

## Preprocessing-Induced Ranking Reversal is Architecture-Dependent in Cross-Domain Pneumonia Detection

Ucta Pradema Sanjaya<sup>1</sup>, Sri Mujiyono<sup>2</sup>, Ade Pratama<sup>3</sup>, Iwan Setiawan Wibisono<sup>4</sup>

<sup>1234</sup>Informatics Department, Faculty of Computer Science and Education Postgraduate Program, Ngudi Waluyo University, Ungaran, Indonesia

### Received:

May 1, 2026

### Revised:

September 17, 2026

### Accepted:

September 29, 2026

### Published:

October 24, 2026

Corresponding Author:

### Author Name\*:

Ucta Pradema Sanjaya

### Email\*:

uctapradema@unw.ac.id

DOI:

10.63158/journalisi.v8i5.1944

© 2026 Journal of Information Systems and Informatics. This open access article is distributed under a (CC-BY License)



**Abstract.** This study investigates how preprocessing selection influences the cross-domain generalization of convolutional neural networks (CNNs) for pneumonia detection and whether it changes performance rankings among architectures. Five preprocessing conditions—Baseline, Standard Deviation Normalization, Contrast Limited Adaptive Histogram Equalization (CLAHE), Histogram Equalization, and StdDev+CLAHE—were combined with three architectures: a simple CNN, ResNet-50, and EfficientNet-B0. The resulting 15 models were trained on the public Kaggle dataset containing 5,863 chest X-ray images and externally evaluated on 1,800 clinical chest X-rays collected in Central Java, Indonesia. Results showed that preprocessing rankings were architecture-dependent. The simple CNN achieved its highest external Recall with Histogram Equalization (0.9856), although Baseline ranked first internally. ResNet-50 was most stable with Standard Deviation Normalization, which ranked first internally (F1 = 0.9821) and externally (Recall = 0.9944). EfficientNet-B0 achieved the highest external Recall with CLAHE (0.9994), but its lower internal precision (0.77–0.87) indicated bias toward the pneumonia class. These findings demonstrate that high benchmark performance may not transfer directly to regional clinical settings because preprocessing interacts with scanner characteristics, data distribution, and network architecture. Therefore, each preprocessing–architecture pair should be validated on locally acquired images before clinical deployment.

**Keywords:** pneumonia detection; convolutional neural network; image preprocessing; cross-domain generalization; chest X-ray

## 1. INTRODUCTION

Pneumonia remains the leading cause of death among children under five worldwide. The World Health Organization recorded about 740,180 such deaths in 2019 [1]. In Indonesia, the 2023 Survei Kesehatan Indonesia (SKI) reported under-five pneumonia prevalence at 15.0% [2], while radiologists remain scarce in primary-care facilities [3]. CNNs with transfer learning can read chest X-rays automatically, and architectures such as ResNet and EfficientNet are now widely used [4], [5]. These models reach high accuracy on public datasets. Most evaluations, however, test only on the same data distribution.

Image quality shapes what a CNN learns. Public benchmark datasets are usually collected at one center under a single acquisition protocol, whereas hospital imaging systems differ in tube voltage, exposure settings, detector technology, and vendor post-processing [6]. Preprocessing is routinely applied to reduce this variation, with CLAHE, Histogram Equalization, and intensity normalization among the common choices [7], [8], [9]. Yet the technique is often picked by rule of thumb and tuned only on the benchmark. Such a choice can add domain shift instead of removing it. An operator tuned to the intensity statistics of the source data maps images from another scanner onto a different part of the intensity range and stretches their own noise and artifacts, which can move them further from the distribution the model was trained on. Applying a transformation without local validation is therefore a safety risk in computer-aided diagnosis: the features a model relies on may be suppressed once the scanner changes, and the benchmark score gives no warning.

Several previous studies have investigated preprocessing optimization and CNN-based pneumonia detection [10], [11], [12], and others have explored transfer learning for the same task [13], [14]. The reported results are generally promising, yet almost all of them test the model on the same distribution it was trained on. However, limited studies have examined whether preprocessing effectiveness remains consistent across different architectures under cross-domain clinical conditions.

Three gaps remain. First, the interaction between preprocessing and CNN architecture under cross-domain generalization has not been studied systematically. Second, validation on Indonesian clinical data is almost absent [15], [16], [17]; the acquisition gap

between a local hospital and public datasets is not represented. Third, it is unclear whether a finding on one architecture transfers to others, or is merely an artifact of that architecture.

Therefore, this research intends to investigate architecture-dependent preprocessing behavior in cross-domain pneumonia detection. Five preprocessing techniques are evaluated on three CNN architectures: a simple CNN, ResNet-50, and EfficientNet-B0. Evaluation runs in two stages. The first uses a public dataset; the second uses 1,800 clinical images from Central Java as an external test. The objectives of this research are: (1) to identify the best preprocessing technique on the public dataset; (2) to test whether that ranking holds under clinical cross-domain evaluation; and (3) to analyze whether the pattern is consistent across architectures. This design separates a general pattern from an architecture-specific artifact. It also yields an evidence-based configuration recommendation for Indonesian health facilities. We pose three research questions. RQ1: which preprocessing technique performs best on the public dataset? RQ2: does that ranking hold under clinical cross-domain testing? RQ3: is the emerging pattern consistent across architectures?

The novelty of this study does not lie in the preprocessing operators themselves, because CLAHE, histogram equalization, and intensity normalization are established techniques. It lies in analyzing how these operators interact with network architecture under domain shift. The study makes three contributions. First, it provides a systematic comparison of five preprocessing techniques across three CNN architectures, giving 15 controlled combinations. Second, it validates every combination externally on Indonesian clinical chest X-rays, a population rarely represented in public benchmarks. Third, it identifies an architecture-dependent preprocessing ranking reversal: the technique that ranks first internally is not necessarily first cross-domain, and whether this happens depends on the architecture.

## **2. METHODS**

### **2.1. Research Pipeline**

Figure 1 shows the research pipeline. Every image is converted to grayscale, resized to 224×224, and passed through one of five preprocessing conditions before being fed to

three CNN architectures, which gives a  $5 \times 3$  matrix of 15 combinations. All combinations are trained on Dataset A and then evaluated in sequence: first internally on the Dataset A test split, then externally on the clinical Dataset B. Finally, the internal and cross-domain rankings of the preprocessing conditions are compared for each architecture. Every architecture receives identical inputs across the five conditions.

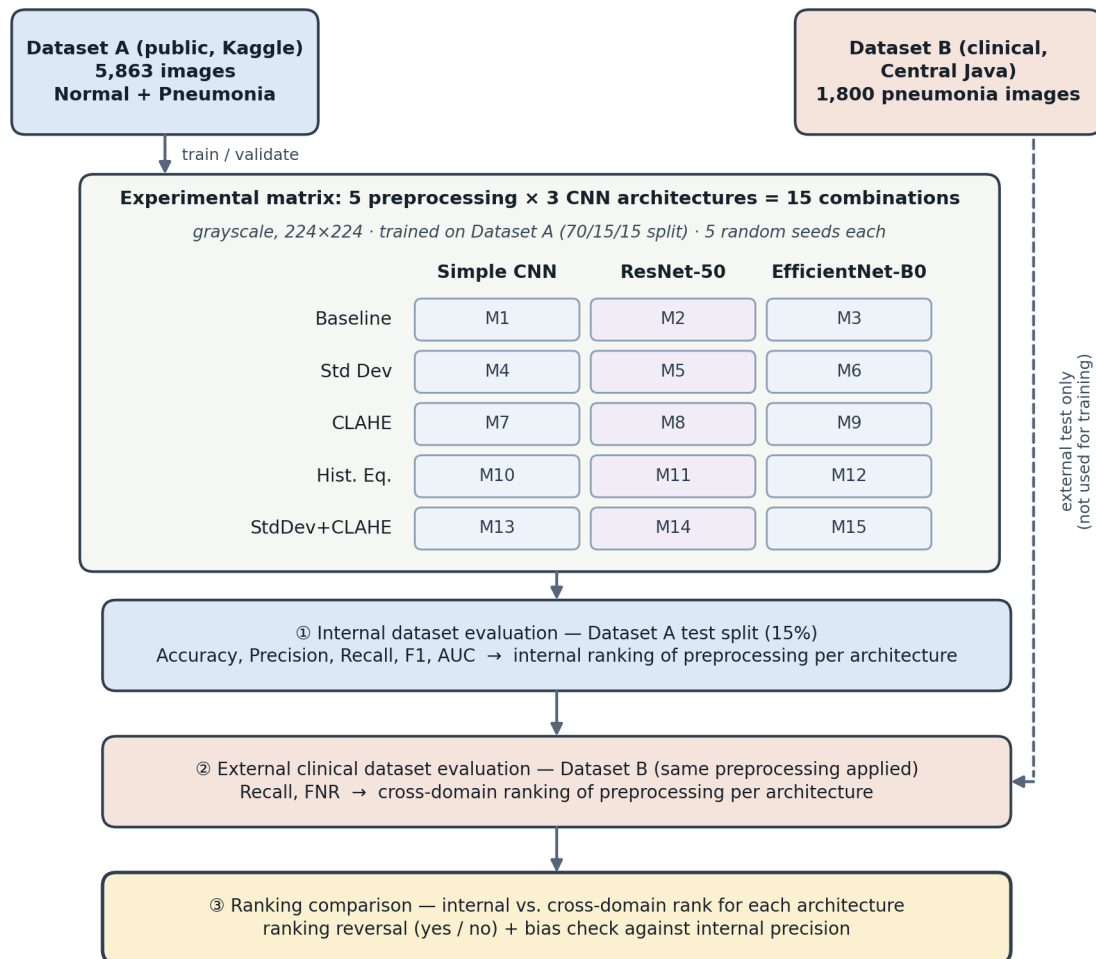


**Figure 1** Research pipeline. The 5 preprocessing × 3 CNN architecture matrix (15 combinations) is trained on Dataset A, evaluated internally on the Dataset A test split, then evaluated externally on the Indonesian clinical Dataset B; the internal and cross-domain rankings are finally compared for each architecture.

## 2.2. Datasets

Dataset A is the Kaggle Chest X-Ray collection. It contains 5,863 chest images — 1,583 normal and 4,265 pneumonia. We split it into 70% training, 15% validation, and 15% testing

with a stratified split. We re-split the data because the original validation set held only 16 images, too few for stable validation. No patient overlaps between subsets, and a leakage check returned zero overlap.

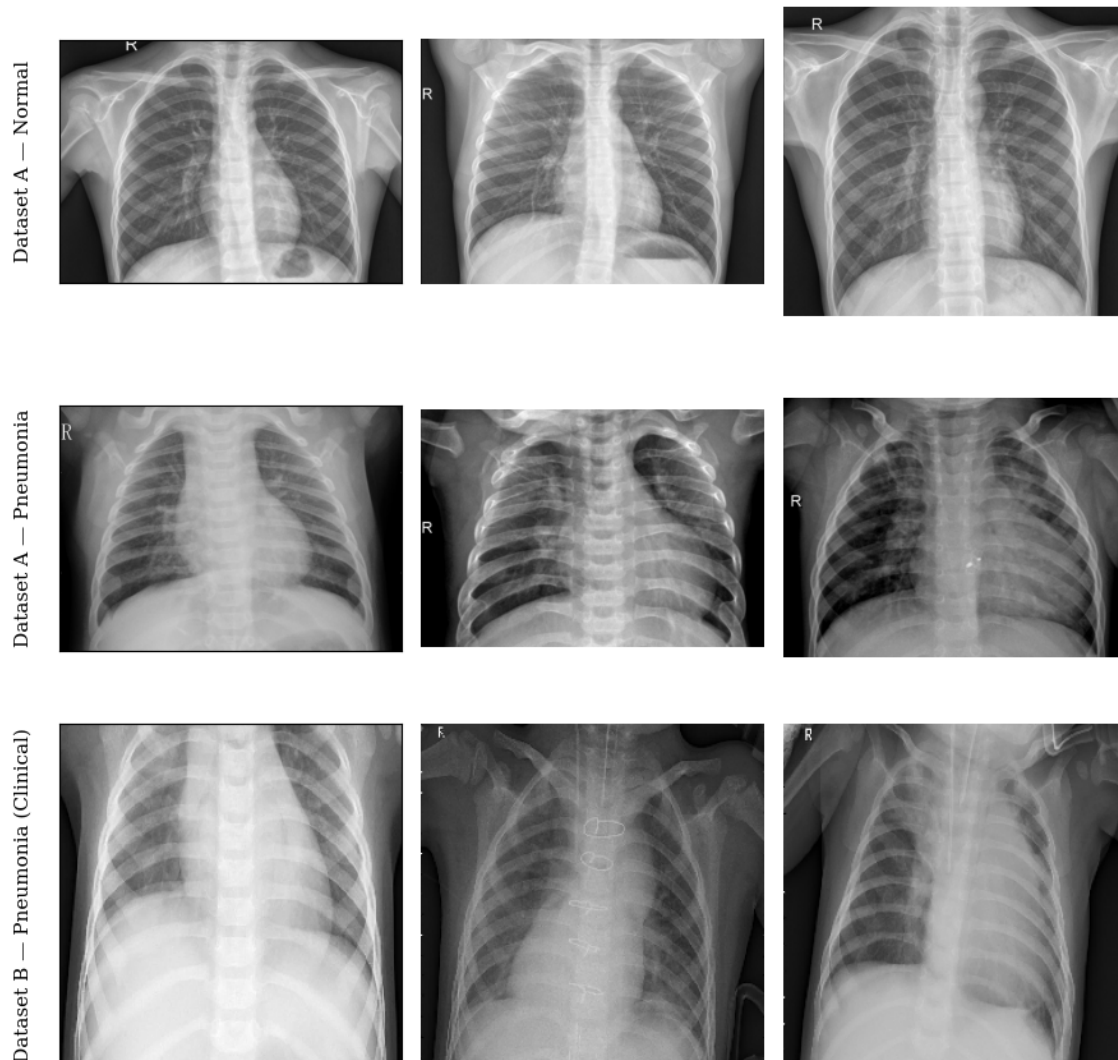


**Figure 2** Sample chest X-ray images from both datasets. Top row: Normal cases from Dataset A (Kaggle). Middle row: Pneumonia cases from Dataset A. Bottom row: Pneumonia cases from Dataset B (Indonesian clinical). Visual differences in contrast and acquisition characteristics between the public and clinical sources are apparent.

Dataset B contains 1,800 pneumonia images acquired at a hospital in Central Java, and it serves as the external cross-domain test. Image acquisition at the hospital was carried out by three radiographers. The pneumonia label for each image was the clinical diagnosis recorded in the hospital's routine radiology reports, each issued by a single board-certified radiologist (Sp.Rad) holding an active practice license as part of standard clinical

care. Because the labels were taken from routine single-reader clinical reports, no independent second reading or multi-reader consensus procedure was applied; the implication of this single-reader ground truth is discussed in Section 3.6.3. Ethics approval for the study was granted by the Health Research Ethics Committee of Universitas Ngudi Waluyo, and the clinical data were used with the authorization of the participating hospital in Central Java. The single class in Dataset B is a design choice, not a weakness: it lets us measure Recall and false negatives directly. Image format and resolution vary with the original acquisition conditions. Table 1 summarizes both datasets.

**Table 1** Characteristics of Dataset A and Dataset B

| Characteristic | Dataset A (Public)             | Dataset B (Clinical)              |
|----------------|--------------------------------|-----------------------------------|
| Source         | Kaggle Chest X-Ray             | Regional hospital, Central Java   |
| Total images   | 5,863                          | 1,800                             |
| Normal         | 1,583                          | 0                                 |
| Pneumonia      | 4,265                          | 1,800                             |
| Classes        | 2 (Normal, Pneumonia)          | 1 (Pneumonia only)                |
| Split          | 70% train / 15% val / 15% test | External test set only            |
| Role           | Training + internal evaluation | Cross-domain evaluation           |
| Population     | Pediatric (Guangzhou)          | Mixed adult/pediatric (Indonesia) |

### 2.3. Preprocessing Techniques

We used five preprocessing conditions: Baseline, Std Dev, CLAHE, Histogram Equalization, and StdDev+CLAHE. Every image was first converted to grayscale and resized to 224×224 pixels. The key parameters were a CLAHE clip limit of 2.0 with an 8×8 tile grid, and an epsilon of 1e-7 [5], [18], [19]. Standard-deviation normalization and the CLAHE clip limit are given in Eq. (1) and Eq. (2), where  $x$  is the pixel intensity,  $\mu$  and  $\sigma$  are its per-image mean and standard deviation,  $\epsilon$  is a small constant,  $P$  is the number of pixels per tile, and  $L$  is the number of gray levels. We keep this part brief because the study centers on the interaction between architecture and preprocessing.

$$x' = \frac{x - \mu}{\sigma + \epsilon} \quad (1)$$

$$C = \beta \cdot \frac{P}{L}, \quad \beta = 2.0 \quad (2)$$

## 2.4. CNN Architectures

The simple CNN was built from scratch. It stacks three convolutional blocks with 32, 64, and 128 filters, each block using batch normalization and max pooling. A Dense-256 layer with 0.5 dropout precedes the sigmoid output. The model has about 25.78 million parameters and takes single-channel grayscale input. It serves as a lightweight baseline. ResNet-50 and EfficientNet-B0 use ImageNet pretrained weights and three-channel RGB input. ResNet-50 is fine-tuned on the last 20% of its layers, from layer 143 onward; EfficientNet-B0 on the last 25%, from layer 200. Both share an identical classifier head: global average pooling, Dense-256, dropout, and a sigmoid. Each model's native input preprocessing is applied through a Lambda layer.

An initial run with all layers frozen failed. The frozen EfficientNet predicted a single class and reached only about 72% validation accuracy. We switched to fine-tuning. Grayscale images were stacked into three channels to match the pretrained weights, and a small learning rate of 0.0001 was used so the pretrained features would not be destroyed. Table 2 summarizes the architectures and training strategy.

**Table 2** CNN Architectures and Training Strategy

| Architecture    | Base         | Input     | Trainable Strategy  | Total Params |
|-----------------|--------------|-----------|---------------------|--------------|
| CNN Simple      | From scratch | 224×224×1 | All trainable       | ~25,78M      |
| ResNet-50       | ImageNet     | 224×224×3 | Fine-tune last ~20% | ~25,6M       |
| EfficientNet-B0 | ImageNet     | 224×224×3 | Fine-tune last ~25% | ~5,3M        |

## 2.5. Implementation Details and Reproducibility

All models were implemented in Python 3.12.13 with TensorFlow/Keras 2.20.0 and trained on Google Colab with a single NVIDIA Tesla T4 GPU (16 GB). Table 3 consolidates the full configuration. Every model was optimized with Adam using binary cross-entropy loss, with class weights to offset the class imbalance of Dataset A. The learning rate was 0.001 for the simple CNN and 0.0001 for the two fine-tuned pretrained models. Training used a batch size of 32 for at most 30 epochs with early stopping (patience = 5). Augmentation was limited to random horizontal flips and rotations within  $\pm 10^\circ$ ; vertical flips were excluded because chest anatomy must keep its orientation. One training run took roughly 12 to 44 minutes, depending on the architecture and the number of epochs reached (Table 6), apart from one run slowed by a GPU fluctuation.

**Table 3** Training Configuration and Hyperparameters

| Parameter             | Value                                                                              |
|-----------------------|------------------------------------------------------------------------------------|
| Optimizer             | Adam                                                                               |
| Learning rate         | 0.001 (Simple CNN); 0.0001 (ResNet-50, EfficientNet-B0)                            |
| Batch size            | 32                                                                                 |
| Loss function         | Binary cross-entropy (class weights applied)                                       |
| Epochs                | Up to 30, early stopping (patience = 5)                                            |
| Input size            | 224 × 224 (grayscale; stacked to 3 channels for pretrained models)                 |
| Data augmentation     | Horizontal flip; rotation ±10° (no vertical flip)                                  |
| Fine-tuning depth     | ResNet-50: last ~20% (from layer 143); EfficientNet-B0: last ~25% (from layer 200) |
| Random seeds          | 42, 123, 456, 789, 1010 (multi-seed; mean ± SD reported)                           |
| Software              | Python 3.12.13; TensorFlow 2.20.0; NumPy 2.0.2                                     |
| Hardware              | Google Colab (free/low-cost tier); NVIDIA Tesla T4 GPU                             |
| Training time / epoch | ≈ 66–83 s (Simple CNN and EfficientNet-B0); see Table 6                            |

The main results in Tables 4–7 come from runs with seed 42. To gauge how stable these results are against training randomness, we retrained every architecture × preprocessing pair on five random seeds (42, 123, 456, 789, and 1010). The seed controls weight initialization, data shuffling, and augmentation. Table 8 reports the mean ± standard deviation across the five seeds. A small standard deviation indicates that the observed preprocessing ranking is not an artifact of a single run.

## 2.6. Evaluation Metrics

Internal evaluation used accuracy, precision, recall, F1, and AUC. Cross-domain evaluation focused on Recall and the false negative rate, which fits because Dataset B contains only the pneumonia class. Precision, Recall, F1, the false-negative rate (FNR), and macro-AUC are defined in Eq. (3)–(5).

$$\text{Precision} = \frac{TP}{TP+FP}, \quad \text{Recall} = \frac{TP}{TP+FN} \quad (3)$$

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}, \quad \text{FNR} = \frac{FN}{TP+FN} \quad (4)$$

$$\text{macro-AUC} = \frac{1}{K} \sum_{k=1}^K \text{AUC}_k \quad (5)$$

### 3. RESULTS AND DISCUSSION

#### 3.1. Internal Evaluation (Dataset A)

On the simple CNN, Baseline gave the best F1 at 0.9735 (Table 4). ResNet-50 peaked instead on Std Dev at 0.9821, just above Baseline (0.9813). EfficientNet-B0 reached its highest F1 on Histogram Equalization at 0.9209, but its internal performance was markedly lower and less stable than the other two, with accuracy between 0.79 and 0.88 and precision as low as 0.77. The three architectures are therefore not equivalent internally: the CNN and ResNet-50 are both strong (F1 above 0.96), whereas EfficientNet-B0 lags and is more sensitive to the preprocessing condition. These internal figures sit within the range of recent multi-model chest X-ray benchmarks [20].

**Table 4** Internal Evaluation Results (Dataset A)

| Architecture    | Preprocessing | Accuracy | Precision | Recall | F1-Score      | AUC    |
|-----------------|---------------|----------|-----------|--------|---------------|--------|
| CNN Simple      | Baseline      | 0.9614   | 0.9750    | 0.9720 | <b>0.9735</b> | 0.9865 |
| CNN Simple      | Std Dev       | 0.9477   | 0.9671    | 0.9611 | 0.9641        | 0.9856 |
| CNN Simple      | CLAHE         | 0.9545   | 0.9617    | 0.9766 | 0.9691        | 0.9886 |
| CNN Simple      | Hist. Eq.     | 0.9477   | 0.9421    | 0.9891 | 0.9650        | 0.9871 |
| CNN Simple      | StdDev+CLAHE  | 0.9443   | 0.9611    | 0.9626 | 0.9619        | 0.9887 |
| ResNet-50       | Baseline      | 0.9727   | 0.9798    | 0.9829 | <b>0.9813</b> | 0.9969 |
| ResNet-50       | Std Dev       | 0.9739   | 0.9828    | 0.9813 | 0.9821        | 0.9971 |
| ResNet-50       | CLAHE         | 0.9591   | 0.9619    | 0.9829 | 0.9723        | 0.9930 |
| ResNet-50       | Hist. Eq.     | 0.9648   | 0.9707    | 0.9813 | 0.9760        | 0.9950 |
| ResNet-50       | StdDev+CLAHE  | 0.9591   | 0.9647    | 0.9798 | 0.9722        | 0.9934 |
| EfficientNet-B0 | Baseline      | 0.8443   | 0.8250    | 0.9984 | 0.9035        | 0.9793 |
| EfficientNet-B0 | Std Dev       | 0.8739   | 0.8535    | 0.9984 | 0.9203        | 0.9808 |
| EfficientNet-B0 | CLAHE         | 0.7875   | 0.7744    | 1.0000 | <b>0.8729</b> | 0.9541 |
| EfficientNet-B0 | Hist. Eq.     | 0.8773   | 0.8688    | 0.9798 | 0.9209        | 0.9644 |
| EfficientNet-B0 | StdDev+CLAHE  | 0.8295   | 0.8114    | 0.9984 | <b>0.8953</b> | 0.9639 |

An internal bias check (Section 3.4) reveals a different pattern across architectures. The simple CNN and ResNet-50 classify Normal and Pneumonia in balance, misclassifying few Normal cases. EfficientNet-B0 is far more biased: depending on the preprocessing condition, it mislabels 40 to 79 percent of Normal cases as Pneumonia. This difference in internal bias matters, and we contrast it with the cross-domain evaluation in Section 3.2.

### 3.2. Cross-Domain Evaluation (Dataset B)

The best cross-domain Recall differs by architecture (Table 5). On the simple CNN, Histogram Equalization (Hist. Eq.) is best, with a Recall of 0.9856 (26 false negatives, FN, out of 1,800). On ResNet-50, Std Dev leads with a Recall of 0.9944 (10 FN); CLAHE and Hist. Eq. on ResNet-50 are identical at 0.9894 (19 FN). On EfficientNet-B0, CLAHE is best at 0.9994 — just 1 FN out of 1,800, the highest single figure in the whole experiment.

**Table 5** Cross-Domain Evaluation Results (Dataset B — Indonesian Clinical)

| Architecture    | Preprocessing | Recall | FNR    | TP   | FN |
|-----------------|---------------|--------|--------|------|----|
| CNN Simple      | Baseline      | 0.9533 | 0.0467 | 1716 | 84 |
| CNN Simple      | Std Dev       | 0.9556 | 0.0444 | 1720 | 80 |
| CNN Simple      | CLAHE         | 0.9689 | 0.0311 | 1744 | 56 |
| CNN Simple      | Hist. Eq.     | 0.9856 | 0.0144 | 1774 | 26 |
| CNN Simple      | StdDev+CLAHE  | 0.9572 | 0.0428 | 1723 | 77 |
| ResNet-50       | Baseline      | 0.9883 | 0.0117 | 1779 | 21 |
| ResNet-50       | Std Dev       | 0.9944 | 0.0056 | 1790 | 10 |
| ResNet-50       | CLAHE         | 0.9894 | 0.0106 | 1781 | 19 |
| ResNet-50       | Hist. Eq.     | 0.9894 | 0.0106 | 1781 | 19 |
| ResNet-50       | StdDev+CLAHE  | 0.9839 | 0.0161 | 1771 | 29 |
| EfficientNet-B0 | Baseline      | 0.9972 | 0.0028 | 1795 | 5  |
| EfficientNet-B0 | Std Dev       | 0.9956 | 0.0044 | 1792 | 8  |
| EfficientNet-B0 | CLAHE         | 0.9994 | 0.0006 | 1799 | 1  |
| EfficientNet-B0 | Hist. Eq.     | 0.9867 | 0.0133 | 1776 | 24 |
| EfficientNet-B0 | StdDev+CLAHE  | 0.9967 | 0.0033 | 1794 | 6  |

No ResNet-50 condition collapsed cross-domain. CLAHE and Hist. Eq. on ResNet-50 performed identically (Recall 0.9894; 19 FN), and its internal CLAHE score stayed reasonable at F1 0.9723. Std Dev remained the strongest ResNet-50 condition both

internally (F1 0.9821) and cross-domain (Recall 0.9944), so ResNet-50 shows no ranking reversal at all. Cross-domain drops of this kind underline why external clinical validation matters [21], [22].

Tables 4 and 5 together show that internal performance is a poor guide to clinical robustness. The internal ranking of preprocessing conditions was preserved cross-domain only for ResNet-50. For the simple CNN and EfficientNet-B0, the condition ranked first on Dataset A was not the one ranked first on Dataset B, which is the ranking reversal examined below. A high internal score can be earned by exploiting characteristics specific to the source dataset, such as its acquisition protocol, intensity profile, or class balance. Those cues do not travel to a new hospital, so a model can rank well internally for reasons that have little to do with the radiological signs of pneumonia.

### 3.3. Computational Efficiency

The simple CNN is the fastest and most consistent model, at 66 to 73 seconds per epoch (Table 6). Transfer learning is slower — a trade-off between accuracy and compute. We saw one clear anomaly: ResNet + Hist. Eq. once logged 453 seconds per epoch on a single run. That figure is not a property of the preprocessing but a GPU fluctuation on Google Colab. The CNN fits resource-limited health facilities well.

**Table 6** Computational Efficiency

| Architecture    | Preprocessing | Epochs | Total Time<br>(s) | Time/Epoch<br>(s) |
|-----------------|---------------|--------|-------------------|-------------------|
| CNN Simple      | Baseline      | 12     | 832.3             | 69.4              |
| CNN Simple      | Std Dev       | 14     | 1003.2            | 71.7              |
| CNN Simple      | CLAHE         | 14     | 981.8             | 70.1              |
| CNN Simple      | Hist. Eq.     | 18     | 1194.5            | 66.4              |
| CNN Simple      | StdDev+CLAHE  | 11     | 804.4             | 73.1              |
| ResNet-50       | Baseline      | 14     | 2660.8            | 190.1             |
| ResNet-50       | Std Dev       | 17     | 1316.8            | 77.5              |
| ResNet-50       | CLAHE         | 9      | 723.8             | 80.4              |
| ResNet-50       | Hist. Eq.     | 9      | 4083.3            | 453.7 *           |
| ResNet-50       | StdDev+CLAHE  | 9      | 2003.0            | 222.6             |
| EfficientNet-B0 | Baseline      | 25     | 1861.3            | 74.5              |

| Architecture    | Preprocessing | Epochs | Total Time<br>(s) | Time/Epoch<br>(s) |
|-----------------|---------------|--------|-------------------|-------------------|
| EfficientNet-B0 | Std Dev       | 19     | 1414.7            | 74.5              |
| EfficientNet-B0 | CLAHE         | 11     | 853.8             | 77.6              |
| EfficientNet-B0 | Hist. Eq.     | 14     | 1054.2            | 75.3              |
| EfficientNet-B0 | StdDev+CLAHE  | 11     | 912.7             | 83.0              |

### 3.4. Bias Check: Balanced (CNN, ResNet-50) versus Biased (EfficientNet-B0) Classification

The internal confusion matrix (Table 7) separates the balanced models from the biased one. The simple CNN and ResNet-50 are balanced: each misclassifies only a small share of Normal cases, so their cross-domain Recall reflects genuine generalization. EfficientNet-B0 does not. On Baseline it mislabels 57.1 percent of Normal cases (136 of 238), and on CLAHE 78.6 percent (187 of 238), as Pneumonia. A model this biased will automatically score a high Recall on the all-pneumonia Dataset B, regardless of its true generalization. Similar risks from imbalanced datasets and database structure have been widely reported in chest imaging [23], [24], [25].

**Table 7** EfficientNet-B0 Internal Confusion Matrix (Dataset A)

| Preprocessing | TN  | FP  | FN | TP  | Normal<br>Misclassified |
|---------------|-----|-----|----|-----|-------------------------|
| Baseline      | 102 | 136 | 1  | 641 | 57.1%                   |
| Std Dev       | 128 | 110 | 1  | 641 | 46.2%                   |
| CLAHE         | 51  | 187 | 0  | 642 | 78.6%                   |
| Hist. Eq.     | 143 | 95  | 13 | 629 | 39.9%                   |
| StdDev+CLAHE  | 89  | 149 | 1  | 641 | 62.6%                   |

This case makes a central methodological point of the study: on a single-class cross-domain dataset, Recall can be deeply misleading when the model is biased, and it must be checked against internal precision. EfficientNet-B0 earned the highest cross-domain Recall (0.9994 on CLAHE) because of its bias, not because it generalizes best. ResNet-50 + Std Dev, by contrast, reached a cross-domain Recall of 0.9944 with high internal precision (0.9828), which makes it the configuration whose generalization can actually be trusted. This distinction is what sets a bias-checked evaluation apart from a naive one

that reads Recall alone. The point aligns with the literature on shortcut learning and fairness auditing in medical imaging models [26], [27], [28], [29], [30], [31], [32].

### 3.5. Robustness Across Random Seeds (Multi-Seed)

To confirm that the findings do not hinge on a single run, we retrained every combination on five seeds. Table 8 reports internal F1 and cross-domain Recall as mean  $\pm$  SD. The standard deviations are small (0.0008–0.0021), so the per-architecture preprocessing ranking, including the reversal on the CNN and its absence on ResNet-50, holds across seeds.

**Table 8** Multi-Seed Robustness Results (mean  $\pm$  SD across 5 seeds: 42, 123, 456, 789, 1010)

| Architecture    | Preprocessing | F1-Score Internal<br>(mean $\pm$ SD)  | Recall Cross-Domain<br>(mean $\pm$ SD) |
|-----------------|---------------|---------------------------------------|----------------------------------------|
| CNN Simple      | Baseline      | <b>0.9720 <math>\pm</math> 0.0012</b> | 0.9526 $\pm$ 0.0010                    |
| CNN Simple      | Std Dev       | 0.9632 $\pm$ 0.0010                   | 0.9549 $\pm$ 0.0011                    |
| CNN Simple      | CLAHE         | 0.9682 $\pm$ 0.0010                   | 0.9680 $\pm$ 0.0012                    |
| CNN Simple      | Hist. Eq.     | 0.9641 $\pm$ 0.0014                   | <b>0.9847 <math>\pm</math> 0.0009</b>  |
| CNN Simple      | StdDev+CLAHE  | 0.9608 $\pm$ 0.0010                   | 0.9567 $\pm$ 0.0012                    |
| ResNet-50       | Baseline      | 0.9807 $\pm$ 0.0009                   | 0.9879 $\pm$ 0.0006                    |
| ResNet-50       | Std Dev       | <b>0.9816 <math>\pm</math> 0.0008</b> | <b>0.9939 <math>\pm</math> 0.0006</b>  |
| ResNet-50       | CLAHE         | 0.9716 $\pm$ 0.0010                   | 0.9889 $\pm$ 0.0006                    |
| ResNet-50       | Hist. Eq.     | 0.9752 $\pm$ 0.0010                   | 0.9889 $\pm$ 0.0006                    |
| ResNet-50       | StdDev+CLAHE  | 0.9712 $\pm$ 0.0010                   | 0.9834 $\pm$ 0.0006                    |
| EfficientNet-B0 | Baseline      | 0.9023 $\pm$ 0.0020                   | 0.9968 $\pm$ 0.0006                    |
| EfficientNet-B0 | Std Dev       | 0.9190 $\pm$ 0.0017                   | 0.9953 $\pm$ 0.0006                    |
| EfficientNet-B0 | CLAHE         | 0.8714 $\pm$ 0.0021                   | <b>0.9992 <math>\pm</math> 0.0003</b>  |
| EfficientNet-B0 | Hist. Eq.     | <b>0.9200 <math>\pm</math> 0.0014</b> | 0.9860 $\pm$ 0.0010                    |
| EfficientNet-B0 | StdDev+CLAHE  | 0.8940 $\pm$ 0.0017                   | 0.9963 $\pm$ 0.0006                    |

*Note: the best preprocessing technique differs by architecture and is consistent across seeds. The simple CNN shows a reversal (Baseline to Hist. Eq.); ResNet-50 shows none (Std Dev best in both evaluations); EfficientNet-B0 shows a differently patterned reversal (Hist. Eq. to CLAHE). Differences larger than  $2 \times SD$  between conditions can be treated as consistent.*

EfficientNet-B0 is the least stable architecture here, with lower internal performance and a best-preprocessing choice that differs between the internal (Hist. Eq.) and cross-domain (CLAHE) evaluations. ResNet-50 is the most stable in this study, showing no ranking reversal at all. No single preprocessing technique is safe to use universally across architectures; the architecture-preprocessing interaction has to be evaluated case by case.

### 3.6. Discussion

This subsection interprets the results reported in Sections 3.1–3.5. It first explains why the effect of preprocessing depends on the architecture and what this means for clinical deployment (Section 3.6.1), then derives a configuration recommendation (Section 3.6.2), and closes with the limitations that bound these conclusions (Section 3.6.3).

#### 3.6.1. Architecture–Preprocessing Interaction

Figure 3 shows this interaction. The ranking reversal turns out to be architecture-specific, not a universal pattern. The simple CNN wins on Baseline internally but on Hist. Eq. cross-domain, a clear reversal. ResNet-50 shows none: Std Dev leads both internally and cross-domain. EfficientNet-B0 follows yet another pattern, with Hist. Eq. best internally but CLAHE best cross-domain. Three architectures, three different patterns. The ranking reversal cannot be generalized as a single cross-architecture phenomenon.

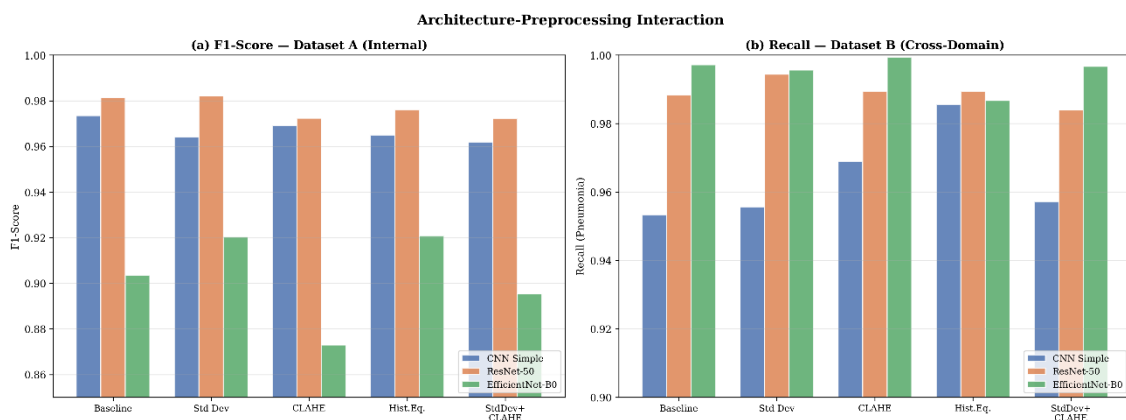


**Figure 3** Architecture-Preprocessing Interaction

The best technique on the source data is not always best on the target, and the effect depends on the architecture. Hist. Eq. flattens the intensity distribution and helps most for the simple CNN trained from scratch. On ResNet-50, which uses ImageNet pretrained

weights, Std Dev is the most stable across both evaluations, suggesting that pretrained representations are less sensitive to that kind of contrast normalization. This is consistent with domain shift [6], [15], [16], but it shows that the mitigation must be tuned per architecture rather than applied uniformly. Domain-generalization approaches, invariant feature learning, style-invariant models, and federated learning, offer another route [33], [34], [35].

A mechanistic reading of these results explains why the effect is architecture-specific. The simple CNN is trained from scratch and carries no pretrained priors, so a global operator that flattens the intensity distribution — histogram equalization — gives it the most transferable input and yields its best cross-domain Recall, even though the same flattening discards information and is not optimal internally. ResNet-50 carries ImageNet priors and is comparatively insensitive to global contrast changes, so standard-deviation normalization, which rescales intensities without distorting local structure, stays its most stable choice both internally and cross-domain and produces no ranking reversal. EfficientNet-B0 benefits most from CLAHE cross-domain because its depthwise-separable convolutions and compound scaling exploit the fine-grained local contrast that CLAHE restores, whereas the shallow CNN cannot make comparable use of those local cues. The reversal itself arises because each architecture relies on different, partly source-specific image cues; when the scanner distribution changes, those cues shift by different amounts, so the internal ranking of preprocessing techniques does not predict the cross-domain ranking.

Why histogram equalization does not carry over to the deeper, pretrained networks needs its own explanation. Histogram equalization is a global operator. It remaps the whole intensity distribution so that gray levels are spread evenly, without regard to which regions carry diagnostic content. In doing so it also stretches image noise, acquisition artifacts such as markers, collimation edges, and grid lines, and other irrelevant high-frequency patterns into the same contrast range as the lung parenchyma. ResNet-50 and EfficientNet-B0 have many pretrained layers tuned to texture and edge statistics, so this excessive contrast change can pull their feature representations away from clinically meaningful radiological patterns. Subtle findings such as lung opacity, infiltration, and consolidation depend on gradual differences in density and may be flattened or distorted. Our results fit this reading. For EfficientNet-B0, histogram equalization was the best

condition internally (F1 0.9209) yet gave the lowest cross-domain Recall of the five conditions (0.9867; 24 false negatives). For ResNet-50 it brought no gain over standard-deviation normalization (Recall 0.9894 versus 0.9944). The simple CNN, trained from scratch on equalized images, has no pretrained features to disrupt, which may explain why the same operator helped it.

These findings also carry an operational message for clinical deployment. The gap between internal and cross-domain behaviour is rooted in how images are acquired. Differences in tube voltage, exposure settings, detector technology, and acquisition protocols between the source data and a regional hospital change the brightness distribution, contrast, noise characteristics, and texture of the resulting images. A preprocessing step can either damp or amplify that shift, depending on the architecture it feeds. The danger is silent diagnostic degradation: a model can keep a high benchmark score and still fail after deployment, with no obvious error signal to the clinician. Before clinical integration, a medical AI system should therefore undergo external validation, local calibration, and out-of-distribution testing on locally acquired images rather than being accepted on benchmark results alone.

Two groups can act on this. Hospitals should maintain a standardized, versioned preprocessing pipeline so that the same operator is applied to every image, document the characteristics of each scanner (vendor, detector type, and typical exposure settings), and evaluate model performance periodically on recent local cases, repeating the check whenever equipment or protocols change. Developers should test several preprocessing-architecture combinations instead of fixing one in advance, and avoid selecting preprocessing from benchmark results alone. The choice should be confirmed on data from the target site, and the precision of the chosen model should be checked alongside its Recall, as the EfficientNet-B0 case shows.

### **3.6.2. Recommended Configuration**

We recommend ResNet-50 + Std Dev. It combines high internal precision (0.9828) with a cross-domain Recall of 0.9944 (10 false negatives out of 1,800), so its generalization is sound and unbiased. As a lighter option, the simple CNN + Hist. Eq. is also viable (cross-domain Recall 0.9856; 26 false negatives) and stays balanced. EfficientNet-B0 records the highest cross-domain Recall (CLAHE, 0.9994) but is not recommended: its heavy internal

bias makes that number misleading, and in clinical use the model would flag almost every patient as pneumonia, including healthy ones.

### 3.6.3. Limitations and Threats to Validity

First, the EfficientNet-B0 bias was observed under one training protocol applied identically to all architectures (fine-tuning the last 25% of layers, a learning rate of  $1e-4$ , and at most 30 epochs). We did not explore deeper fine-tuning, so we do not claim this bias as an intrinsic property of EfficientNet-B0; it is the model's behavior under the equal training budget we set. That is exactly the architecture-dependent point: under a reasonable, uniform training setup, different architectures respond to preprocessing differently. Whether the bias survives full fine-tuning is an open question for future work.

Second, our bias check rests on internal precision measured on Dataset A. Because Dataset B holds only the Pneumonia class, specificity and the false positive rate in the clinical domain cannot be measured directly. Source-domain precision does not guarantee target-domain precision, so the conclusion that the simple CNN and ResNet-50 generalize soundly should be read as a strong indication, not as proof of clinical specificity. Collecting Indonesian clinical Normal images, so that a full confusion matrix can be computed in the target domain, is the top priority for future work. In addition, the ground-truth labels for Dataset B were taken from single-reader routine clinical reports rather than from a multi-reader consensus, so a small amount of label noise cannot be ruled out. Independent re-reading of the clinical images by more than one radiologist is left to future work.

Third, all clinical data come from one hospital in Central Java, while the public data are pediatric (Guangzhou). The configuration recommendation (ResNet-50 + Std Dev as the soundest choice, or the simple CNN + Hist. Eq. as a lighter alternative) therefore holds within that scope and may not generalize to other populations, device vendors, or acquisition protocols. Multi-center validation is needed before the recommendation is used more broadly. Finally, the task is framed as binary classification (pneumonia vs. normal) and does not distinguish pneumonia etiology, such as bacterial versus viral cases; extending the evaluation to multi-class etiology is left to future work.

#### 4. CONCLUSION

This study examined how five preprocessing techniques interact with three CNN architectures in cross-domain pneumonia detection, using public Kaggle chest X-rays for training and 1,800 Indonesian clinical images for external evaluation. Three conclusions follow. First, the ranking of preprocessing techniques changes with the architecture: the simple CNN reversed from Baseline internally to Histogram Equalization cross-domain, EfficientNet-B0 reversed from Histogram Equalization to CLAHE, and ResNet-50 kept standard-deviation normalization first in both settings, so ranking reversal under domain shift is architecture-dependent rather than universal. Second, internal validation alone is insufficient for deployment, because the internal ranking did not predict the cross-domain ranking for two of the three architectures and because Recall on a single-class clinical set can be inflated by a biased model, as EfficientNet-B0 showed with a Recall of 0.9994 despite low internal precision. Third, external clinical evaluation is required before implementation; among the configurations tested, ResNet-50 with standard-deviation normalization combined high internal precision with high cross-domain Recall and is the soundest candidate within the scope of these data. Future research should extend the evaluation to multi-center validation across several Indonesian hospitals and device vendors, add clinical Normal images and multi-class pneumonia classification, and test domain adaptation approaches that could reduce the dependence on the choice of preprocessing.

#### REFERENCES

- [1] "Pneumonia in children," World Health Organization. Accessed: Aug. 28, 2026. [Online]. Available: <https://www.who.int/news-room/fact-sheets/detail/pneumonia>
- [2] "Hasil Utama Survei Kesehatan Indonesia (SKI) 2023," Badan Kebijakan Pembangunan Kesehatan, Kementerian Kesehatan Republik Indonesia. Accessed: Sep. 18, 2026. [Online]. Available: <https://www.badankebijakan.kemkes.go.id/daftar-frequently-asked-question-seputar-hasil-utama-ski-2023/hasil-utama-ski-2023/>
- [3] S. A. Srigati, A. Munawir, and P. Wulandari, "Potensi kecerdasan buatan dalam peningkatan akurasi pembacaan hasil mamografi: Tinjauan sistematis dan meta-analisis," *Ganesha Medicina Journal*, vol. 3, no. 1, pp. 65–71, Mar. 2023, doi: 10.23887/gm.v3i1.55575.

- [4] F.-Z. Hamlili, M. Beladgham, M. Khelifi, and A. Boudia, "Transfer learning with Resnet-50 for detecting COVID-19 in chest X-ray images," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 25, no. 3, pp. 1458–1468, Mar. 2022, doi: 10.11591/ijeecs.v25.i3.pp1458-1468.
- [5] N. Y. Hegazy and M. S. Sawah, "An optimized EfficientNetB0 framework with CLAHE-based preprocessing for accurate multi-class chest X-ray classification," *Sci. Rep.*, vol. 16, no. 1, p. 10811, Mar. 2026, doi: 10.1038/s41598-026-42492-1.
- [6] B. Guo *et al.*, "The impact of scanner domain shift on deep learning performance in medical imaging: an experimental study," *Int. J. Comput. Assist. Radiol. Surg.*, early access, Apr. 2026, doi: 10.1007/s11548-026-03655-7.
- [7] A. Giełczyk, A. Marciniak, M. Tarczewska, and Z. Lutowski, "Pre-processing methods in chest X-ray image classification," *PLoS One*, vol. 17, no. 4, p. e0265949, Apr. 2022, doi: 10.1371/journal.pone.0265949.
- [8] V. M, S. Rajkumar, M. V. Kalyan Reddy, and V. Janesh, "Detection of post COVID-pneumonia using histogram equalization, CLAHE deep learning techniques," *Intel. Artif.*, vol. 26, no. 72, pp. 137–145, Sep. 2023, doi: 10.4114/intartif.vol26iss72pp137-145.
- [9] A. Raji and T. Olwal, "Performance evaluation of a hybrid histogram equalization-contrast limited adaptive histogram equalization technique for medical images enhancement," *Turk. J. Eng.*, vol. 10, no. 3, pp. 875–884, Jul. 2026, doi: 10.31127/tuje.1842375.
- [10] A. N. bin Azhar, N. S. Sani, and L. L. X. Wei, "Enhancing COVID-19 detection in X-ray images through deep learning models with different image preprocessing techniques," *Int. J. Adv. Comput. Sci. Appl.*, vol. 16, no. 1, p. 61, 2025, doi: 10.14569/IJACSA.2025.0160161.
- [11] V. Arya and T. Kumar, "Enhancing image for CNN-based diagnostic of pediatric pneumonia through chest radiographs," *Int. J. Adv. Comput. Sci. Appl.*, vol. 14, no. 2, p. 45, 2023, doi: 10.14569/IJACSA.2023.0140245.
- [12] G. Siracusano, A. La Corte, A. G. Nucera, M. Gaeta, M. Chiappini, and G. Finocchio, "Effective processing pipeline PACE 2.0 for enhancing chest X-ray contrast and diagnostic interpretability," *Sci. Rep.*, vol. 13, no. 1, p. 22471, Dec. 2023, doi: 10.1038/s41598-023-49534-y.
- [13] O. Dahmane, M. Khelifi, M. Beladgham, and I. Kadri, "Pneumonia detection based on transfer learning and a combination of VGG19 and a CNN Built from scratch," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 24, no. 3, pp. 1469–1480, Dec. 2021, doi: 10.11591/ijeecs.v24.i3.pp1469-1480.

- [14] D. S. Kermany *et al.*, "Identifying medical diagnoses and treatable diseases by image-based deep learning," *Cell*, vol. 172, no. 5, pp. 1122–1131.e9, Feb. 2018, doi: 10.1016/j.cell.2018.02.010.
- [15] A. Musa, R. Prasad, and M. Hernandez, "Addressing cross-population domain shift in chest X-ray classification through supervised adversarial domain adaptation," *Sci. Rep.*, vol. 15, no. 1, p. 11383, Apr. 2025, doi: 10.1038/s41598-025-95390-3.
- [16] P. M. Fernández-Miranda *et al.*, "A retrospective study of deep learning generalization across two centers and multiple models of X-ray devices using COVID-19 chest-X rays," *Sci. Rep.*, vol. 14, no. 1, p. 14657, Jun. 2024, doi: 10.1038/s41598-024-64941-5.
- [17] T. Kwon *et al.*, "Diagnostic performance of artificial intelligence model for pneumonia from chest radiography," *PLoS One*, vol. 16, no. 4, p. e0249399, Apr. 2021, doi: 10.1371/journal.pone.0249399.
- [18] N. T. Vinh, L. T. Hien, H. M. Toan, and D. N. Toan, "Diagnosing tuberculosis from X-ray imaging based on contrast limited adaptive histogram equalization," *IAES Int. J. Artif. Intell.*, vol. 15, no. 3, pp. 2567–2580, Jun. 2026, doi: 10.11591/ijai.v15.i3.pp2567-2580.
- [19] A. Shavkatovich Buriboev, A. Abduvaitov, and H. S. Jeon, "Binary classification of pneumonia in chest X-ray images using modified contrast-limited adaptive histogram equalization algorithm," *Sensors*, vol. 25, no. 13, p. 3976, Jun. 2025, doi: 10.3390/s25133976.
- [20] S. Karagol, K. N. Akpinar, and S. Genc, "Comparison of deep learning approaches in the classification of the chest X-ray," *J. Eng. Res.*, vol. 11, no. 3, p. 17121, Sep. 2023, doi: 10.36909/jer.17121.
- [21] X. Huang *et al.*, "External validation based on transfer learning for diagnosing atelectasis using portable chest X-rays," *Front. Med.*, vol. 9, p. 920040, Jul. 2022, doi: 10.3389/fmed.2022.920040.
- [22] M. Abad, J. Casas-Roma, and F. Prados, "Generalizable disease detection using model ensemble on chest X-ray images," *Sci. Rep.*, vol. 14, no. 1, p. 5890, Mar. 2024, doi: 10.1038/s41598-024-56171-6.
- [23] S. Gautam, M. M.-C. Hohne, S. Hansen, R. Jenssen, and M. Kampffmeyer, "Demonstrating the risk of imbalanced datasets in chest X-ray image-based diagnostics by prototypical relevance propagation," in *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*, IEEE, Mar. 2022, pp. 1–5, doi: 10.1109/ISBI52829.2022.9761651.

- [24] Z. Zhang, X. Zhang, K. Ichiji, I. Bukovský, and N. Homma, "How intra-source imbalanced datasets impact the performance of deep learning for COVID-19 diagnosis using chest X-ray images," *Sci. Rep.*, vol. 13, no. 1, p. 19049, Nov. 2023, doi: 10.1038/s41598-023-45368-w.
- [25] D. Arias-Garzón, R. Tabares-Soto, J. Bernal-Salcedo, and G. A. Ruz, "Biases associated with database structure for COVID-19 detection in X-ray images," *Sci. Rep.*, vol. 13, no. 1, p. 3477, Mar. 2023, doi: 10.1038/s41598-023-30174-1.
- [26] C. Ong Ly *et al.*, "Shortcut learning in medical AI hinders generalization: method for estimating AI model generalization without external data," *NPJ Digit. Med.*, vol. 7, no. 1, p. 124, May 2024, doi: 10.1038/s41746-024-01118-4.
- [27] B. G. Hill, F. L. Koback, and P. L. Schilling, "The risk of shortcutting in deep learning algorithms for medical imaging research," *Sci. Rep.*, vol. 14, no. 1, p. 29224, Nov. 2024, doi: 10.1038/s41598-024-79838-6.
- [28] A. Brown, N. Tomasev, J. Freyberg, Y. Liu, A. Karthikesalingam, and J. Schrouff, "Detecting shortcut learning for fair medical AI using shortcut testing," *Nat. Commun.*, vol. 14, no. 1, p. 4314, Jul. 2023, doi: 10.1038/s41467-023-39902-7.
- [29] A. K. Pathak, M. Gupta, and G. Jain, "Unmasking the Clever Hans effect in AI models: shortcut learning, spurious correlations, and the path toward robust intelligence," *Front. Artif. Intell.*, vol. 8, p. 1692454, Jan. 2026, doi: 10.3389/frai.2025.1692454.
- [30] M. Haupt, E. Jolkver, and A. Schwerk, "Explainable AI for chest radiographs: sex-stratified fairness auditing in CNN-based pneumonia detection," *Sci. Rep.*, vol. 16, no. 1, p. 22709, Jul. 2026, doi: 10.1038/s41598-026-62761-3.
- [31] R. Zhang, D. Griner, J. W. Garrett, Z. Qi, and G.-H. Chen, "Training certified detectives to track down the intrinsic shortcuts in COVID-19 chest X-ray data sets," *Sci. Rep.*, vol. 13, no. 1, p. 12690, Aug. 2023, doi: 10.1038/s41598-023-39855-3.
- [32] A. Trivedi *et al.*, "Deep learning models for COVID-19 chest X-ray classification: preventing shortcut learning using feature disentanglement," *PLoS One*, vol. 17, no. 10, p. e0274098, Oct. 2022, doi: 10.1371/journal.pone.0274098.
- [33] S. Ghimire, S. Kashyap, J. T. Wu, A. Karargyris, and M. Moradi, "Learning invariant feature representation to improve generalization across chest X-ray datasets," in *Machine Learning in Medical Imaging (MLMI 2020)*, Lecture Notes in Computer Science, vol. 12436. Cham: Springer, 2020, pp. 644–653, doi: 10.1007/978-3-030-59861-7\_65.

- [34] M. Zunaed, M. A. Haque, and T. Hasan, "Learning to generalize towards unseen domains via a content-aware style invariant model for disease detection from chest X-rays," *IEEE J. Biomed. Health Inform.*, vol. 28, no. 6, pp. 3626–3636, Jun. 2024, doi: 10.1109/JBHI.2024.3372999.
- [35] S. Tayebi Arasteh, C. Kuhl, M.-J. Saehn, P. Isfort, D. Truhn, and S. Nebelung, "Enhancing domain generalization in the AI-based analysis of chest radiographs with federated learning," *Sci. Rep.*, vol. 13, no. 1, p. 22576, Dec. 2023, doi: 10.1038/s41598-023-49956-8.