E-ISSN: 2963-4946
Vol. 1 No. 5 Februaray 2026, pp. 206-223
DOI: 10.59888/ajosh.v4i10.736

Asian Journal of
Social and Humanities

Comparison of VGG16 and EfficientNetB3 for Pneumonia Classification on Chest X-Ray Images Using Grad-CAM

Satrya Permana
Sekolah Tinggi Manajemen Informatika dan Komputer (STMIK) IKMI Cirebon, Indonesia
Corresponding Author: satryapermana9@gmail.com

Keywords:
pneumonia; chest X-ray; VGG16; EfficientNetB3; Grad-CAM

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

Copyright © 2026 by Author. Published by
Pelopor Publikaasi Akademika

Abstract

Background: High accuracy in chest X-ray (CXR) pneumonia classification does not by itself show which image regions drive a model's decisions, and public paediatric datasets are vulnerable to data leakage and shortcut learning.
Objective: This study compares VGG16 and EfficientNetB3 for binary Normal–Pneumonia classification and examines their outputs using Gradient-weighted Class Activation Mapping (Grad-CAM).
Method: A quantitative comparative experiment used ImageNet weights and 224 × 224-pixel inputs. A pool of 5,242 image entries was stratified per image into 3,669 training, 786 validation, and 787 internal test images. Four single-seed scenarios compared frozen backbones with backbones unfrozen under a last-20-layer rule. Wilson confidence intervals and bounded McNemar tests were derived from the confusion matrices.
Results: VGG16 fine-tuning achieved 97.84% accuracy (95% CI 96.57–98.65%), 97.26% recall, 99.51% specificity, 98.53% F1-score, and 0.9977 ROC-AUC, with 16 false negatives and one false positive. EfficientNetB3 fine-tuning achieved 95.30% accuracy and 0.9963 ROC-AUC, with 37 false negatives and no false positives; the accuracy difference remained significant under all compatible discordance patterns ($p \leq 0.009$). Several Grad-CAM heatmaps showed strong activations at image borders.
Conclusion: VGG16 fine-tuning performed best in this preliminary experiment, but border activations and per-image splitting indicate possible shortcut learning. Patient-level splitting, the original test set, multiple seeds, and external validation are required.

INTRODUCTION

Pneumonia remains a leading infectious cause of death in early childhood. The World Health Organization, (2022) reported that pneumonia killed 740,180 children under five years of age in 2019, accounting for 14% of all deaths in this age group. UNICEF, (2019) estimated that approximately 19,000 Indonesian children under five died of pneumonia in 2018, placing Indonesia among the countries with the highest absolute number of child pneumonia deaths. Chest radiography is commonly used to support the diagnosis of pneumonia, and standardized criteria for interpreting paediatric chest radiographs have

been established under the coordination of the World Health Organization (Cherian et al., 2005).

However, the interpretation of these images requires trained readers whose availability is uneven. Indonesia had 2,161 registered radiologists for a population of about 276 million people living across more than 17,000 islands, with radiologists concentrated in Java and Sumatra and in short supply in remote provinces (Yunus, 2024). This gap motivates research on automated CXR classification as a potential decision-support tool, provided that its performance and reasoning are evaluated rigorously.

Chest X-ray (CXR) images are widely used in pneumonia classification research because they allow the development of models that group images according to visual patterns. Jain et al. (2020) demonstrated the application of several convolutional neural network (CNN) architectures to Normal and Pneumonia classification. However, success in separating labels on a particular dataset needs to be assessed in a measurable way. High accuracy does not explain all characteristics of a model, particularly when the class distribution is imbalanced or when the research data differ from the intended conditions of use.

Public datasets allow experiments to be documented and compared with previous studies. A paediatric CXR dataset with Normal and Pneumonia classes was also used by Saboo et al. (2023). The dataset was originally released by Kermany et al. (2018), together with a separate test partition of 624 images from patients not included in the training partition. Reusing data is relevant for comparative evaluation, but the contribution of a study must lie in a clear research question and evaluation design. Merging folders and re-splitting the data are evaluation-protocol decisions; they do not create a new dataset and do not by themselves constitute evidence of novelty.

VGG16 and EfficientNetB3 were selected because of their different architectural designs. Shaikh et al. (2023) described the stacked 3 × 3 convolutions of VGG16 and the use of the EfficientNet family in pneumonia classification. The VGG16 architecture itself was introduced by Simonyan and Zisserman (2015) as a deep network built from small stacked convolutional filters. Chowdhury et al. (2021) discussed compound scaling in EfficientNet, which balances network depth, width, and resolution. The compound scaling principle and the EfficientNet family, including EfficientNetB3, were originally proposed by Tan and Le (2019). VGG16 provides a comparator with a regular convolutional structure, whereas EfficientNetB3 represents a family designed with parameter efficiency in mind.

These characteristics form the basis for empirical testing without assuming the superiority of either model in advance. Transfer learning allows weights trained on ImageNet to be used as a starting point. Rahman et al. (2020) tested the transfer of pretrained features to pneumonia classification, whereas Avola et al., (2022) evaluated the transfer-learning capability of a range of CXR architectures. Shaikh et al. (2023) described freezing the backbone and training new layers for the target task.

This study compares frozen and unfrozen backbone configurations in a preliminary experiment; differences in results are interpreted according to how the models were

actually trained rather than assumed to be the effect of a single isolated factor. A predicted class and its output score do not reveal which parts of an image support the model's decision. Grad-CAM uses the gradients of the target score with respect to the feature maps to produce a map of spatial contributions, as described by Chowdhury et al. (2021).

The original formulation of Grad-CAM was introduced by Selvaraju et al. (2017). Mahmud et al. (2020) combined a multi-dilation CNN with Grad-CAM for pneumonia classification, whereas Saporta et al. (2022) compared several interpretability techniques. Visual explanations support the analysis of model behaviour, but activations that appear to lie within the lungs do not in themselves prove that a lesion has been localized correctly. Three closely related studies define the position of this research. Jain et al. (2020) compared VGG16, VGG19, ResNet50, and InceptionV3; Mahmud et al. (2020) examined a multi-dilation CNN with Grad-CAM; and Chowdhury et al. (2021) evaluated EfficientNet variants from B0 to B5. None of these studies focused on the VGG16–EfficientNetB3 pair under the four scenarios used here.

Other developments include an explainable transformer ensemble (Ukwuoma et al., 2023) and Vision Transformers for pneumonia (Okolo et al., 2022). The position of this study is limited to the selected comparison and is not claimed to be the first such comparison worldwide.A further concern is that many studies on this paediatric dataset report near-perfect accuracy, which should be interpreted with caution. Deep networks are prone to shortcut learning, that is, the use of spurious cues that perform well on test data drawn from the same distribution but fail to transfer to new conditions (Geirhos et al., 2020).

In CXR analysis, Zech et al. (2018) showed that pneumonia models could exploit hospital-specific image characteristics and generalized less well to external sites, whereas DeGrave et al., (2021) showed that radiographic classifiers relied on laterality markers, text annotations, and regions outside the lung fields rather than on pathological signal. Data leakage creates a related bias: when images from the same patient are distributed across training and test partitions, the test set no longer measures generalization to new patients. Accordingly, the novelty of this study does not lie in a new dataset or architecture, but in a transparent comparison of the VGG16–EfficientNetB3 pair under four freezing scenarios that reports error counts, confidence intervals, training behaviour, and Grad-CAM evidence of possible shortcut cues together with the limitations of the implementation.

This study aims to compare the performance of VGG16 and EfficientNetB3 under frozen and unfrozen backbone configurations and to examine their patterns of visual contribution using Grad-CAM. The analysis uses the outputs of a preliminary experiment with one seed and one fixed split. Its contribution consists of a comparison of metrics, error counts, training behaviour, and visual inspection, reported together with the limitations of the implementation. The study is restricted to binary classification on a paediatric dataset and does not address stand-alone diagnosis, pathogen identification, or the assessment of pneumonia severity.

METHOD

Experimental Design

This study used a quantitative approach with a comparative experimental design. The treatment variables were the backbone architecture and the layer-freezing configuration. The outcomes compared were classification metrics, error counts, the actual number of epochs, and the recorded training time. The Normal label was coded as 0 and Pneumonia as 1. All four scenarios used the same split with seed 42. Results come from a single run of each scenario; no averaging across seeds or significance testing was performed. Significance testing was not part of the original training pipeline; post hoc confidence intervals and bounded McNemar tests were subsequently derived from the stored confusion matrices, as described in the Evaluation and Experiment Records subsection.

Data Source and Image Records

The data were obtained from the Chest X-Ray Images (Pneumonia) dataset uploaded by Paul Mooney to Kaggle (https://www.kaggle.com/datasets/paultimothymooney/chest-xray-pneumonia). The origin of the images, namely paediatric anterior–posterior radiographs of children aged one to five years from Guangzhou Women and Children's Medical Center, is also described by Saboo et al. (2023). The dataset was originally published by Kermany et al. (2018). In the source repository, Pneumonia file names encode a patient identifier (e.g., personNNN_bacteria_NNNN), so a patient-grouped split is technically feasible; this identifier, however, was not used in the preliminary experiment reported here. The link is given to identify the data source, not as a new publication.

The notebook collected JPG, JPEG, and PNG paths from the Normal and Pneumonia folders. The reported counts follow the notebook output and do not represent the number of unique patients. The output recorded 5,242 entries from the source train and val folders and 624 entries from the original test folder, which were kept separate. In total, 5,866 entries were counted, which differs from the figure of 5,863 given in the proposal and the source description. The notebook did not store a comprehensive audit that could explain this difference. No evidence of duplicate-hash checking, label-conflict checking, or patient separation was available for this experiment.

The counts are therefore treated as numbers of image entries in the output and are not claimed to be numbers of audited unique images. A comparison with the folder counts of the source repository (train: 1,341 Normal and 3,875 Pneumonia; val: 8 Normal and 8 Pneumonia; 5,232 images in total) shows that the Normal count of the pool (1,349) matches exactly, whereas the Pneumonia count (3,893) exceeds the expected 3,883 by 10 entries. Plausible sources are hidden system files (e.g., __MACOSX copies) or a nested duplicate folder captured by the recursive path search. This could not be verified retrospectively because no hash-based duplicate audit was stored; the 10 surplus entries (0.19% of the pool) are therefore reported as an unresolved data-integrity issue.

Data Splitting in the Preliminary Experiment

The notebook merged the source train and val folders and then applied a two-stage train_test_split with label stratification and random_state 42. The first stage separated 70% of the data for training and 30% as temporary data. The second stage divided the temporary data equally into validation and internal test sets. The result was 3,669 training images, 786 validation images, and 787 internal test images (Table 1). Stratification kept the class proportions approximately equal across subsets, but it does not guarantee the separation of patient identities. The 624 images from the original test folder were collected separately.

The evaluation function that produced the metrics reported in this article used X_test_split, namely the internal test set of 787 images. Therefore, the result tables are not reported as an evaluation on the original 624-image test set or as the result of re-merging all source train, val, and test partitions. The training data were used to optimise model parameters, the validation data to monitor training and select checkpoints, and the internal test data to compute classification results. No patient-group manifest was available in this output.

Re-splitting all partitions and repeating the experiment with several seeds are planned developments and are not presented as procedures that produced the preliminary results. Because the original 624-image test folder was drawn from patients who are not represented in the source training partition (Kermany et al., 2018), it is the appropriate primary held-out test for the repeated experiment. In contrast, the internal test set of the present run may contain images from patients who also contributed training images, which can inflate all reported metrics.

Table 1. Distribution of image entries in the source train+val pool

Partition	Normal	Pneumonia	Total	Share of pool
Training	944	2,725	3,669	69.99%
Validation	202	584	786	14.99%
Internal test	203	584	787	15.01%
Total pool	1,349	3,893	5,242	100%

Source: Data Processed

Preprocessing and Architecture

Images were read as three-channel JPEGs and resized to 224 × 224 pixels. The code used vgg_preprocess for VGG16 and eff_preprocess for EfficientNetB3; in the EfficientNetB3 model, rescaling is built into the called architecture. These details are part of the implementation configuration of this study. Both models received inputs of the same size, so the results do not represent the optimal resolution of each architecture. A shared input size was also used in the CXR experiments of Shaikh et al. (2023). The notebook applied augmentation only to the training data, comprising horizontal RandomFlip, RandomRotation(0.08), RandomZoom(0.08), RandomContrast(0.1), and RandomBrightness(0.15). In Keras, a rotation factor of 0.08 is a fraction of a full turn,

that is, approximately ±28.8°, not ±8°. The code applied model preprocessing before per-batch augmentation.

This order is recorded as a limitation because the brightness transformation, with its default value range of 0–255, was not adjusted to the value space produced by VGG preprocessing. The validation and test data were not augmented. Because Keras VGG16 preprocessing converts images to mean-centred BGR values, whereas EfficientNetB3 preprocessing passes pixel values through unchanged and rescales them inside the model, this ordering affected the VGG16 scenarios more than the EfficientNetB3 scenarios. In the repeated experiment, augmentation should be applied to raw pixel values before model-specific preprocessing, and the rotation factor should be reduced to a radiographically plausible range (e.g., 0.028, approximately ±10°).

Because the Pneumonia class dominated the training set, class weights were computed from the training data using the balanced weighting in Equation (1), where N is the number of training images, K is the number of classes, and n_c is the number of training images in class c. With N = 3,669, K = 2, n_{Normal} = 944, and $n_{Pneumonia}$ = 2,725, this gives weights of 1.9433 for Normal and 0.6732 for Pneumonia. The weights increase the contribution of the minority Normal class to the binary cross-entropy loss without changing the number of images.

$$w_c = N / (K \times n_c) \qquad (1)$$

Both architectures used ImageNet weights without their original classifiers. The head consisted of Global Average Pooling, Batch Normalization, a Dense layer of 256 units with ReLU, Dropout 0.5, a Dense layer of 128 units with ReLU, Dropout 0.3, and a single sigmoid output neuron. In the transfer-learning (TL) scenarios, the whole backbone was frozen. In the fine-tuning (FT) scenarios, the backbone was made trainable and all layers before the last 20 layers were frozen again. Because the VGG16 backbone has fewer than 20 layers, this rule unfroze the entire VGG16 backbone; its coverage is therefore not equivalent to the last 20 layers of EfficientNetB3. Batch Normalization layers in the backbone were not specifically frozen in the FT scenarios. The four scenarios are summarized in Table 2.

Table 2. Training scenarios compared

Scenario	Architecture	Treatment
S1	VGG16	Backbone frozen; head trained from scratch.
S2	VGG16	New model with ImageNet weights; the rule of unfreezing the last 20 layers covers the entire backbone.
S3	EfficientNetB3	Backbone frozen; head trained from scratch.
S4	EfficientNetB3	New model with ImageNet weights; the last 20 backbone layers unfrozen.

Source: Data Processed

Training Configuration

Each of the four scenarios called the model-building function separately. S2 did not load the S1 checkpoint and S4 did not load the S3 checkpoint, so the term FT in this

article refers to training with an unfrozen backbone starting from ImageNet weights. The results are not interpreted as the effect of continuing training from the same TL model. All models were optimized using Adam and binary cross-entropy, and checkpoint selection monitored the highest validation AUC. The full configuration is presented in Table 3. Staged fine-tuning, in which the FT model is initialized from the TL checkpoint and the backbone Batch Normalization layers are kept frozen, was not performed in this run and is reserved for the repeated experiment.

Table 3. Experimental configuration

Component	Configuration in the preliminary experiment
Seed	42; one run per scenario
Batch size and optimizer	32; Adam
Loss function	Binary cross-entropy; class weights from the training set
Maximum epochs	TL 20; FT 30; models built separately
Initial learning rate	TL 0.001; FT 0.0001
Checkpoint selection	Highest validation AUC
EarlyStopping	Monitor val_auc; patience 5; restore best weights
ReduceLROnPlateau	Monitor val_loss; patience 3; factor 0.5; minimum 0.0000001
Classification threshold	0.5

Source: Data Processed

Evaluation and Experiment Records

Evaluation used the same 787 internal test images for all scenarios, with Pneumonia as the positive class. A true positive (TP) is a pneumonia image predicted as pneumonia; a true negative (TN) is a normal image predicted as normal; a false positive (FP) is a normal image predicted as pneumonia; and a false negative (FN) is a pneumonia image predicted as normal. The sigmoid output was thresholded at 0.5. This score is not treated as a calibrated clinical probability.

Accuracy measures the proportion of all correct predictions. Precision is the proportion of TP among all positive predictions. Recall, or sensitivity, is the proportion of TP among all pneumonia images. Specificity is the proportion of TN among all normal images. The F1-score is the harmonic mean of precision and recall for the Pneumonia class. These metrics are defined in Equations (2)–(6).

$$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN) \quad (2)$$

$$\text{Precision} = TP / (TP + FP) \quad (3)$$

$$\text{Recall} = TP / (TP + FN) \quad (4)$$

$$\text{Specificity} = TN / (TN + FP) \quad (5)$$

$$F1 = 2 \times \text{Precision} \times \text{Recall} / (\text{Precision} + \text{Recall}) \quad (6)$$

ROC-AUC was computed from the continuous scores using scikit-learn; this value is distinguished from the Keras AUC approximation used to monitor training. Metrics are

reported as single-run values for each scenario rather than as mean ± standard deviation. Accuracy, precision, recall, specificity, and F1-score are expressed as percentages, and ROC-AUC on a 0–1 scale. Differences between scenarios are discussed descriptively in percentage points. No significance testing or conclusions about stability across seeds were made.Two post hoc analyses were added to quantify uncertainty using only the stored prediction counts.

First, 95% confidence intervals for accuracy, precision, recall, and specificity were computed with the Wilson score method (Wilson, 1927), which remains valid for proportions close to 100%. Second, paired differences in accuracy between scenarios were assessed with the exact McNemar test (McNemar, 1947). Because the per-image agreement between two models cannot be recovered from aggregated confusion matrices, the number of discordant images was bounded: if model A made d more correct predictions than model B, the number of images misclassified only by A could range from zero to the smaller of the total errors of A and the total errors of B minus d.

The exact McNemar p-value was computed for every admissible discordance pattern, and a difference was considered robust only when the largest compatible p-value was below the Bonferroni-adjusted threshold of 0.0167 for three planned comparisons. These analyses quantify the uncertainty of a single run on the internal test set and do not replace multi-seed or patient-level evaluation. The experiment was run on Google Colab using Python and TensorFlow/Keras. The output showed TensorFlow 2.20.0 and the device /physical_device:GPU:0. The Python version, Keras version, physical GPU model, and execution date were not printed in the log and are therefore not specified. The reported times are the durations of model.fit as measured by the notebook, not inference benchmarks or measures of efficiency across devices.

Grad-CAM Analysis

Grad-CAM was computed on block5_conv3 for VGG16 and on top_conv for EfficientNetB3. The gradients were spatially averaged and then used to weight the feature maps, followed by ReLU and map normalization, in line with the Grad-CAM description in Chowdhury et al. (2021). The procedure follows the original formulation of Selvaraju et al. (2017). For a feature map A^k of layer k and the target score y^c, the channel weight and the class activation map are computed as in Equations (7) and (8), where Z is the number of spatial positions in the feature map.

$$\alpha_k^c = (1/Z) \Sigma_i \Sigma_j \partial y^c / \partial A_{ij}^k \quad (7)$$

$$L_{\text{Grad-CAM}}^c = \text{ReLU}(\Sigma_k \alpha_k^c A^k) \quad (8)$$

The notebook implementation always used the single sigmoid channel predictions[:, 0], that is, the Pneumonia score. Consequently, maps for samples predicted as Normal still display contributions to the Pneumonia score and are not described as Normal-target maps. Visualization used models S2 and S4 after weight selection based on validation. The samples were the first two Normal images and the first two Pneumonia images in the

internal test order; they were neither random samples nor a selection covering all TP, TN, FP, and FN groups.

Both models were applied to the same four images. For the visual panels, images were resized using PIL, whereas the evaluation pipeline used TensorFlow; panel scores are therefore not assumed to be identical to the last digit with those of the evaluation pipeline. The maps were overlaid on the images with a coefficient of 0.4. The analysis discusses the location of activations and the possible contribution of image borders or markers. Intensities were normalized per map, so colour is not used to measure absolute strength across models. No lesion annotations or physician assessments measuring the accuracy of activation locations were available. The Grad-CAM assessment is therefore qualitative.

RESULTS AND DISCUSSION

Data Distribution and Completeness of Results

The pool of 5,242 entries comprised 1,349 Normal and 3,893 Pneumonia images. The Pneumonia class accounted for about 74.27% of the pool and the Normal class for about 25.73%. A similar distribution was maintained in each subset, as shown in Table 1. The Normal class received a larger weight because it contained fewer images. Applying class weights changes the contribution to the loss during training and does not change the number of original images.

All four scenarios had stored test results. S1 ran for six epochs, S2 for 14 epochs, S3 for 20 epochs, and S4 for 21 epochs. Weights were restored from epochs 1, 9, 15, and 16, respectively, based on validation AUC. These outputs come from the completed preliminary experiment, which differs from a later multi-seed revision experiment that was interrupted. No additional evaluation on the original 624-image test set is included in the summarized results.

Comparison of Classification Performance

Table 4 shows the test results on 203 Normal and 584 Pneumonia images. VGG16 FT obtained the highest accuracy, recall, F1-score, and ROC-AUC. EfficientNetB3 TL and FT reached 100% precision and specificity on this test set because they produced no FP. These 100% values apply to the composition of this test set and do not imply error-free performance in other populations.

Table 4. Results on the Internal Test Set of 787 Images with Seed 42

Metric	VGG16 TL	VGG16 FT	B3 TL	B3 FT
Accuracy (%)	91.99	97.84	93.65	95.30
Precision (%)	94.38	99.82	100.00	100.00
Recall (%)	94.86	97.26	91.44	93.66
Specificity (%)	83.74	99.51	100.00	100.00
F1 Pneumonia (%)	94.62	98.53	95.53	96.73
ROC-AUC	0.9719	0.9977	0.9937	0.9963

Note. TL = transfer learning with a frozen backbone; FT = fine-tuning with an unfrozen backbone; B3 = EfficientNetB3.

VGG16 FT was 5.85 percentage points higher in accuracy and 3.91 points higher in F1-score than VGG16 TL. EfficientNetB3 FT was 1.65 points higher in accuracy and 1.20 points higher in F1-score than EfficientNetB3 TL. Compared with EfficientNetB3 FT, VGG16 FT was 2.54 points higher in accuracy, 3.60 points higher in recall, and 1.80 points higher in F1-score. Conversely, the specificity of VGG16 FT was 0.49 points lower. The results show a trade-off between recognizing pneumonia and producing false positives within the same run, without demonstrating a statistically significant advantage.

Table 5. Prediction counts on the internal test set

Scenario	TN	FP	FN	TP
VGG16 TL	170	33	30	554
VGG16 FT	202	1	16	568
EfficientNetB3 TL	203	0	50	534
EfficientNetB3 FT	203	0	37	547

Source: Data Processed

Compared with VGG16 TL, VGG16 FT reduced the number of FN from 30 to 16 and the number of FP from 33 to one. EfficientNetB3 FT reduced FN from 50 to 37 compared with EfficientNetB3 TL, while FP remained zero. When the two FT models are compared, VGG16 recognized 21 additional Pneumonia images at the cost of one additional FP. This difference explains why the accuracy and recall of VGG16 FT were higher, even though the specificity of EfficientNetB3 FT was higher.

In a screening-oriented reading of these results, false negatives are usually the more consequential error because a missed pneumonia image would not be flagged for further review. From this perspective, the reduction of FN in VGG16 FT is the most practically relevant difference between the two fine-tuned models. However, this interpretation depends on the intended use, the decision threshold, and the prevalence of pneumonia in the target population. Because all scores were thresholded at 0.5 and were not calibrated, the observed balance between FN and FP should not be generalized to other thresholds or settings.

Training Behaviour and Computational Requirements

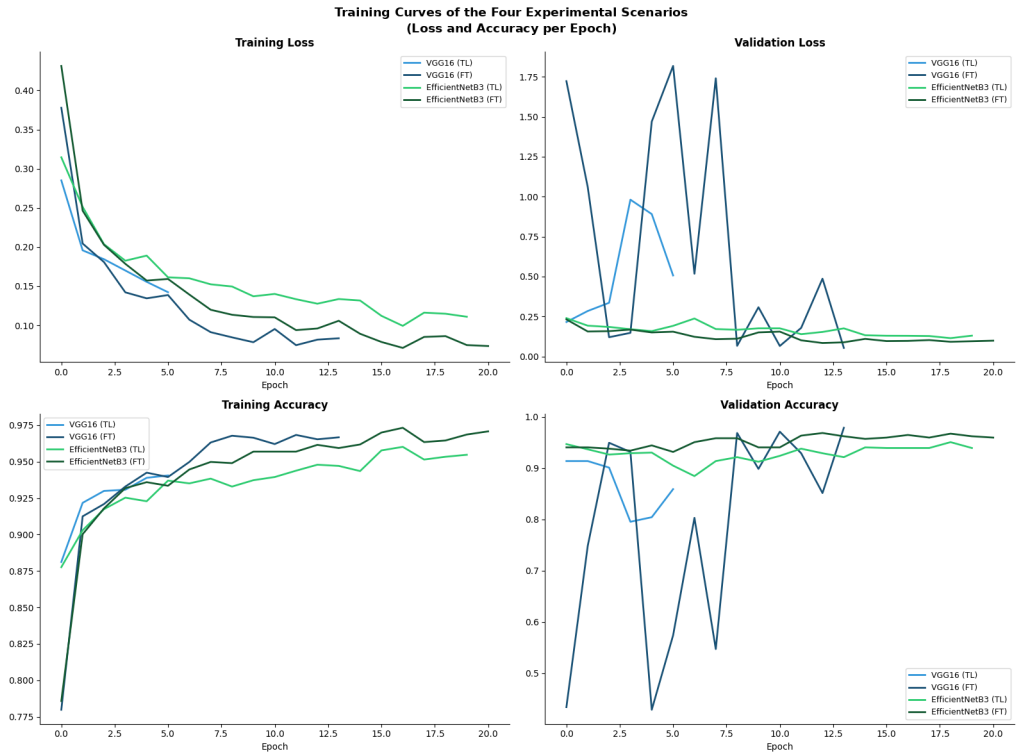


Figure 1. Stored training curves of the four scenarios. The epoch axis in the figure starts from index 0, whereas epochs in the text are counted from 1

The training curves (Figure 1) show decreasing training loss in all four scenarios, whereas the validation curves of VGG16 fluctuated more strongly. In S2, validation accuracy moved from 43.38% in the first epoch, fell to 42.88% in epoch 5, and then reached 96.82% in epoch 9. Although epoch 14 had a validation accuracy of 97.84%, the checkpoint was restored from epoch 9 because the selection criterion was validation AUC rather than validation accuracy. In S1, epoch 1 was the best checkpoint and training stopped at epoch 6. This situation shows the importance of distinguishing between the last epoch and the checkpoint that is actually evaluated.

The S2 validation accuracy of 43.38% in the first epoch is below the majority-class rate of the validation set (584/786 = 74.30%), indicating that the model temporarily assigned most validation images to the Normal class. Together with the selection of epoch 1 as the best S1 checkpoint, this pattern points to training instability rather than gradual convergence. Three implementation factors may have contributed: the brightness augmentation applied after VGG16 preprocessing, which could shift mean-centred inputs outside the range seen by the pretrained filters; the class weight of 1.9433 assigned to the Normal class, which amplifies the loss of minority-class errors while the head is still random; and the full unfreezing of the VGG16 backbone.

In the EfficientNetB3 scenarios, preprocessing leaves pixel values in their original range, so the brightness transformation operated on the intended scale, which is consistent with their smoother validation curves. The relative contribution of these factors cannot be

isolated from a single run and requires the repeated experiment with the corrected augmentation order. The large validation fluctuations of S2 are consistent with fully unfreezing a deep backbone while a randomly initialized head is trained at the same time, without freezing Batch Normalization statistics. Under these conditions, early updates can disturb pretrained representations before the head has stabilized. EfficientNetB3, in which only the last 20 layers were unfrozen, showed smoother validation curves. These observations are descriptive and do not isolate the effect of any single factor, because the scenarios differ simultaneously in the number of trainable layers, learning rate, and epoch budget.

The TL model summaries recorded 14,881,089 total parameters for VGG16, of which 165,377 were trainable, and 11,216,176 parameters for EfficientNetB3, of which 429,569 were trainable. The fit durations of S1, S2, S3, and S4 were 1,399.9; 1,529.1; 1,709.0; and 1,781.6 seconds, respectively. These durations follow different actual numbers of epochs and may have been affected by data access and environment initialization. The timing results are therefore insufficient to determine which architecture is faster under equivalent conditions.

Visual Interpretation and Relation to Previous Studies

All four visualization samples were classified according to their labels by both models. For the two Normal images, the VGG16 Pneumonia scores were 0.13 and 0.03, whereas those of EfficientNetB3 were 0.00 and 0.04 after rounding to two decimals. For the two Pneumonia images, the VGG16 scores were 1.00 and 0.96, whereas those of EfficientNetB3 were 1.00 and 0.76. These values are the scores shown in the visualization panels, not measures of diagnostic certainty. The samples contain only correct predictions and do not represent all error patterns.

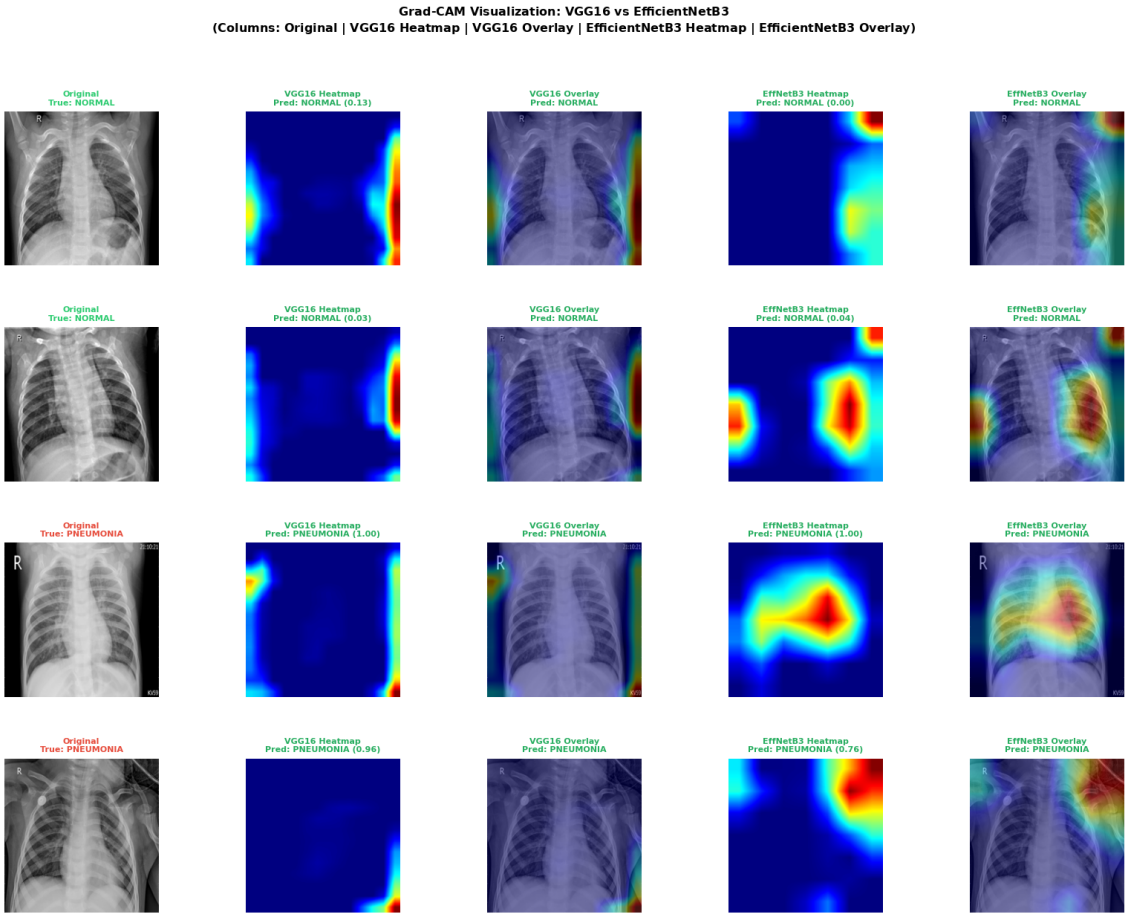


Figure 2. Original notebook visualization for two Normal and two Pneumonia samples. All heatmaps target the Pneumonia score, including those for Normal predictions. Colours are retained for reading the heatmaps

In the VGG16 panels (Figure 2), strong activations appear at the lateral edges and corners of the images, including in Pneumonia samples. EfficientNetB3 showed a more central area in one Pneumonia sample, but also activations at the edges or corners of other samples. Attention to the background is relevant to the experiment of Maguolo and Nanni (2021), which assessed the influence of image background on COVID-19 pneumonia identification. Panwar et al. (2020) developed lesion visualization based on Grad-CAM, whereas Saporta et al. (2022) examined the gap between classification performance and the localization reliability of saliency maps. The higher classification results in this study therefore do not prove more accurate localization.

These border and corner activations are regarded here as a central finding rather than a minor qualification. Comparable patterns have been documented as shortcut cues in CXR models: Zech et al. (2018) found that pneumonia classifiers exploited hospital-specific image features, and DeGrave et al., (2021) showed that radiographic classifiers relied on laterality markers, text annotations, and regions outside the lung fields. In the present experiment, the combination of per-image splitting and border-focused activations means that part of the high internal accuracy may derive from acquisition-related cues rather than parenchymal findings. Because only four correctly classified

images were inspected, this effect cannot be quantified from the current output. A systematic assessment, such as measuring the proportion of Grad-CAM energy inside lung masks across all test images or retraining with lung-masked inputs, is required in the repeated experiment.

Activations at borders and corners may reflect non-anatomical cues such as laterality markers, text overlays, or differences in image framing between classes. Such shortcut cues can inflate internal test performance without corresponding to the radiological signs of pneumonia, which is one reason why background effects have been examined explicitly in previous work (Maguolo & Nanni, 2021). Lung segmentation is a possible way to restrict the input to the anatomical region of interest, but it represents a different pipeline from the one evaluated here: Liu et al. (2022) tested an improved U-Net for CXR lung segmentation, whereas Teixeira et al. (2021) tested the effect of lung segmentation on CXR classification and its explanations. Neither approach was applied in this experiment, and neither is assumed to have improved the reported results.

The comparison with Jain et al. (2020) focuses on VGG16 as a CNN comparator; that with Mahmud et al. (2020) on Grad-CAM analysis; and that with Chowdhury et al. (2021) on EfficientNet backbone comparison and Grad-CAM computation. Avola et al., (2022) also reported comparisons of architectures, execution time, and visualization. Numerical differences across studies must take into account the split, data audit, preprocessing, and evaluation protocol. Hwang et al. (2019) showed the importance of external validation against a reference standard in clinical evaluation; the public-dataset results reported here are not equivalent to such clinical validation.

The results also need to be read against the original benchmark. Kermany et al. (2018) reported an accuracy of 92.8% for pneumonia detection on the patient-separated 624-image test set. The higher accuracy of VGG16 FT in this study (97.84%) was obtained on a per-image internal split and is therefore not directly comparable; the gap is consistent with the optimistic bias expected when images from the same patient can appear in both training and test partitions (Zech et al., 2018). Likewise, studies on this dataset that report accuracies close to 100% should be interpreted in light of their splitting protocol and of the shortcut cues described by (DeGrave et al., 2021). Evaluation on the original test folder is therefore necessary before the ranking of the two architectures can be regarded as a generalizable finding.

Limitations

This study used a single paediatric data source and a single training seed. Splitting was performed per image without verification of patient separation, so the presence of duplicates or the same patient across partitions cannot be ruled out. The discrepancy between the number of entries and the source description has not been resolved. The results are therefore not yet a patient-based estimate of generalization or an external validation, and they cannot be generalized to adult patients or other facilities.

The preliminary implementation also has limitations: augmentation was applied after model preprocessing; a rotation factor of 0.08 corresponds to approximately ±28.8°; FT models were built separately rather than continued from TL checkpoints; and the rule of

unfreezing the last 20 layers has a different scope in each architecture and does not specifically lock backbone Batch Normalization. In FT, the head was trained from scratch at the same time as the unfrozen backbone layers. Differences in results may have been influenced by a combination of these factors, the learning rate, and the epoch budget. Grad-CAM evaluated only four correct predictions with the Pneumonia score as the target, without lesion annotations or expert assessment.

CONCLUSION

This study compared VGG16 and EfficientNetB3 under frozen and unfrozen backbone configurations for paediatric CXR pneumonia classification. On the internal test set, VGG16 FT performed best, with 97.84% accuracy, 97.26% recall, 98.53% F1-score, and 0.9977 ROC-AUC, and its accuracy advantage over EfficientNetB3 FT remained significant under all compatible McNemar bounds. EfficientNetB3 FT produced no false positives but 37 false negatives. Fine-tuned configurations outperformed frozen-backbone configurations for both architectures, although each model was trained separately rather than through staged fine-tuning. Grad-CAM showed that correct predictions were not always supported by lung-centred activations, and border activations suggest possible shortcut learning. The findings therefore constitute preliminary evidence, not clinical readiness. Future work should apply patient-grouped splitting, evaluate the original 624-image test set, correct the augmentation order, and repeat training with several random seeds.

FUNDING

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors. The computational experiments were conducted on the freely available Google Colab platform using a publicly accessible dataset, and no external financial support was received for the implementation of the study or the publication of this article.

CONFLICT OF INTEREST

The author declares that this research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest. The author has no affiliation with the developers of the deep learning frameworks, pretrained models, computing platform, or dataset used in this study.

AUTHOR CONTRIBUTION

Satrya Permana: Conceptualization, Methodology, Software, Data curation, Formal analysis, Investigation, Validation, Visualization, Writing – original draft, and Writing – review & editing. The author has read and approved the final version of the manuscript and takes full responsibility for the integrity of the data and the accuracy of the analysis.

GENERATIVE ARTIFICIAL INTELLIGENCE USAGE STATEMENT

During the preparation of this manuscript, the author used generative artificial intelligence (AI) tools to improve language clarity and grammatical structure and to assist with reference management, including formatting citations and identifying candidate peer-reviewed sources. All AI-assisted outputs were critically reviewed and edited by the author, and every cited source was checked against its original publication. Generative AI was not used to generate, analyse, or interpret the experimental data, to produce the results, or to create the figures. This use complies with applicable ethical guidelines on the use and disclosure of generative AI in scientific research. The author takes full responsibility for the content and integrity of the published article.

REFERENCES

Avola, D., Bacciu, A., Cinque, L., Fagioli, A., Marini, M. R., & Taiello, R. (2022). Study on transfer learning capabilities for pneumonia classification in chest-x-rays images. *Computer Methods and Programs in Biomedicine*, *221*, 106833. https://doi.org/10.1016/j.cmpb.2022.106833

Cherian, T., Mulholland, E. K., Carlin, J. B., Ostensen, H., Amin, R., de Campo, M., Greenberg, D., Lagos, R., Lucero, M., Madhi, S. A., O'Brien, K. L., Obaro, S., & Steinhoff, M. C. (2005). Standardized interpretation of paediatric chest radiographs for the diagnosis of pneumonia in epidemiological studies. *Bulletin of the World Health Organization*, *83*(5), 353–359. https://doi.org/10.1590/S0042-96862005000500011

Chowdhury, N. K., Kabir, M. A., Rahman, M. M., & Rezoana, N. (2021). ECOVNet: A highly effective ensemble based deep learning model for detecting COVID-19. *PeerJ Computer Science*, *7*, e551. https://doi.org/10.7717/peerj-cs.551

DeGrave, A. J., Janizek, J. D., & Lee, S.-I. (2021). AI for radiographic COVID-19 detection selects shortcuts over signal. *Nature Machine Intelligence*, *3*(7), 610–619. https://doi.org/10.1038/s42256-021-00338-7

Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., & Wichmann, F. A. (2020). Shortcut learning in deep neural networks. *Nature Machine Intelligence*, *2*(11), 665–673. https://doi.org/10.1038/s42256-020-00257-z

Hwang, E. J., Park, S., Jin, K.-N., Kim, J. I., Choi, S. Y., Lee, J. H., Goo, J. M., Aum, J., Yim, J.-J., Cohen, J. G., Ferretti, G. R., & Park, C. M. (2019). Development and validation of a deep learning–based automated detection algorithm for major thoracic diseases on chest radiographs. *JAMA Network Open*, *2*(3), e191095. https://doi.org/10.1001/jamanetworkopen.2019.1095

Jain, R., Nagrath, P., Kataria, G., Kaushik, V. S., & Hemanth, D. J. (2020). Pneumonia detection in chest X-ray images using convolutional neural networks and transfer learning. *Measurement*, *165*, 108046. https://doi.org/10.1016/j.measurement.2020.108046

Kermany, D. S., Goldbaum, M., Cai, W., Valentim, C. C. S., Liang, H., Baxter, S. L., McKeown, A., Yang, G., Wu, X., Yan, F., Dong, J., Prasadha, M. K., Pei, J., Ting,

M. Y. L., Zhu, J., Li, C., Hewett, S., Dong, J., Ziyar, I., … Zhang, K. (2018). Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning. *Cell*, *172*(5), 1122-1131.e9. https://doi.org/10.1016/j.cell.2018.02.010

Liu, W., Luo, J., Yang, Y., Wang, W., Deng, J., & Yu, L. (2022). Automatic lung segmentation in chest X-ray images using improved U-Net. *Scientific Reports*, *12*, 8649. https://doi.org/10.1038/s41598-022-12743-y

Maguolo, G., & Nanni, L. (2021). A critic evaluation of methods for COVID-19 automatic detection from X-ray images. *Information Fusion*, *76*, 1–7. https://doi.org/10.1016/j.inffus.2021.04.008

Mahmud, T., Rahman, M. A., & Fattah, S. A. (2020). CovXNet: A multi-dilation convolutional neural network for automatic COVID-19 and other pneumonia detection from chest X-ray images with transferable multi-receptive feature optimization. *Computers in Biology and Medicine*, *122*, 103869. https://doi.org/10.1016/j.compbiomed.2020.103869

McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, *12*(2), 153–157. https://doi.org/10.1007/BF02295996

Okolo, G. I., Katsigiannis, S., & Ramzan, N. (2022). IEViT: An enhanced vision transformer architecture for chest X-ray image classification. *Computer Methods and Programs in Biomedicine*, *226*, 107141. https://doi.org/10.1016/j.cmpb.2022.107141

Panwar, H., Gupta, P. K., Siddiqui, M. K., Morales-Menendez, R., Bhardwaj, P., & Singh, V. (2020). A deep learning and grad-CAM based color visualization approach for fast detection of COVID-19 cases using chest X-ray and CT-Scan images. *Chaos, Solitons & Fractals*, *140*, 110190. https://doi.org/10.1016/j.chaos.2020.110190

Rahman, T., Chowdhury, M. E. H., Khandakar, A., Islam, K. R., Islam, K. F., Mahbub, Z. B., Kadir, M. A., & Kashem, S. (2020). Transfer learning with deep convolutional neural network (CNN) for pneumonia detection using chest X-ray. *Applied Sciences*, *10*(9), 3233. https://doi.org/10.3390/app10093233

Saboo, Y. S., Kapse, S., & Prasanna, P. (2023). Convolutional Neural Networks (CNNs) for Pneumonia Classification on Pediatric Chest Radiographs. *Cureus*. https://doi.org/10.7759/cureus.44130

Saporta, A., Gui, X., Agrawal, A., Pareek, A., Truong, S. Q. H., Nguyen, C. D. T., Ngo, V.-D., Seekins, J., Blankenberg, F. G., Ng, A. Y., Lungren, M. P., & Rajpurkar, P. (2022). Benchmarking saliency methods for chest X-ray interpretation. *Nature Machine Intelligence*, *4*(10), 867–878. https://doi.org/10.1038/s42256-022-00536-x

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *2017 IEEE International Conference on Computer Vision (ICCV)*, 618–626. https://doi.org/10.1109/ICCV.2017.74

Shaikh, M., Siddiqui, I. F., Arain, Q., Koo, J., Unar, M. A., & Qureshi, N. M. F. (2023). MDEV model: A novel ensemble-based transfer learning approach for pneumonia

classification using CXR images. *Computer Systems Science and Engineering*, *46*(1), 287–302. https://doi.org/10.32604/csse.2023.035311

Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. *Proceedings of the 3rd International Conference on Learning Representations (ICLR 2015)*. https://doi.org/10.48550/arXiv.1409.1556

Tan, M., & Le, Q. V. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. *Proceedings of the 36th International Conference on Machine Learning*, *97*, 6105–6114.

Teixeira, L. O., Pereira, R. M., Bertolini, D., Oliveira, L. S., Nanni, L., Cavalcanti, G. D. C., & Costa, Y. M. G. (2021). Impact of lung segmentation on the diagnosis and explanation of COVID-19 in chest X-ray images. *Sensors*, *21*(21), 7116. https://doi.org/10.3390/s21217116

Ukwuoma, C. C., Qin, Z., Heyat, M. B. Bin, Akhtar, F., Bamisile, O., Muaad, A. Y., Addo, D., & Al-antari, M. A. (2023). A hybrid explainable ensemble transformer encoder for pneumonia identification from chest X-ray images. *Journal of Advanced Research*, *48*, 191–211. https://doi.org/10.1016/j.jare.2022.08.021

UNICEF. (2019, November). *One child dies of pneumonia every 39 seconds, agencies warn* [Press release].

Wilson, E. B. (1927). Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, *22*(158), 209–212. https://doi.org/10.1080/01621459.1927.10502953

World Health Organization. (2022, November). *Pneumonia in children*.

Yunus, R. E. (2024). Radiology loading and coverage hours in Indonesia. *Korean Journal of Radiology*, *25*(7), 597–599. https://doi.org/10.3348/kjr.2024.0267

Zech, J. R., Badgeley, M. A., Liu, M., Costa, A. B., Titano, J. J., & Oermann, E. K. (2018). Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. *PLOS Medicine*, *15*(11), e1002683. https://doi.org/10.1371/journal.pmed.1002683