International Journal of Engineering Trends and Technology
ISSN: 2231-5381/ https://doi.org/10.14445/22315381/IJETT-V74I3P102

Volume 74 Issue 3, 11-24, March 2026

Original Article

A Deep Transfer Learning Model for Robust Pneumonia Detection from Medical Imaging

Adel Rajab

Department of Computer Science, College of Computer Science and Information Systems, Najran University, Najran, Saudi Arabia.

Corresponding Author : adrajab@nu.edu.sa

Received: 09 December 2025 Revised: 24 January 2026 Accepted: 27 January 2026 Published: 28 March 2026

Abstract *- Pneumonia is a serious infection of the lungs brought about by several types of bacteria and viruses. It is usually difficult to detect and treat with chest X-rays because its visual patterns often overlap those of other pulmonary diseases. In recent years, deep learning has shown considerable success in medical image processing, such as the fully automated detection and classification of diseases with high accuracy. These models can identify complex patterns from large datasets, hence finding their perfect applications in radiology. This work proposes a deep learning-based pneumonia detection approach using transfer learning models. Images were obtained from an updated, publicly available version of the Paul Pulmonary Chest X-ray dataset. To extract meaningful features, a pretrained DenseNet121 network was utilized. Various transfer learning architectures, such as VGG16, ResNet50, InceptionV3, and DenseNet121, were trained and tested in this work. All architectures employed a unified MLP classification head to refine the extracted features and generate the final prediction. Model performance metrics include accuracy, precision, recall, F1 score, confusion matrix, AUC, and loss curves. Among them, DenseNet121 produced the highest accuracy of 89% and yielded an AUC of 0.96. The findings have shown that deep learning models, especially DenseNet121, can effectively detect pneumonia from chest X-rays, hence providing a very important tool for the radiologist and healthcare professional to improve both the speed and accuracy of diagnosis.*

Keywords *- Pneumonia, Chest X-ray, Transfer Learning Models, Feature Extractions.*

1. Introduction

Pneumonia is a serious infection of the lungs that can develop when bacteria, viruses, or fungi invade the respiratory system. Reports from the World Health Organization indicate the disease claims close to four million lives each year and remains a major cause of mortality, mainly among young children and older adults [1]. Depending on the responsible organism, pneumonia may be classified as viral, bacterial, or fungal, and primarily affects the alveoli, small air sacs responsible for oxygen exchange in the lungs. The illness becomes particularly dangerous to people who already have chronic problems breathing, weakened immunity, or who are very young or elderly. Late diagnosis can easily become deadly for ventilated patients or those with various other underlying health issues [2]. Chest X-rays remain the most common method of imaging in diagnosing pneumonia because they are low-cost and widely available. However, these images are not always interpreted easily. The signs of pneumonia can closely resemble other lung issues, such as fluid accumulation, inflammation, or masses, which makes manual interpretation prone to delays and errors, even by experienced radiologists. Classic diagnostic methods might therefore lack uniformity and precision [3]. Advances in Artificial Intelligence (AI), particularly deep learning, have revolutionized the analysis of medical images. CNNs can automatically learn relevant visual patterns from raw chest X-ray images themselves without requiring explicit feature engineering. Transfer learning facilitates high-performance pneumonia detectors even when small annotated medical datasets are available since it reuses models already trained on large datasets. This has recently been a more popular strategy in medical imaging because it enhances accuracy and accelerates decision-making procedures [4, 5]. These models offer many advantages to medical image processing. For instance, it drastically reduces the need for large, annotated datasets to collect in healthcare environments. However, the key discriminative features are already learned by pre-trained models, which requires considerably less computational effort and training time. Transfer learning can further improve the accuracy and stability of the model, specifically on small or imbalanced datasets, by reducing overfitting and making the model concentrate on clinically important features. It follows that the practice of transfer learning has found broad applications in radiology for the development of reliable diagnostic tools [6, 7]. Unlike previous works, which focus on the individual deep learning models one by one, this paper

introduces a comprehensive framework incorporating transfer learning-based feature extraction, selection, and MLP-based classification. A comparison across the four most widely used pre-trained models, such as VGG16, ResNet50, InceptionV3, and DenseNet121, is also performed to ensure a fair comparison across all models. A critical comparison of their performance would provide more insightful information regarding the drawbacks and strengths of the models regarding the pneumonia diagnosis application, which remains to be addressed by previous research.

While transfer learning provides many benefits, it has also presented its own challenges. Large, pretrained models are typically trained on natural image databases such as ImageNet, which possess very dissimilar structure and texture compared to medical images. This mismatch can affect knowledge transfer. Additionally, fine-tuning deep models requires parameters to be selected carefully to avoid overfitting, especially when working with small datasets. Performance can also be compromised if medical datasets are imbalanced, making it more difficult for models to detect conditions represented in lower proportions than others, such as pneumonia. These problems highlight the need for careful model evaluation and optimization when applying to medical imaging [8, 9].

Past research has shown that the implementation of the concept of transfer learning has a significant effect on improving the accuracy of the detection of pneumonia as compared to training CNN models for the purpose of accurate detection. However, the observed outcomes are not entirely uniform. The proposed study presents a deep learning framework integrating feature extraction, feature selection, and several well-known CNN models, namely VGG16, ResNet50, InceptionV3, and DenseNet121, together with multilayer perceptron (MLP) classifiers for the automatic identification of pneumonia.

This research is significant because it addresses major challenges associated with automated pneumonia diagnosis by providing a reliable comparative deep learning approach for helping radiologists make informed decisions using chest X-ray images. Early recognition and accurate diagnosis of pneumonia can reduce the mortality rate and improve the quality of patient care in hospitals or clinics where expert radiologists are not available or in limited-resource areas. Furthermore, this research contributes to the development of clinical decision-support tools that can improve diagnostic accuracy while lessening the burden on radiology teams.

Pneumonia detection from chest X-rays using automated systems effectively necessitates a careful assessment of the performance of different deep learning models in terms of accuracy, robustness, and reliability. While transfer learning has been useful in medical image analysis, it still remains a systematic evaluation that needs to be carried out on multiple pretrained architectures to check which models perform better in real-world clinical applications. The study was designed following two Research Questions (RQs):

- RQ-1: Which DL transfer-learning model performs best for the classification of pneumonia using chest X-ray images?
- RQ-2: How do different models be compared in terms of various metrics such as accuracy, precision, recall, f1 score, and AUC for pneumonia detection?

The originality of this study lies in its integrated approach, as it involves the fusion of transfer learning for feature extraction, feature selection, and MLP-based classification across multiple pretrained models. In contrast to most previous studies that commonly explore only a single architecture, this work systematically investigates the performance of four popular CNN models under the same experimental conditions, thus granting a more accurate idea of their different strengths and weaknesses in pneumonia detection. Moreover, based on a balanced dataset and using extensive evaluation metrics, such as ROC-AUC, confusion matrices, and loss analysis, this study conducts a much more robust and comprehensive investigation of model performance than most of the previous studies. The key contributions of this work are:

1. Improving dataset quality by balancing training and validation using a curated pneumonia X-ray dataset.
2. Selecting the most important image features to boost prediction accuracy.
3. Comparing several pre-trained deep-learning models to identify the best-performing architecture.
4. Evaluating the models using accuracy, ROC-AUC, and confusion matrices to demonstrate their potential in real-world clinical decision support.

The rest of the paper is organized as follows: Section 2 describes the research gap identified from existing work. Section 3 presents the step-by-step working pipeline. Sections 4 and 5 interpret the key findings and their discussions. Finally, Section 6 shows the research conclusion and future work direction.

2. Related Work

Chest X-ray imaging remains one of the most widespread approaches to the diagnosis of disorders in the chest cavity. The objective of the [10] was to assist healthcare professionals in selecting practical, real-time diagnostic methods based on the analysis of existing datasets and the interpretation of the results presented in the literature, to provide an overview of the usability and scale of currently available open-access chest X-ray datasets, and to discuss the main strategies adopted so far to overcome the class imbalance problem. Current studies constantly confirm that deep learning methods exhibit the best

results for pneumonia detection, with some models reaching an accuracy of approximately 98.7%, a sensitivity of 0.99, and a specificity of 0.98. Early diagnosis of pneumonia is essential to avoid further complications due to delay, which also casts a considerable burden of mortality worldwide. While both machine learning and deep learning approaches are applicable to pneumonia prediction, the latter has become the preferred choice because it eliminates the need for manual feature engineering while frequently yielding more dependable results. This study [11] comprehensively discusses a large number of recent deep learning-based pneumonia detection methods, focusing on convolutional neural networks, transfer learning models, and ensemble-based methods. It synthesizes findings related to the choice of hyperparameters, performance metrics, and fine-tuning strategies, and demonstrates the higher robustness of ensemble-based detection systems compared to single models.

With the fast development of deep learning, pneumonia detection by automata from chest radiographs has gone through impressive improvements. Recent models (VGG, ResNet, and ViT) and methodologies for mitigating the negative effects of highly imbalanced datasets. The study utilizes multiple publicly referenced datasets, such as the Pneumonia Chest X-Ray set, BRAX, and CheXpert. Transfer learning using weights pre-trained on ImageNet is studied to improve model generalization. Experimental results demonstrate that techniques such as transfer learning, image augmentation, and class-balanced training significantly improve model reliability in skewed data distributions [12].

Pneumonia, along with a few other chest diseases like cardiomegaly and atelectasis, remains difficult to diagnose in environments where radiology expertise is limited. One study introduced a simplified VGG-style network that had significantly fewer parameters. To deal with the low contrast of many X-ray images, Dynamic Histogram Enhancement was applied during the preprocessing stage. The final model had drastically fewer parameters compared to VGG-16, ResNet-50, Xception, and DenseNet121, but still achieved strong results of 96.07% accuracy with an AUC of 0.99107 [13].

Considering the fact that manual assessment of chest radiographs usually faces inconsistencies, another study developed a deep learning system incorporating spatial attention, recurrent modules for learning temporal patterns, and biologically inspired spiking-based processing to boost noise resistance. Testing showed the proposed architecture yielded 99.35% accuracy with high precision, recall, and F1-score consistently. These attested facts support the idea that such models might act as reliable diagnostic assistants, mainly when the healthcare context is below optimal standards [14]. Pneumonia remains one of the significant burdens among infectious diseases worldwide; thus, its rapid and accurate diagnosis is paramount. One of the attempts proposed for a publicly available Kaggle dataset relies on a Vision Transformer-based architecture that is targeted at extracting global contextual relationships from radiographic images using self-attention mechanisms. This model reached 97.61% accuracy, while also achieving 95% sensitivity and 98% specificity, outperforming various CNN-based methods and proving the advantage of transformer-based processing for complex medical imagery [15].

Another comparative study compared the performance of VGG16, ResNet, InceptionNet, DenseNet, and a custom CNN for pneumonia classification. Based on the results obtained using mean absolute error as the evaluation metric, VGG16 had the best performance. To improve the computational power, the researchers uses TPU-based distributed training provided by TensorFlow and reduced the training time for their CNN by over half when compared to GPU-based training and by over two-thirds when compared to CPU training a different line of research presented an ensemble framework, known as PELM (Pneumonia Ensemble Learning Model), which combines feature outputs from InceptionV3, VGG16, ResNet50, and ViT. A dataset of 50,000 X-ray images from multiple well-known repositories, ensuring balanced representation of pneumonia and non-pneumonia cases. PELM achieved 96% accuracy, 99% precision, 95% F1-score, and an AUC of 0.91, outperforming its individual components and showing impressive generalization [16].

A hybrid architecture combining CNN layers with modified Swin Transformer blocks was investigated in another related work. The CNN components capture localized patterns, while the transformer units model long-range dependencies through a mechanism of window-based attention. Following preprocessing of images using CLAHE and data augmentation via flips, rotations, and zooming, the hybrid model was trained on data from Guangzhou Women and Children's Medical Center, resulting in 98.72% accuracy and a loss of 0.064 on unseen test data. This substantially outperforms a standard CNN baseline [17].

By focusing on the interpretability of AI-driven medical diagnoses, this research tries to solve this important problem. Its main objective is therefore to unpack the decision-making processes of complex artificial intelligence systems for pneumonia detection. This shall be achieved by deploying explainable AI models strategically, with the focus being on LIME. The reason LIME will be considered is because of its ability to approximate the behavior of complex models in a manner that is interpretable around a given instance, hence giving clear insights into their predictions. The research study involves a critical comparison and evaluation of a number of AI models based on their interpretability as far as diagnostic accuracy is concerned. In doing so, the study hopes to bridge the gap between the need to have healthcare professionals understand and trust the decision-making processes of AI models and the potent predictive capabilities of these models.

The output is expected to improve pneumonia detection techniques, ensuring that AI models perform accurately and produce interpretable and reliable insights [18].

The diagnosis of pneumonia from the observation of images of the lungs by different medical professionals using the X-ray technique might result in different diagnoses. In an effort to facilitate the diagnosis of pneumonia, in this research work, the approach of SE-MobileNet has been proposed. The performance of the proposed approach, that is, the proposed SE-MobileNet approach, has been compared with the default version of the MobileNetV2 approach, along with the concept of transfer learning. Using the publicly available Kaggle dataset, it has been noted that the proposed approach, that is, the proposed SE-MobileNet approach, provided a 97.4% level of accuracy for a chosen set of test images against the 96.4% level of accuracy provided by the default version of the MobileNetV2 approach.

Moreover, using 10-fold cross-validation, the proposed approach provided an average level of accuracy of 95.92% against the 92.35% level of accuracy provided by the default version of the MobileNetV2 approach. Moreover, to enhance the performance of the study, the robustness test has been incorporated, where the FGSM method is used to produce the adversarial images. Finally, in order to justify the relevance of the model's training, the techniques proposed in Explainable AI (XAI) have been used [19]. Pneumonia was responsible for respiratory failure in COVID-19 patients. Symptoms of COVID-19 are very diverse and robust, ranging from asymptomatic to severe respiratory failure. Present detection techniques of COVID-19 are time-consuming, less precise, less accurate, and very costly. To address such issues, a model for COVID-19 and Pneumonia detection using multi-deep learning algorithms, followed by a deployment model, was proposed.

Transfer learning architectures such as VGG-19, ResNet-50, Inception V3, and Xception, on two different datasets of CT scan and X-ray images (COVID/Non-COVID), to determine which of them are best-suited, were evaluated. The findings show an efficient range of percentage accuracies between 86% and 99% for COVID-19 detection using the proposed model, depending on the model and dataset. Moreover, a Flask app was designed to deploy the proposed approach and display the results for detected COVID and non-COVID images. The result of this study would be significant to develop an AI-assisted automated tool to make cost-effective and fast detection and effective management possible for COVID-19 patients [20].

3. Research Methodology

A deep learning model for pneumonia detection is proposed in this work by using chest X-ray images. The proposed study comprises steps such as data collection, preprocessing steps, feature extraction, and model development and evaluation. The step-by-step methodology is shown in Figure 1.

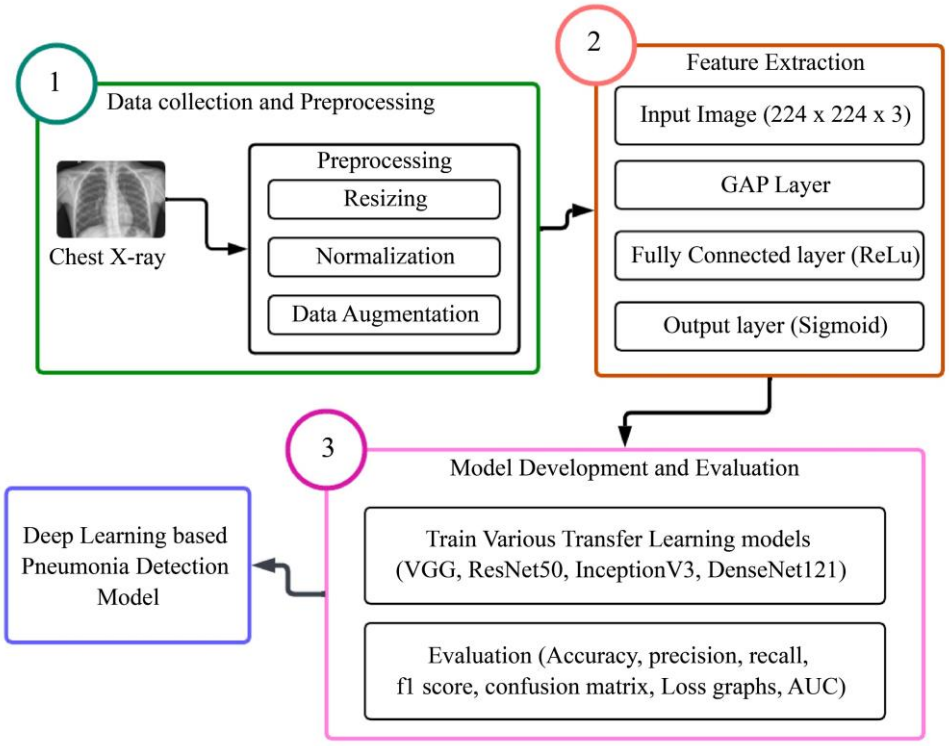


Fig.1 Research methodology

3.1. Data Collection

To detect CXR, the updated version [21] of Paul Mooney's Chest X-Ray Images (Pneumonia) dataset was selected and reorganized so that the training and validation sets are more evenly balanced for machine-learning work. The dataset consists of 5,856 X-ray images in total. Table 1 gives an overview of the dataset used in this work. It includes two classes: Normal and Lung Opacity. The training set is made up of 1,082 normal images and 3,110 lung opacity images, while the validation set contains 267 normal and 773 lung opacity images, with the testing set having 234 normal and 390 lung opacity images. Throughout all subsets, the lung opacity cases are more numerous, reflecting a clear class imbalance within the dataset.

Undersampling was intentionally not used because sample removal might result in the loss of some clinically critical lung opacity cases. Oversampling methods, including random duplication and SMOTE, were not implemented since the inclusion of unrealistic or artificial patterns in medical images might introduce problems to model reliability.

In this respect, we rearranged the dataset based on stratified splitting using the updated version of the dataset that has been proposed in [21]. This approach enhanced the balance between the training and validation sets while preserving the original data distribution. In addition, class weighting was considered during training, with a higher loss weight assigned to the minority class, Normal, to reduce the model's bias toward the majority class. In comparison with resampling-based methods, this strategy does not alter the authenticity of the medical data, while it effectively addresses class imbalance during model learning.

Table 1. Dataset description

	Normal cases	Lung opacity cases
Training	1082	3110
Validation	267	773
Testing	234	390

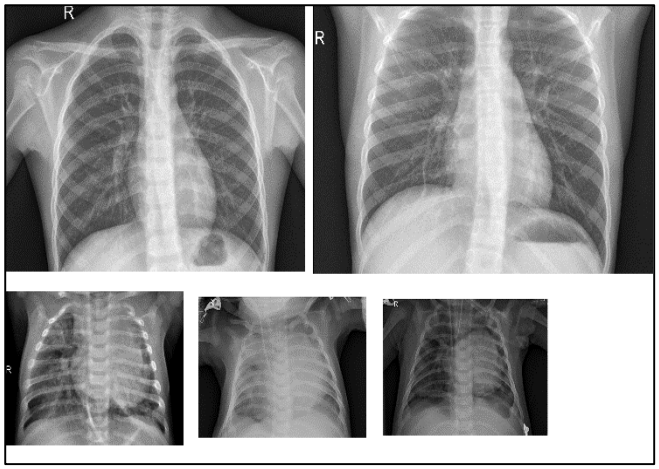


Fig. 2 Samples of the dataset

Figure 2 depicts some samples from the dataset. The first row is showing normal cases, while the second row is showing pneumonia cases.

3.2. Preprocessing and Analysis

We applied the following image processing techniques:

3.2.1. Image Resizing and Normalization

All images were resized to 224 × 224 pixels so that they would be the same size as those expected by common ImageNet-trained CNN models such as VGG16, ResNet50, InceptionV3, and DenseNet121. Using a consistent image size keeps the data compatible with these pretrained networks, helps lower the amount of computation needed, and still maintains the important visual details needed for accurate analysis [22].

3.3. Data Augmentation

The ImageDataGenerator with several enhancement steps was used to avoid overfitting and enhance the model's ability to generalize. This includes rotations up to 20°, horizontal and vertical shifts of 20%, horizontal flips, zoom level changes, and changes in brightness. Chest X-ray datasets tend to have limited variability because many images are taken under similar circumstances. By augmenting this data, such differences can be simulated that may appear in real clinical settings, thereby increasing the robustness of the model against changes in patient positioning or lighting conditions [23].

3.4. Feature Extraction using Pretrained models

Feature extraction is the process of converting the raw data, in this study, chest X-ray images, into a meaningful and compact representation that emphasizes the most useful visual patterns for accurate predictions. A DenseNet121, pre-trained on ImageNet, was used to extract useful features and robust capability of capturing rich and layered features from large collections of natural images.

It froze the convolutional layers so that the learned filters would not change, while the input chest X-ray images (224 × 224 × 3) were passed through it to generate deep, high-level feature maps. A Global Average Pooling (GAP) on these maps for distilling each channel into a single scalar, yielding a fixed-length feature vector, was utilized, which was then fed into a fully connected layer with ReLU activation, followed by dropout ($p = 0.5$) against overfitting.

Finally, a sigmoid output node provided the predicted probability of pneumonia. The aim of selecting DenseNet121 architecture is due to its dense connections; DenseNet121 can be relatively deep without requiring an excessive number of parameters. This aids in maintaining high representational capacity while minimizing the risk of threatened gradients.

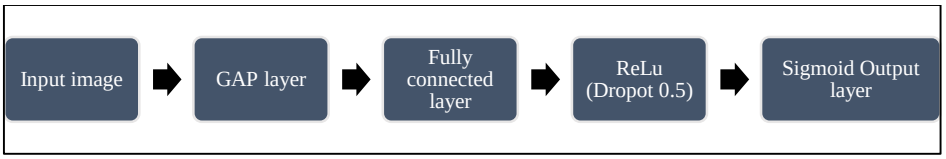


Fig. 3 Feature extraction block

Figure 3 represents the feature extraction block.

Algorithm-1: Deep Feature Extraction

Input: Chest X-ray Image $\boldsymbol{I} \in \boldsymbol{R^{224\,x\,224\,x3}}$

Output: Predicted pneumonia Probability $\boldsymbol{Y'} \in [\mathbf{0}, \mathbf{1}]$

Start:

1. Preprocessing
   - Image Resizing:
     $$\boldsymbol{I'} = \boldsymbol{Resiz}(\boldsymbol{I}, \mathbf{224x224x3})$$
   - Normalization to pixel intensity:
     $$\boldsymbol{I_n} = \frac{\boldsymbol{I'}}{\mathbf{255}}$$
2. DenseNet121 Feature Extraction (Frozen Backbone)
   - Load pre-trained DenseNet121 on ImageNet and freeze all conventional layers:
     $$\boldsymbol{\theta_{DenseNet}}\ \boldsymbol{Frozen}$$
   - Forward-propagate the image to extract feature maps:
     $$\boldsymbol{F} = \boldsymbol{I_{DenseNet}}\ (\boldsymbol{I_n}) \quad : \boldsymbol{F} \in \boldsymbol{R^{H\,x\,W\,x\,C}}$$
3. Global Average Pooling
   - Convert feature maps into a fixed-size feature vector:
     For each channel $\boldsymbol{C}$:
     $$\mathbf{g_C} = \frac{\mathbf{1}}{\mathbf{HW}} \sum_{\mathbf{i=1}}^{\mathbf{H}} \sum_{\mathbf{j=1}}^{\mathbf{W}} \mathbf{F_{i,j,C}}$$
   - Resulting Features
     $$\mathbf{g} = [\mathbf{g_1}, \mathbf{g_2}, \ldots\ldots, \mathbf{g_C}]$$
4. Fully connected Layers
   - Apply a dense layer with Relu activation:
     $$\mathbf{h} = \mathbf{Relu}\left(\mathbf{W_{1g}} + \mathbf{b_1}\right)$$
   - Apply dropout to reduce overfitting:
     $$\boldsymbol{h'} = \boldsymbol{Dropout}\ (\boldsymbol{h},\ \boldsymbol{p} = \mathbf{0.5})$$
5. Output Layer:
   - Generate pneumonia probability using a sigmoid neuron:
     $$\mathbf{Y'} = \boldsymbol{\sigma}(\mathbf{W_2 h'} + \mathbf{b_2}) \quad : \boldsymbol{\sigma}(\boldsymbol{z}) = \frac{\mathbf{1}}{\mathbf{1}+\boldsymbol{e}^{-\boldsymbol{z}}}$$

Algorithm 1 shows the stepwise working process of feature extraction from chest X-ray images. The algorithm accepts chest X-ray images as input. Feature aggregation via GAP and final classification through a lightweight MLP.

The final output layer used a sigmoid as an activation function. Thus, the algorithm provides an efficient and robust deep feature extraction mechanism centered on DenseNet121, concluding in a compact and discriminative prediction head for pneumonia detection.

3.5. Model Development and Training Parameters

This research is based on the following transfer learning architecture to detect accurate pneumonia.

3.5.1. VGG16 Model

VGG16 is a CNN, which is constructed by stacking 3×3 convolutional layers upon each other. This architecture allows it to be simple and consistent, thus reliable for extracting meaningful features from medical images; this model provides

a very strong baseline for many tasks, such as detecting pneumonia [24].

ResNet50

ResNet50 introduces skip connections around some of its layers that permit information to bypass certain layers. This architecture prevents problems such as vanishing gradients, which enables the network to learn deeper and more complex patterns. This is very helpful for finding small details in X-ray images [25].

InceptionV3

InceptionV3 combines several convolutional filters of varying sizes in one layer. The multi-scale nature allows the model to identify both small and large features within chest X-rays, enhancing its ability to analyze complicated patterns in medical images [26].

DenseNet121

DenseNet121 connects each layer to all previous layers; thus, every layer has the opportunity to access all previously calculated features. This dense connectivity promotes feature reuse, improves gradient flow, and makes the network efficient, especially in tasks involving detailed radiographic images [27].

The rationale for the selection of these model is their simple and widely used CNN architectures: VGG16, ResNet50, InceptionV3, and DenseNet121, because of their good performance in previous medical imaging tasks and the availability of pretraining weights. Although transformer-based models and ensemble approaches are the latest developments, the study focused on CNNs that are more computationally efficient, reproducible, and suitable for clinical deployment, providing a solid baseline for potential future research that might wish to incorporate such a more advanced architecture.

To maintain consistent classification across all pretrained models, a lightweight MLP head was integrated with every backbone. After preprocessing, frozen conventional layers were passed to extract feature maps into the GAP layer.

The dense, fully connected layer consists of 512 ReLU units with a dropout rate of 0.5. For the final output layer, the sigmoid neuron is used to produce a probability. This classifier head is kept identical across all four pretrained architectures.

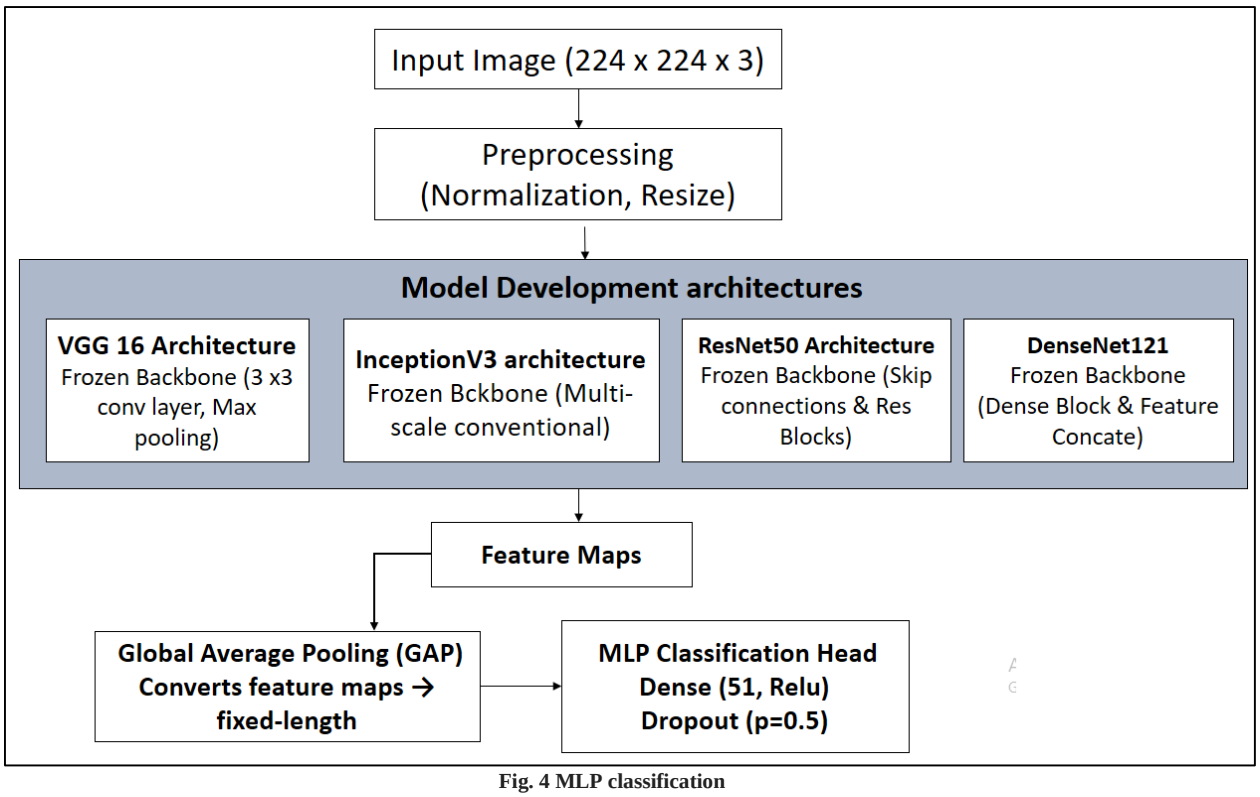


Fig. 4 MLP classification

Figure 4 demonstrates the various model development architecture parameters, and finally, the MLP classification head with a dense unit of 512, ReLU activation. The framework starts with a 224×224×3 chest X-ray image and performs a simple preprocessing on it, which includes resizing of the image and normalizing its pixel values. Subsequently, the image is fed into four off-the-shelf convolutional models: VGG16, InceptionV3, ResNet50, and DenseNet121. The

models are kept frozen and used only for feature extraction purposes. Each network produces its own set of deep feature maps, capturing different spatial patterns and visual cues from the X-ray. Once these feature maps are obtained, they are reduced to fixed-size vectors with Global Average Pooling.

All models' pooled features go through a shared MLP classifier. It consists of a Dense layer with ReLU activation and a dropout rate of 0.5 and finishes with a single sigmoid unit for the prediction of pneumonia likelihood.

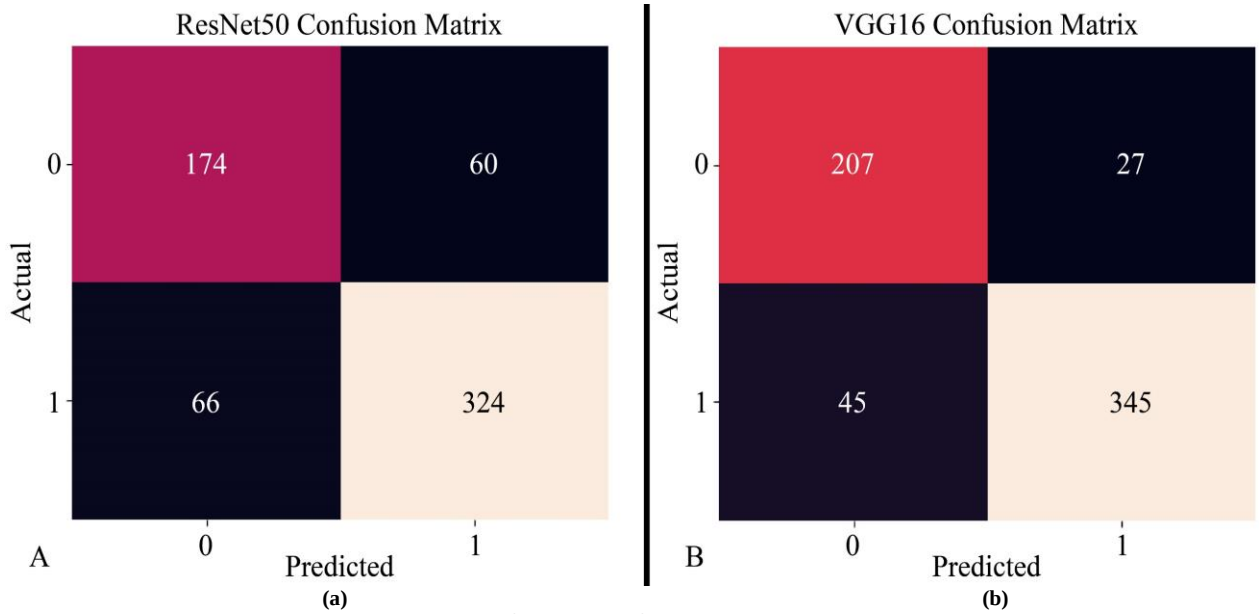
4. Results and Discussion

The models were trained on Windows 10 with a Kaggle notebook using the Python programming language. Hyperparameters such as learning rate, batch size, number of epochs, and optimizer used were determined and settled upon through some initial testing and validation on a validation set of data. A split of 70% of the data for training, 20% for testing, and 10% used for validation.

This attention to detail ensures that images of a given patient do not both train and test. A batch size of 32 and 30 epochs with a binary loss function was used. An Adam optimizer and a Relu activation function are used. The performance analysis of transfer learning models is evaluated as depicted in Table 2. The highest performance across all metrics is accuracy (89%), precision (86%), recall (80%), and F1 score (82%), which is obtained by DenseNet121, signifying a strong balance between sensitivity and specificity. The VGG16 obtained 88% accuracy and 86% precision, but with a slightly lower recall than DenseNet121. InceptionV3 has the lowest accuracy with 83%, suggesting it was less capable of extracting relevant features from this dataset. While other performances were satisfactory enough, ResNet50 recorded the lowest levels of precision and F1 score at 79%, which is indicative of a higher rate of false positives and hence less reliable classification. Overall, this shows that the architecture of DenseNet121 is more appropriate for capturing robust features and generalizing well for this task.

Table 2. Performance evaluation of DL models

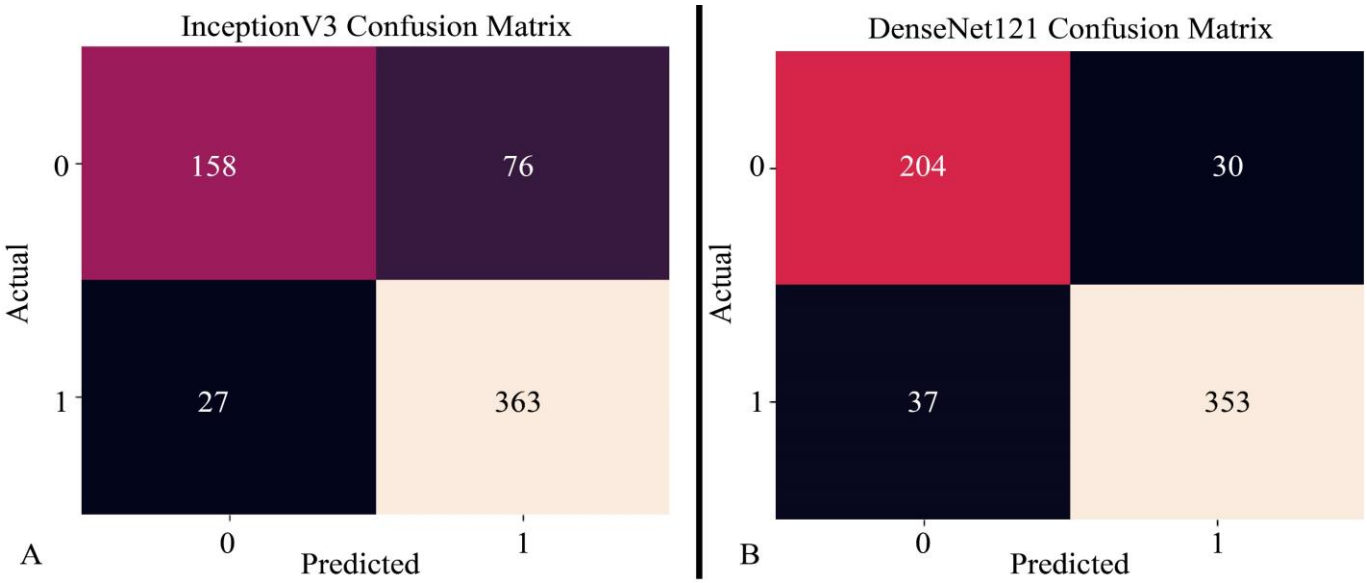
Models	Accuracy	Precision	Recall	F1 score
VGG	88%	86%	87%	86%
ResNet50	80%	79%	80%	79%
DenseNet	89%	86%	80%	82%
InceptionV3	83%	82%	81%	81%



(a) (b)

Fig. 5 Confusion matrix for VGG and ResNet50

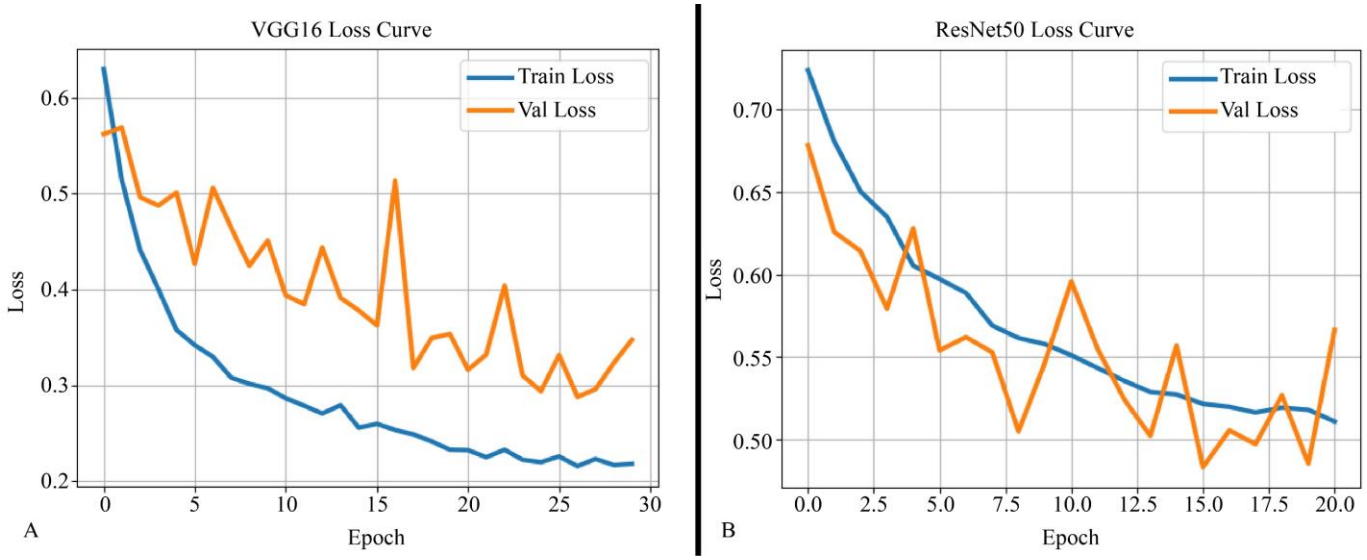
Figure 5 shows the confusion matrix for the ResNet50 and VGG16 models over unseen test data samples. In Figure 5 (a), the ResNet50 analysis is shown, where the model correctly classifies 498 samples out of a total of 624 samples, while 126 samples were misclassified. Figure 5 (b) depicts the confusion matrix for the VGG16 model, where 552 samples were correctly predicted out of 624 samples with 72 predictions. The magnitude of this error suggests that although ResNet50 learns valuable patterns, its decision boundaries do not generalize to all pneumonia and non-pneumonia cases in the test set. By contrast, VGG16 has a lower error rate, indicating it handles the dataset more effectively since it extracts features that better match the visual traits found in chest X-ray images. Overall, the comparison illustrates that VGG16 is more consistent and reliable when it comes to correctly identifying pneumonia in the evaluation set.



(a) **(b)**

Fig. 6 Confusion matrix for densenet121 and inception

Figure 6 shows the confusion matrix for InceptionV3 and DenseNet121 models over unseen test data samples. In Figure 6 (a), the InceptionV3 analysis is shown, where the model correctly classifies 521 samples out of a total of 624 samples, while 103 samples were misclassified. Figure 6 (b) shows the confusion matrix for the DeseNet121 model, where the model impressively classified 557 samples out of a total of 624; only 67 samples were misclassified.



(a) **(b)**

Fig. 7 Loss graph for VGG16 and ResNet50

Figure 7 represents the training and validation losses of the VGG16 and ResNet50 models for 30 epochs. The blue curve shows the training loss, while the orange curve shows the validation loss. Figure 7 (a) illustrates the loss of the VGG16. During initial epochs, the training loss drops while the validation loss overlaps at a loss value of 0.4 and fluctuates. Figure 7 (b) shows the loss curves of the ResNet50 model, where the training loss rapidly decreases in a straight fashion while the validation loss increases/decreases as the number of epochs increases. The trends in the two training curves speak volumes about just how dissimilar these architectures are. VGG16's strictly stacked layers tend to learn much more slowly and are more prone to overfitting, particularly when the dataset is small. This helps to explore the unpredictable behavior in its validation loss. In contrast, ResNet50 employed residual ways that keep gradients stable as the network deepens.

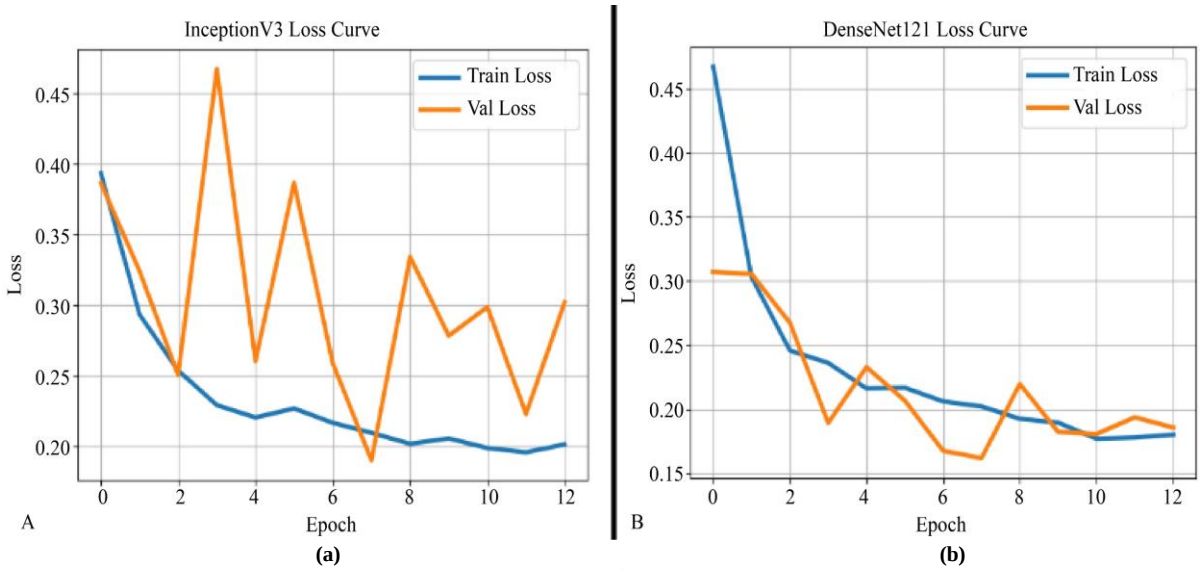


Fig. 8 Loss graph for DL models

Figure 8 represents the training and validation losses of the InceptionV3 and DesNet121 models for 30 epochs. The blue curve shows the training loss, while the orange curve shows the validation loss.

Figure 8 (a) depicts the loss curves of the InceptionV3 model, the training loss starts dropping from a loss value of 0.38 and moving towards a loss value of 0 as the number of epochs increases, while validation loss drops during initial epochs; however, reaching epoch 2, it suddenly increases and then drops. Figure 8 (b) shows the loss curves of the DenseNet121 model, where the training loss rapidly decreases in a straight fashion, while the validation loss starts dropping from a loss value of 0.32 and fluctuates continuously. The DenseNet121 model shows much higher fluctuation in the validation performance during training compared with the InceptionV3 model.

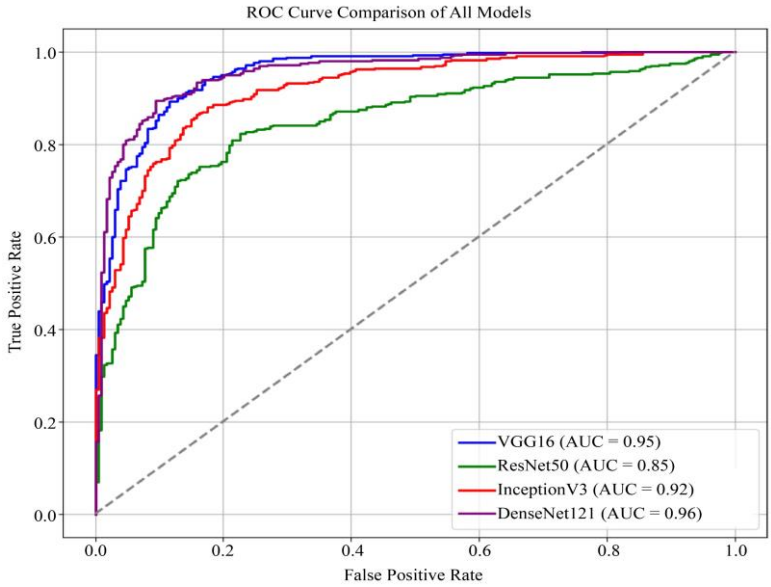


Fig. 9 ROC curve analysis

Figure 9 illustrates the ROC analysis of deep learning models: the best result in terms of AUC was by DenseNet121 with an AUC of 0.96, depicting a very strong ability in class discrimination. VGG16 and InceptionV3, though slightly lower than DenseNet121, also attained good performance with AUCs of 0.95 and 0.92, respectively. ResNet50 recorded the lowest AUC of 0.85, depicting less reliability in class classification. This therefore suggests that the architecture of DenseNet121 provides better feature extraction and generalization compared to the rest.

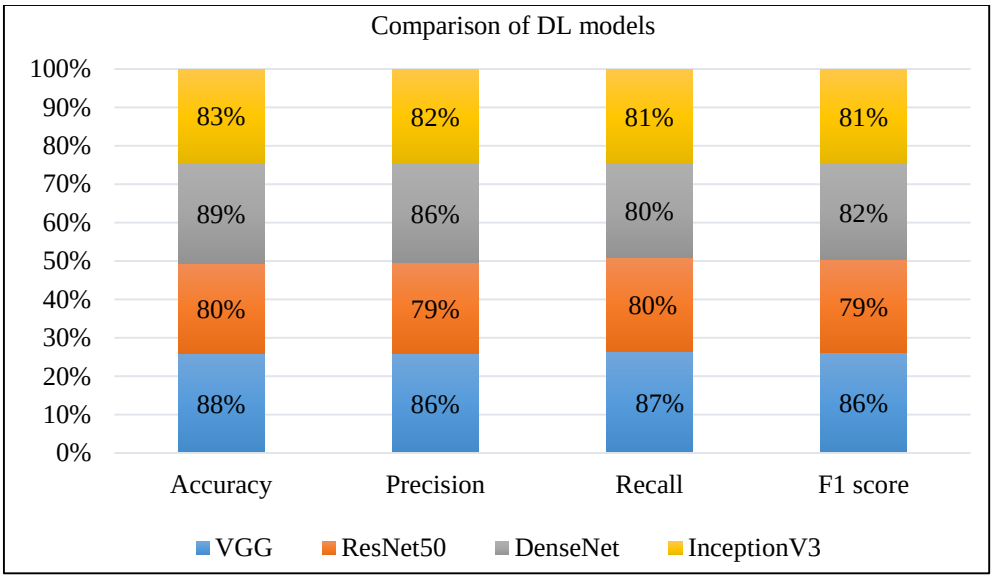


Fig. 10 Comparative analysis

Figure 10 presents a comparison of the performance of the models evaluated on aspects such as accuracy, precision, recall, and F1 score. DenseNet121 had the highest overall accuracy with 89%, followed by VGG16 with 88%, whereas ResNet50 had the lowest, with 80%, meaning a weaker predictive performance. On the other hand, DenseNet121 and VGG16 have recorded the highest precision with 86%, indicating that their predictions of the positive class are more reliable. This shows that the architecture of DenseNet121 offers better feature representation and generalization than ResNet50, which is more prone to misclassifications when dealing with class imbalance.

5. Discussion

Pneumonia is one of the most significant burdens to health worldwide. As such, early recognition and prompt treatment of this disease are essential in the prevention of complications and the improvement of patient outcomes [28]. Recent advances in DL have translated into automated tools that can help doctors to diagnose cases more rapidly and with higher accuracy [29]. A pneumonia detection system based on a deep learning approach using four transfer-learning models was proposed. This architecture is pre-trained on capturing detailed and discriminative features from the medical images. Among all the tested models, the best overall accuracy was 89% by DenseNet121, while the weakest was by ResNet50, which achieved an accuracy of 80%. Additionally, the DenseNet121 and ResNet50 achieved the highest precision rate (86%). The DenseNet121 model attained the impressive AUC score of 0.96, while the lowest AUC is 0.85, which is obtained by ResNet50. In addition to evaluating AUC, we also analyzed sensitivity and specificity in terms of confusion matrices to further grasp clinical relevance. The sensitivity of DenseNet121 is high, indicating its strength in truly predicting pneumonia cases. The specificity also shows its strength in predicting non-pneumonia cases. Though not in its calibration curve form, metrics such as these, along with the results of ROC-AUC, provide a reliable assessment of its performance.

Although formal statistical significance testing was not done, noticeable differences across the architectures in performance, especially in AUC, precision, and sensitivity, hint at significant variation in the behavior of the models. Further studies should include statistical analyses, such as paired t-tests or McNemar's test, to determine whether these performance differences are statistically significant.

The study demonstrates that among models, the DenseNet121 performs best for pneumonia detection due to its dense connectivity and higher feature representation. Across all metrics, DenseNet121 had the strongest performance, whereas ResNet50 showed the lowest reliability. It is noted that these tests were performed on a single split of the training and testing data without using k-fold cross-validation. Cross-validation should be used in future tests.

5.1. State-of-the-Art Comparison

In comparison to other recent state-of-the-art approaches, the suggested framework has proved to be better performing in terms of accuracy, ROC-AUC, and specificity and sensitivity. This may be attributed to the combination of feature selection techniques and MLP classifiers, together with balancing and transfer learning approaches from CNNs, which improve the discriminative capability and relevance to the task, respectively. The findings and results from this work have proved to be significant and insightful not only to researchers involved in this area but also to medical professionals regarding model performance and successful approaches.

5.2. Error / Bias Analysis

Analysis for errors indicated misclassification errors were found in such cases where there is an overlap between the radiographic findings and patients in early-stage pneumonia. The sample size limitations in the dataset did not allow for the analysis of stratified variables such as age, sex, and type of imaging device. Sensitivities and specificities were determined from the confusion matrices. Even without calibration curves, the obtained ROC-AUC values ensured that performance is trustworthy.

5.3. Explainability / Interpretability

This study did not incorporate any explainability techniques, such as Grad-CAM; however, future work should incorporate the Grad-CAM technique to enable radiologists to recognize regions on chest X-ray images that contribute to model predictions. This would lead to greater confidence in the outcomes.

5.4. External / Prospective Validation

This particular work utilized only one publicly accessible chest X-ray dataset. The validation in external datasets or multi-institutional datasets, as well as cross-validation, was also not done due to the unavailability of a labeled dataset.

5.5. Clinical Impact, Ethics, and Regulation

The study did not involve the incorporation of healthcare workflows, regulation, or ethics. Future studies should concentrate on the acquisition of ethics approval, ensuring privacy to patients, model validation in real-world scenarios, and following the regulations of the deployment of AI.

5.6. Scalability / Resource Analysis Response

The existing system is founded on CNN architectures that have been pretrained, which are more efficient than the original method of building the model. Nonetheless, the efficiency of resource use, computational cost, or scalability for implementation within different hospitals for the proposed systems is yet to be determined. The feasibility of implementing the systems for the analysis of the image should be considered.

5.7. Practical Implication

The proposed model can be useful in the real world to facilitate clinicians by emphasizing the diagnosis of pneumonia earlier and with great confidence, which in turn helps to improve the treatment methods and patient outcomes. Moreover, the workload of radiologists can be reduced with these advanced automated detection techniques.

5.8. Limitations

The study relies on only four transfer-learning models and is a single architecture; DenseNet121 was used to extract features. Although there was consistency in the model's performance, it would be valuable to confirm the model's external validity and robustness using external datasets and multi-institution trials. Detection accuracy can be improved by using various DL architectures for feature extraction. Accuracy and generalizability might be improved with multiple architectures and k-fold cross-validation and external data sets. In this study, advanced methods such as federated learning, self-supervised learning, and aspects regarding clinical integration and ethical considerations, and future studies regarding AI-assisted analysis have not been explored. It is pertinent to investigate such questions in future studies to make AI-based systems for pneumonia detection clinically applicable and ethically compliant.

While this research predominantly underscores model performance, on closer consideration, error analysis, subgroup analysis, and calibration analysis have not been considered for inclusion in this research work. The reason for this is that these components play a significantly important part in developing more comprehensive ideas related to model accuracy and applicability in practical medical terms. In subsequent research work, the error analysis on a more prominent scale would be considered. Applying the techniques of transfer learning from pre-trained models for natural images may not fully capture the textures/structure observed within the domain of interest for medical images. Future studies should explore domain adaptation methods to boost the relevance of the features for the task of detecting pneumonia.

6. Conclusion

This paper proposes a DL model for identifying normal and lung opacities from chest x-ray images. The dataset was obtained from an open-access repository. The images were preprocessed by resizing, normalizing, and generating additional images to improve the quality and reduce overfitting. In this study, the researchers used a pretrained DenseNet121 network for feature extraction, while training and comparing four transfer-learning models: VGG16, InceptionV3, ResNet50, and DenseNet121. Among the compared models, DenseNet121 produced the best accuracy of 89% and AUC of 0.96. Transfer learning enables the design of solid medical-imaging models even when the available data is limited because these models borrow feature representations learned from large image collections.

The dense connectivity pattern of DenseNet121 ensures stronger information flow throughout the network and reduces the chance of threatened gradients, allowing it to capture refined radiographic details more effectively. These steps provide additional support for the model by reducing noise and increasing the diversity of training samples, both of which are important for reliable classification performance in medical imaging. Future work will investigate even more advanced architectures, including ViT and other state-of-the-art deep learning models, which could further this performance. The long-term objective is to develop a practical tool to assist clinicians, radiologists, and other healthcare professionals in a real-world medical setting.

Reference

[1] Shagun Sharma, and Kalpna Guleria, "A Systematic Literature Review on Deep Learning Approaches for Pneumonia Detection using Chest X-Ray Images," *Multimedia Tools and Applications*, vol. 83, no. 8, pp. 24101-24151, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[2] Abdullahi Umar Ibrahim et al., "Pneumonia Classification using Deep Learning from Chest X-Ray Images During COVID-19," *Cognitive Computation*, vol. 16, no. 4, pp. 1589-1601, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[3] Sangapu Sreenivasa Chakravarthi et al., "Pneumonia Detection using Chest X-Rays: A Comprehensive Review," *International Conference on Computational Intelligence in Data Science*, Chennai, India, vol. 2, pp. 292-305, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[4] Davide Ippolito et al., "Artificial Intelligence Applied to Chest X-ray: A Reliable Tool to Assess the Differential Diagnosis of Lung Pneumonia in the Emergency Department," *Diseases*, vol. 11, no. 4, pp. 1-13, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[5] Rezaul Haque et al., "A Scalable Solution for Pneumonia Diagnosis: Transfer Learning for Chest X-Ray Analysis," *2024 7th International Conference on Contemporary Computing and Informatics (IC3I)*, Greater Noida, India, pp. 255-262, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[6] Chanhoe Gu, and Minhyeok 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, pp. 1-21, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[7] Zaenab Alammar et al., "Deep Transfer Learning with Enhanced Feature Fusion for Detection of Abnormalities in X-Ray Images," *Cancers*, vol. 15, no. 15, pp. 1-36, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[8] Vivek Yadav, Saksham Shrivastava, and Ajay Pal Singh, "Enhanced Pneumonia Detection using Deep Learning Techniques on Chest X-Rays," *2024 2nd International Conference on Advances in Computation, Communication and Information Technology (ICAICCIT)*, Faridabad, India, pp. 923-929, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[9] Ahmad Waleed Salehi et al., "A Study of CNN and Transfer Learning in Medical Imaging: Advantages, Challenges, Future Scope," *Sustainability*, vol. 15, no. 7, pp. 1-28, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[10] Wasif Khan, Nazar Zaki, and Luqman Ali, "Intelligent Pneumonia Identification from Chest X-Rays: A Systematic Literature Review," *IEEE Access*, vol. 9, pp. 51747-51771, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[11] Rakhmonalieva Farangis Oybek Kizi, Tagne Poupi Theodore Armand, and Hee-Cheol Kim, "A Review of Deep Learning Techniques for Leukemia Cancer Classification Based on Blood Smear Images," *Applied Biosciences*, vol. 4, no. 1, pp. 1-32, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[12] Faisal Alshanketi et al., "Pneumonia Detection from Chest X-Ray Images using Deep Learning and Transfer Learning for Imbalanced Datasets," *Journal of Imaging Informatics in Medicine*, vol. 38, no. 4, pp. 2021-2040, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[13] Dejun Zhang et al., "Pneumonia Detection from Chest X-Ray Images based on Convolutional Neural Network," *Electronics*, vol. 10, no. 13, pp. 1-17, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[14] Houmem Slimi et al., "Trustworthy Pneumonia Detection in Chest X-Ray Imaging through Attention-Guided Deep Learning," *Scientific Reports*, vol. 15, no. 1, pp. 1-15, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[15] Sukhendra Singh et al., "Efficient Pneumonia Detection using Vision Transformers on Chest X-Rays," *Scientific Reports*, vol. 14, no. 1, pp. 1-17, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[16] Erdem Yanar, Fırat Hardalaç, and Kubilay Ayturan, "PELM: A Deep Learning Model for Early Detection of Pneumonia in Chest Radiography," *Applied Sciences*, vol. 15, no. 12, pp. 1-28, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[17] Benzorgat Mustapha et al., "Enhanced Pneumonia Detection in Chest X-Rays Using Hybrid Convolutional and Vision Transformer Networks," *Current Medical Imaging*, vol. 21, no. 1, pp. 1-23, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[18] Mitul Ambaliya et al., "Enhancing Pneumonia Detection Transparency: Exploring Explainable AI Model," *International Conference on Power Engineering and Intelligent Systems (PEIS)*, Srinagar, India, vol. 2, pp. 465-478, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[19] Somak Saha et al., "Explainable SE-MobileNet for Pneumonia Detection Integrated with Robustness Assessment using Adversarial Examples," *Smart Health*, vol. 33, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[20] Nahid Islam et al., "COVID-19 and Pneumonia Detection and Web Deployment from CT Scan and X-Ray Images using Deep Learning," *PLOS One*, vol. 19, no. 7, pp. 1-21, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[21] Paulo Breviglieri, Pneumonia X-Ray Images, Kaggle, 2020. [Online]. Available: https://www.kaggle.com/datasets/pcbreviglieri/pneumonia-xray-images

[22] Li-Heng Chen et al., "Estimating the Resize Parameter in End-to-End Learned Image Compression," *Signal Processing: Image Communication*, vol. 135, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[23] Alhassan Mumuni, and Fuseini Mumuni, "Data Augmentation with Automated Machine Learning: Approaches and Performance Comparison with Classical Data Augmentation Methods," *Knowledge and Information Systems*, vol. 67, no. 5, pp. 4035-4085, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[24] Chittathuru Himala Praharsha, and Alwin Poulose, "CBAM VGG16: An Efficient Driver Distraction Classification using CBAM Embedded VGG16 Architecture," *Computers in Biology and Medicine*, vol. 180, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[25] Pouya Bohlol, Soleiman Hosseinpour, and Mahmoud Soltani Firouz, "Improved Food Recognition using a Refined ResNet50 Architecture with Improved Fully Connected Layers," *Current Research in Food Science*, vol. 10, pp. 1-16, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[26] Md Abu Hanif, Md Khadimul Islam Zim, and Harpreet Kaur, "ResNet vs Inception-v3 vs SVM: A Comparative Study of Deep Learning Models for Image Classification of Plant Disease Detection," *2024 IEEE International Conference on Interdisciplinary Approaches in Technology and Management for Social Innovation (IATMSI)*, Gwalior, India, pp. 1-6, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[27] M. Kavitha et al., "DenseNet-121 Architecture for Plant Leaf Disease Classification," *AIP Conference Proceedings, International Virtual Conference on Machine Learning Applications in Applied Sciences and Mathematics: IVCMASM2022*, Chennai, India, vol. 2802, no. 1, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[28] Dias Nessipkhanov et al., "Deep CNN for the Identification of Pneumonia Respiratory Disease in Chest X-Ray Imagery," *International Journal of Advanced Computer Science and Applications (IJACSA)*, vol. 14, no. 10, pp. 652-661, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[29] Mudasir Ali et al., "Pneumonia Detection using Chest Radiographs with Novel EfficientNetV2L Model," *IEEE Access*, vol. 12, pp. 34691-34707, 2024. [CrossRef] [Google Scholar] [Publisher Link]