Natural Resources for Human Health
eISSN: 2583-1194

Original Research

Received 27 May 2026
Revised 24 June 2026
Accepted 16 July 2026

Natural Resources for Human Health 2026, Vol. 6 (10s): 897–911

KEYWORDS:
Chest X-ray images; transfer learning; deep learning; medical image analysis.

A Deep Ensemble Transfer Learning for Pneumonia Detection from Chest X-Ray Images to Improve Health Outcome

*Manab Kumar Das[1, 2], Prasenjit Das[3], Chiranjib Chakrabarty[4], Payel Roy[4], Priti Deb[5], Indrajit De[6]

[1]University of Engineering and Management, Kolkata, 700160, India
*Mail ID: manabdas.aec@gmail.com, ORCID: 0009-0004-3669-2720
[2]Department of Information Technology, MCKVIE, Howrah, 7112024, India
[3]Department of Computer Science, RMVCC, Rahara, 700118, West Bengal, India
[4]Department of Computer Application and Digital Marketing, iLead, Kolkata,700046 India
[5]Department of Computer Application, IEM-UEM, Kolkata, 700091, India
[6]Department of CSE (AIML), IEM-UEM, Kolkata, 700091, India

ABSTRACT: Pneumonia is a serious respiratory issue that requires timely and accurate diagnosis to reduce the mortality rate. Chest X-ray images are a common diagnostic tool for pneumonia assessment. Early detection of pneumonia is crucial for prompt treatment and improved patient outcomes. The conventional diagnosis of pneumonia is highly challenging because of limited medical expertise or poor radiograph (X-ray) image quality. The recent advancement of deep learning for image analysis might be resolved the problem. The proposed study ensemble has two pre-trained Convolutional Neural Networks (CNNs), including EfficientNetB3 and DenseNet121. Pre-processing is the crucial step for this investigation, where lung region cropping, Contrast Limited Adaptive Histogram Equalization (CLAHE), and augmentation are employed to enhance the image visibility. The primary aim is more focused on important areas during training, leading to improved classification performance. Transfer learning is utilized, where EfficientNetB3 is trained using a warm-up training followed by fine-tuning. The DenseNet121 is trained for end-to-end processing. Predictions from both the models are combined using a soft average ensemble technique to enhance the efficiency of the system. The experimental results achieve an accuracy of 98.21% and an Area under the Curve (AUC) score of 0.9983.The outcomes indicate that the proposed system can serve as an effective computer-aided screening tool for early pneumonia diagnosis.

1. INTRODUCTION

Recently numerous cutting-edge technologies, such as Internet of Things (IoT) and Machine Learning (ML), have gained popularity to address a variety of healthcare issues [1]. In this context, ML-based healthcare approaches offer the solution to identify various diseases at an early stage. Pneumonia is one of the most prevalent causes of death and disability among children globally. The World Health Organization (WHO) reports 80,000 children under the age of five are suffering from pneumonia [2]. Preventing respiratory infections at an early stage to reduce the fatality rate is extremely challenging for poor-income counties [3, 4]. In this regard, inadequate healthcare infrastructure and delayed medical intervention play the key role. ML-based pneumonia identification become popular during COVID-19 pandemic to reduce the mortality rate [5]. Prior and accurate diagnosis of pneumonia is crucial to minimize the death rate and improve treatment outcomes. The Deep Learning (DL)-based advanced technology could resolve several issues in the contemporary medical system. In this context, this cutting-edge technology has the potential to significantly improve classification and diagnostic efficiency through automated chest X-ray

image analysis [6, 7]. Models based on DL have been employed in numerous investigations to identify pneumonia automatically from chest X-ray images [8, 9]. Ensemble learning has been explored to improve classification performance. Several DL techniques are combined to increase model robustness by minimizing model bias [10].

The suggested study utilized an ensemble deep learning framework by integrating two pre-trained CNNs. This study incorporates lung region cropping, CLAHE, and augmentation during the data pre-processing stage to enhance image quality. Here high-quality images are essential to concentrate for diagnosis. The proposed system aims to deliver dependable and scalable computer-aided pneumonia detection using a soft average ensemble technique. **Figure 1** shows the key element (chest radiograph images) of our research investigation of both normal and pneumonia cases.

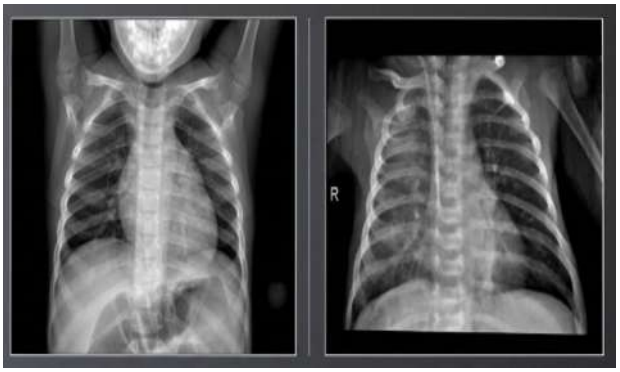

Figure 1. Chest X-rays image for normal and pneumonia case

2. LITERATURE REVIEW

Numerous research study highlight the adoption of machine learning or deep learning for pneumonia detection. This section includes several machine learning-based research articles for pneumonia detection.

Automated pneumonia detection has been extensively explored using deep learning techniques. Davinder Paul Singh et al. [11] published a review article that analysed various hybrid deep learning approaches for automated pneumonia detection. The Multi-Context Shifted Window Attention (McSWA) framework was implemented to archive an accuracy of 98.76%. Mohammad Farukh Hashmi et al. [12] developed a deep learning based hybrid technique for pneumonia detection. The ResNet-50 model was employed to extract features while integrating the wavelet transform. Finally, the SVM classifier produced a remarkable accuracy of 97%. A transfer learning based method for pneumonia identification using an unbalanced dataset was developed by Faisal Alshanketi et al. [13]. Here the researchers evaluated the performances of various DL approaches that explore semi-supervised learning techniques. The researcher's primary objective was to show how ML might be revaluated for pneumonia diagnosis. Neeta et al. [14] created an unsupervised learning method (federated) to detect pneumonia. Additionally, the auto encoder neural network architecture was employed to enhance the system performance. Hanan et al. [15] examined several alternative CNN architectures for an experimental evaluation of pneumonia diagnosis. The study considered 6,939 X-ray images for individual algorithms; out of them *VGGNet* produced a superior result, with an overall accuracy of 97%. A CNN-based framework for pneumonia identification was designed by T.S. Arulananth et al. [16]. DenseNet-121 played a vital role in this study that employed chest X-ray images. The authors focus on only one DL approach to measure the performance, although it produces an outstanding accuracy of 97.03%. In another study, Nigus Wereta Asnake et al. [17] implement a YOLOv3 detector to identify pneumonia from chest radiograph images. The authors concentrated on several performance matrices without focusing on the AUC score. A deep learning based novel approach was developed by Mudasir Ali et al. [18]. Here six individual algorithms were employed to check the accuracy along with the Adam optimizer. The model was trained by 5856 image data, where EfficientNetV2L produced the highest accuracy of 94.02%. Raheel Siddiqi et al. [19] introduced a DL based comprehensive study for pneumonia detection. The authors considered the limitations and effectiveness of specific COVID-19-related approaches during 2020 and 2023. Prerna Agrawal et al. [20] designed a smart pneumonia identification technique by evaluating ResNet50, DenseNet121, and NASNet. DenseNet121 demonstrated outstanding performance with a recall value of 0.9795, whereas the AUC score was 0.9808. Erdem Yanar et al. [21] used a large data set to develop a novel DL-based architecture. This framework integrates various DL models, including Vision Transformer (ViT), InceptionV3, ResNet50, and VGG16. The system produce an outstanding performance with accuracy of 96% and 0.91 AUC score. Mohammad Farukh

Hashmi et al. [22] introduced a hybrid DL approach for pneumonia detection. In this study ResNet-50 played a crucial role along with wavelet transform for deep feature extraction. Support Vector Machine was employed as a classifier to produces the accuracy of 97%. The major drawback of this study was the author did not concentrate on the AUC score

3. MATERIALS AND METHODS

Data pre-processing, model training, evaluation, and ensemble learning are the four fundamental phases of the proposed methodology. Image pre-processing is the initial step followed by model training using EfficientNetB3 and DenseNet121. Accuracy and the AUC score are two common technique that are employed to measure the model performance. Finally, each model is blended using a soft average ensemble technique to enhance overall stability.

3.1 Dataset

The dataset contains a total of 5856 lung images from both cases, including 1583 normal and 4273 pneumonia cases [23]. Daniel S. Kermany et al. [24] developed the dataset for their research investigation. The researchers collected chest X-ray images from paediatric patients at the Guangzhou Women and Children's Medical Centre in Guangzhou, China, between the ages of one and five.

The dataset is originally distributed into predefined training, validation, and test folders. The dataset is split into a 70:15:15 ratio to ensure that every image associated with a single patient is allocated to a single subset. It resolves the data leakage problem and assures a realistic evaluation scenario. A fixed random seed of 42 is utilized to ensure consistency across the experiments. The dataset exhibits class imbalance, with pneumonia cases constituting approximately 73% of the total samples. This distribution is preserved with training, validation, and test sets to maintain consistent class proportions and ensure unbiased model evaluation. This strategy guarantees a fair and accurate assessment of model performance by eliminating data leakage.

3.2 Model Selection

This study employed two pre-trained CNN architectures, such as EfficientNetB3 and DenseNet121 to create a reliable pneumonia detection framework. These models have been chosen because of their strong performance for medical image analysis and classification. EfficientNetB3 facilitates efficient learning for compound scaling strategy that balances network depth, width, and input resolution. DenseNet121 introduces dense connections between layers' architecture to enhance gradient propagation and maximize feature reuse across the network. The suggested model is helpful to learn characteristics and improve classification accuracy that is depicted in **Algorithm 1**. Subsequently, an ensemble approach is employed to improve overall diagnostic performance for pneumonia detection.

3.3 Proposed Algorithm

Algorithm 1: Average Ensemble Model of EfficientNetB3 and DenseNet121

Require: D_{raw} (Raw dataset), W_E, W_D (Weights for EfficientNetB3 and DenseNet121)

Ensure: Y_{final} (Final predictions)

1: **procedure** ClassifyChestXRay (D_{raw}, W_E, W_D)

2: D_{train}, D_{val}, D_{test} ←GroupShuffleSplit (Merge (D_{raw}), 0.70, 0.15, 0.15)

3: **for each** split S ∈ { D_{train}, D_{val}, D_{test} } **do** ▷Pre-processing & Augmentation

4: **for** X ∈ S **do**

5: X_E, X_D ←BackbonePreprocess (ApplyCLAHE(Crop(X)))

6: **if** S = D_{train} **then**

7: X_E, X_D ←Augment(X_E, X_D)

8: **end if**

9: **end for**

10: **end for**

11: ▷Independent Model Training

12: M_E ←TrainEfficientNetB3 (D_{train}, D_{val}, W_E)

13: M_D ←TrainDenseNet121(D_{train}, D_{val},, W_D)

14: ▷Ensemble Threshold Optimization (Validation)

15: $Y_{ens}^{(val)}$ ←½ (M_E .predict (D_{val}) + M_D. predict (D_{val}))

16: T ←ThresholdOptimization ($Y_{ens}^{(val)}$, [0.0, 1.0], 0.01)

17: ▷Inference & Binary Classification (Test)

18: Y_{final} ← ∅

19: **for** $X_i \in D_{test}$ **do**

20: $Y_i^{(ens)}$ ←½ (M_E .predict(X_i) + M_D .predict(X_i))

21: **if** $Y_i^{(ens)} > T$ **then**

22: Y_{final} ← $Y_{final} \cup \{1\}$

23: **else**

24: Y_{final} ← $Y_{final} \cup \{0\}$

25: **end if**

26: **end for**

27: **return** Y_{final}

28: **end procedure**

3.4 Data Pre-processing

Prior to model training, every chest X-ray image is pre-processed to enhance pertinent visual attributes and ensure deep learning architecture compatibility. Region-of-interest (ROI) cropping, contrast improvement, scaling, normalization, and data augmentation included in the pre-processing pipeline.

A) Region of Interest Cropping: A simple heuristic-based ROI cropping strategy is applied to reduce irrelevant regions such as image borders, textual annotations, and non-informative background areas. The cropping boundaries are defined proportionately based on the measurements of the input image. The system removes the top 8% of the image's height and the left 5% of its width, whereas the bottom boundary retains the image's entire height, and the right border adjusts to 95% of its width. Therefore, the cropped region is the area between 0.08h to 1h along the vertical axis and between 0.05w to 0.95w along the horizontal axis.

B) CLAHE for Contrast Enhancement: CLAHE is employed to enhance the visibility of sensitive anatomical structures and pathological patterns. The images are ensured to be in 8-bit unsigned integer format by clipping pixel values between the range [0, 255] for proper intensity scaling. The individual image is then transformed from the RGB color space to the LAB color matrix, where intensity information is encoded through the luminance channel (L). CLAHE is applied with a tile grid size of 8 × 8 and a clip limit of 2.0, to this luminance channel. This step reduces noise amplification through contrast clipping and improves local contrast on small regions. The boosted L is subsequently recombined with the original chromatic channels (A and B), and the image is converted back to RGB format. This method enhances the visibility of fine-grained characteristics including consolidations and opacities as well as overall structural image integrity.

C) Image Resizing and Backbone Pre-processing: The images are resized in accordance with the individual CNN architecture's input requirements. The image's pixels have been scaled to 224 × 224 for DenseNet121 and 300 × 300 for

EfficientNetB3. This procedure allows for effective training and ensures compatibility with pre-trained ImageNet weights. Additionally, Images are analyzed using the backbone-specific pre-processing functions to align input images with the expectations of the pre-trained CNN backbones. This phase ensures that input data distributions match with individual model training to facilitating efficient transfer learning and steady convergence.

D) Data Augmentation: The augmentation technique is applied to reduce overfitting in the training dataset using the Keras ImageDataGenerator. These transformations simulate realistic variations observed during radiographic acquisition while preserving diagnostic features. In this context, random rotations up to 10° are used to simulate variations in patient positioning. Horizontal and vertical translations up to 5% of the image dimensions. The model can learn features at various spatial scales via random zoom transformations up to 10%. Pixel intensity variations between [0.8, 1.2] are used to simulate alterations in imaging conditions. Maximum distortion up to 0.05 is required for geometric variability through shear transformation. Random horizontal mirroring is crucial to boost data diversity via horizontal flipping. The reflective approach is applied to fill newly introduced pixels during transformations.

The augmentation method is exclusively applied to training datasets when combined with the corresponding model-specific preprocessing step. The pre-processing strategy involves ROI extraction utilizing lung cropping and contrast enhancement using CLAHE to increase visibility of the input image. Diverse augmented samples are generated by applying various random seeds to create unique transformation instances (variations in rotation and zoom). The image pre-processing is depicted in **Figure 2**. The deterministic preprocessing techniques such as ROI cropping, CLAHE, backbone preprocessing are employed to prepare the validation and test datasets.

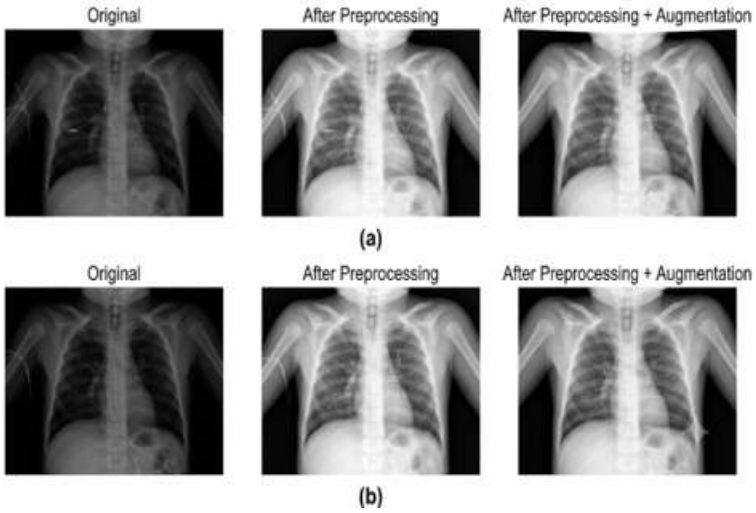

Figure 2. Pneumonia image pre-processing before and after (a) EfficientNetB3 and (b) DenseNet121

3.5 CNN Architectures and Ensemble Strategy

The proposed framework utilizes two advanced CNN architectures such as EfficientNetB3 and DenseNet121. Both networks are initialized with weights pretrained on the large-scale ImageNet dataset for effective transfer learning. The CNN models are independently trained using various random seeds to learn discriminative radiological features from the chest X-ray dataset. Finally, the predictions of individual models are combined for an average ensemble strategy to enhance overall performance.

A) EfficientNetB3 Fine-Tuned Model:

EfficientNetB3 is selected for a compound scaling strategy that uniformly scales network depth, width, and resolution to improve performance with fewer parameters. The architecture effectively balances computational efficiency for medical image analysis. The model architecture used in this study is represented as:

$$X \rightarrow EfficientNetB3\ Base\ (without\ top) \rightarrow GAP \rightarrow Dropout(0.4) \rightarrow Dense(1, sigmoid) \quad \text{--- (1)}$$

In **equation (1)** *EfficientNetB3Base* represents pretrained model initialization with weights for primary feature extraction. The network employs Mobile Inverted Bottleneck Convolution (MBConv) blocks that enable efficient multi-scale feature learning.

Global Average Pooling (*GAP*) converts the spatial feature maps into a compact one-dimensional feature vector via a convolutional backbone. A dropout layer rate of 0.4 is introduced to mitigate overfitting by randomly disabling neurons during training. The final dense layer with sigmoid activation outputs the probability score indicating the presence of pneumonia.

Training Strategy:

I) Random Seed: EfficientNetB3 is trained with seed 101 to control weight initialization as well as dataset shuffling, ensuring reproducibility and introducing diversity for ensemble learning.

II) Batch Size: The batch size is set to 16 during training.

III) Warm-up Training: The pretrained EfficientNetB3 backbone is frozen, and only the newly added classification layers are trained for 5 epochs. This allows the classifier to maintain the generic visual features acquired from the ImageNet dataset while adapting to the pneumonia dataset. The model is optimized using the Adam optimizer with a learning rate of 1×10^{-3}.

IV) Fine-Tuning: After the warm-up stage, the top 180 layers of the EfficientNetB3 backbone are unfrozen, while the remaining lower layers remained frozen. These layers are then jointly trained with the classification head using a reduced learning rate of 3×10^{-5} . It refines high-level feature representations for the pneumonia classification with 30 epochs.

V) Loss Function: Binary Focal Cross-Entropy loss is utilized to optimize the models and address potential class imbalance in the dataset. The focal loss is widely adopted for imbalanced classification tasks. The Binary Focal Cross-Entropy loss is defined in **equation (2)**.

$$L = -\frac{1}{N}\sum_{i=1}^{N} \alpha(1-\hat{y}_i)^{\gamma} y_i \log(\hat{y}_i) + (1-\alpha)\hat{y}_i^{\gamma}(1-y_i)\log(1-\hat{y}_i) \quad \text{--- (2)}$$

Where, y_i represents the ground-truth label, $\hat{y}_i$ denotes the predicted probability, $\alpha(0.25)$ balances class importance, and $\gamma(2.0)$ controls the focusing effect on hard samples.

VI) Callback Mechanisms: Early stopping automatically halted whenever the validation AUC failed to improve for five consecutive epochs. The weights corresponding to the best validation AUC are restored to guarantee optimal model performance. The learning rate is dynamically reduced by ReduceLROnPlateau when the validation AUC stops improving. This rate is multiplied by a factor of 0.5 if no improvement in validation AUC is observed for three consecutive epochs. The model weights achieving the highest validation AUC during training are automatically saved through Model Checkpoint. It ensures that the best-performing model is preserved for subsequent evaluation.

VII) Optimizer: Model parameters are optimized using the Adam optimizer. It adjusts learning rates for individual parameters based on first and second-order gradient moments. The parameter update rule for Adam optimization is shown in **equation (3)**.

$$\theta_{t+1} = \theta_t - \eta \cdot \frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \epsilon} \quad \text{--- (3)}$$

Where θ_t represents model parameters, η the learning rate, and $\hat{m}_t$, $\hat{v}_t$ are bias-corrected first and second moment estimates of gradients and ϵ is a small constant added for numerical stability.

VIII) Threshold Optimization: The optimal classification threshold is determined using the validation set with a grid search over the range [0.0, 1.0] with a step size of 0.01. Predicted probabilities are transformed into binary labels for individual threshold values t_i and the corresponding classification accuracy is computed. The threshold that maximized the validation accuracy is determined as the optimal decision boundary.

B) DenseNet121 End-to-End Model

The DenseNet121 architecture is employed as the second CNN model due to its dense connectivity pattern. Here, an individual layer receives feature maps from all preceding layers. This strategy encourages feature reuse, improves gradient flow, and enhances parameter efficiency. The model structure can be described as:

$$X \rightarrow DenseNet121\ Base\ (without\ top) \rightarrow GAP \rightarrow Dropout(0.35) \rightarrow Dense(1, sigmoid) \quad \text{--- (4)}$$

The model is initialized with weights using the *DenseNet121Base* function for primary feature extraction, defined in **equation (4)**. It employs dense blocks, where each convolutional layer receives feature maps from previous layers. The DenseNet121 framework assists in learning stronger feature characterizations and facilitates information flow. *GAP* represents a compact one-dimensional feature vector from the convolutional backbone. This operation significantly reduces the number of trainable parameters and enhances model extension. A dropout layer with a rate of 0.35 is utilized to reduce overfitting through randomly deactivating neurons during training. The final dense layer with sigmoid activation outputs a probability value indicating the presence of pneumonia.

Training Strategy:

I) Random Seed: DenseNet121 is trained with seed 42 to ensure independent initialization. In this phase data is shuffled and promotes complementary behavior for the ensemble.

II) Batch Size: The batch size is set to 16 during training.

III) End-to-End Training: The DenseNet121 is trained using an end-to-end training strategy. Here individual layers of the network are allowed to update during training for 30 epochs. The pretrained convolutional backbone is not frozen, unlike EfficientNetB3. End-to-end strategy enables the model to adapt its feature representations specifically for the pneumonia classification.

IV) Loss Function: Binary Cross-Entropy loss is used to optimize classification.

V) Optimizer and Learning Rate: The parameters are optimized using the Adam optimizer with a learning rate of 1×10^{-4}.

VI) Callback: The callbacks employed to maximize convergence are the same as EfficientNetB3: early stopping, ReduceLROnPlateau, and Model Checkpoint.

VII) Threshold Optimization: The optimal classification threshold is determined using the validation set by performing a grid search over the range [0.0, 1.0] with a step size of 0.01. Predicted probabilities are transformed into binary labels for each threshold value t_i, and corresponding classification accuracy is measured. The threshold determined the maximized validation accuracy and optimal decision boundary.

C) Average Ensemble and Threshold Optimization

An average ensemble approach is adopted to take the strengths of both the CNNs. For an input image X_i, the predicted probabilities from each model are as follows:

$i)\ \hat{y}_i^{(EfficientNetB3)}$

$$ii)\ \hat{y}_i^{(DenseNet121)}$$

These predictions are first generated on the validation set, and then ensemble probability is computed. The ensemble threshold optimization technique is used to obtain the averaged ensemble probabilities on the validation set that is depicted in **equation (5)**.

$$\hat{y}_i^{(ensemble,val)} = \frac{1}{2}(\hat{y}_i^{(EfficientNetB3,val)} + \hat{y}_i^{(DenseNet121,val)}) \quad \text{----- (5)}$$

A grid search is performed over the range [0.0, 1.0] with a step size of 0.01 to determine an optimal decision threshold i.e. T. The threshold yielding the highest validation accuracy is selected. After determining the optimal threshold, predictions are generated on the test set using both models. These test probabilities are then averaged using the same ensemble formulation that is shown in **equation (6)**.

$$\hat{y}_i^{(ensemble,test)} = \frac{1}{2}(\hat{y}_i^{(EfficientNetB3,test)} + \hat{y}_i^{(DenseNet121,test)}) \quad \text{----- (6)}$$

Finally, the optimized threshold T (obtained from the validation set) is applied to the ensemble test probabilities to obtain binary predictions, illustrated in **equation (7)**:

---- (7)

$$\hat{y}_i^{(final)} = \begin{cases} 1, & if\ \hat{y}_i^{(ensemble,test)} > T \\ 0, & otherwise \end{cases}$$

These final predictions are then used for performance evaluation on the test dataset.

4. RESULTS AND DISCUSSION

This section describes the research findings of DenseNet121, EfficientNetB3, and their average ensemble for pneumonia detection. A single pre-processing procedure that includes lung ROI cropping, applying CLAHE, using a specific model backbone, and adding augmentations is utilized to train both CNNs. ROI extraction focuses on relevant anatomical regions, while CLAHE enhances local contrast to improve feature visibility. EfficientNetB3 is trained using a warm-up and fine-tuning strategy, whereas DenseNet121 is trained using end-to-end processing. Experiments are conducted with random seed initialization.

Performance evaluation is carried out using accuracy, confusion matrices, classification reports, and ROC curves. Training and validation curves are examined to evaluate convergence and overfitting. Optimal decision thresholds for individual models are determined on the validation set via grid search to increase the accuracy. These thresholds are used for separate model evaluation. The final predictions are obtained by averaging the estimated raw probabilities, followed by a separate threshold optimization.

4.1 EfficientNetB3 Fine-Tuned Model Performance

This section evaluates the performance of EfficientNetB3. The model is initialized with ImageNet-pretrained weights and fine-tuned on the target dataset. The validation and test sets are processed using deterministic preprocessing steps without any augmentation. It ensures that model evaluation is conducted on routinely processed but non-augmented data for fair and realistic assessment.

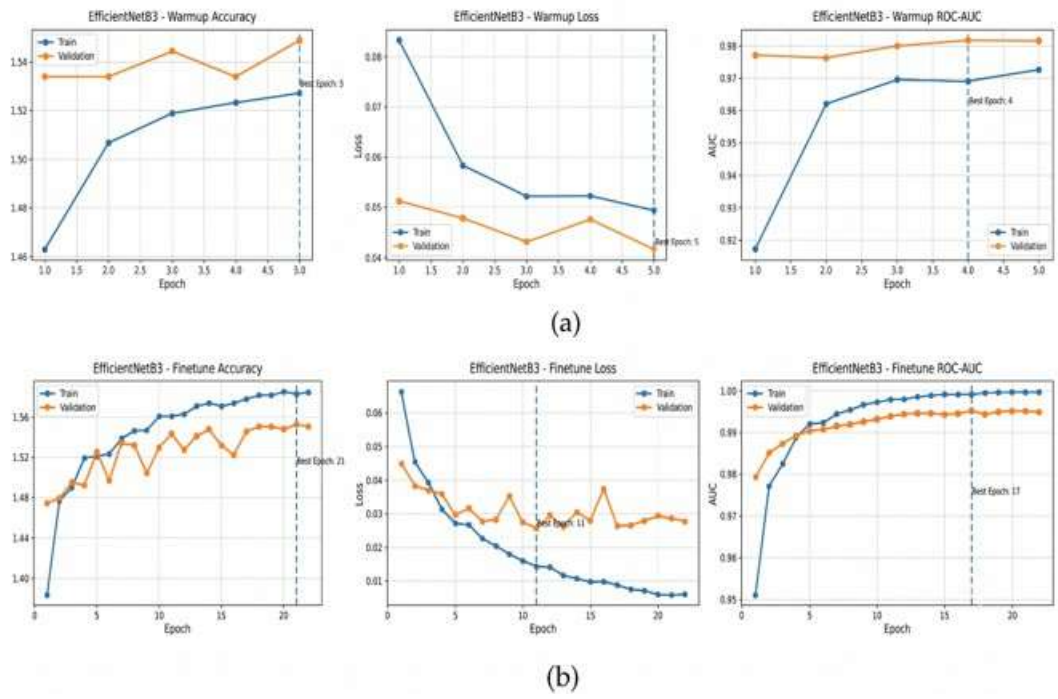

Figure 3. Accuracy, Loss and AUC of EfficientNetB3 for (a) Warm-up Training (b) Fine-Tuning

Figure 3(a) illustrates the training and validation performance, including accuracy, loss, and ROC-AUC of the EfficientNetB3 during the warm-up phase. The model is trained for five epochs without the use of any callbacks. It allows the newly initialized classification layers to adjust the domain-specific features of chest X-ray images. **Figure 3(b)** depicted the performance of the model during the fine-tuning phase, where deeper layers are unfrozen for further optimization. The training is scheduled for

30 epochs, and early stopping is applied to prevent overfitting. All callbacks are configured to monitor the validation ROC-AUC It ensures the model optimization is guided by discriminative performance rather than accuracy alone. The training is terminated after 22 epochs, with the best-performing weights obtained at epoch 17. The model achieved optimal generalization before completing the full training schedule.

Prior to evaluating the model, the optimal decision threshold is determined using the validation data. A threshold of 0.43 is chosen instead of the default threshold (0.5) to obtain best validation accuracy. This optimized threshold is subsequently applied for generating the confusion matrix and classification report on the test set. The ROC-AUC is computed independently of the threshold. **Table 1** presents the classification reports with the best threshold evaluation. The model demonstrates strong performance across both the classes by achieving an overall test accuracy of 97.54%.

Table 1. EfficientNetB3 classification report with best threshold

Class	Precision	Recall	F1-score	Support
Normal	0.9444	0.9675	0.9558	246
Pneumonia	0.9875	0.9784	0.9829	647
Accuracy	—	—	0.9754	893
Macro Avg.	0.9660	0.9729	0.9694	893
Weighted Avg.	0.9757	0.9754	0.9755	893

Figure 4(a) demonstrates the confusion matrix, where the model correctly classified 238 normal samples out of 246. 633 samples are properly identified as pneumonia from 647 images. These results indicate that the model achieves high classification accuracy with a relatively low number of misclassifications across both classes. The ROC-AUC score of 0.9971 for the fine-tuned EfficientNetB3 model is shown in **Figure 4(b)**. This high AUC value indicates the model's strong ability to distinguish between normal and pneumonia cases.

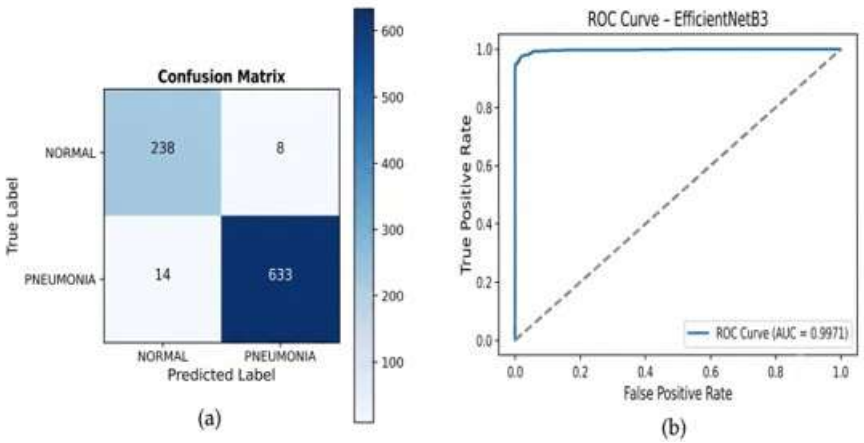

Figure 4. EfficientNetB3 (a) Confusion Matrix, and (b) ROC Curve

4.2 DenseNet121 End-to-End Model Performance

This section examines the end-to-end performance of the DenseNet121 model. The model is initialized with ImageNet-pretrained weights, where every layer is trained jointly on the target dataset. The data augmentation strategy is utilized on the training set to enhance generalization. This strategy ensures a fair and realistic evaluation of the model's generalization capability on unseen data.

Natural Resources for Human Health
eISSN: 2583-1194

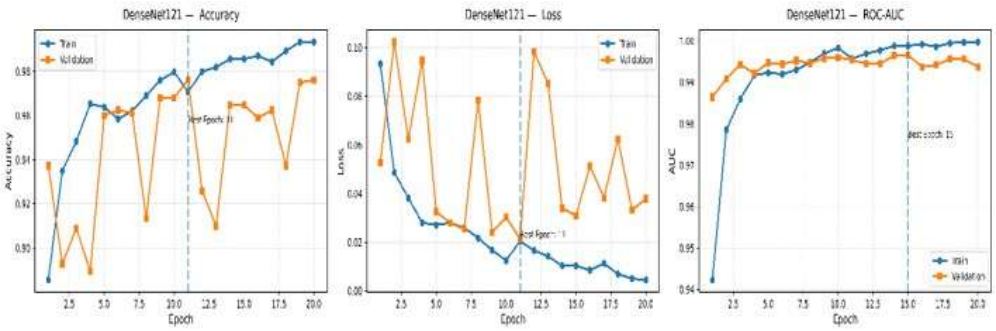

Figure 5. DenseNet121 model Training Accuracy, Loss and AUC

Figure 5 illustrates the DenseNet121 training and validation performance in terms of accuracy, loss, and ROC-AUC value. Early stopping is employed during the model's 30-epoch training period to avoid overfitting. Individual callbacks are configured to monitor the validation ROC-AUC. This procedure ensures that model optimization is guided by its discriminative performance rather than accuracy alone. As an outcome, training is halted after 20 epochs, with the best performance obtained at epoch 15.

The optimal decision threshold is determined using the validation data before evaluating the model on the test set. A threshold of 0.24 is selected instead of the default threshold to obtain the highest validation accuracy. **Table 2** represents the classification report with the best threshold on the test dataset. The model demonstrates strong performance across both classes with an accuracy of 97.98%. The model effectively distinguishes normal and pneumonia cases with minimal misclassification.

Table 2. DenseNet121 classification report with best threshold

Class	Precision	Recall	F1-score	Support
Normal	0.9711	0.9553	0.9631	246
Pneumonia	0.9831	0.9892	0.9861	647
Accuracy	—	—	0.9798	893
Macro Avg.	0.9771	0.9722	0.9746	893
Weighted Avg.	0.9798	0.9798	0.9798	893

The confusion matrix on the test set is presented in **Figure 6(a)**, where 235 normal and 640 pneumonia cases are classified properly out of 246 and 647 images. These results indicate that the model achieves the classification accuracy of 97.98% with a low number of misclassifications. The ROC curve is shown in **Figure 6(b)** with an ROC-AUC score of 0.9983. This exceptionally high AUC value demonstrates the model's strong ability to discriminate between normal and pneumonia cases across a wide range of decision thresholds.

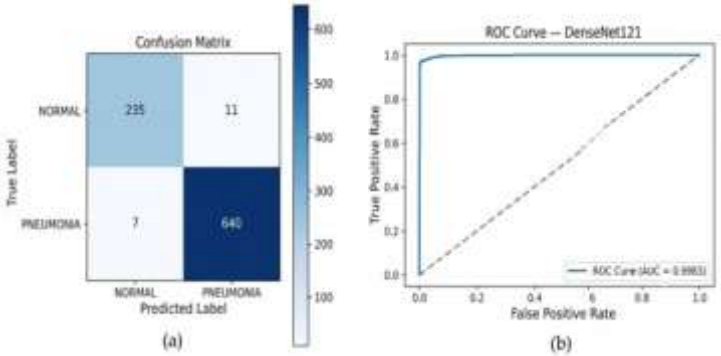

Figure 6. DenseNet121 (a) Confusion Matrix, and (b) ROC Curve

4.3 Average Ensemble Model Performance

The performance and efficiency of the proposed ensemble model are evaluated in this section. It combines the predictions of the fine-tuned EfficientNetB3 and the end-to-end trained DenseNet121 models using an average probability fusion strategy. This ensemble technique attempts to reduce model-specific biases and leverage the individual strengths by aggregating the outcomes of two CNN architectures. Exclusively deterministic pre-processing pipelines are required for validation and test sets without data augmentation. The final prediction score is computed through averaging the predicted probability from each model. The suggested approach enhances robustness and stability, leading to improved generalization performance compared to individual models.

The validation data is utilized for determining the optimum decision threshold prior to assessing the model. A threshold of 0.32 is selected rather than the default threshold. Subsequently, this optimized threshold is applied during the computation of the classification report, whereas the ROC-AUC metric is evaluated independently. The classification report for the average ensemble with the optimal threshold is illustrated in **Table 3.** The suggested model demonstrates a strong performance by achieving an overall test accuracy of 98.21%.

Table 3. Average ensemble classification report with best threshold

Class	Precision	Recall	F1-score	Support
Normal	0.9675	0.9675	0.9675	246
Pneumonia	0.9876	0.9876	0.9876	647
Accuracy	—	—	0.9821	893
Macro Avg.	0.9776	0.9776	0.9776	893
Weighted Avg.	0.9821	0.9821	0.9821	893

The average ensemble confusion matrix on the test set is presented in **Figure 7(a)**. 238 normal and 639 pneumonia samples are accurately classified out of 246 and 647 images. The outcome indicates that the model achieves high classification accuracy with a relatively low number of misclassifications against both classes. **Figure 7(b)** depicted the ROC-AUC score of 0.9983. This AUC value demonstrates the model's strong ability to discriminate pneumonia from a wide range of datasets.

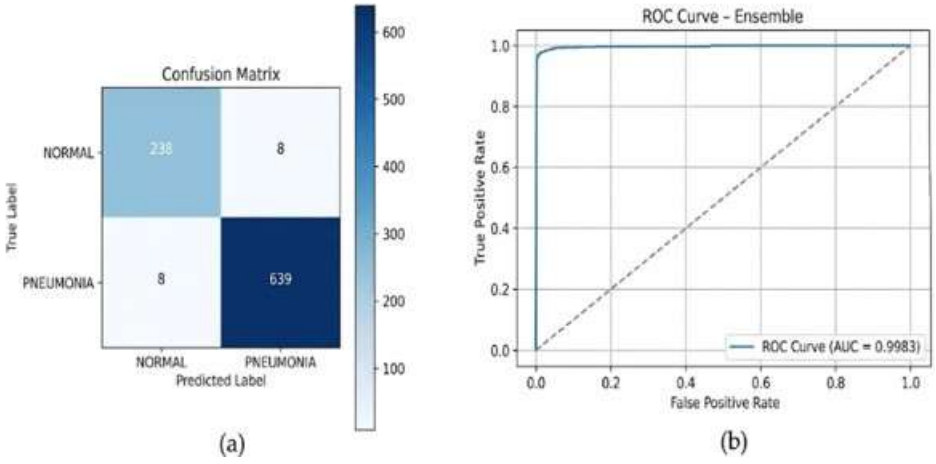

Figure 7. Average Ensemble (a) Confusion Matrix, and (b) ROC Curve

4.4 Grad-CAM Visualization

Identifying the areas of the input images for model predictions is equally important in this study. The model concentrates on

clinically relevant areas to ensure interpretability and validation through Grad-CAM visualization. Grad-CAM heatmaps are determined separately for every model to create a single attention map. Subsequently, they are combined using an average fusion strategy to produce a unified attention map. The Grad-CAM visualizations of individual CNNs are depicted in **Figure 8(a)** and **8(b)**.

The ensemble Grad-CAM is generated by aggregating the contributions of individual models. The predicted probabilities from both models are averaged to obtain the final ensemble prediction score. The proposed approach provides a comprehensive visualization through capturing the complementary attention patterns of each model. **Figure 8(c)** illustrates the Grad-CAM visualizations for the proposed ensemble approach. The highlighted regions indicate the areas that most strongly influence the ensemble's predictions.

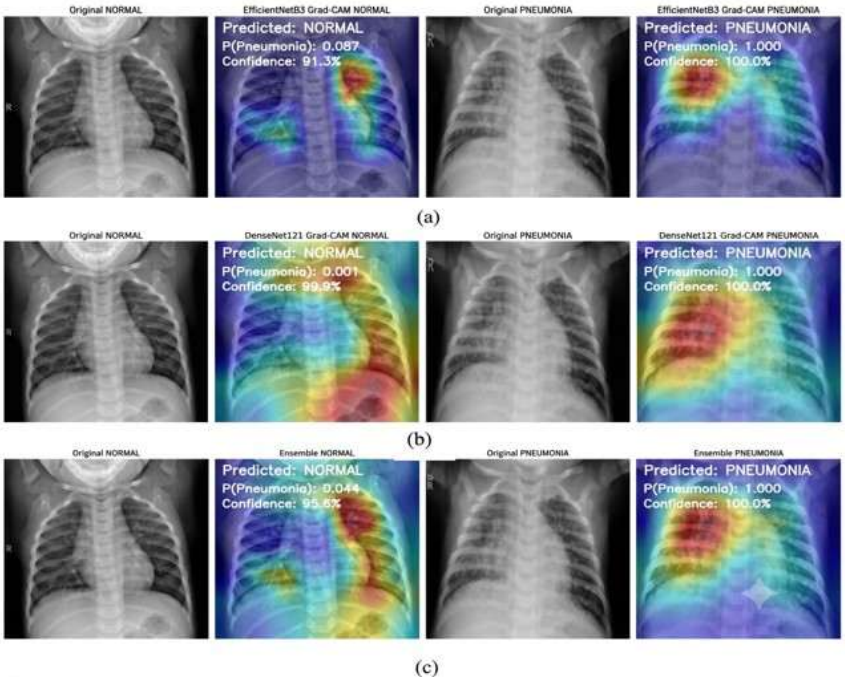

Figure 8. Grad-CAM Visualization of (a) EfficientNetB3, (b) DenseNet121, and (c) Average Ensemble model for pneumonia detection

4.5 Comparative Evaluation

Table 4 demonstrates comparisons of EfficientNetB3, DenseNet121, and the average ensemble model employing the same test dataset. This comparative analysis provides the strengths of individual architectures along with suggested ones. Both CNN models produce a strong classification result. The EfficientNetB3 accuracy is 97.54%, whereas DenseNet121 slightly outperforms with an accuracy of 97.98%. The average ensemble model outperforms the individual CNNs. The model produces an accuracy of 98.21% with an AUC score of 0.9983. **Figure 9** shows the performance analysis of individual CNNs along with the average ensemble model.

Table 4. Comparative performance of individual CNN with average ensemble model

Model	Accuracy (%)	Precision	Recall	F1-score	ROC-AUC
EfficientNetB3	97.54	0.9875	0.9784	0.9829	0.9971
DenseNet121	97.98	0.9831	0.9892	0.9861	0.9983
Proposed	98.21	0.9876	0.9876	0.9876	0.9983

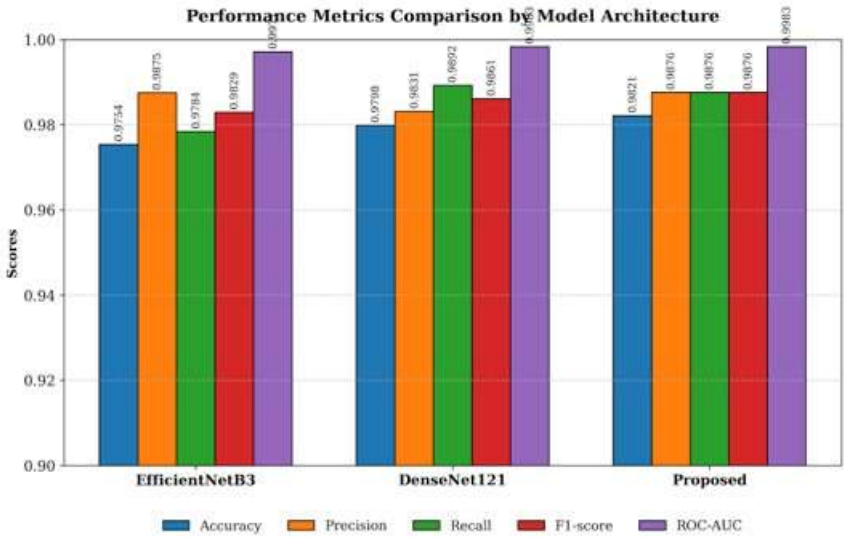

Figure 9. Performance analysis of individual CNN and average ensemble

5. COMPARATIVE ANALYSIS

Table 5 demonstrates a detailed comparative analysis of our proposed method with other existing methods. The comparison highlights various deep learning-based frameworks for pneumonia identification along with accuracy.

Table 5. Comparative Analysis

Reference No.	Publication Year	Model (s)	Accuracy (%)	Purpose
[6]	2024	Xception	95.06	Pneumonia prediction using deep learning
[7]	2024	VGG16	91	Pneumonia detection from using Deep Learning and Transfer Learning
[8]	2025	LRP	91	Interpretable Deep Learning for Pneumonia Detection
[9]	2025	Custom CNN	85.74	CNN-based pneumonia detection
[10]	2022	Ensemble Learning (VIT, MobileNetV2, and DenseNet169)	93.91	Pneumonia detection using Ensemble of Deep Convolutional Neural Networks
Proposed	-------	Average Ensemble CNNs (EfficientNetB3, DenseNet121) Predictions	98.21	Ensemble meta model for pneumonia detection using Transfer Learning

The ensemble-based approach demonstrates a higher diagnostic accuracy with other existing methods. The performance shows our suggested system's efficacy and clinical implementation for pneumonia detection.

6. CONCLUSION

This research presents a deep learning-based technique for automated pneumonia detection. The study focuses on the fine-tuning and end-to-end training strategy of pretrained models instead of manual feature extraction. The pretrained architectures are retrained on the pneumonia dataset by applying transfer learning to identify visual patterns. In comparison to a number of contemporary CNN approaches, the proposed method achieved an impressive accuracy of 98.21% with the ROC-AUC score of 0.9983.

In the future, a web-based application such as a Flask API may be employed for real-time pneumonia predictions. The integration of APIs makes the system more perfect and user-friendly. Future work may focus on using larger and more diverse datasets. This study explores alternative deep learning architectures for further improvement by enhancing accuracy.

ACKNOWLEDGEMENTS

The authors are very thankful to IEM-UEM, Kolkata, MCKV Institute of Engineering, Howrah and iLead, Kolkata jointly for motivating this research work. Additionally, we are delighted to offer it to other faculty members of these institutes who either directly or indirectly assist in completing this research work.

AUTHOR CONTRIBUTIONS

Manab Kumar Das: Principal Investigation, Research Gap Identification, Define Algorithm

Prasenjit Das: Methodology, Implementation.

Chiranjib Chakrabarty: Data Analysis and Pre-processing.

Payel Roy: Result Analysis, Comparative Study.

Priti Deb: Supervision, Writing—Review & Editing.

Indrajit De: Supervision, Writing—Review & Editing.

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest.

REFERENCES

[1] Das, M.K., Deb, P. and De, I., 2024. De I Remote Patient Health Monitoring Frameworks using IoT and ML: A Comparative Study. *International Journal of Intelligent Systems and Applications in Engineering*, *12*(13s), p.367.

[2] World Health Organization, 2022. *WHO operational handbook on tuberculosis. Module 5: management of tuberculosis in children and adolescents*. World Health Organization.

[3] Kitsios, G.D., Restrepo, M.I., Karsies, T.W., Solomon, M., Marciniuk, J., Avila de Benedictis, L., Arancibia, F., Colares Monteiro, J.T., Ribeiro Ferreira da Silva-Filho, L.V., Celedón, J.C. and McGrath-Morrow, S.A., 2026. Community-acquired pneumonia in vulnerable populations across the Americas: ongoing barriers to prevention, diagnosis, and treatment. *Annals of the American Thoracic Society*, *23*(2), pp.153-160.

[4] Kayembe, J.M.N. and Kayembe, H.C.N., 2025. Challenges of Early Diagnosis and Effective Treatment of Pneumonia in Resource-Limited Settings. In *Viral Infectious Diseases Annual Volume 2025*. IntechOpen.

[5] Abedin, M.J., Nisha, J.R., Nahyan, N., Akter, S., Hridi, T.E.F. and Ahammed, M.T., 2025, July. Enhancing chest X-ray image analysis for pneumonia, normal, and COVID classification using deep learning. In *2025 International Conference on Quantum Photonics, Artificial Intelligence, and Networking (QPAIN)* (pp. 1-6). IEEE.

[6] Maniruzzaman, M., Sami, A., Hoque, R. and Mandal, P., 2024. Pneumonia prediction using deep learning in chest X-ray Images. *International Journal of Science and Research Archive*, *12*(1), pp.767-773. doi: 10.30574/ijsra.2024.12.1.0880.

[7] Alshanketi, F., Alharbi, A., Kuruvilla, M., Mahzoon, V., Siddiqui, S.T., Rana, N. and Tahir, A., 2025. Pneumonia detection from chest x-ray images using deep learning and transfer learning for imbalanced datasets. *Journal of imaging informatics in medicine*, *38*(4), pp.2021-2040. doi: 10.1007/s10278-024-01334-0.

[8] Colin, J. and Surantha, N., 2025. Interpretable deep learning for pneumonia detection using chest x-ray images. *Information*, *16*(1), p.53. doi: 10.3390/info16010053.

[9] Lakshmi, T.S.S. and Aluru, R.R., 2025. Pneumonia Detection Using Deep Learning: A CNN-Based Approach. *Review of Computer Engineering Studies*, *12*(2). doi: 10.18280/rces.120201.

[10] Mabrouk, A., Diaz Redondo, R.P., Dahou, A., Abd Elaziz, M. and Kayed, M., 2022. Pneumonia detection on chest X-ray images using ensemble of deep convolutional neural networks. *Applied Sciences*, *12*(13), p.6448. doi: 10.3390/app12136448.

[11] Singh, D.P., Banerjee, T., Durai C, A.D., Kumar J, S., Dutta, C., Dass, P., Tanwar, V.S., Kaushal, R.K., Singh, R.M., Narayan, Y. and Swain, D., 2026. A comprehensive study of various hybrid deep learning models for automated and explainable pneumonia detection in the pulmonary alveolar region: current insights and future directions. *Archives of Computational Methods in Engineering*, *33*(4), pp.5401-5433.

[12] Hashmi, M.F., Husain, A.A., Kotambkar, D.M. and Kumar, K.D., 2025. Hybrid deep learning and handcrafted feature fusion for pneumonia detection in chest X-rays. *Journal of Electrical Systems and Information Technology*, *12*(1), p.94.

[13] Alshanketi, F., Alharbi, A., Kuruvilla, M., Mahzoon, V., Siddiqui, S.T., Rana, N. and Tahir, A., 2025. Pneumonia detection from chest x-ray images using deep learning and transfer learning for imbalanced datasets. *Journal of imaging informatics in medicine*, *38*(4), pp.2021-2040.

[14] Rana, N. and Marwaha, H., 2025. Pneumonia detection from X-ray images using federated learning–an unsupervised learning approach. *Measurement: Sensors*, *37*, p.101410.

[15] Aljuaid, H., Adlan, H., Alkebsi, B., Alfurhood, B.S., Liotta, A. and Cavallaro, L., 2026. An experimental comparison of deep learning models for pneumonia classification from chest X-ray images. *Biomedical Signal Processing and Control*, *112*, p.108742.

[16] Arulananth, T.S., Prakash, S.W., Ayyasamy, R.K., Kavitha, V.P., Kuppusamy, P.G. and Chinnasamy, P., 2024. Classification of paediatric pneumonia using modified DenseNet-121 deep-learning model. *IEEE Access*, *12*, pp.35716-35727.

[17] Asnake, N.W., Salau, A.O. and Ayalew, A.M., 2024. X-ray image-based pneumonia detection and classification using deep learning. *Multimedia Tools and Applications*, *83*(21), pp.60789-60807.

[18] Ali, M., Shahroz, M., Akram, U., Mushtaq, M.F., Altamiranda, S.C., Obregon, S.A., Díez, I.D.L.T. and Ashraf, I., 2024. Pneumonia detection using chest radiographs with novel efficientnetv2l model. *IEEE Access*, *12*, pp.34691-34707.

[19] Siddiqi, R. and Javaid, S., 2024. Deep learning for pneumonia detection in chest x-ray images: A comprehensive survey. *Journal of imaging*, *10*(8), p.176.

[20] Agrawal, P. and Gandhi, S., 2025, July. PneumoSense: Smart pneumonia detection using deep learning. In *International Conference on ICT for Sustainable Development* (pp. 13-27). Cham: Springer Nature Switzerland.

[21] Yanar, E., Hardalaç, F. and Ayturan, K., 2025. PELM: a deep learning model for early detection of pneumonia in chest radiography. *Applied Sciences*, *15*(12), p.6487.

[22] Hashmi, M.F., Husain, A.A., Kotambkar, D.M. and Kumar, K.D., 2025. Hybrid deep learning and handcrafted feature fusion for pneumonia detection in chest X-rays. *Journal of Electrical Systems and Information Technology*, *12*(1), p.94.

[23] P. Mooney, "Chest X-Ray Images (Pneumonia)", Kaggle, 2018. [Online]. Available: https://www.kaggle.com/paultimothymooney/chest-xray-pneumonia

[24] Kermany, D.S., Goldbaum, M., Cai, W., Valentim, C.C., Liang, H., Baxter, S.L., McKeown, A., Yang, G., Wu, X., Yan, F. and Dong, J., 2018. Identifying medical diagnoses and treatable diseases by image-based deep learning. *cell*, *172*(5), pp.1122-1131.