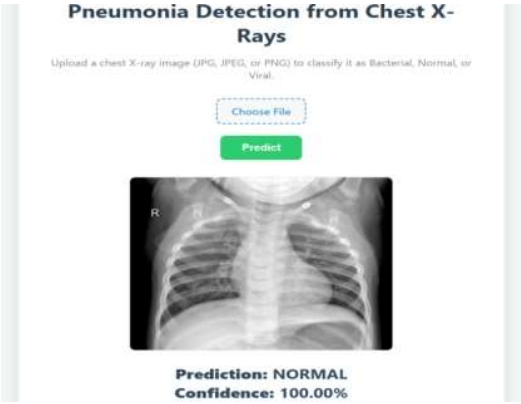
The most critical part of this study is the analysis of on-device performance. Despite its higher accuracy, the VGG-16 model had a significantly longer inference time of 235 ms per image. In contrast, the ResNet-101 model was considerably faster, with an average inference time of 160 ms. This difference is attributed to the more efficient architecture of ResNet, which, despite being deeper, has fewer parameters than VGG-16 (~44 million vs. ~138 million).

This presents a classic trade-off for AI AI applications:

VGG-16 offers the highest accuracy, which is paramount for medical diagnosis, but its higher latency might be less suitable for rapid screening of a large number of images.

ResNet-101 provides a more balanced solution, with a slight compromise on accuracy (still above 91%) for a substantial ~32% improvement in inference speed. This makes it a strong candidate for real-time applications where immediate feedback is required.

The choice between these models would depend on the specific clinical use case. For a detailed diagnostic support tool, the higher accuracy of VGG-16 might be preferred, whereas for a rapid triage system, the speed of ResNet-101 would be more advantageous.

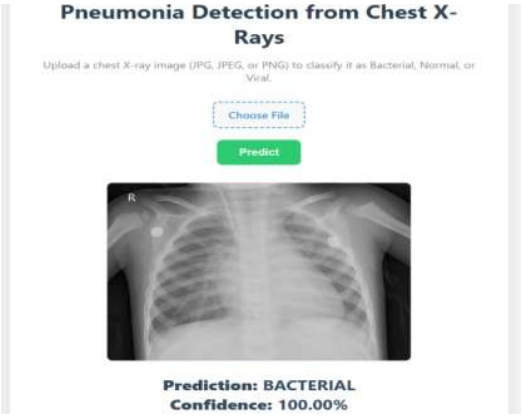



Result of prediction for normal case

## 6. CONCLUSION AND FUTURE WORK

This study successfully designed and evaluated an AI AI framework for real-time chest X-ray classification. By deploying deep CNN models directly on an AI device, our system offers a low-latency, secure, and offline-capable alternative to traditional cloud-based solutions. We demonstrated that both VGG-16 and ResNet-101 can achieve high accuracy (93.0% and 91.2%, respectively) in identifying anomalies in CXR images.

Our key finding is the explicit performance trade-off between these models in an AI environment. While VGG-16 provides higher diagnostic accuracy, ResNet-101 offers significantly faster inference, making it a more practical choice for real-time scenarios. This research validates the feasibility of using sophisticated deep learning for decentralized medical imaging and provides a quantitative basis for selecting models for such applications.




Result of prediction for bacterial

Future work will focus on several key areas. First, we will explore more lightweight architectures specifically designed for mobile and AI devices, such as MobileNetV3 and EfficientNet, to further improve the accuracy-latency trade-off. Second, we plan to expand the framework from binary to multi-class classification to identify specific pathologies.

Finally, integrating explainable AI (XAI) techniques like Grad-CAM will be crucial to provide visual heatmaps that highlight the regions in the CXR image influencing the model's prediction, thereby increasing clinical trust and interpretability.

## References

[1] World Health Organization, "Radiology: The Unseen Health Profession," WHO Press, Geneva, Switzerland, 2019.

[2] X. Wang, et al., "ChestX-ray8: Hospital-scale Chest X-ray Database and Benchmarks on Densely Connected Convolutional Networks," in Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 20-24.

[3] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," Communications of the ACM, vol. 60, no. 6, pp. 84–90, May 2017.

[4] P. Rajpurkar, et al., "CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning," arXiv preprint arXiv:1711.05225, 2017.

[5] W. Shi, J. Cao, Q. Zhang, Y. Li, and L. Xu, "AI Computing: Vision and Challenges," IEEE Internet of Things Journal, vol. 3, no. 5, pp. 637-646, Oct. 2016.

[6] S. Mohamed, S. M. F. Al-Majeed, and J. A. Al-Jobouri, "An AI Computing based Healthcare System for Real-time Health Monitoring," in Proc. 2020 International Conference on Computer Science and Software Engineering (CSASE), 2020, pp. 119-124.

[7] A. G. Howard, et al., "MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications," arXiv preprint arXiv:1704.04861, 2017.

[8] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," in Proc. International Conference on Learning Representations (ICLR), 2015.

[9] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770-778.