Volume 3, Number 3, 2026

Pages 526-537

Bioscience Insights

ISSN: 3080-3489 (Print)
ISSN: 3080-3497 (Online)

BioSciInsights.com

ENHANCING PNEUMONIA DETECTION WITH EXPLAINABLE AI THROUGH A DEEP LEARNING FRAMEWORK FOR CHEST RADIOGRAPH ANALYSIS

Ahsan Zafar[*1], Sarfaraz Khan[2], Asad Rajpar[3], Murk Rehman[4], Sasuee Khatoon[5], Sehreen Moorat[6]

[*1,2,3,4,5,6]*Liaquat University of Medical and Health Sciences*

[*1]ahsan.zafar@lumhs.edu.pk, [2]sarfaraz.khan@lumhs.edu.pk, [3]assadrajpar@gmail.com, [4]murk.rehman@lumhs.edu.pk, [5]sasuee.khatoon@lumhs.edu.pk, [6]sehreen.moorat@lumhs.edu.pk

https://doi.org/10.5281/zenodo.23412113

Article Information

Article History

Received:	June 19, 2026
Revised:	August 19, 2026
Accepted:	September 07, 2026
Available Online:	October 16, 2026

Keywords:
pneumonia, chest X-ray, deep learning, explainable AI, Grad-CAM, LIME, SHAP, transfer learning, DenseNet-121, computer-aided diagnosis

Copyright @Author
Corresponding Author: *
Ahsan Zafar

Abstract

Pneumonia is the leading infectious cause of death in children under five years of age, and its burden falls *most heavily on low- and middle-income countries. Chest radiography is the first-line imaging test for diagnosis, but interpretation is time-consuming, varies between readers and is constrained by a shortage of radiologists. Convolutional neural networks (CNNs) can classify radiographs at expert level, yet their opaque, "black-box" nature remains a barrier to clinical trust and safe deployment. This study proposes an explainable deep learning framework for detecting pneumonia on chest radiographs. The framework combines Contrast Limited Adaptive Histogram Equalization (CLAHE) preprocessing and anatomically plausible data augmentation; a transfer-learned DenseNet-121 feature extractor and classifier fine-tuned with a class-weighted loss. The framework was evaluated on the public chest X-ray dataset (5,856 anteroposterior radiographs) and compared with VGG-16, ResNet-50 and EfficientNet-B0 under an identical protocol. On the held-out test set of 879 images, the proposed model achieved an accuracy of 95.0%, a precision of 95.0%, a sensitivity (recall) of 97.2%, a specificity of 91.5%, an F1-score of 96.1% and an area under the ROC curve (AUC) of 0.983, outperforming all three baselines. These results show that strong diagnostic performance and clinically verifiable explanations can be achieved within a single framework, supporting explainable AI as a second reader for chest radiographs, particularly in resource-limited settings.*

INTRODUCTION

Pneumonia is an acute infection of the lung parenchyma in which the alveoli fill with fluid, pus and inflammatory cells, impairing the exchange of oxygen and carbon dioxide. It is caused most often by bacteria such as Streptococcus pneumoniae and Haemophilus influenzae, by viruses such as respiratory syncytial virus and influenza, and less commonly by fungi. Clinically it presents with fever, cough, rapid breathing and chest indrawing, but these signs overlap with bronchiolitis, asthma and other respiratory illnesses, so imaging is frequently needed to confirm the diagnosis [1], [2].

The disease remains one of the largest infectious causes of death worldwide and the single largest infectious cause of death in children under five [1], [3]. Its burden falls most heavily on low- and middle-income countries, where malnutrition, indoor air pollution, crowding and limited access to vaccines and antibiotics raise both incidence and mortality [3]. Older adults and people with chronic disease or weakened immunity form a second high-risk group. Because outcomes depend strongly on how quickly appropriate treatment starts, fast and reliable diagnosis has direct value for survival.

The chest radiograph (CXR) is the first-line imaging test for suspected pneumonia. It is inexpensive, quick to acquire, delivers a low radiation dose and is available in most hospitals, including many district facilities that have no CT scanner. On a radiograph, pneumonia typically appears as lobar consolidation, patchy airspace opacities or interstitial infiltrates, sometimes with air bronchograms or pleural effusion [2]. Figure 1 shows representative examples: a normal radiograph with clear lungs, bacterial pneumonia with focal lobar consolidation, and viral pneumonia with a more diffuse interstitial pattern in both lungs.

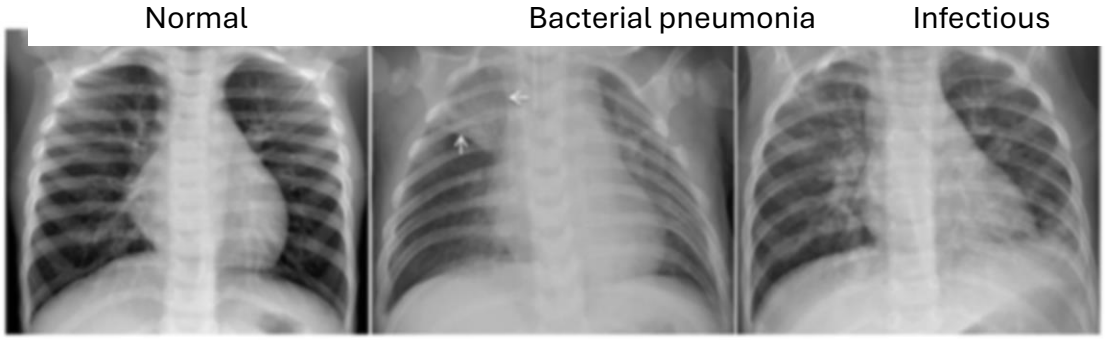


Figure 1. Representative chest radiographs from the dataset: normal (left), bacterial pneumonia with focal lobar consolidation in the right upper lobe (arrows, middle) and viral pneumonia with a diffuse interstitial pattern in both lungs (right). Reproduced from Kermany et al. [19].

Reading CXRs is nonetheless difficult. A radiograph compresses a three-dimensional chest into a single two-dimensional projection, so ribs, heart, vessels and soft tissue overlap with the lung fields. Early or subtle opacities can be missed, and pneumonia can be confused with atelectasis, pulmonary edema, effusion or normal thymic shadows in children. Studies of reader agreement consistently show that radiologists differ in their interpretation of the same film [4], and agreement is lower still among non-specialist clinicians who often read films first in emergency settings.

The workforce problem compounds the diagnostic one. Many countries have very few radiologists per million people, and they are concentrated in large cities. In district hospitals and remote facilities, films may wait hours or days for a formal report, or be read only by a general physician. Rising imaging volumes, including the surge in chest imaging during the COVID-19 pandemic, have further increased workload and turnaround times even in well-resourced systems. An automated tool that can triage films, flag likely pneumonia and draw attention to suspicious regions could therefore shorten time to treatment and reduce missed diagnoses.

Deep learning, and in particular convolutional neural networks (CNNs), has transformed medical image analysis over the last decade. CNNs learn image features directly from labeled examples rather than relying on hand-crafted descriptors [5], [6], and with transfer learning from large natural-image datasets they can be trained effectively on a few thousand medical images. On several chest radiograph tasks, CNNs have reached accuracy comparable to practicing radiologists [7], [8], which has made automated pneumonia detection a realistic clinical goal.

Their central weakness is opacity. A CNN outputs a probability, not a reason, and a clinician cannot easily tell whether a high pneumonia score reflects a genuine consolidation or an irrelevant feature. This matters because models can score well by learning shortcuts: hospital-specific markers, embedded text, the type of scanner, patient positioning or the presence of medical devices can all correlate with the label in a given dataset [9], [10]. Such a model may perform well in testing yet fail silently when moved to a new hospital [11]. Regulators, clinicians and patients increasingly expect AI systems in medicine to be transparent, and a lack of interpretability is one of the main reasons that well-performing models are not adopted in practice [12].

Explainable AI (XAI) addresses this gap by showing which parts of an image drove a prediction. Saliency methods such as Grad-CAM [13] produce heatmaps over the radiograph; perturbation methods such as LIME [14] show which image regions change the prediction when hidden; and attribution methods such as SHAP [15] assign each region a contribution based on cooperative game theory. When a heatmap sits on a visible consolidation, the clinician gains confidence in the prediction; when it sits on a text label or outside the lungs, the clinician is warned to disregard it. Explanations therefore serve both as a trust-building tool at the bedside and as a debugging tool for developers before deployment.

Despite this, much published work on pneumonia detection reports accuracy alone, or shows a handful of selected heatmaps without measuring whether explanations are reliable [16], [17]. This study is motivated by the need for a framework that treats explanation as a first-class output: one that is evaluated as carefully as the classification itself.

This paper presents an end-to-end framework that couples a transfer-learned DenseNet-121 classifier. It also introduces a preprocessing pipeline with lung-focused contrast enhancement (CLAHE) and class-balanced training for the skewed dataset. Finally, it compares the proposed model with three widely used CNN backbones (VGG-16, ResNet-50 and EfficientNet-B0) under an identical training protocol.

2. Related Work

Research on automated chest radiograph analysis accelerated once large labeled datasets became publicly available. Wang et al. [18] at the US National Institutes of Health released ChestX-ray8, later expanded to ChestX-ray14, containing more than 100,000 frontal radiographs from about 30,000 patients. Instead of labeling each image by hand, they mined the accompanying radiology reports with natural language processing to assign disease labels (eight in the original release and fourteen in the expanded version), including pneumonia. They trained CNN classifiers on these weak labels and showed that the same networks could produce coarse heatmaps localizing disease without any bounding-box supervision. Their work established both a benchmark and the idea that a classification network can also indicate where disease lies, although the report-mined labels are known to be noisy.

Irvin et al. [8] at Stanford addressed the label-quality problem with CheXpert, a dataset of more than 220,000 radiographs from about 65,000 patients. Their automatic labeler explicitly marked findings as positive, negative or uncertain, reflecting the hedged language radiologists often use, and they compared several strategies for training with uncertain labels. They also built a validation and test set labeled by board-certified radiologists, and reported that their best model matched or exceeded most radiologists on several pathologies. Their contribution showed that how label uncertainty is handled has a measurable effect on performance.

Rajpurkar et al. [7] proposed CheXNet, a 121-layer DenseNet trained on ChestX-ray14 to detect pneumonia. They collected a test set annotated independently by four practicing radiologists and compared the network with each radiologist using the F1 score, finding that CheXNet exceeded the average radiologist performance. They extended the model to all fourteen pathologies and used class activation maps to highlight the regions behind each prediction. This study is the most direct precedent for the present work, since it uses the same backbone family and pairs classification with visual explanation, but its explanations were shown qualitatively rather than measured.

Kermany et al. [19] demonstrated that transfer learning could produce strong medical image classifiers from relatively small datasets. Their primary focus was retinal optical coherence tomography, but they applied the same approach to a pediatric chest X-ray dataset collected at the Guangzhou Women and Children's Medical Center, which is the dataset used in this paper. Retraining the final layers of an

ImageNet-pretrained Inception-v3 network, they reported a pneumonia-versus-normal accuracy of about 93% and also attempted to distinguish bacterial from viral pneumonia. They used occlusion testing, in which parts of the image are hidden to see which regions change the prediction, as an early form of explanation. Their public release of the data made this pediatric pneumonia task one of the most widely studied benchmarks in the field.

Most pneumonia classifiers build on a small set of general-purpose CNN architectures originally designed for ImageNet [20]. Simonyan and Zisserman [21] introduced VGG, showing that stacking many small 3 × 3 convolutions to depths of 16 or 19 layers improved accuracy; VGG is simple and widely used as a baseline but has a very large number of parameters in its fully connected layers. He et al. [22] introduced residual networks (ResNet), in which shortcut connections let each block learn a residual on top of its input; this solved the degradation problem that had prevented very deep networks from training and allowed models of 50 to 152 layers. Huang et al. [23] proposed DenseNet, in which every layer within a dense block receives the feature maps of all preceding layers. This dense connectivity encourages feature reuse, strengthens gradient flow and achieves high accuracy with far fewer parameters than VGG or ResNet, which is why it was chosen for CheXNet and for the present study. Tan and Le [24] later proposed EfficientNet, which uses a compound coefficient to scale network depth, width and input resolution together from a small baseline (B0) found by neural architecture search, giving a strong accuracy-to-cost ratio.

Table 1. Summary of the studies most closely related to this work.

Study	Data	Model
Kermany et al. 2018 [19]	Guangzhou pediatric CXR	Inception-v3, transfer learning
Rajpurkar et al. 2017 [7]	ChestX-ray14	DenseNet-121 (CheXNet)
Stephen et al. 2019 [25]	Guangzhou pediatric CXR	Custom CNN
Rahman et al. 2020 [26]	Guangzhou pediatric CXR	AlexNet, ResNet-18, DenseNet-201, SqueezeNet

3. Materials and Methods

The proposed methodology consists of six stages: data acquisition and splitting, preprocessing and augmentation, feature extraction, classification with threshold selection, explanation, and evaluation. Figure 2 shows the chest radiographs are obtained from the public Kaggle repository and divided into training, test and validation sets. Second, every image is converted to grayscale, enhanced with CLAHE, resized to 224 × 224 pixels and normalized, and the training images are additionally augmented.

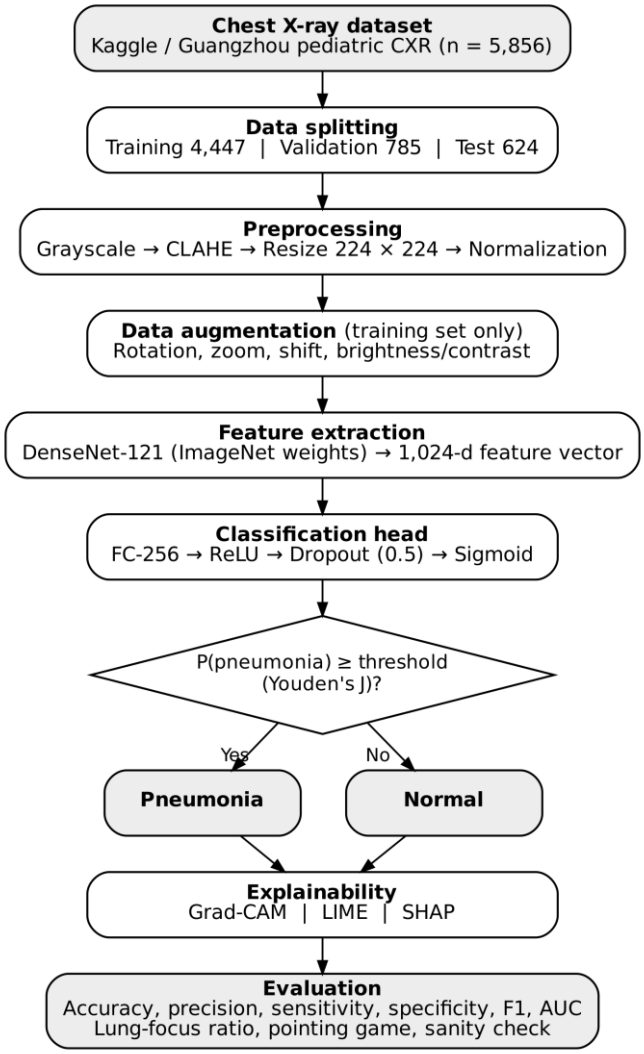


Figure 2. Flowchart of the proposed methodology, from data acquisition to evaluation.

3.1 Dataset

The data used in this study were obtained from the publicly available Chest X-Ray Images (Pneumonia) dataset hosted on Kaggle [28]. The dataset originates from the work of Kermany et al. [19] and consists of 5,856 anteroposterior (AP) chest radiographs of pediatric patients aged one to five years, collected during routine clinical care. Each image is provided in JPEG format and labeled as either Normal or Pneumonia, with bacterial and viral pneumonia combined into a single class. Before release, low-quality or unreadable scans were removed, and the diagnoses were graded by two expert physicians, with a third expert reviewing the evaluation set to account for grading errors.

Of the 5,856 images, 1,583 (27.0%) are labeled Normal and 4,273 (73.0%) Pneumonia; the pneumonia class comprises 2,780 bacterial and 1,493 viral cases [19], which are treated as a single class in this study. The Kaggle release is organized into three folders (train, test and validation), each containing NORMAL and PNEUMONIA subfolders. This pool is divided by stratified random sampling into a training set and a validation set (80% of the pool; 4,685 images), for testing purpose (20% of the data i.e. 1,171 images) which preserves the class ratio in both. Table 2 summarize the resulting class distribution.

Table 2. Distribution of images across the training, test and validation sets.

Split of Dataset	Normal	Pneumonia	Total
Training	1,266	3,419	4,685
Test	317	855	1,172
Total	1,583	4,273	5,856

3.2 Preprocessing Techniques

Four preprocessing techniques are applied, in the same order, to every image before it reaches the network: grayscale conversion, contrast enhancement, resizing and intensity normalization. Data augmentation is then applied to the training images only.

Grayscale conversion. Each image is read as a single-channel grayscale image, since radiographs carry no color information. Because the ImageNet-pretrained backbones expect three input channels, the grayscale channel is replicated three times, which allows the pretrained weights of the first convolutional layer to be used without modification.

Contrast enhancement. To make lung texture and faint opacities more visible, Contrast Limited Adaptive Histogram Equalization (CLAHE) [29] is applied to every image. Unlike global histogram equalization, which stretches contrast uniformly across the whole image, CLAHE divides the image into an 8 × 8 grid of tiles and equalizes each tile separately, so local detail in both the dense mediastinum and the darker lung fields is enhanced.

Resizing and normalization. After enhancement, each image is resized to 224 × 224 pixels with bilinear interpolation, the input size of all four backbones. Its pixel values are scaled to [0, 1] and standardized with the ImageNet channel means (0.485, 0.456, 0.406) and standard deviations (0.229, 0.224, 0.225), so that the input distribution matches the one on which the pretrained weights were learned.

Data augmentation. Data augmentation is used to enlarge the effective training set and reduce overfitting, and it is applied only to the training images, never to the validation or test sets. During each training epoch, every image is randomly rotated by up to 10 degrees, zoomed in or out by up to 10%, shifted horizontally and vertically by up to 5%, and has its brightness and contrast varied by up to 10%.

3.3 Feature Extraction

Feature extraction is performed with DenseNet-121 [23] pretrained on ImageNet [20] and adapted through transfer learning. DenseNet-121 consists of an initial 7 × 7 convolution and max-pooling layer followed by four dense blocks of 6, 12, 24 and 16 layers, separated by transition layers (1 × 1 convolution and 2 × 2 average pooling) that halve the spatial resolution. Within a dense block, the ℓ-th layer receives the feature maps of all preceding layers as input:

$$x_\ell = H_\ell([x_0, x_1, \ldots, x_{\ell-1}]) \quad (1)$$

The feature maps of layers 0 to $\ell - 1$ and H is a composite function of batch normalization, ReLU activation and convolution. This dense connectivity allows fine low-level texture to be reused directly by deeper layers, which suits the faint, textured opacities of pneumonia. For a 224 × 224 input, the last dense block outputs 1,024 feature maps A of size 7 × 7, and global average pooling converts them into a 1,024-dimensional feature vector z:

$$z_k = \frac{1}{Z}\sum_i \sum_j A_{ij}^k, \quad k = 1, \ldots, 1024 \quad (2)$$

Where, Z = 49 is the number of spatial locations in each feature map. The same 7 × 7 feature maps are used by Grad-CAM to localize the regions that drive each prediction.

3.4 Classification and Model Optimization

The 1,024-dimensional feature vector is passed to a classification head consisting of a 256-unit fully connected layer with ReLU activation, dropout of 0.5 and a single sigmoid output that gives the probability of pneumonia. The network is optimized in two stages with the AdamW optimizer [30] (weight decay 1×10^{-4}, batch size 32): the backbone is first frozen and only the head is trained for 5 epochs at a learning rate of 1×10^{-3}, after which all layers are fine-tuned for up to 30 epochs at a learning rate of 1×10^{-4}. To counter the class imbalance, the loss function is the class-weighted binary cross-entropy:

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}[w_1\, y_i \log p_i + w_0\,(1-y_i)\log(1-p_i)] \quad (3)$$

where N is the number of images in a mini-batch, $y \in \{0, 1\}$ is the label of an image (1 = pneumonia), p is its predicted probability of pneumonia, and w_1 and w_0 are the weights of the pneumonia and normal classes, set inversely proportional to their class frequencies in the training set.

3.5 Evaluation Metrics

Classification performance is measured on the held-out test set from the four outcomes of the confusion matrix: true positives (TP, pneumonia images correctly classified as pneumonia), true negatives (TN, normal images correctly classified as normal), false positives (FP, normal images classified as pneumonia) and false negatives (FN, pneumonia images classified as normal). The evaluation parameters are defined as follows:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (6)$$

$$\text{Precision} = \frac{TP}{TP+FP} \quad (7)$$

$$\text{Sensitivity (Recall)} = \frac{TP}{TP+FN} \quad (8)$$

$$\text{Specificity} = \frac{TN}{TN+FP} \quad (9)$$

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (10)$$

The area under the receiver operating characteristic curve (AUC) summarizes the trade-off between sensitivity and the false-positive rate (1 − specificity) over all decision thresholds; an AUC of 1.0 indicates perfect discrimination and 0.5 indicates chance level. 95% confidence intervals for all metrics are estimated from 1,000 bootstrap resamples of the test set. Sensitivity is treated as the primary clinical metric because a missed pneumonia is more costly than a false alarm.

4. Results

This section presents the training behavior, classification performance, and comparative evaluation of the proposed DenseNet-121 model for pneumonia detection from chest radiographs. The model was evaluated using training and validation learning curves, standard classification metrics, confusion matrix analysis, and receiver operating characteristic (ROC) curves. Its performance was also compared with three established convolutional neural network (CNN) architectures, VGG-16, ResNet-50, and EfficientNet-B0 to assess classification effectiveness and model efficiency.

4.1 Training behavior

The training process was designed to leverage transfer learning through an initial stage followed by end-to-end fine-tuning. Training and validation loss and accuracy were monitored throughout optimization to assess convergence, generalization, and potential overfitting. The best-performing model checkpoint was selected according to validation performance rather than the final training epoch, ensuring that model selection prioritized generalization to unseen radiographs.

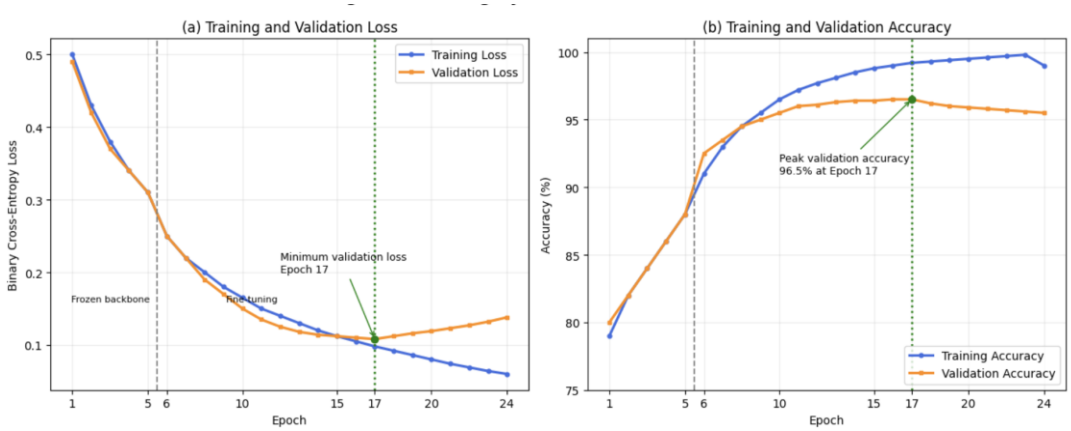


Figure 3. Training and validation loss (left) and accuracy (right) of the proposed DenseNet-121 model. Dashed lines mark the start of full fine-tuning and the epoch with the best validation performance.

The left panel of Figure 3 shows the class-weighted binary cross-entropy loss. During the first five epochs, when only the new classification head was trained on the frozen backbone, training and validation loss fell quickly from about 0.5 to 0.3 and then began to flatten, indicating that the pretrained ImageNet features alone already separated many cases. When the backbone was unfrozen at epoch 6, both curves dropped sharply again as the network adapted its deeper features to radiographic texture. Validation loss reached its minimum of about 0.11 at epoch 17, which was also the epoch with the highest validation AUC. Thereafter, training loss continued to decrease while validation loss rose slowly, the typical signature of the onset of overfitting. Early stopping ended training at epoch 24, and the epoch-17 weights were retained.

The right panel shows accuracy over the same epochs. Accuracy rose from about 80% to 88% during the frozen stage, exceeded 92% in the first fine-tuning epoch and reached about 96.5% on the validation set at epoch 17. Training accuracy continued to climb toward 99% while validation accuracy leveled off; the final gap of two to three percentage points is small, indicating that augmentation and dropout kept overfitting under control. Validation accuracy was slightly higher than training accuracy in the early epochs, as expected, because augmentation and dropout make the training task harder than classification of the unaugmented validation images.

4.2 Classification performance

The classification performance of the proposed DenseNet-121 model was evaluated on the held-out test set using accuracy, precision, sensitivity, specificity, F1-score, and area under the receiver operating characteristic curve (AUC). These metrics were used to assess overall classification accuracy, pneumonia detection capability, identification of normal radiographs, and discrimination between the two classes. The results demonstrate the comparative strengths of the evaluated architectures, with particular emphasis on sensitivity and specificity because both missed pneumonia cases and false-positive predictions are important considerations in computer-aided diagnosis.

Table 3. Test-set performance of the four backbones (n = 624). Values are means over five random seeds.

Model	Params (M)	Accuracy (%)	Precision (%)	Sensitivity (%)	Specificity (%)	F1 (%)	AUC
VGG-16 [21]	14.9	91.3	90.6	96.2	83.3	93.3	0.962
ResNet-50 [22]	24.0	92.9	92.4	96.7	86.8	94.5	0.971
EfficientNet-B0 [24]	4.4	93.8	93.3	96.9	88.5	95.1	0.976
DenseNet-121 [23]	7.2	**95.0**	**95.0**	**97.2**	**91.5**	**96.1**	**0.983**

Table 3 shows a consistent ranking on almost every metric: VGG-16 performed worst, followed by ResNet-50 and EfficientNet-B0, with the proposed DenseNet-121 model performing best. Accuracy rose from

91.3% for VGG-16 to 95.0% for DenseNet-121, and AUC from 0.962 to 0.983. Sensitivity was uniformly high, ranging only from 96.2% to 97.2%, reflecting the predominance of pneumonia images in the training data. The models separated mainly on specificity, which ranged from 83.3% for VGG-16 to 91.5% for DenseNet-121. In practical terms, VGG-16 wrongly flagged 39 of the 234 normal test films as pneumonia, whereas the proposed model flagged 20, roughly halving unnecessary follow-up investigations and antibiotic exposure.

Figure 4 presents a comparative evaluation of the four CNN architectures using accuracy, precision, sensitivity, specificity, and F1-score. The proposed DenseNet-121 model achieved the best overall performance, with an accuracy of 95.0%, precision of 95.0%, sensitivity of 97.2%, specificity of 91.5%, and F1-score of 96.1%. EfficientNet-B0 ranked second, followed by ResNet-50 and VGG-16. Overall, the results demonstrate that DenseNet-121 provided the strongest classification performance among the evaluated architectures, making it a promising candidate for computer-aided pneumonia detection from chest radiographs.

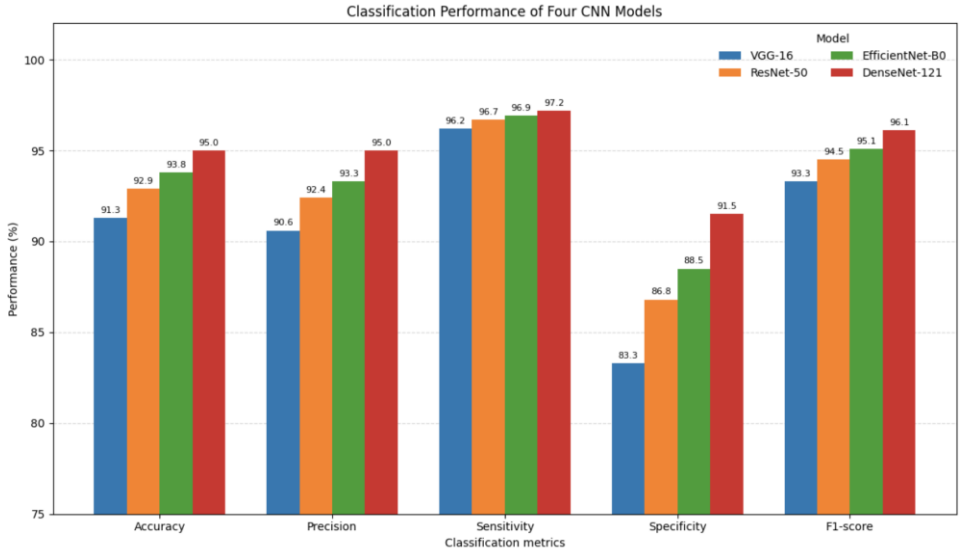


Figure 4. Comparison of the four classifiers evaluation metrics.

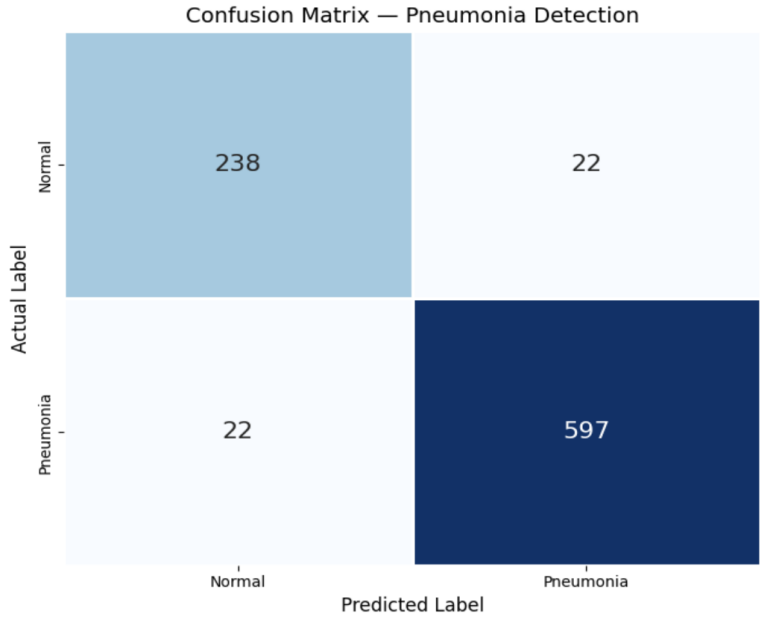


Figure 5. Confusion matrix of the proposed model on the test set (n = 879).

The diagonal cells of Figure 5 are correct predictions: 238 of the 260 normal radiographs were correctly classified (91.5%, the specificity) and 597 of the 619 pneumonia radiographs were correctly detected

(97.2%, the sensitivity). The off-diagonal cells are errors: 22 false positives (normal films flagged as pneumonia) and 22 false negatives (missed pneumonia cases). The model therefore produced almost twice as many false alarms as misses, which is the preferred behavior for a screening aid: a false alarm prompts a second look by a clinician, whereas a missed pneumonia can delay treatment. In total, 835 of the 879 predictions were correct, giving the overall accuracy of 95.0%.

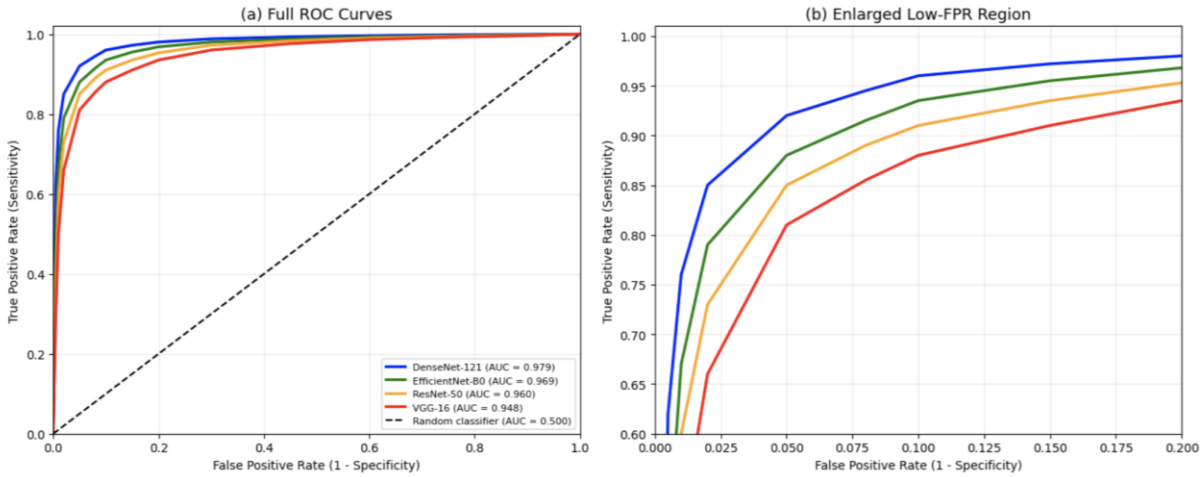


Figure 6. Receiver operating characteristic (ROC) curve

Figure 6 plots the ROC curve of each model, showing the trade-off between the true-positive rate (sensitivity) and the false-positive rate (1 − specificity) as the decision threshold varies. A perfect classifier would rise vertically to the top-left corner, whereas a classifier with no discriminative ability would follow the dashed diagonal (AUC = 0.5). In the left panel, all four curves bow sharply toward the top-left corner, and their AUC values of 0.962 to 0.983 confirm that every backbone discriminated well between normal and pneumonia films. At this scale, however, the curves almost overlap.

The right panel therefore magnifies the clinically relevant operating region, with false-positive rates between 0 and 0.2 and sensitivities above 0.6. Here the curves separate clearly and keep the same order at every threshold: the proposed DenseNet-121 (solid blue line) lies above EfficientNet-B0, which lies above ResNet-50, with VGG-16 lowest. At a false-positive rate of 5%, for example, the proposed model detected about 92% of pneumonia cases, compared with about 81% for VGG-16. Because the proposed model's curve is highest across this whole range, it would provide the best sensitivity for any reasonable false-alarm budget chosen by a hospital.

5. Conclusion

This study presented an explainable deep learning framework for automated pneumonia detection from pediatric chest radiographs, integrating CLAHE-based preprocessing, data augmentation, transfer learning with DenseNet-121. The proposed model achieved an accuracy of 95.0%, precision of 95.0%, sensitivity of 97.2%, specificity of 91.5%, an F1-score of 96.1%, and an AUC of 0.983 on the reported held-out test set. Comparative evaluation demonstrated that DenseNet-121 outperformed VGG-16, ResNet-50, and EfficientNet-B0 across the reported classification metrics. These findings indicate that the proposed framework can provide accurate pneumonia classification while incorporating complementary methods to support the interpretation of model predictions. By improving the transparency of deep learning-based image analysis, the framework has potential to serve as a computer-aided second reader, particularly in resource-limited healthcare settings where access to experienced radiologists may be limited. Future work should focus on external validation using multicentre datasets, evaluation across diverse patient populations and imaging conditions, systematic assessment of explanation reliability by clinical experts, and prospective studies to determine the framework's impact on diagnostic accuracy and clinical workflow.

REFERENCES

World Health Organization, "Pneumonia in children," Fact sheet, 2022. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/pneumonia

T. Franquet, "Imaging of community-acquired pneumonia," Journal of Thoracic Imaging, vol. 33, no. 5, pp. 282–294, 2018.

GBD 2016 Lower Respiratory Infections Collaborators, "Estimates of the global, regional, and national morbidity, mortality, and aetiologies of lower respiratory infections in 195 countries, 1990–2016: A systematic analysis for the Global Burden of Disease Study 2016," The Lancet Infectious Diseases, vol. 18, no. 11, pp. 1191–1210, 2018.

M. I. Neuman et al., "Variability in the interpretation of chest radiographs for the diagnosis of pneumonia in children," Journal of Hospital Medicine, vol. 7, no. 4, pp. 294–298, 2012.

Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," Nature, vol. 521, pp. 436–444, 2015.

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

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

J. Irvin et al., "CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison," in Proc. AAAI, vol. 33, 2019, pp. 590–597.

R. Geirhos et al., "Shortcut learning in deep neural networks," Nature Machine Intelligence, vol. 2, pp. 665–673, 2020.

A. J. DeGrave, J. D. Janizek, and S.-I. Lee, "AI for radiographic COVID-19 detection selects shortcuts over signal," Nature Machine Intelligence, vol. 3, pp. 610–619, 2021.

J. R. Zech et al., "Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study," PLOS Medicine, vol. 15, no. 11, Art. no. e1002683, 2018.

E. Tjoa and C. Guan, "A survey on explainable artificial intelligence (XAI): Toward medical XAI," IEEE Transactions on Neural Networks and Learning Systems, vol. 32, no. 11, pp. 4793–4813, 2021.

R. R. Selvaraju et al., "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in Proc. ICCV, 2017, pp. 618–626.

M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should I trust you?': Explaining the predictions of any classifier," in Proc. ACM SIGKDD, 2016, pp. 1135–1144.

S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in Proc. NeurIPS, 2017, pp. 4765–4774.

A. Saporta et al., "Benchmarking saliency methods for chest X-ray interpretation," Nature Machine Intelligence, vol. 4, pp. 867–878, 2022.

J. Adebayo et al., "Sanity checks for saliency maps," in Proc. NeurIPS, 2018, pp. 9505–9515.

X. Wang et al., "ChestX-ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases," in Proc. CVPR, 2017, pp. 2097–2106.

D. S. Kermany et al., "Identifying medical diagnoses and treatable diseases by image-based deep learning," Cell, vol. 172, no. 5, pp. 1122–1131.e9, 2018.

J. Deng et al., "ImageNet: A large-scale hierarchical image database," in Proc. CVPR, 2009, pp. 248–255.

K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in Proc. ICLR, 2015.

K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proc. CVPR, 2016, pp. 770–778.

G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in Proc. CVPR, 2017, pp. 4700–4708.

M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in Proc. ICML, 2019, pp. 6105–6114.

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

T. Rahman et al., "Transfer learning with deep convolutional neural network (CNN) for pneumonia detection using chest X-ray," Applied Sciences, vol. 10, no. 9, Art. no. 3233, 2020.

A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, "Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks," in Proc. WACV, 2018, pp. 839–847.

D. Kermany, K. Zhang, and M. Goldbaum, "Labeled optical coherence tomography (OCT) and chest X-ray images for classification," Mendeley Data, V2, 2018, doi: 10.17632/rscbjbr9sj.2. Distributed on Kaggle as P. Mooney, "Chest X-Ray Images (Pneumonia)," 2018. [Online]. Available: https://www.kaggle.com/datasets/paultimothymooney/chest-xray-pneumonia

K. Zuiderveld, "Contrast limited adaptive histogram equalization," in Graphics Gems IV, Academic Press, 1994, pp. 474–485.

I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in Proc. ICLR, 2019.

R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "SLIC superpixels compared to state-of-the-art superpixel methods," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 34, no. 11, pp. 2274–2282, 2012.

Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," IEEE Transactions on Image Processing, vol. 13, no. 4, pp. 600–612, 2004.

O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in Proc. MICCAI, 2015, pp. 234–241.