Jurnal Teknik Informatika (JUTIF)
P-ISSN: 2723-3863
E-ISSN: 2723-3871

Vol. 6, No. 6, December 2025, Page. 5837-5853
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.6.5404

An Interpretable Deep Learning Framework for Multi-Class Lung Disease Diagnosis Using ConvNeXt Architecture

Muhammad Khalidin Basyir*[1], Mhd Furqan[2], Aulia Fadlan[3]

[1,2]Department of Computer Science, Universitas Islam Negeri Sumatera Utara, Indonesia
[3]Master of Data Science Programme, Faculty of Computer Science and Information Technology, Universiti Malaya, Malaysia

Email: [1]muhammadkhalidinbasyir@gmail.com

Received : Oct 28, 2025; Revised : Nov 18, 2025; Accepted : Nov 18, 2025; Published : Dec 31, 2025

Abstract

Lung diseases remain a major global health challenge, requiring accurate and interpretable diagnostic systems to support timely detection and treatment. This study proposes a high-fidelity deep learning approach using the ConvNeXt architecture for automated multi-class classification of chest X-ray (CXR) images into five categories: Bacterial Pneumonia, Viral Pneumonia, COVID-19, Tuberculosis, and Normal. The methodology involved preprocessing 10.095 Kaggle-sourced images (normalization, CLAHE, augmentation, resizing) and training a ConvNeXt model for 70 epochs with the Adam optimizer. The model achieved strong performance with 92.66% validation accuracy, 86.32% test accuracy, a macro-average F1-score of 0.86, and a macro-average AUC of 0.99. Grad-CAM visualizations demonstrated the model's consistent focus on clinically relevant lung regions, significantly improving interpretability and clinical applicability. This study contributes to advancing interpretable AI methods for clinical decision support in medical imaging, offering a reliable and transparent framework for automated lung disease diagnosis.

Keywords : *Chest X-Ray, ConvNeXt, Deep Learning, Grad-CAM, Lung Disease*

1. INTRODUCTION

Lung diseases remain a significant global health problem, with high morbidity and mortality rates and a widespread impact on patients' quality of life. According to the latest WHO report, the Eastern Mediterranean region accounted for approximately 8.7% of global tuberculosis (TB) cases, with 936.000 new cases and nearly 86.000 deaths in 2023 [1]. Besides TB, pneumonia is also a leading cause of death, particularly among children. UNICEF (2025) reported that pneumonia kills more than 700.000 children under the age of five annually, including about 190.000 newborns, equivalent to 2.000 child deaths per day [2]. This data underscores the critical importance of early detection and accurate diagnosis of lung diseases.

A chest X-ray (CXR) serves as a highly utilized diagnostic method for identifying pulmonary conditions. Beyond its ready availability and modest cost, CXR imaging offers a visual assessment of both the lungs and adjacent structures. Despite its efficacy, the manual analysis of CXR scans is a laborious process contingent upon the interpreter's skill, thereby indicating a need for a more streamlined, automated methodology [3]

Significant progress in artificial intelligence (AI), especially in deep learning techniques, has created novel possibilities within medical image analysis. A highly notable approach used is the Convolutional Neural Network (CNN), which has proven effective in classification and object detection tasks in medical images [4]. Convolutional Neural Networks (CNNs) have been demonstrated to be effective in extracting spatial features from two-dimensional images like X-rays [5]. Furthermore, CNNs can automatically recognize patterns and important features in images that are often difficult for human observers to identify [6].

Based on the journal introduction above, several relevant studies have been conducted. Souid et al. (2021) highlighted the limitations of radiologists in diagnosing lung diseases through X-ray images and proposed the use of MobileNet V2 with a transfer learning approach on the ChestX-ray14 dataset. Their results showed an accuracy of over 90% with an average AUC value of 0.811, confirming the effectiveness of this lightweight model for implementation on IoT devices [7]. Hasanah et al. (2023) conducted a comparative study of ResNet-50, ResNet-101, and ResNet-152 architectures for identifying pneumonia, COVID-19, and lung opacities using 21.885 CXR images. They reported that ResNet-152 delivered the best performance with an F1-score of 94%, higher than ResNet-50 (91%) and ResNet-101 (93%), thus recommending it for CXR-based lung disease classification [8]. Meanwhile, Shamrat et al. (2022) developed LungNet22, a customized model based on VGG16, for the multi-class classification of 10 lung diseases, including COVID-19, TB, and pneumonia, using a dataset of over 80.000 X-ray images. This model achieved a very high accuracy of 98.89%, with other evaluation metrics such as precision, recall, F1-score, and ROC-AUC reinforcing its reliability [9]. These findings confirm that the application of various CNN architectures, both lightweight and deeper models, can deliver superior performance in the automated detection of lung diseases from X-ray images.

Similarly, Izdihar et al. (2024) compared VGG16 and ResNet50 for pneumonia detection, showing that ResNet50 outperformed VGG16 in accuracy and processing time [10]. Jiang et al. (2021) proposed an improved VGG13 model (IVGG13) with data augmentation to address imbalanced datasets, achieving better precision, recall, and F1 scores compared to standard CNN models [11]. Recently, Bundea and Danciu (2024) used the DenseNet architecture (DenseNet121, DenseNet169, and DenseNet201), reporting an accuracy rate of 92% for normal cases and 97% for pneumonia cases, highlighting the robustness and efficiency of DenseNet in clinical decision support [12]. These findings collectively demonstrate that diverse CNN architectures, ranging from lightweight models to deeper models, can significantly improve the automated diagnosis of lung diseases from CXR images.

The ConvNeXt architecture, introduced by Liu et al. (2022), represents a cutting-edge development in the field of computer vision. Developed in response to the success of Transformer-based models, ConvNeXt successfully adapts Transformer design principles into a conventional CNN architecture while delivering significant performance improvements. The advantage of ConvNeXt lies in its ability to achieve performance competitive with the latest Transformer models, but with greater computational efficiency and ease of implementation within existing CNN ecosystems. This study posits that ConvNeXt offers a distinct advantage over widely used architectures like ResNet and DenseNet for medical image analysis, particularly in achieving a superior balance between accuracy and inherent interpretability. While ResNet and DenseNet excel through feature reuse and depth, their complex, highly interconnected feature maps can be challenging to interpret. In contrast, ConvNeXt's modernized design—featuring large kernel depthsise convolutions and a simplified, stage-based structure—generates cleaner and more spatially coherent feature representations. This architectural clarity naturally facilitates the generation of more precise and clinically meaningful visual explanations using techniques like Grad-CAM, a crucial factor for building trust in clinical decision support systems. In the context of medical image analysis, architectures like ConvNeXt have the potential to offer better accuracy and efficiency in disease classification tasks, including the classification of lung diseases from X-ray images

[13].

This investigation seeks to formulate and assess the effectiveness of the ConvNeXt deep learning framework for categorizing pulmonary ailments using radiographic images. Additionally, this research endeavors to pinpoint elements impacting the precision of lung disease classification and provide recommendations for the development of automated diag-nosis systems based on X-ray images.

2. METHOD

The research framework of this study consists of several systematic stages, starting from dataset collection obtained from Kaggle, followed by pre-processing which includes normalization, CLAHE, augmentation, and resizing to prepare the images for model input. The processed data is then trained using the ConvNeXt architecture, which incorporates convolutional, normalization, activation, and downsampling blocks, with the final layer using softmax to produce classification probabilities. The training phase of the model utilized the Adam optimizer and categorical cross-entropy as the loss function, with the training process spanning 70 epochs. Model performance was subsequently assessed using metrics such as accuracy, precision, recall, F1-score, a confusion matrix, and the ROC-AUC. Following its training, the model underwent prediction testing on novel datasets. To elucidate the model's decision-making rationale, Grad-CAM visualization techniques were applied, specifically focusing on identifying and emphasizing pertinent areas within the lung scans.

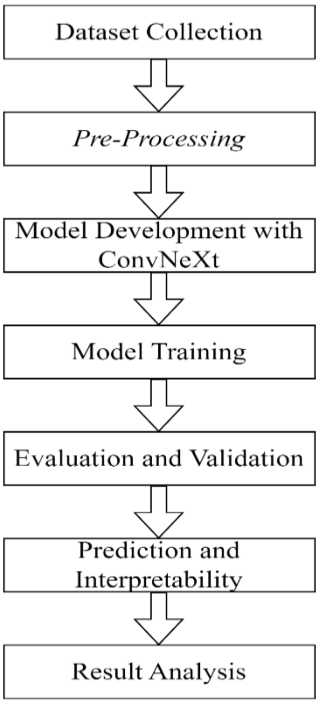

Dataset Collection
Pre-Processing
Model Development with ConvNeXt
Model Training
Evaluation and Validation
Prediction and Interpretability
Result Analysis

Figure 1. Research Framework

2.1. Dataset

The dataset used was obtained from the Kaggle website, titled "Lungs Disease Dataset (4 types)," published by Omkar Manohar Dalvi. The lung disease data set used in this study was divided into five classes: Bacterial Pneumonia, Viral Pneumonia, Coronavirus Disease, Tuberculosis, and Normal. The images collected have varying dimensions and are in JPEG format. These images are used separately in the training, validation, and testing processes. The amount of training data, testing data, and validation data can be seen in Table 1.

Table 1. Number of Dataset

Class	Training	Testing	Validation
Bacterial Pneumonia	1205	403	401
Viral Pneumonia	1204	403	401
Corona Virus Disease	1218	407	406
Tuberculosis	1220	408	406
Normal	1207	404	402
Total	6054 Figure	2025 Figure	2016 Figure

2.2. Pre-Processing

The Before being used as input in training, the image goes through a pre-processing stage that aims to make it easier for the ConvNeXt model to train and recognize features in the input image [14]. Before the CNN algorithm can process the image, a number of image pre-processing steps are performed. The following are the stages of image pre-processing:

1. Image Normalization

This normalization helps reduce excessive pixel value variation and improves deep learning model convergence during training [15]. Normalization is performed by changing the image pixel value range from [0, 255] to [-1, 1].

2. CLAHE (Contrast Limited Adaptive Histogram Equalization)

CLAHE is used to enhance local contrast in images, especially in areas with low lighting or uneven contrast [16]. This technique divides the image into several small regions (tiles) and performs histogram equalization on each region, with a clip limit to prevent excessive noise increase. With CLAHE, details in dark or bright areas can be more visible, making it easier for CNN to recognize patterns and important features in the image.

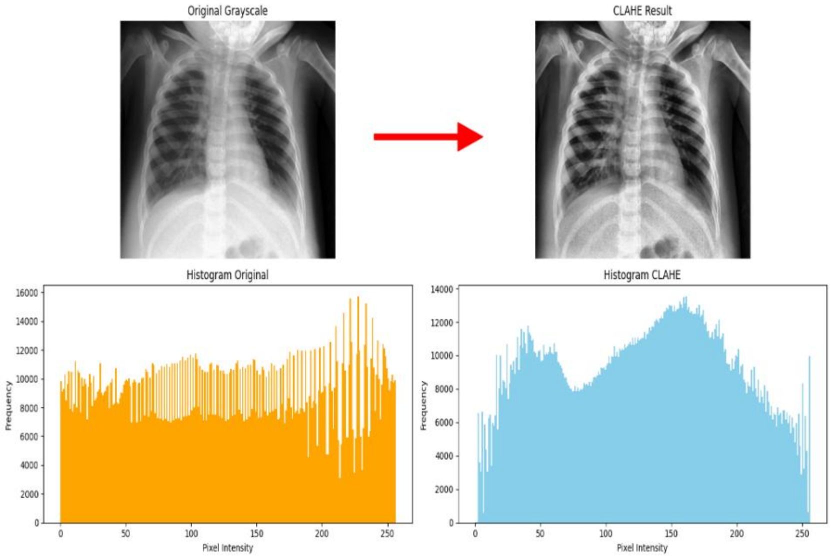


Figure 2. CLAHE

3. Image Augmentation

Image augmentation to help models become stronger and generalize better to new images that have never been seen before, which can reduce overfitting [17]. The following table shows the image augmentation process.

Table 2. Image Augmentation

Method	Default	Adjustment
Rotation	0	30
Width Shift	0	0.2
Height Shift	0	0.2
Shear	0	0.2
Zoom	-	0.2
Horizontal Flip	None	True
Fill Mode	None	Nearest

4. Image Resizing

Resizing images in preprocessing for Convolutional Neural Networks (CNN) is an important step that aims to convert images of different sizes into the same size so that they can be processed consistently by CNN. This process adjusts the pixel size of the entire dataset to 224 × 224, in accordance with the input standard used in the ConvNeXt architecture for training on the ImageNet dataset, so that the model can optimize visual feature representation consistently [18].

2.3. ConvNeXt Architecture

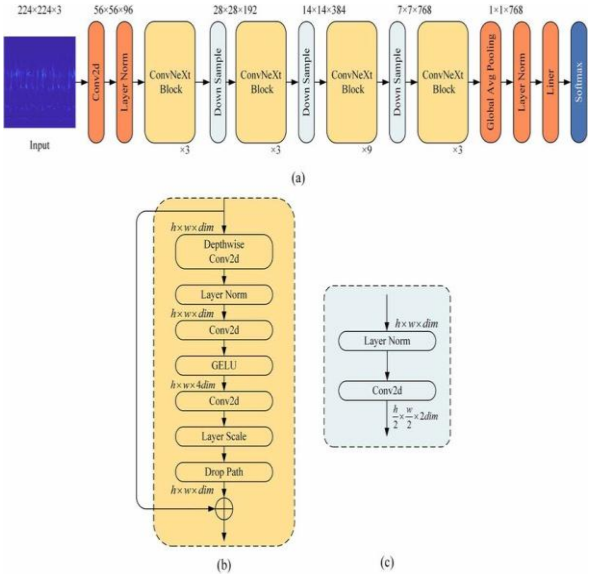


Figure 3. ConvNeXt Architecture

2.3.1. Conv2D

Conv2d in ConvNeXt is a two-dimensional convolution operation that serves to convert input images into more informative, efficient, and structured feature representations. In the initial layer, Conv2d acts as a patch embedding with a specific kernel and stride to reduce image resolution while increasing the number of channels, thereby transforming raw images into initial feature maps.

$$y_{i,j,k_{out}} = \sum_{m=-1}^{+1} \sum_{n=-1}^{+1} \sum_{k_{in}=1}^{+1} W_{m,n,k_{in},k_{out}} \cdot X_{i+m,j+n,k_{in}} \qquad (1)$$

2.3.2. Layer Normalization

Layer Normalization after Conv2d in the ConvNeXt architecture serves to stabilize the activation distribution by normalizing the feature map output values at each spatial position based on the mean and variance in the feature dimension. This process helps maintain the stability of activation values so that they are not too large or too small, speeds up convergence during training, and makes the model more robust against data variations. Unlike Batch Normalization, which is highly dependent on batch size, Layer Normalization is more flexible because it works per layer, so it remains effective even with small batch sizes.

$$\hat{x} = \frac{x_i - \mu}{\sqrt{\sigma^2 + \in}} \tag{2}$$

2.3.3. ConvNeXt Block

1. Depthwise Conv2d

Depthwise Conv2d processes each input channel separately, rather than mixing all channels together like regular Conv2d. The goal is to reduce the number of parameters and computational complexity, while still capturing the spatial patterns of each feature channel for greater efficiency.

$$Y_{i,j,c} = \sum_{m=0}^{k_h-1} \sum_{n=0}^{k_w-1} w_{m,n,c} \cdot X_{i+m,j+n,c} \tag{3}$$

2. Layer Normalization

Layer Normalization normalizes activation values in feature dimensions to stabilize data distribution. This helps accelerate convergence during training, reduces dependence on weight initialization, and maintains gradient flow stability, making the model easier to train.

$$\hat{x} = \frac{x_i - \mu}{\sqrt{\sigma^2 + \in}} \tag{4}$$

3. Conv2d

Conv2d is used at this stage to project the depthwise conv feature results into a higher-dimensional space. This operation combines information between channels, thereby enriching the feature representation, which is more complex than just spatial per channel.

$$Y_{i,j,f} = \sum_{c=1}^{C_{in}} W_{c,f} \cdot X_{i,j,c} + b_f \tag{5}$$

4. GELU

GELU (Gaussian Error Linear Unit) is a non-linear activation function that is smoother than ReLU. GELU suppresses negative values with probability, rather than cutting them off immediately, thereby helping models learn richer representations and improving performance in various vision tasks [19].

$$GELU(x) = 0.5 \cdot x \cdot \left(tanh \left[\sqrt{\frac{2}{\pi}} \ (x + 0.044715x^3 \right] \right) \tag{6}$$

5. Conv2d

The second Conv2d serves to restore the feature dimensions to their original form after they have been enlarged (usually 4×). In this way, the block can capture complex non-linear representations in high-dimensional space while remaining efficient by restoring the output size to match the input.

$$Z_{i,j,f} = \sum_{c=1}^{4 \cdot C_{in}} W'_{c,f} \cdot Y_{i,j,c} + b'_f \tag{7}$$

6. Layer Scale

Layer Scale is a trainable scalar parameter used to balance the output contribution of blocks. It is initially initialized with a small value to make the network more stable at the beginning of training, then its value can evolve according to the model's needs.

$$Y = X + \gamma \cdot F(X) \qquad (8)$$

7. Drop Path

Drop Path is a form of regularization in the form of stochastic depth, where certain residual paths are randomly "removed" during training. This technique prevents overfitting, improves generalization, and makes the model more robust by training several alternative information paths.

$$Drop\ Path(x) = \begin{cases} 0, & \text{with probability p} \\ \frac{\text{x}}{1-\text{p}}, & \text{with probability } 1-\text{p} \end{cases} \qquad (9)$$

2.3.4. Down Sample

1. Layer Normalization

Layer Normalization serves to normalize feature activation values in each channel, so that the distribution of input values to the next layer becomes more stable. This process helps reduce internal covariate shift, speeds up convergence during training, and makes the network more resistant to data variation. In the Downsample Block, Layer Norm is performed before the convolution process so that the data entering Conv2d is already on a consistent scale. Layer Normalization serves to normalize feature activation values on each channel, so that the distribution of input values to the next layer becomes more stable. This process helps reduce internal covariate shift, speeds up convergence during training, and makes the network more resistant to data variation. In the Downsample Block, Layer Norm is performed before the convolution process so that the data entering Conv2d is already on a consistent scale.

$$\hat{x} = \frac{x_i - \mu}{\sqrt{\sigma^2 + \epsilon}} \qquad (10)$$

2. Conv2d

Conv2d is used here with a stride of 2 to perform downsampling, which reduces the spatial resolution of features (height and width are halved) while increasing the number of channels (from dim to 2×dim). This aims to reduce computational complexity, enlarge the receptive field, and extract deeper feature representations so that the model is able to capture more complex patterns in the next stage.

$$Y(i,j,c_{out}) = \sum_{m=0}^{k_{h-1}} \sum_{n=0}^{k_{h-1}} \sum_{c_{in}=0}^{C_{in-1}} X(i \cdot s + m - p, j \cdot s + n - p, C_{in}) \cdot K(m,n,C_{in},C_{out}) \qquad (11)$$

2.3.5. Global Average Pooling

Global Average Pooling (GAP) is a pooling technique that serves to reduce the dimensions of the feature map by calculating the average value of all elements in each channel, so that each channel produces a single value.

$$\boldsymbol{GAP\ (c)\ = \frac{1}{H \times W} + 1 \sum_{i=1}^{H} \sum_{j=1}^{W} x_{i,j,c}} \qquad (12)$$

2.3.6. Layer Normalization

Layer Normalization after Global Average Pooling serves to normalize the distribution of feature values from the pooling result vector so that each feature has a balanced scale, making the model more stable during training and reducing sensitivity to variations in value between features.

$$\hat{x} = \frac{x_i - \mu}{\sqrt{\sigma^2 + \epsilon}} \qquad (13)$$

2.3.7. Linear

The Linear layer after the Normalization Layer serves to perform the final mapping of the feature representation extracted by the CNN to the desired class output space. After the Norm Layer normalizes the feature distribution to make it stable and balanced, the Linear layer (fully connected layer) will convert feature vectors of a certain dimension into logit scores for each class.

$$Z = W \cdot x + b \tag{14}$$

2.3.8. Softmax

Softmax after the Linear Layer serves to convert the output from the Linear Layer, which is in the form of logit values (unbounded real numbers), into a probability distribution for each class. The Linear Layer produces raw scores for each class, then Softmax normalizes them using an exponential function so that all output values are in the range of 0–1 and their total sum is equal to 1. In this way, Softmax makes it easier to interpret the results as prediction probabilities, so that the class with the highest probability can be selected as the final output in the classification process.

$$Softmax(z_i) = \frac{e^{zi}}{\sum_{j=1}^{n} e^{zj}} \tag{15}$$

2.4. Training

The model training process was carried out with predetermined parameters to achieve maximum results. The researchers used 70 epochs, which was considered sufficient to improve model performance [20]. The learning rate used was set at 0.0001, which is a common value often used in deep learning model training. The optimizer chosen was Adam, which is known to be effective in handling large and complex datasets and can adapt dynamically to different learning rates. Adam works by combining the advantages of two previous methods, namely RMSprop and Stochastic Gradient Descent (SGD) with momentum [21]. Adam utilizes adaptive estimates of the first moment (average gradient) and second moment (average gradient squared) to improve network weights.

To prevent model overfitting during the training phase, the hold-out validation methodology can be employed. This approach entails partitioning the available dataset into two primary segments: a training set and a validation set. The model undergoes training utilizing the training data, and its subsequent performance is assessed against the validation data, which remains segregated from the training procedures [22]. By employing this method, one can effectively evaluate the model's capabilities and confirm its proficiency in generalizing to novel, unencountered data.

The loss function used is Categorical Crossentropy, as this is a multi-class classification problem with more than two classes [23]. The evaluation metrics monitored during training include accuracy and loss, which provide an initial overview of the model's performance [24]. We also utilize GPUs to accelerate the training process, given the high computational load of CNN models. The use of GPUs not only reduces training time but also allows for experimentation with larger batch sizes, improving the stability of the training process.

2.5. Model Evaluation

Upon conclusion of the training phase, the subsequent action involves assessing the model's performance against the pre-segregated test dataset. This assessment will utilize several key performance indicators, including accuracy, precision, recall, the F1-score, and a confusion matrix. The primary objective is to ascertain the model's efficacy in adapting to previously unseen data [25]. The precise mathematical definitions for each evaluation metric employed in this research endeavor are delineated in the subsequent formulae.

$$\boldsymbol{Accuracy} = \frac{\boldsymbol{TP+TN}}{\boldsymbol{TP+TN+FP+FN}} \quad (16)$$

Accuracy is a comparative value obtained from the amount of data predicted correctly against the total data.

$$Precission = \frac{TP}{TP+ FP} \quad (17)$$

Precision is the ratio of the total number of positive data divided by the correct data.

$$Recall = \frac{TP}{TP+ \mathrm{FN}} \quad (18)$$

Recall is the result of the total number of positive data divided by the correct data and incorrect data.

$$F1 - Score = 2 \times \frac{Precission \times Recall}{Precission + Recall} \quad (19)$$

F1-Score is the overall total that includes a combination of recall and precision.

In addition, this study also evaluated the model's performance using Receiver Operating Characteristic - Area Under the Curve (ROC AUC). ROC AUC analysis was performed using the One-vs-Rest (OvR) approach, in which each class was compared with the combination of all other classes [26]. The AUC calculation results per class were then averaged using the macro-averaging method so that each class had the same weight even though the amount of data differed.

2.6. Prediction

In the prediction stage, the ConvNeXt model that has undergone the training process is used to classify lung X-ray images into five classes, namely Bacterial Pneumonia, Viral Pneumonia, Coronavirus Disease, Tuberculosis, and Normal. The test images used as input have undergone pre-processing steps, including normalization, Contrast Limited Adaptive Histogram Equalization (CLAHE), image augmentation, and resizing to a size of 224 × 224 pixels.

In order to generate a prediction, the test images were processed by the model, which subsequently produced a probability vector via the softmax function. The class corresponding to the highest probability value was then identified as the model's determined outcome.

To improve the interpretability of the prediction results, this study utilized the Gradient-weighted Class Activation Mapping (Grad-CAM) method. Grad-CAM is used to generate a heatmap that visualizes the areas in the image that contribute most to the model's decision-making [27]. This process is performed by calculating the gradient of the feature map at the last convolutional layer, then combining it into an activation map that shows the importance of each area [28]. The heatmap is then overlaid on the original image, allowing the model's focus areas to be visually observed.

3. RESULT

3.1. Training Result

Based on this visualization, it can be observed that the model shows a significant increase in accuracy during training, with training and validation accuracy approaching each other. The training results show that the ConvNeXt model achieves an accuracy of 99.67% on train accuracy and 92.66%

on validation accuracy. Meanwhile, the loss value also shows a steady decrease, with the ConvNeXt model's loss decreasing to 0.97% on train loss and 39.65% on validation loss. This indicates that the ConvNeXt model performs well in learning. Although there are some minor fluctuations, it appears that the model does not experience significant overfitting

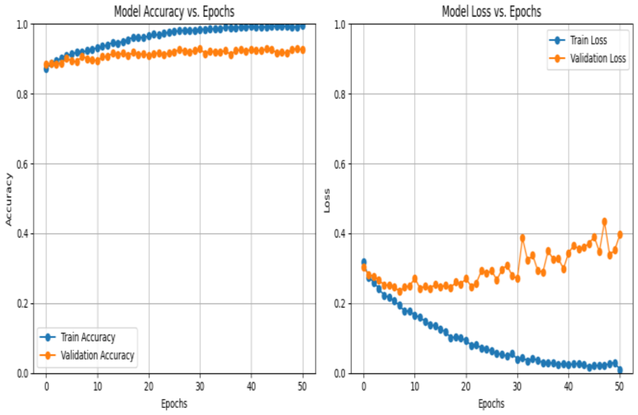


Figure 4. Accuracy and Loss Results

3.2. Performance Evaluation

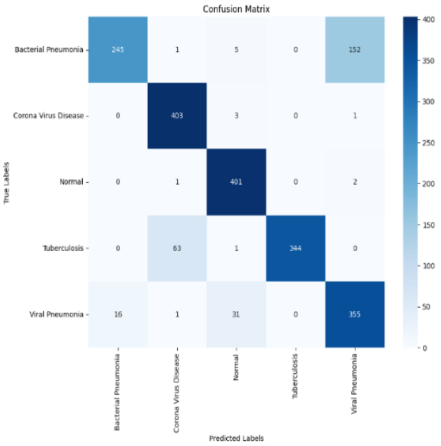


Figure 5. Confusion Matrix

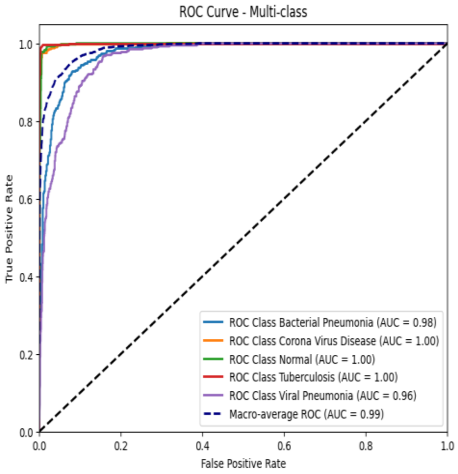


Figure 6. ROC AUC Curve

Vol. 6, No. 6, December 2025, Page. 5837-5853
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.6.5404

Based on the model performance evaluation results in the Evaluation Metrics per Class table, the model shows an overall accuracy rate of 86.32% with a macro average F1-score of 0.8599 and a weighted average F1-score of 0.8602. Model performance varies across classes. The Tuberculosis class achieved the best performance with high precision and recall (1.000 and 0.8431), resulting in an F1-score of 0.9149. The Normal and Coronavirus Disease classes also showed very good results, with recall values of 0.9926 and 0.911, respectively, and F1-scores above 0.92. Conversely, the Bacterial Pneumonia class had relatively low recall (0.608) despite high precision (0.9387), indicating that there are still a significant number of Bacterial Pneumonia cases not detected by the model (false negatives). The lowest performance was observed in the Viral Pneumonia class, with precision of 0.6961 and an F1-score of 0.7777, indicating significant prediction errors in this class.

ROC curve analysis in Figure ROC Curve - Multi-class confirms that the model has excellent discriminative ability across most classes, with AUC values approaching 1.0. The Normal and Tuberculosis classes achieve perfect AUC (1.00), followed by Coronavirus Disease (1.00) and Bacterial Pneumonia (0.98). The Viral Pneumonia class has the lowest AUC (0.96), consistent with the lower precision and F1-score results in the evaluation table. The macro-average AUC value of 0.99 indicates that, overall, the model has very high multi-class classification performance.

These findings suggest that while the model can recognize most classes with high accuracy, there are still challenges in distinguishing between Bacterial Pneumonia and Viral Pneumonia, possibly due to the similarity in visual patterns on X-ray images of the two diseases. Further optimization, such as increasing the amount of data in classes with low performance or applying more varied augmentation techniques, has the potential to improve the model's performance in those classes.

Table 3. Evaluation Result

Class	Precision	Recall	F1-Score	AUC
Bacterial Pneumonia	93%	60%	73%	98%
Viral Pneumonia	69%	88%	77%	96%
Corona Virus Disease	85%	99%	92%	100%
Tuberculosis	100%	84%	91%	100%
Normal	90%	99%	94%	100%

To establish a performance benchmark, we conducted a comparative analysis against two widely-used baseline models—ResNet-50 and DenseNet-121—under identical dataset and training conditions. As quantitatively demonstrated in Table 4, our ConvNeXt model consistently outperformed both baselines across all key metrics, achieving the highest test accuracy, macro F1-score, and macro AUC, which substantiates its superior capability for lung disease classification from CXR images.

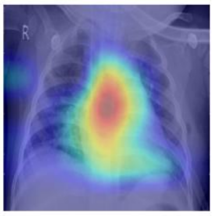
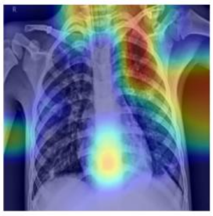
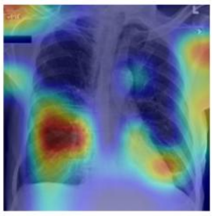
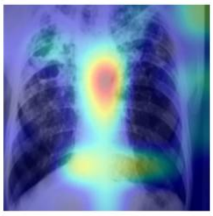
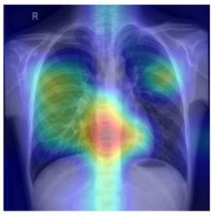
Table 4. Comparative Performance of Different Architectures

Model	Test Accuracy	Macro F1-Score	Macro AUC
ResNet-50	82.15%	0.8112	0.97
DenseNet-121	84.60%	0.8395	0.98
ConvNeXt (Ours)	86.32%	0.8599	0.99

3.3. Prediction Result and Grad-CAM Visualization

Prediction results show that the model has a high level of confidence in classifying images. From the test results, the model was able to classify images into three classes, namely Coronavirus Disease (COVID-19), Tuberculosis, and Normal, with a perfect probability of 100% on the test samples displayed. Meanwhile, for the Bacterial Pneumonia and Viral Pneumonia classes, the model also provides excellent prediction results with a probability of 99%. This indicates that the important features of X-ray images in all five classes are successfully recognized strongly and consistently by the model.

Table 5. Prediction Result

Testing	Class	Probability
	Bacterial Pneumonia	99%
	Viral Pneumonia	99%
	Corona Virus Disease	100%
	Tuberculosis	100%
	Normal	100%

To strengthen the analysis, this study also uses Gradient-weighted Class Activation Mapping (Grad-CAM) as an interpretability method. Grad-CAM generates activation maps that highlight the areas of the image most contributing to the classification decision. Grad-CAM visualizations show that areas with red intensity are focused on the lung regions relevant to disease indications. In cases of correct predictions, the model's focus is precisely on the abnormal lung areas, providing visual justification for the decisions made.

However, in some cases of misclassification, the Grad-CAM distribution appears less directed and spreads to areas outside the lungs or insignificant parts of the image. This indicates limitations in the model's ability to extract certain features under complex image conditions.

4. DISCUSSIONS

The findings of this research indicate that the ConvNeXt architecture exhibits strong capability in classifying multiple types of lung disease based on chest X-ray (CXR) images. With a test accuracy of 86.32%, macro-average F1-score of 0.8599, and macro-average AUC of 0.99, the model demonstrates high robustness and discriminative ability across classes. These results prove that ConvNeXt, which combines convolutional operations with modern architectural principles inspired by Vision Transformers, can effectively extract spatial and textural features relevant to disease classification.

When compared with previous studies, the performance achieved by the proposed ConvNeXt model is competitive and, in certain aspects, superior. A critical analysis reveals that while some prior works report higher raw accuracy, the present study offers a more balanced and interpretable approach. Souid et al. (2021) [7] applied MobileNetV2 for lung disease classification using the ChestX-ray dataset and achieved an average accuracy above 90% with an AUC of 0.811. Although MobileNetV2 performed efficiently on lightweight devices, its interpretability and robustness were limited. In contrast, the ConvNeXt model in this study not only achieved a substantially higher AUC value (0.99), indicating excellent class separation capability, but also provided superior visual interpretability through Grad-CAM analysis, making it more suitable for clinical decision support where justifying a diagnosis is paramount.

Hasanah et al. (2023) [8] compared ResNet50, ResNet101, and ResNet152 architectures and found that ResNet152 achieved the best F1-score of 0.94. Although ConvNeXt produced a slightly lower macro F1-score (0.8599), its training convergence and generalization ability were more stable across multiple classes, with lower overfitting tendencies as shown in the loss curves. This suggests that ConvNeXt can maintain consistent performance even with heterogeneous image distributions, which is a common challenge in real-world medical datasets where image quality and patient demographics vary widely. The stability of ConvNeXt, derived from its modern design elements like Layer Normalization and GELU activation, is a significant advantage over deeper but potentially more unstable networks like ResNet152.

Meanwhile, Shamrat et al. (2022) [9] introduced LungNet22, a modified VGG16-based model, which achieved a very high accuracy of 98.89% in classifying up to 10 lung disease categories. However, a critical limitation of their study was the potential for dataset redundancy and overlapping features, which could inflate the reported accuracy without guaranteeing generalizability. In contrast, the present study employed a more balanced dataset with five distinct disease categories and maintained high accuracy while ensuring interpretability through Grad-CAM, a feature absent in LungNet22. This emphasis on a robust evaluation framework and model transparency is a key differentiator of our work.

Izdihar et al. (2024) [10] and Jiang et al. (2021) [11] demonstrated that ResNet50 and improved VGG models (IVGG13) can enhance pneumonia detection accuracy through deep architectures and data augmentation. The ConvNeXt model builds upon these advancements by incorporating depthwise convolutions, GELU activation, and Layer Normalization, resulting in improved feature extraction and computational stability. The results obtained in this study corroborate the findings of those earlier works, showing that deeper and well-regularized architectures yield higher performance and better generalization on medical imaging tasks. However, our work extends this by systematically evaluating a modern CNN architecture that bridges the performance gap with Transformers while retaining computational efficiency, a crucial factor for clinical deployment.

Bundea and Danciu (2024) [12] employed the DenseNet architecture for pneumonia classification and achieved accuracy rates between 92% and 97%, depending on class. While DenseNet demonstrated strong feature reuse and parameter efficiency, the ConvNeXt model in this research achieved comparable test accuracy (86.32%) with a considerably simpler design and faster training process.

Moreover, the Grad-CAM visualizations in this study confirmed that ConvNeXt could accurately focus on the pathological lung regions, providing a clear interpretative advantage over DenseNet-based approaches that lack explicit attention visualization. This combination of competitive accuracy, efficiency, and built-in interpretability positions ConvNeXt as a more practical solution for developing computer-aided diagnosis (CAD) systems in resource-constrained environments.

From an analytical perspective, the ConvNeXt model's interpretability offers an important contribution to clinical AI applications. The Grad-CAM visualization showed that the model consistently concentrated on relevant pulmonary regions, such as areas affected by opacities or tissue abnormalities. This visual evidence aligns with clinical reasoning patterns and supports the potential use of ConvNeXt as an assistive diagnostic tool for radiologists. The ability to "show its work" is not merely a technical feature but a fundamental requirement for building trust with medical practitioners and facilitating the integration of AI into clinical workflows. However, misclassification cases, particularly between Bacterial Pneumonia and Viral Pneumonia, highlight an ongoing challenge due to the radiographic similarity between the two conditions—a limitation similarly reported by Souid et al. (2021) [7] and Shamrat et al. (2022) [9]. This recurring issue across multiple studies underscores a fundamental limitation of CXR imaging for this specific diagnostic task and suggests that future work might need to integrate clinical metadata or multi-modal data to achieve a definitive differentiation.

In addition, the preprocessing strategies implemented in this study, including normalization, CLAHE, augmentation, and resizing, proved effective in enhancing image contrast and improving model generalization. This supports the conclusions drawn by Stojnev et al. (2020) and Goceri (2020), who emphasized that proper preprocessing can significantly improve CNN performance in medical image analysis. Our study validates these established practices within the context of a modern architecture like ConvNeXt, demonstrating their continued importance.

Overall, this research contributes to the advancement of interpretable deep learning in medical imaging. The primary scientific significance of this work lies in the demonstration that ConvNeXt, a pure CNN architecture, can achieve Transformer-level performance on a complex medical image classification task while offering superior computational efficiency and inherent interpretability. The ConvNeXt model successfully balances accuracy, interpretability, and computational efficiency, demonstrating its potential for deployment in computer-aided diagnostic (CADx) systems. The practical implication is the provision of a reliable, transparent, and efficient tool that can assist radiologists in screening and diagnosing lung diseases, potentially reducing workload and improving diagnostic consistency, especially in regions with a shortage of expert radiologists. Future studies should focus on expanding the dataset, improving class balance, and integrating attention mechanisms or Transformer-based hybrids to further enhance differentiation between visually similar diseases and strengthen model generalization across diverse clinical datasets.

5. CONCLUSION

This study demonstrated the effectiveness of the ConvNeXt architecture in multi-class lung disease classification using chest X-ray images. By employing a dataset consisting of five classes (Bacterial Pneumonia, Viral Pneumonia, Coronavirus Disease, Tuberculosis, and Normal) with comprehensive preprocessing techniques such as normalization, CLAHE, augmentation, and resizing, the proposed model achieved high performance. The key findings conclusively show a training accuracy of 99.67%, a validation accuracy of 92.66%, and a test accuracy of 86.32%, confirming the model's strong learning capability and generalization. The model's robustness is further validated by a macro-average F1-score of 0.8599 and an exceptional macro-average ROC-AUC of 0.99. Additionally, the integration of Grad-CAM visualizations proved to be a significant advantage, as it provided transparent and clinically plausible explanations for the model's decisions by consistently focusing on pathological lung regions, thereby enhancing trust and interpretability for potential clinical users.

The primary impact of this research lies in its contribution to the development of accurate and interpretable AI-driven diagnostic systems in pulmonology. By achieving high fidelity in classifying multiple diseases, including COVID-19 and Tuberculosis, this work presents a viable tool for assisting radiologists, reducing diagnostic time, and improving early detection rates in both routine screenings and urgent care scenarios. The use of a computationally efficient architecture like ConvNeXt also underscores the potential for deploying such systems in settings with limited resources, thereby making advanced diagnostic support more accessible.

Despite these promising results, the model still encountered difficulties in distinguishing between Bacterial Pneumonia and Viral Pneumonia, largely due to their similar radiographic characteristics. To address this limitation and further advance this research, several directions are proposed for future work. First, expanding the dataset with more samples from underperforming classes and incorporating images from diverse demographic and clinical sources could enhance model generalization. Second, exploring hybrid models that combine the strengths of ConvNeXt with attention mechanisms or Vision Transformers may improve feature discrimination for visually similar diseases. Third, technical strategies such as advanced data augmentation tailored for medical images, cost-sensitive learning to handle class imbalance, and hyperparameter tuning could yield further accuracy gains. Finally, the most critical next step is to transition from a prototype to a clinical implementation, which involves developing a user-friendly software interface and conducting real-world validation trials in collaboration with healthcare institutions to assess the model's practical utility, workflow integration, and overall impact on diagnostic decision-making.

CONFLICT OF INTEREST

The authors declares that there is no conflict of interest between the authors or with research object in this paper.

ACKNOWLEDGEMENT

The author would like to thank the State Islamic University of North Sumatra (UINSU) for its support and funding of this research through the Basic Research Assistance Scheme.

REFERENCES

[1] R. O. for the E. M. World Health Organization, “World TB Day 2025,” World Health Organization (WHO). Accessed: Aug. 28, 2025. [Online]. Available: https://www.emro.who.int/world-tb-day/2025/index.html

[2] UNICEF, “Pneumonia,” UNICEF. Accessed: Aug. 28, 2025. [Online]. Available: https://data.unicef.org/topic/child-health/pneumonia/

[3] H. Iqbal, A. Khan, N. Nepal, F. Khan, and Y. K. Moon, “Deep Learning Approaches for Chest Radiograph Interpretation: A Systematic Review,” *Electron.*, vol. 13, no. 23, pp. 1–24, 2024, doi: 10.3390/electronics13234688.

[4] G. Kourounis, A. A. Elmahmudi, B. Thomson, J. Hunter, H. Ugail, and C. Wilson, “Computer image analysis with artificial intelligence: a practical introduction to convolutional neural networks for medical professionals,” *Postgrad. Med. J.*, vol. 99, no. 1178, pp. 1287–1294, 2023, doi: 10.1093/postmj/qgad095.

[5] S. Tiwari, G. Jain, D. K. Shetty, M. Sudhi, J. M. Balakrishnan, and S. R. Bhatta, “A Comprehensive Review on the Application of 3D Convolutional Neural Networks in Medical Imaging,” *Eng. Proc.*, vol. 59, no. 1, pp. 1–9, 2023, doi: 10.3390/engproc2023059003.

[6] N. M. Pillai, S. Manimala, A. S. Rongali, A. Gugnani, A. S. Kumar, and G. V Sriramakrishnan, “Automated Classification of Medical Images Using Convolutional Neural Networks,” in *2024*

15th International Conference on Computing Communication and Networking Technologies (ICCCNT), 2024, pp. 1–6. doi: 10.1109/ICCCNT61001.2024.10725642.

[7] A. Souid, N. Sakli, and H. Sakli, "Classification and predictions of lung diseases from chest x-rays using mobilenet v2," *Applied Sciences*, vol. 11, no. 6, 2021, doi: 10.3390/app11062751.

[8] S. A. Hasanah, A. A. Pravitasari, A. S. Abdullah, I. N. Yulita, and M. H. Asnawi, "A Deep Learning Review of ResNet Architecture for Lung Disease Identification in CXR Image," *Applied Sciences*, vol. 13, no. 24, 2023, doi: 10.3390/app132413111.

[9] F. M. Javed Mehedi Shamrat *et al.*, "LungNet22: A Fine-Tuned Model for Multiclass Classification and Prediction of Lung Disease Using X-ray Images," *J. Pers. Med.*, vol. 12, no. 5, 2022, doi: 10.3390/jpm12050680.

[10] N. Izdihar, S. B. Rahayu, and K. Venkatesan, "Comparison Analysis of CXR Images in Detecting Pneumonia Using VGG16 and ResNet50 Convolution Neural Network Model," *Int. J. Informatics Vis.*, vol. 8, no. 1, pp. 326–332, 2024, doi: 10.62527/joiv.8.1.2258.

[11] Z. P. Jiang, Y. Y. Liu, Z. E. Shao, and K. W. Huang, "An improved VGG16 model for pneumonia image classification," *Applied Sciences*, vol. 11, no. 23, 2021, doi: 10.3390/app112311185.

[12] M. Bundea and G. M. Danciu, "Pneumonia Image Classification Using DenseNet Architecture," *Inf.*, vol. 15, no. 10, 2024, doi: 10.3390/info15100611.

[13] Z. Liu, H. Mao, C. Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A ConvNet for the 2020s," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2022-June, pp. 11966–11976, 2022, doi: 10.1109/CVPR52688.2022.01167.

[14] S. I. A. Stojnev D., "Preprocessing Image Data for Deep Learning," in *Sinteza 2020 - International Scientific Conference on Information Technology and Data Related Research*, 2020, pp. 312–317. doi: 10.15308/Sinteza-2020-312-317.

[15] S. Albert *et al.*, "Comparison of Image Normalization Methods for Multi-Site Deep Learning," *Applied Sciences*, vol. 13, no. 15, pp. 1–13, 2023, doi: 10.3390/app13158923.

[16] R. Srikanth, B. Nagarjuna, K. Varshith, G. Supriya, and S. P. Reddy, "Enhancing of Night Time Vehicle Images using CLAHE Method," in *2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, 2024, pp. 1–8. doi: 10.1109/ICCCNT61001.2024.10725464.

[17] E. Goceri, "Image Augmentation for Deep Learning Based Lesion Classification from Skin Images," in *2020 IEEE 4th International Conference on Image Processing, Applications and Systems (IPAS)*, 2020, pp. 144–148. doi: 10.1109/IPAS50080.2020.9334937.

[18] H. Talebi and P. Milanfar, "Learning to Resize Images for Computer Vision Tasks," *Proc. IEEE Int. Conf. Comput. Vis.*, pp. 487–496, 2021, doi: 10.1109/ICCV48922.2021.00055.

[19] C. Urrea, Y. Garcia-Garcia, and J. Kern, "Improving Surgical Scene Semantic Segmentation through a Deep Learning Architecture with Attention to Class Imbalance," *Biomedicines*, vol. 12, no. 6, 2024, doi: 10.3390/biomedicines12061309.

[20] K. Solanki, S. P. Singh, A. Yadav, V. Sharma, S. Awasthi, and S. Vats, "Enhancing Ophthalmic Diagnostics: CNNs in Cataract Detection," in *2024 2nd International Conference on Disruptive Technologies (ICDT)*, 2024, pp. 842–846. doi: 10.1109/ICDT61202.2024.10489576.

[21] S. Haji and A. Abdulazeez, "COMPARISON OF OPTIMIZATION TECHNIQUES BASED ON GRADIENT DESCENT ALGORITHM: A REVIEW PJAEE, 18 (4) (2021) COMPARISON OF OPTIMIZATION TECHNIQUES BASED ON GRADIENT DESCENT ALGORITHM: A REVIEW Comparison Of Optimization Techniques Based On Gradient Descent Al," *PalArch's J. Archaeol. Egypt/ Egyptol.*, vol. 18, pp. 2715–2743, Jun. 2021.

[22] G. Maillard, S. Arlot, and M. Lerasle, "Aggregated Hold-Out," *J. Mach. Learn. Res.*, vol. 22, no. 20, pp. 1–55, 2021, [Online]. Available: http://jmlr.org/papers/v22/19-624.html

[23] E. Kristiani, Y. T. Tsan, P. Y. Liu, N. Y. Yen, and C. T. Yang, "Binary and Multi-Class Assessment of Face Mask Classification on Edge AI Using CNN and Transfer Learning," *Human-centric Comput. Inf. Sci.*, vol. 12, 2022, doi: 10.22967/HCIS.2022.12.053.

[24] B. J. Erickson and F. Kitamura, "Magician's Corner: 9. Performance Metrics for Machine Learning Models," *Radiol. Artif. Intell.*, vol. 3, no. 3, p. e200126, May 2021, doi: 10.1148/ryai.2021200126.

[25] A. Mehra, Y. Zhang, and J. Hamm, "Test-time Assessment of a Model's Performance on Unseen Domains via Optimal Transport," *IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Work.*, no. Ml, pp. 173–182, 2024, doi: 10.1109/CVPRW63382.2024.00022.

[26] S. F. Hussain and M. M. Ashraf, "A novel one-vs-rest consensus learning method for crash severity prediction," *Expert Syst. Appl.*, vol. 228, p. 120443, 2023, doi: https://doi.org/10.1016/j.eswa.2023.120443.

[27] D. Song *et al.*, "A new xAI framework with feature explainability for tumors decision-making in Ultrasound data: comparing with Grad-CAM," *Comput. Methods Programs Biomed.*, vol. 235, p. 107527, 2023, doi: https://doi.org/10.1016/j.cmpb.2023.107527.

[28] S. Sattarzadeh, M. Sudhakar, K. N. Plataniotis, J. Jang, Y. Jeong, and H. Kim, "Integrated Grad-Cam: Sensitivity-Aware Visual Explanation of Deep Convolutional Networks Via Integrated Gradient-Based Scoring," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 1775–1779. doi: 10.1109/ICASSP39728.2021.9415064.