Jurnal Informatika: Jurnal pengembangan IT

Vol. 11, No. 2, 2026
ISSN : 2477-5126, DOI:10.30591/jpit.v11i2.10373

Comparative Evaluation of VGG16, MobileNetV2, and ResNet50 for Pediatric Pneumonia Classification Using Grad-CAM

Rendra Gunawan [1], Cinantya Paramita [2], Suryanti [3]
[1,2] Faculty of Computer Science, Universitas Dian Nuswantoro, Indonesia
[2] Dinus Research Group for Al in Medical Science (DREAMS), Universitas Dian Nuswantoro, Indonesia
[3] Faculty of Medicine, Universitas Dian Nuswantoro, Indonesia
111202214235@mhs.dinus.ac.id, cinantya.paramita@dsn.dinus.ac.id, suryanti@dsn.dinus.ac.id

Info Artikel

Riwayat Artikel:

Received 2026-03-02
Revised 2026-04-30
Accepted 2026-05-13

Abstract **-** Pneumonia remains one of the leading causes of mortality among children worldwide, particularly in developing countries, where early and accurate diagnosis is crucial. This study aims to evaluate and compare the performance of three deep learning architectures, namely VGG16, MobileNetV2, and ResNet50, for pediatric pneumonia classification using chest X-ray images. The dataset consists of pediatric chest radiographs (ages 1-5 years) obtained from Guangzhou Women and Children's Medical Center, which were preprocessed through normalization and data augmentation techniques to improve model generalization. The classification task involves three categories: normal, bacterial pneumonia, and viral pneumonia. Model performance was evaluated using accuracy, precision, recall, specificity, F1-score, G-Mean, and AUC. The experimental results show that all models achieve competitive performance, with accuracy ranging from 79% to 82%, where VGG16 outperforms the other architectures. Furthermore, Grad-CAM is applied to enhance interpretability by visualizing important regions in X-ray images that influence model decisions. The results demonstrate that Grad-CAM provides meaningful visual explanations, supporting the reliability of deep learning models in assisting clinical diagnosis.

Keywords: Pediatric Pneumonia, Image Classification, VGG16, MobileNetV2, ResNet50.

Corresponding Author:
Cinantya Paramita
Email:
cinantya.paramita@dsn.dinus.ac.id

This is an open access article under the CC BY 4.0 license.

Abstrak *- Pneumonia merupakan salah satu penyakit saluran pernapasan yang paling berbahaya, terutama pada anak-anak. Diagnosis cepat dan akurat sangat penting guna mempercepat penanganan serta mencegah komplikasi lebih lanjut. Penelitian ini membahas penerapan tiga arsitektur deep learning paling populer yaitu VGG16, MobileNetV2, dan ResNet50 dalam klasifikasi citra X-ray dada untuk deteksi pneumonia dengan tiga kategori: normal, pneumonia bakteri, dan pneumonia virus. Dataset yang digunakan merupakan citra X-ray anak yang berusia 1-5 tahun dari Guangzhou Hospital, yang telah melalui tahapan preprocessing, normalisasi, dan augmentasi guna meningkatkan generalisasi model. Evaluasi dilakukan menggunakan matrik akurasi, presisi, recall, specificity, F1-score, G-Mean, dan AUC untuk memberikan penilaian komprehensif tiap model. Hasil menunjukkan akurasi model berkisar antara 79% hingga 82%, dengan VGG16 memiliki akurasi tertinggi, diikuti ResNet50 dan MobileNetV2. Implementasi teknik interpretabilitas Grad-CAM berhasil memvisualisasikan area utama fokus model pada citra X-ray, sehingga dapat memberikan informasi tambahan bagi tim medis dalam memahami letak infeksi yang terdeteksi oleh model. Studi ini memberi kontribusi nyata terhadap pengembangan sistem pendukung diagnosis pneumonia berbasis citra X-ray, khususnya untuk deteksi jenis pneumonia bakteri dan virus secara otomatis dengan tingkat akurasi yang tinggi serta visualisasi yang membantu validasi diagnostik.*

Kata Kunci*: Pneumonia Anak, Klasifikasi Citra, VGG16, MobileNetV2, ResNet50.*

I. INTRODUCTION

According to the World Health Organization (WHO), pneumonia constitutes an acute respiratory infection primarily triggered by bacteria, fungi, parasites, or viruses[1], [2]. Pneumonia occurs as a result of an infection that induces inflammation of the alveoli (air sacs in the lungs), which may then become filled with fluid or pus and consequently impair normal respiratory function[3], [4]. This disease can lead to a wide spectrum of clinical manifestations, ranging from mild symptoms to life-threatening conditions[5]. WHO statistics alongside multiple clinical investigations identify pneumonia as the primary cause of mortality in children under five, especially within low, and middle income nations[6], [7], [8].

In Indonesia, the 2018 Basic Health Research (Riskesdas) reported that the prevalence of pneumonia across all age groups reached 2.21%. This figure increases among older adults, with proportions of 2.5% in the 44-64 year age group, 3.0% in the 64-74 year group, and 2.9% in individuals aged 75 years and above. Furthermore, according to National Health Insurance (JKN) statistics from 2014-2018, pneumonia is among the top ten diseases with the highest number of inpatient cases in healthcare facilities[9]. The National Center for Health Statistics reported that approximately 1.2 million emergency department visits in the United States were attributed to pneumonia caused by

infectious organisms as the primary diagnosis, based on data from the National Hospital Ambulatory Medical Care Survey: 2022 Emergency Department Summary Tables[10], [11].

One of the most widely used approaches for pneumonia detection is chest X-ray image analysis, which enables faster and more accurate identification of radiological signs of pneumonia[12], [13]. However, chest X-ray examination generally requires interpretation by professional radiologists[12], [14]. To address this limitation, numerous researchers have turned to deep learning techniques, which have demonstrated substantial potential for medical image detection and analysis[15].

Previous studies have highlighted research opportunities, such as the work in[16], which employed a deep learning approach based on Convolutional Neural Networks (CNNs) using three architectures: VGG16, achieving an accuracy of 94.22%; MobileNetV2, with an accuracy of 96.36%; and ResNet50, with an accuracy of 82.20% for pneumonia detection. Among the many CNN architectures available for medical image classification, VGG16, MobileNetV2, and ResNet50 were selected for this study because they represent three distinct paradigms of network design that are highly relevant to the pneumonia detection context. VGG16 is a deep sequential architecture known for its strong feature extraction capability using stacked 3×3 convolutional layers, making it well-suited for capturing fine-grained radiological patterns in chest X-rays. MobileNetV2, designed with inverted residual blocks and depthwise separable convolutions, represents the lightweight paradigm suitable for resource-constrained or mobile deployment scenarios in clinical settings. ResNet50, on the other hand, introduces skip connections to address the vanishing gradient problem, enabling stable training of deeper networks with strong generalization capacity.

The dataset used consisted of chest X-ray images obtained from a public repository [17]. In the preprocessing stage, the dataset categories will be refined into three classes: Normal, Bacterial Pneumonia, and Viral Pneumonia. The evaluation metrics employed include Accuracy, Precision, Recall, F1-score, G-mean, and AUC. Furthermore, model interpretability will be incorporated using Grad-CAM (Gradient-weighted Class Activation Mapping), as implemented in several previous studies. Prior studies confirm that Grad-CAM is an effective post-hoc interpretability method for pneumonia diagnosis. Research by Ennab and Mcheick [18], Şahin et al. [19], and Paramita et al. [20] shows that gradient-based heatmaps can highlight clinically relevant lung lesion areas, improving model transparency without reducing classification performance. These findings from previous studies indicate that Grad-CAM generates clinically meaningful activation maps focused on lung lesion areas, thereby directly enhancing radiologists' understanding[18], [19], [20].

After identifying this research gap, the objective of the present study is to evaluate the performance of three CNN architectures which is VGG16, MobileNetV2, and ResNet50 in addressing the problem of pneumonia classification on chest X-ray images into three class: Normal, Bacterial Pneumonia, and Viral Pneumonia. Although numerous studies have explored X-ray-based pneumonia classification, model accuracy and effectiveness still vary considerably, influenced by architectural complexity and dataset characteristics. Therefore, this research focuses on a comparative evaluation of these three CNN models using the Guangzhou Hospital dataset. In addition to assessing their performance, this study also applies Grad-CAM (Gradient-weighted Class Activation Mapping) to visualize the model's regions of interest on X-ray images, thereby providing visual interpretability that aids in understanding the model's decision-making process.

II. METHODOLOGY

The stages of this study are presented in Figure 1. The research begins with obtaining the dataset, which is then further processed in the preprocessing stage. After preprocessing, the study proceeds to the modeling stage using transfer learning algorithms, namely VGG16, MobileNetV2, and ResNet50. Following the modeling stage, the models are evaluated using performance metrics, and the final stage of this research involves interpretability by applying Grad-CAM.

Figure 1. Research Methodology

A. Dataset

The dataset used in this study consists of chest X-ray images obtained from Guangzhou Women and Children's Medical Center. The chest X-ray images were collected from pediatric patients aged 1-5 years for research purposes. The dataset contains a total of 5,856 images divided into two categories, namely Normal and Pneumonia[17]. The detailed distribution of the dataset is presented in Table 1 below.

TABLE 1
DATASET

Classes	Train	Test	Val
Normal	234	1341	8
Pneumonia	390	3875	8

The chest X-ray image dataset with Normal and Pneumonia categories can be seen in Figure 2 and Figure 3 below.

Normal

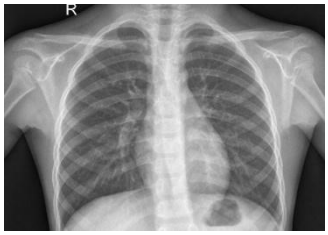

Figure 2. *Chest X-ray* Normal

Pneumonia

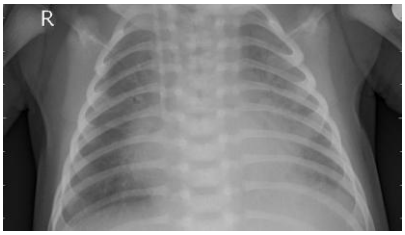

Figure 3. *Chest X-ray Pneumonia*

B. Preprocessing

Preprocessing data serves as a foundational phase in constructing machine learning models for statistical analysis[21]. Preprocessing in this work incorporates normalization alongside augmentation techniques like image resizing, rotations, flips (horizontal/vertical), shearing, and zooming. Before resizing, the class labels are expanded into three categories: Normal, Viral Pneumonia, and Bacterial Pneumonia to provide more detailed class separation for analysis. This modification is intended to facilitate model interpretation in the visualization stage using the Grad-CAM technique, enabling clearer identification of class-specific regions of interest. The distribution of the dataset after the category expansion is presented in Table 2 below.

TABLE 2
DATASET AFTER PREPROCESSING

Classes	Train	Test	Val
Normal	1266	158	159
Pneumonia Bacteria	1194	150	149
Pneumonia Virus	2224	278	278

At the resizing stage, the dimensions of all images are standardized to match the input size required by the Convolutional Neural Network (CNN) architecture, with each image resized to a resolution of 224 × 224 pixels. Each image pixel is then normalized from the original range of 0-255 to 0-1 using rescaling, which helps accelerate model convergence and improves training stability.

The final stage is data augmentation, which applies several transformations, including rotations of up to 20°, horizontal and vertical shifts of 10%, a shear factor of 0.1, and zooming of up to 15%. In addition, horizontal flipping and brightness adjustments in the range of 0.7-1.3 are used to simulate variations in imaging conditions in radiographic devices, enabling the model to learn image patterns more effectively.

C. Model Selection

1) VGG16: The VGG16 architecture used in this study employs stacked convolutional blocks with 3 × 3 kernels and padding that preserves spatial dimensions, where each block consists of two to three Conv2D layers followed by ReLU activation, and the number of filters gradually increases from 64, 128, 256 to 512 to extract hierarchical features. After each block, MaxPooling2D is applied to reduce spatial resolution, then GlobalAveragePooling2D is used instead of flatten to reduce dimensionality while preserving feature distribution, followed by Batch Normalization to stabilize activations and accelerate convergence, a Dense layer with 512 units and Dropout for regularization, and a final Dense layer for classification output. Overall, this structure maintains the core principles of VGG16 (repeated small convolutions with progressive pooling) while incorporating modern optimizations to improve training stability and performance. VGG16 was selected in this study due to its proven capacity for rich hierarchical feature learning through deep sequential convolutions, which is particularly advantageous for detecting subtle radiological patterns in chest X-ray images such as consolidations and infiltrates associated with pneumonia.

2) MobileNetV2: The MobileNetV2 architecture in this study implements a series of inverted residual bottleneck blocks with three main stages: 1 × 1 channel expansion, followed by computationally efficient 3 × 3 depthwise convolution, and a final 1 × 1 pointwise convolution to project the channels back to a lower dimension. Residual connections are applied only when the input and output dimensions match to maintain stable information

flow, while the linear bottleneck preserves essential feature representations with low complexity; at the end of the network, GlobalAveragePooling and a Dense layer are used to produce the classification output, yielding a lightweight and fast model that still provides competitive feature extraction quality on resource-constrained devices. MobileNetV2 was chosen because its lightweight design with depthwise separable convolutions offers a balance between computational efficiency and classification performance, making it a relevant candidate for deployment in resource-constrained clinical environments.

3) ResNet50: The ResNet50 architecture used in this research consists of a 50-layer residual network based on a series of bottleneck residual blocks, starting with a 7 × 7 Conv2D stem layer (stride 2) followed by MaxPooling. The blocks are organized into four stages (3, 4, 6, 3) with progressively increasing channel depth, resulting in high-dimensional final feature maps; each convolution operation is followed by Batch Normalization and ReLU activation, while shortcut connections perform element-wise addition between the block input and output to preserve gradient flow. This residual mechanism enhances training stability and mitigates the vanishing gradient problem, enabling effective training of very deep networks; feature downsampling is performed using stride in specific blocks, and at the end of the network, Global Average Pooling reduces the spatial dimensions before passing the features to a fully connected layer with softmax output, yielding a model with around 25 million parameters and highly expressive feature representations for classification tasks. ResNet50 was selected in this study due to its high feature representation capacity, achieved through bottleneck residual blocks organized into four stages with progressively increasing channel depth. This strong generalization capability is particularly relevant for chest X-ray image classification, where radiological patterns across classes (Normal, Bacterial Pneumonia, and Viral Pneumonia) can be highly subtle.

D. Model Training and Fine Tuning

After selecting the model architectures, the next stage is training. Given the limited amount of data, transfer learning is applied by leveraging pretrained architectures from libraries such as torchvision and Keras, including models that have previously been trained for pneumonia detection on chest X-ray images. The main advantage of transfer learning lies in its ability to reduce the required amount of training data while improving model accuracy.

- Feature extraction: In this stage, the pretrained models are used to identify important visual characteristics in chest X-ray images, such as anatomical structures, organ contours, and texture patterns associated with pneumonia.
- Fine-tuning: The original final layers of the pretrained models are removed and replaced with fully connected layers tailored to the pneumonia classification task. In addition, selected layers are unfrozen and their weights are adjusted to enhance performance in multi-class classification.
- Loss function and optimization process: Optimization algorithms such as Adam or SGD are employed to minimize the loss function, enabling the model to achieve optimal performance in distinguishing the classes in the chest X-ray dataset.

E. Evaluation Metrics

This study uses seven evaluation metrics, the metrics used include:

$$\boldsymbol{Accuracy} = \frac{\boldsymbol{TP+TN}}{\boldsymbol{TP+TN+FP+FN}} \quad (1)$$

Accuracy is calculated based on the number of true negatives (TN), true positives (TP), false positives (FP), and false negatives (FN). It describes how closely the model's predictions correspond to the true or reference labels[22].

$$\boldsymbol{Recall} = \boldsymbol{Sensitivity} = \frac{\boldsymbol{TP}}{\boldsymbol{TP+FN}} \quad (2)$$

Sensitivity, or recall, is defined as the ratio between the number of correctly predicted positive cases (true positives) and the total number of actual positive cases in the dataset[7].

$$\boldsymbol{Specificity} = \frac{\boldsymbol{TN}}{\boldsymbol{TN+FP}} \quad (3)$$

Specificity is used to measure the percentage of negative cases that are correctly classified, indicating the model's ability to minimize misclassification of instances that are truly negative[23].

$$\boldsymbol{Precision} = \frac{\boldsymbol{TP}}{\boldsymbol{TP+FP}} \quad (4)$$

This metric evaluates the proportion of true positive cases out of all positive predictions, thereby reflecting the model's ability to minimize the occurrence of false positives[23].

$$\boldsymbol{F1 - Score = \frac{2*TP}{(2*TP)+FP+FN}} \quad (5)$$

The F1-score is the harmonic mean of precision and recall, and this metric is used to evaluate the accuracy of a model, particularly when the model operates on more than one task or category[24].

$$\boldsymbol{G - mean = \sqrt{Sensitivity * Specificity}} \quad (6)$$

The G-mean metric provides a balanced overview of model performance because it directly penalizes low values of both the true positive rate (TPR, sensitivity) and the true negative rate (TNR, specificity) within a single integrated score. G-mean can proportionally evaluate classifier performance by considering the model's ability to correctly recognize both positive and negative classes[25], [26].

$$\boldsymbol{AUC = \int_0^1 TPR(FPR)d(FPR)} \quad (7)$$

The AUC formula is used to represent the area under the ROC curve, which reflects the model's ability to distinguish between positive and negative classes across different threshold values. In this context, the True Positive Rate (TPR) and False Positive Rate (FPR) are defined as TPR = TP/(TP + FN) and FPR = FP/(FP + TN), respectively[23].

F. Result Interpretability (Grad-CAM)

To visualize pneumonia regions, the Grad-CAM technique is employed. Its implementation loads the trained model, extracts feature maps from the last convolutional layer, and computes the gradients of the class score with respect to each feature map using tf.GradientTape(). The gradients are then spatially averaged to obtain contribution weights, which are combined with the feature maps to generate an activation map that is passed through ReLU, upsampled to the original image size, normalized, and visualized as a heatmap and as an alpha-blended overlay on the original image. The implementation supports both multi-class outputs (via argmax over softmax) and binary outputs (sigmoid converted into a two-class vector) for selecting the target class[27].

III. RESULT AND DISCUSSIONS

A. Dataset

The implementation of this study is carried out using the Kaggle platform, which provides a cloud-based environment for training and evaluating deep learning models on medical imaging datasets. The hyperparameters used in the experiments, such as learning rate, batch size, number of epochs, and optimizer configuration, are summarized in Table 3.

TABLE 3
EXPERIMENTAL SETUP

Parameter	Value
Img Size	224*224
Batch Size	32
Random State	42
FE Epoch	80
FT Epoch	20
Learning Rate FE	0.0001
Learning Rate FT	0.00001
Early Stopping	Patience=8 (monitoring val-loss)
Activation	ReLU
Optimizer	Adam
Dropout	0.5

The hyperparameter configuration in this study includes an image size of 224 × 224 pixels, a batch size of 32, and a random state of 42, chosen to balance training efficiency and ensure experimental reproducibility. The training process consists of 80 epochs for feature extraction and 20 epochs for fine-tuning. The learning rates are set to 0.0001 for the feature extraction phase and 0.00001 for the fine-tuning phase to help mitigate overfitting by allowing more controlled weight updates. Early stopping with a patience of 8 based on validation loss is applied to automatically halt training when validation performance no longer improves. The ReLU activation function is used to enable the model to learn non-linear patterns in chest X-ray images, while the Adam optimizer provides adaptive gradient updates

during training. Additionally, a dropout rate of 0.5 is employed to enhance model generalization by reducing reliance on specific features.

B. Proposed Model Evaluation

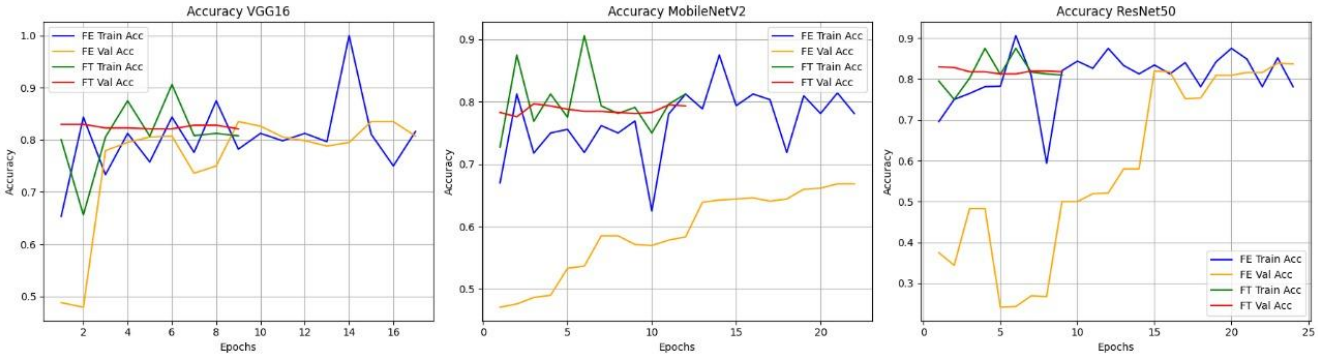

Figure 4. Accuracy Model

In the graph shown in Figure 4, the blue curve represents the training accuracy during feature extraction, the orange curve represents the validation accuracy during feature extraction, the green curve represents the training accuracy during fine-tuning, and the red curve represents the validation accuracy during fine-tuning. Each model and training scenario may be stopped early once the validation performance no longer improves, so the curve lengths and stopping points differ for each color and for each model.

All three models exhibit stable learning behavior, with validation accuracy converging around 0.8. VGG16 attains consistent performance from the early epochs without clear signs of overfitting, while MobileNetV2 shows gradual improvement with minor fluctuations in the initial phase before stabilizing above 0.8; ResNet50 displays a similar pattern between its training and validation accuracy curves, indicating good generalization capability for CXR-based pneumonia classification.

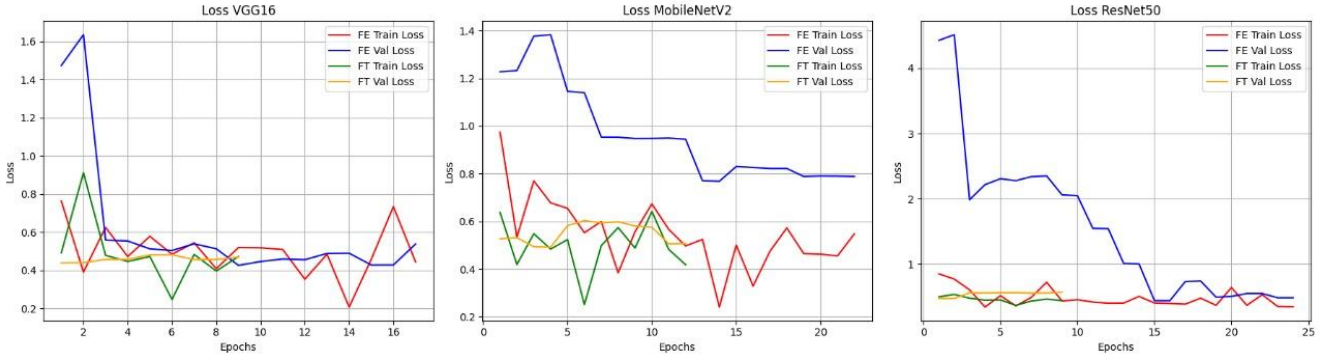

Figure 5. Loss Model

Figure 5 presents the loss curves for the three models under two training scenarios, namely feature extraction (FE) and fine-tuning (FT), for both training and validation data. The blue curve denotes FE training loss, the orange curve denotes FE validation loss, the green curve denotes FT training loss, and the yellow curve denotes FT validation loss. The length of each curve varies across panels because each model or scenario may be trained with a different number of epochs or stopped early once the validation performance has converged.

In terms of loss behavior shown in Figure 5, VGG16 exhibits a rapid decrease from the early epochs and stabilizes around 0.4-0.6, indicating fast learning without a significant increase in loss toward the end of training. MobileNetV2 also shows a consistent reduction in training loss to below 0.8 with persistently low validation loss, reflecting an efficient learning process. Meanwhile, ResNet50 initially starts with a higher loss value but undergoes a sharp decline and stabilizes below 0.5 after approximately the 10th epoch, demonstrating good convergence during training.

C. Evaluation Metrics

The experimental results show that all three models achieve accuracies in the range of 79-82%. VGG16 attains the highest accuracy of 0.821, followed by ResNet50 with 0.817 and MobileNetV2 with 0.793. The detailed evaluation metrics for these models are presented in Table 4. From a theoretical perspective, VGG16 outperforms the other models due to its deep architecture with uniformly stacked 3×3 convolutional layers, which effectively capture fine-grained features in chest X-ray images such as ground-glass opacities and consolidation patterns. Unlike MobileNetV2,

which prioritizes efficiency over representational capacity, and ResNet50, which may be overly deep for the dataset size, VGG16 achieves a better balance between feature richness and overfitting control. Additionally, the use of Global Average Pooling, Batch Normalization, and Dropout enhances model regularization. These factors contribute to VGG16's superior performance, reflected in higher recall and specificity, particularly for Normal (98.7% recall, 92.9% specificity) and Bacterial Pneumonia (86.6% recall), leading to better overall G-Mean and F1-score.

TABLE 4
EVALUASI MODEL

Classes	Accuracy	Precision	Recall	Specificity	F1-Score	G-Mean
			VGG16			
Normal	94.5	83.8	98.7	92.9	90.7	95.8
Pneumonia Bacteria	84.8	82.2	86.6	83.1	84.4	84.8
Pneumonia Virus	84.8	78.5	56	94.7	65.3	72.8
			MobileNetV2			
Normal	92.8	79.9	98.1	90.8	88	94.4
Pneumonia Bacteria	82.7	81.7	82	83.4	81.8	82.7
Pneumonia Virus	83.4	73.4	55.3	93.1	63.1	71.7
			ResNet50			
Normal	93.5	82.3	96.8	92.2	88.9	94.5
Pneumonia Bacteria	82.9	82.3	81.6	84.1	81.9	82.8
Pneumonia Virus	83.9	72.6	60	92.2	65.7	74.3

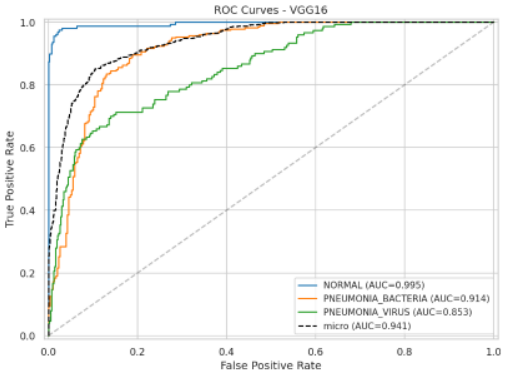

Figure 6. ROC Curves VGG16

In Figure 6, the VGG16 model demonstrates excellent performance for the NORMAL class, with an AUC of 0.995, indicating that the model is almost perfect at distinguishing NORMAL images from pneumonia cases. However, its ability to detect the PNEUMONIA_BACTERIA class is slightly lower, with an AUC of approximately 0.914, and the lowest performance is observed for the PNEUMONIA_VIRUS class, with an AUC of around 0.854. These results suggest that VGG16 has more difficulty discriminating viral pneumonia from bacterial pneumonia and normal chest X-rays.

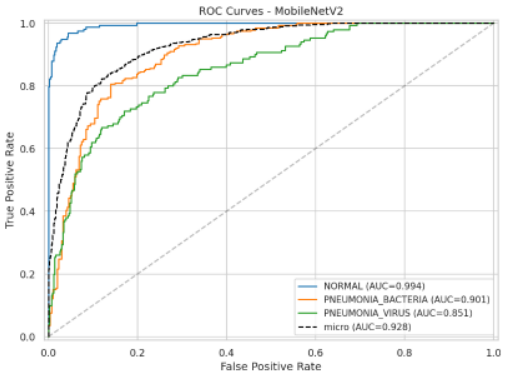

Figure 7. ROC Curves MobileNetV2

MobileNetV2 yields an AUC of approximately 0.994 for the NORMAL class, indicating strong capability in identifying healthy chest X-ray images. For the PNEUMONIA_BACTERIA and PNEUMONIA_VIRUS classes, the model attains AUC values of around 0.901 and 0.851, respectively, which reveals limitations in distinguishing between different pneumonia types, as illustrated in Figure 7.

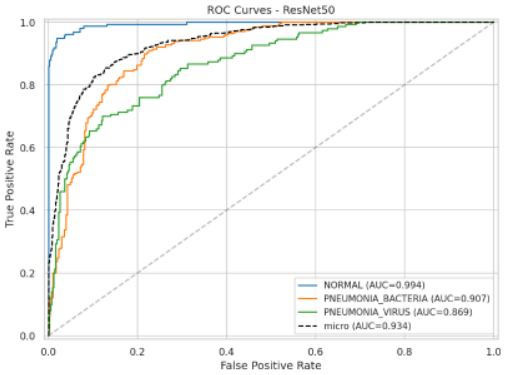

Figure 8. ROC Curves ResNet50

As shown in Figure 8, ResNet50 exhibits the most stable and balanced AUC performance across all three classes. The AUC for the NORMAL class is approximately 0.994, comparable to the other two models, while for the PNEUMONIA_BACTERIA class it is around 0.907. The model outperforms VGG16 and MobileNetV2 on the PNEUMONIA_VIRUS class, achieving an AUC of about 0.869, which indicates that ResNet50 has better generalization capability in distinguishing visual patterns associated with viral pneumonia.

D. Result Interpretability (Grad-CAM)

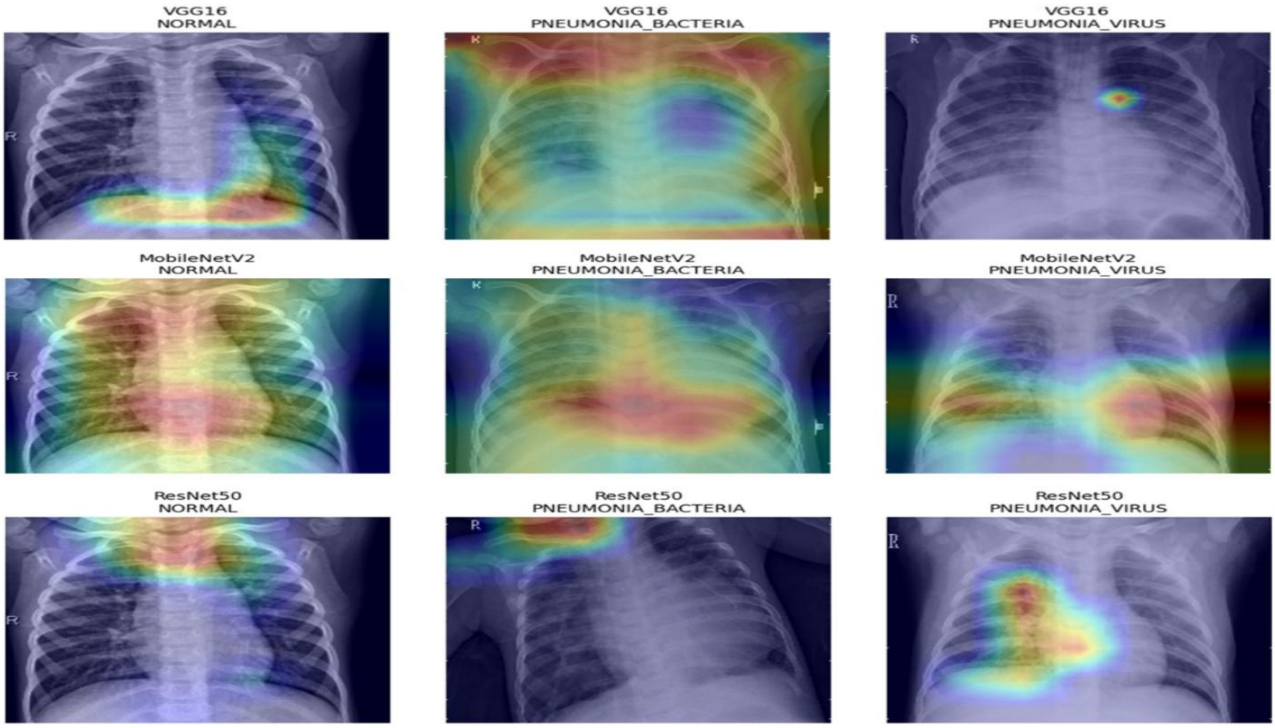

Figure 9 Grad-CAM Result

As shown in Figure 9, differences in the Grad-CAM heatmap distribution arise because each model learns distinct feature patterns from the data, and this is most evident in the metrics of VGG16 as the best-performing model. To contextualize the contribution of Grad-CAM in this study, it is important to consider the baseline scenario without Grad-CAM. When models are trained and evaluated without any interpretability mechanism, the classification outputs (accuracy, F1-score, AUC) remain identical, but clinicians and researchers have no means to assess whether the model is focusing on clinically relevant lung regions or spurious artifacts such as body contours, bone structures, or imaging labels. In this baseline (non-Grad-CAM) setting, a model achieving 82% accuracy might still be attending to irrelevant regions, reducing trust and clinical validity. The introduction of Grad-CAM in this study adds a visual validation layer:

it confirms that VGG16's quantitative superiority is accompanied by more diagnostically meaningful feature localization.

In VGG16, metrics such as high recall for the Normal (98.7%) and Bacterial (86.6%) classes, together with high specificity across all classes (92.9%), indicate that the model consistently focuses on truly relevant lung regions. Consequently, VGG16 heatmaps tend to be more concentrated on parenchymal lesions or infiltrates (for example, radiopaque patches in pneumonia) while largely suppressing activation on bones, body contours, and background areas.

In contrast, MobileNetV2 and ResNet50, which achieve slightly lower G-Mean and F1-score values especially for the Viral class (for instance, a G-Mean of 71.7 for MobileNetV2 versus 72.8 for VGG16) more often display heatmaps that are more diffuse or extend into less relevant regions, reflecting residual uncertainty in distinguishing subtle differences between bacterial pneumonia, viral pneumonia, and normal images. In summary, the better the quantitative metrics (as in VGG16), the sharper and more targeted the Grad-CAM focus on clinically relevant lung structures, whereas slightly lower performance tends to manifest as less selective activations in the heatmaps.

IV. CONCLUSION

This study aims to evaluate and compare the performance of three CNN architectures VGG16, MobileNetV2, and ResNet50 for classifying chest X-ray images of children aged 1-5 years into Normal, Bacterial Pneumonia, and Viral Pneumonia classes, using accuracy, precision, recall, specificity, F1-score, G-Mean, and AUC as evaluation metrics, while also providing visual interpretability through Grad-CAM. The experimental results show that all three models achieve overall accuracies in the range of 79-82%, with VGG16 achieving the best performance at 82%. This is supported by more balanced per-class metrics, particularly for the Normal and Bacterial Pneumonia classes, where VGG16 attains superior combinations of recall, specificity, F1-score, and G-Mean compared with MobileNetV2 and ResNet50. For the Viral Pneumonia class, all models still face challenges, as reflected by relatively lower recall and AUC values. Grad-CAM visualizations confirm that VGG16 tends to produce more focused heatmaps over clinically relevant lung regions, leading to the conclusion that VGG16 is the most promising candidate for development as a chest X-ray based computer-aided diagnosis system for pediatric pneumonia. Despite these contributions, this study has several limitations that should be acknowledged. First, the dataset originates from a single institution (Guangzhou Women and Children's Medical Center), which may limit the generalizability of the trained models to other clinical settings or imaging equipment. Second, the class imbalance between Viral Pneumonia and other categories may have contributed to the relatively lower performance on that class, as standard augmentation may not fully compensate for this imbalance. Third, the study did not implement a quantitative comparison between models with and without Grad-CAM in terms of clinical acceptance or radiologist evaluation, which would further validate the interpretability benefit. Future research can focus on improving detection performance for Viral Pneumonia and validating the models on larger and more diverse multi-center datasets to enhance robustness and readiness for deployment in real clinical settings.

REFERENCES

[1] S. Phine, "Pneumonia Classification Using Deep Learning VGG19 Model," in *IEEE International Conference on Control and Automation, ICCA*, IEEE Computer Society, 2023, pp. 67–71. doi: 10.1109/ICCA51723.2023.10181954.

[2] A. F. Nurjannah, A. S. D. Kurniasari, Z. Sari, and Y. Azhar, "Pneumonia Image Classification Using CNN with Max Pooling and Average Pooling," *Jurnal RESTI*, vol. 6, no. 2, pp. 330–338, Apr. 2022, doi: 10.29207/resti.v6i2.4001.

[3] E. Shakeri and B. Far, "Advancing Pneumonia Classification and Detection: Comparative Analysis of Deep Learning Models Using Convolutional Neural Networks," in *Proceedings - 2024 IEEE International Conference on Information Reuse and Integration for Data Science, IRI 2024*, Institute of Electrical and Electronics Engineers Inc., 2024, pp. 108–113. doi: 10.1109/IRI62200.2024.00033.

[4] P. Dhungana, M. Roy, R. Aryal, S. Chaudhary, T. R. Mahesh, and P. Dhungana, "Ensemble Deep Learning Approach for Pneumonia Detection Using DenseNet, MobileNet, and EfficientNet with Transfer Learning," in *2025 International Conference on Data Science and Business Systems, ICDSBS 2025*, Institute of Electrical and Electronics Engineers Inc., 2025. doi: 10.1109/ICDSBS63635.2025.11031996.

[5] H. Bysani, S. Garg, A. Danda, T. Singh, C. Jyotsna, and P. Duraisamy, "Detection of Pneumonia in Chest X-Ray Using Ensemble Learners and Transfer Learning with Deep Learning Models," in *2023 14th International Conference on Computing Communication and Networking Technologies, ICCCNT 2023*, Institute of Electrical and Electronics Engineers Inc., 2023. doi: 10.1109/ICCCNT56998.2023.10307035.

[6] Y. Xu, G. He, and D. Chen, "A ResNet Based Pneumonia Classification Algorithm for Chest X-Ray Images," in *2024 IEEE 6th International Conference on Power, Intelligent Computing and Systems, ICPICS 2024*, Institute of Electrical and Electronics Engineers Inc., 2024, pp. 1588–1592. doi: 10.1109/ICPICS62053.2024.10796977.

[7] T. S. Arulananth, S. W. Prakash, R. K. Ayyasamy, V. P. Kavitha, P. G. Kuppusamy, and P. Chinnasamy, "Classification of Paediatric Pneumonia Using Modified DenseNet-121 Deep-Learning Model," *IEEE Access*, vol. 12, pp. 35716–35727, 2024, doi: 10.1109/ACCESS.2024.3371151.

[8] F. Ojo, S. Guntupalli, and O. Adewunmi, "A Two-Phase Pneumonia Detection and Subclassification Model Using Hybridized EfficientNet and Random Forest," in *Proceedings - 2024 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2024*, Institute of Electrical and Electronics Engineers Inc., 2024, pp. 5032–5037. doi: 10.1109/BIBM62325.2024.10822155.

[9] Kementerian Kesehatan Republik Indonesia, "Keputusan Menteri Kesehatan Republik Indonesia Nomor HK.01.07/MENKES/2147/2023 tentang Pedoman Nasional Pelayanan Kedokteran Tata Laksana Pneumonia pada Dewasa," Jakarta, Nov. 2023.

[10] National Center for Health Statistics, "Number of visits to emergency departments with pneumonia due to infectious organism as the primary diagnosis." Accessed: Apr. 30, 2026. [Online]. Available: https://www.cdc.gov/nchs/fastats/pneumonia.htm

[11] N. Center for Health Statistics, "National Hospital Ambulatory Medical Care Survey: 2022 Emergency Department Summary Tables," 2022. [Online]. Available: https://www.cdc.gov/nchs/nhamcs/documentation/index.html#cdc_research_or_data_

[12] M. Shenghan and T. T. Toe, "High-Precision Detection and Classification of Pneumonia Cases Based on Optimized ResNet," in *Proceedings of the IEEE International Conference on Computer and Communications, ICCC*, Institute of Electrical and Electronics Engineers Inc., 2024, pp. 2090–2095. doi: 10.1109/ICCC62609.2024.10942079.

[13] Y. Lu and Y. Zheng, "A Lightweight Residual Network for Pneumonia Classification," in *Proceedings - 2024 4th Asia-Pacific Conference on Communications Technology and Computer Science, ACCTCS 2024*, Institute of Electrical and Electronics Engineers Inc., 2024, pp. 597–601. doi: 10.1109/ACCTCS61748.2024.00111.

[14] C. V. Narayana, Y. Neha, C. V. Kumar, V. G. Manasa, and T. C. Teja, "Classification of Pneumonia in Chest Radiographs using Deep Learning Techniques," in *Proceedings of 8th International Conference on Inventive Computation Technologies, ICICT 2025*, Institute of Electrical and Electronics Engineers Inc., 2025, pp. 104–110. doi: 10.1109/ICICT64420.2025.11004923.

[15] S. Pappula, T. Nadendla, N. B. Lomadugu, and S. Revanth Nalla, "Detection and Classification of Pneumonia Using Deep Learning by the Dense Net-121 Model," in *2023 9th International Conference on Advanced Computing and Communication Systems, ICACCS 2023*, Institute of Electrical and Electronics Engineers Inc., 2023, pp. 1671–1675. doi: 10.1109/ICACCS57279.2023.10113110.

[16] S. Shi, "A comparison of MobileNetV2, VGG16, and ResNet50 for classifying brain tumors," in *2024 IEEE 2nd International Conference on Electrical, Automation and Computer Engineering, ICEACE 2024*, Institute of Electrical and Electronics Engineers Inc., 2024, pp. 176–180. doi: 10.1109/ICEACE63551.2024.10898757.

[17] Paul Mooney, "Chest X-Ray Images (Pneumonia)," Kaggle. [Online]. Available: https://www.kaggle.com/datasets/paultimothymooney/chest-xray-pneumonia/data

[18] M. Ennab and H. Mcheick, "Enhancing Pneumonia Diagnosis Through AI Interpretability: Comparative Analysis of Pixel-Level Interpretability and Grad-CAM on X-ray Imaging With VGG19," *IEEE Open Journal of the Computer Society*, vol. 6, pp. 1155–1165, 2025, doi: 10.1109/OJCS.2025.3582726.

[19] E. Şahin, A. Çelikten, S. Demirel, A. Akpulat, K. Budak, and H. Karataş, "Pneumonia Detection in Chest X-Ray Images Using Deep Learning and Grad-CAM++ Visualization," in *2024 Innovations in Intelligent Systems and Applications Conference, ASYU 2024*, Institute of Electrical and Electronics Engineers Inc., 2024. doi: 10.1109/ASYU62119.2024.10756959.

[20] C. Paramita, S. Rakasiwi, P. N. Andono, G. F. Shidik, Shier Nee Saw, and M. I. Rafsanjani, "Deep Learning-Based Eye Disorder Classification: A K-Fold Evaluation of EfficientNetB and VGG16 Models," *Scientific Journal of Informatics*, vol. 12, no. 3, pp. 441–452, Sep. 2025, doi: 10.15294/sji.v12i3.26257.

[21] E. Mustika Sari, C. Sabila, R. Fakhrizal Adam, and R. Kurniawan, "Analisis dan Prediksi Indeks Kualitas Udara Jakarta: Penerapan Algoritma XGBoost," *Jurnal Nasional Teknologi dan Sistem Informasi*, vol. 11, no. 2, pp. 161–169, Sep. 2025, doi: 10.25077/teknosi.v11i2.2025.161-169.

[22] B. Bryant and M. R. Abid, "COVID-19 and Pneumonia X-ray Classification via Custom Convolutional Neural Network," in *2022 IEEE 13th Annual Information Technology, Electronics and Mobile Communication Conference, IEMCON 2022*, Institute of Electrical and Electronics Engineers Inc., 2022, pp. 239–245. doi: 10.1109/IEMCON56893.2022.9946540.

[23] M. Fu, C. Tantithamthavorn, and T. Le, "DAViT: A Domain-Adapted Vision Transformer for Automated Pneumonia Detection and Explanation Using Chest X-Ray Images," *IEEE Access*, vol. 13, pp. 103033–103044, 2025, doi: 10.1109/ACCESS.2025.3579314.

[24] J. Yang, H. Li, L. Zhang, L. Zhao, S. Jiang, and H. Xie, "Multilabel Concrete Bridge Damage Classification Using Graph Convolution," *IEEE Trans. Instrum. Meas.*, vol. 73, 2024, doi: 10.1109/TIM.2024.3400345.

[25] "G-Mean Score (GMS)," Permetrics. Accessed: Nov. 16, 2025. [Online]. Available: https://permetrics.readthedocs.io/en/latest/pages/classification/GMS.html

[26] C. Paramita, C. S. Simbolon, A. S. Pamungkas, J. M. Triono, E. P. Widi Utomo, and E. R. Subhiyakto, "Analisis Pengaruh SMOTE terhadap Kinerja Model KNN untuk Prediksi Risiko Stroke," *Jurnal Informatika: Jurnal Pengembangan IT*, vol. 10, no. 4, pp. 978–988, Sep. 2025, doi: 10.30591/jpit.v10i4.8809.

[27] M. Wildan Akhya, C. Paramita, P. N. Andono, and A. History, "Jurnal Teknologi dan Manajemen Informatika Market Value Tier Classification of Indonesian Football Players Using Ensemble Machine Learning and SHAP Analysis Article Info ABSTRACT," vol. 11, no. 2, pp. 212–223, 2025, [Online]. Available: http://http://jurnal.unmer.ac.id/index.php/jtmi