Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control
Journal homepage: http://kinetik.umm.ac.id
ISSN: 2503-2267
Vol. 10, No. 3, August, Pp. 411-430

Efficient thoracic abnormalities detection using mobile deep learning models

Achmad Bauravindah*[1], Dhomas Hatta Fudholi[2], Rima Tri Wahyuningrum[3]
Master Program in Informatics, Faculty of Industrial Technology, Islamic University of Indonesia, Yogyakarta, Indonesia[1]
Department of Informatics, Faculty of Industrial Technology, Islamic University of Indonesia, Yogyakarta, Indonesia[2]
Department of Informatics Engineering, Faculty of Engineering, Universitas Trunojoyo Madura, Bangkalan, Indonesia[3]

Article Info

Keywords:
Artificial Intelligence, Healthcare, Chest X-Ray Image, Mobile Deep Learning, Lung Abnormalities

Article history:
Received: March 04, 2025
Accepted: June 02 2025
Published: August 31, 2025

Cite:
A. Bauravindah, D. H. Fudholi, and R. T. Wahyuningrum, "Efficient Thoracic Abnormalities Detection Using Mobile Deep Learning Models", *KINETIK*, vol. 10, no. 3, Aug. 2025. https://doi.org/10.22219/kinetik.v10i3.2268

*Corresponding author.
Achmad Bauravindah
E-mail address:
23917016@students.uii.ac.id

Abstract

Indonesia faces a critical shortage of radiologists, with only 1.2 radiologists per 100,000 individuals. This shortage leads to delays in diagnosing thoracic abnormalities such as pneumothorax, cardiomegaly, nodule/mass, consolidation, and infiltration. Chest X-ray (CXR) interpretation remains challenging due to overlapping radiological features, necessitating AI-assisted solutions. This study evaluates three lightweight deep learning models—MobileNetV2, ShuffleNetV2, and EfficientNetB0—for automated thoracic abnormality detection using the ChestX-ray8 dataset. We assessed model performance using accuracy, precision, recall, F1-score, and AUC-ROC, selecting the best model based on the highest per-fold F1-score. EfficientNetB0 emerged as the top-performing model, achieving a macro-average F1-score of 0.556 and AUC-ROC of 0.765, outperforming MobileNetV2 (0.494, 0.719) and ShuffleNetV2 (0.481, 0.713). Grad-CAM analysis revealed strong localization for pneumothorax and consolidation but misclassifications in cardiomegaly and nodule/mass detection due to poor feature differentiation. The findings highlight EfficientNetB0's potential as an AI-assisted diagnostic tool for low-resource settings while also underscoring the need for segmentation-based pretraining and multi-scale feature extraction to enhance detection accuracy. Future work should focus on optimizing sensitivity to subtle abnormalities and ensuring clinical trust through improved interpretability techniques.

1. Introduction

Indonesia faces significant public health challenges, with a high prevalence of both communicable and non-communicable diseases, notably tuberculosis (TB) and pneumonia. According to the Global TB Report 2023, Indonesia ranks second globally in TB cases, with an estimated 1,060,000 cases and 134,000 deaths annually—equating to about 15 deaths every hour [1]. Pneumonia remains the leading cause of infectious death among children under five in Indonesia, affecting approximately half a million children annually and resulting in about 10,000 deaths [2]. These diseases, along with chronic obstructive pulmonary disease (COPD), lung cancer, heart failure, and interstitial lung diseases, contribute significantly to the high morbidity and mortality rates [3][4][5][6]. Prompt and accurate diagnosis is crucial for effective treatment and control; yet, it remains a challenge in Indonesia due to resource limitations and a shortage of radiologists. The country has only 2,161 registered radiologists, which equates to approximately 1.2 radiologists per 100,000 individuals, and faces a shortage of 31,481 specialized doctors, including radiologists, across its healthcare system [7][8]. This shortage, combined with the uneven distribution of healthcare professionals across provinces, particularly in rural areas, overwhelms the healthcare system, causing delays in diagnosis and treatment and exacerbating the health burden [7].

Chest X-ray (CXR) is a fundamental imaging tool for detecting a variety of thoracic conditions, including tuberculosis (TB), pneumonia, and other critical diseases [9][10][11]. As a non-invasive, cost-effective, and widely available modality, CXR plays a crucial role in diagnosing respiratory diseases. However, interpreting CXR images requires substantial expertise, as abnormalities may present subtly or overlap, complicating accurate diagnosis—particularly in resource-limited settings [12][13].

Tuberculosis (TB) often presents on chest imaging as nodules, masses, and consolidation, aiding diagnosis; nodules or masses may indicate active infection, granulomas, or lung cancer, with tuberculomas occurring in about 5% of post-primary TB cases, sometimes forming cavities [14][15]. Pneumonia, especially in children, manifests as consolidation from alveolar exudate and infiltration, seen as patchy or diffuse opacities and often linked to viral infections like influenza [16][17][18]. COPD typically shows pneumothorax, causing lung collapse and respiratory distress, and infiltration, indicating inflammatory airway changes and lung damage [19][20][21][22][23]. Lung cancer appears as nodules or masses on imaging, with malignancy suspected in lesions over 3 cm or those with irregular growth patterns

[24][25]. Heart failure impacts both cardiovascular and pulmonary systems, leading to pulmonary congestion and cardiomegaly, identified on chest X-rays by a cardiothoracic ratio above 50%, signaling left ventricular dysfunction or pericardial effusion [26][27][28][29][30][31]. These abnormalities—pneumothorax, cardiomegaly, nodules/masses, consolidation, and infiltration—are key indicators of severe thoracic diseases, but accurate interpretation requires expert radiological skills, posing a challenge in resource-limited settings.

Deep learning offers a promising solution to challenges in medical imaging, particularly in detecting these specific abnormalities. Instead of diagnosing diseases directly, deep learning models provide a more detailed analysis by identifying signs indicative of various conditions. For example, a nodule on CXR might suggest different diagnoses, such as TB or lung cancer, depending on other clinical factors. By quantifying these abnormalities, deep learning models can assist radiologists in making faster and more accurate assessments, which is invaluable in resource-limited settings.

The use of Artificial Intelligence (AI) tools for pre-screening and highlighting potential abnormalities can significantly reduce the time required for radiologists to produce diagnostic reports [32][33][34][35][36], allowing them to focus more on areas requiring attention. This is especially beneficial in high-volume environments, such as large hospitals or during disease outbreaks, where rapid turnaround is crucial. However, emerging gaps include limitations on available hardware, particularly in areas with limited resources or in rural health facilities [37]. Many hospitals and clinics in Indonesia lack access to advanced computing devices, such as servers or computers with high GPU capabilities [38], due to the high costs involved in providing such technology [39]. Thus, there is an urgent need to develop deep learning models that are not only accurate but also can run efficiently on devices with low computing power.

Mobile models, such as MobileNet, ShuffleNet, and EfficientNet, provide efficient solutions for medical imaging challenges by optimizing computational speed without sacrificing accuracy [40][41][42]. These models are highly efficient and enable deployment on mobile devices and low-power systems, making them ideal for healthcare facilities with limited technology. Their accessibility benefits regions with poor infrastructure, as smartphones or tablets can facilitate fast and accurate diagnoses in rural or developing areas [37]. Additionally, their speed allows radiologists to produce diagnostic reports quickly, which is vital during high patient volumes or health emergencies [43]. Economically, mobile models reduce costs by minimizing hardware needs, enabling hospitals to implement AI without expensive infrastructure investments [40]. Their scalability allows integration into various healthcare settings, from clinics to telemedicine platforms, ensuring wider deployment and remote usability [44]. MobileNet, with depthwise separable convolutions, reduces computational load, making it efficient for mobile use [40]. ShuffleNet employs pointwise group convolutions and channel shuffling to maintain accuracy while lowering computational costs [42]. EfficientNet uses multi-scale compound scaling to balance network dimensions, achieving efficient performance with fewer parameters [41]. Together, these models offer fast, scalable, and cost-effective diagnostic solutions suitable for resource-limited healthcare environments.

In addition to their architectural efficiency, these models have shown strong empirical performance in chest X-ray classification tasks. MobileNetV2, in particular, has demonstrated exceptional accuracy and speed, making it highly suitable for medical image analysis. For instance, MobileNetV2 achieved 98.65% accuracy and 98.15% recall in pneumonia and COVID-19 detection [45], and was ranked highest among 11 Convolutional Neural Networks (CNNs) for both accuracy and speed [46]. [47] Velu (2023) further fine-tuned MobileNetV2 for accurate and rapid COVID-19 detection from chest X-rays, achieving 92.5% training accuracy and 93.75% validation accuracy, outperforming both scratch-trained CNNs (81.4%) and fine-tuned ResNet50 (80.6%). Additionally, a systematic review by Iqbal et al. (2024) reported a 94% accuracy rate of MobileNetV2 on the CXR-14 dataset [48], while Gu and Lee (2024) demonstrated the model's effectiveness with transfer learning, achieving 90.9% accuracy in pneumonia detection [49].

ShuffleNetV2, another lightweight and efficient model, has also shown promising results. Gu and Lee (2024) reported that ShuffleNetV2 achieved 91.2% accuracy using transfer learning for pneumonia detection, highlighting its potential for fast and accurate disease classification [49]. An et al. (2022) emphasized its low parameter count and model weight, making it ideal for embedded applications, which are critical in resource-limited healthcare settings [50].

EfficientNetB0 has also demonstrated outstanding performance, balancing accuracy and computational efficiency. It reached 99% classification accuracy, outperforming deeper models like ResNet-50 and VGG-19 in both efficiency and accuracy [51]. Furthermore, Kansal et al. (2024) reported that EfficientNetB0 outperformed ResNet-50 with a testing accuracy of 99.62% on the Kaggle dataset and 99.78% on the Mendeley dataset for multi-centric lung abnormality classification [52]. Iqbal et al. (2024) corroborated these findings, noting EfficientNetB0's 98% accuracy in COVID-19 detection [48]. Additionally, An et al. (2024) demonstrated that combining EfficientNetB0 with DenseNet121, enhanced by attention mechanisms, achieved a high diagnostic accuracy of 95.19%, with enhanced precision (98.38%) and F1 score (96.06%) [53].

These findings from prior research validate the selection of MobileNetV2, ShuffleNetV2, and EfficientNetB0 for this study and reinforce their suitability for real-world clinical applications in underserved regions. Their balance of accuracy, computational efficiency, and adaptability through fine-tuning and transfer learning makes them prime candidates for diagnostic tasks where both speed and accuracy are crucial. All the significant literature discussed is summarized in Table 1, providing a concise overview of the models' performance and key findings across various studies.

To enhance the interpretability of these models in medical applications, techniques such as Gradient-weighted Class Activation Mapping (Grad-CAM) have been employed [32][54][55]. Grad-CAM provides a mechanism to visualize which areas of the input image contribute most to the model’s prediction, allowing clinicians to understand and trust the diagnostic decisions made by the model. In the context of thoracic abnormality detection, Grad-CAM can help highlight areas of the lungs affected by disease, providing a visual tool for localizing pathological areas in CXR. This capability is invaluable in clinical environments where affordability and transparency of AI models are essential for gaining acceptance and trust from healthcare professionals.

Evaluating deep learning models in medical imaging often involves rigorous testing procedures to ensure robustness and generalization. Techniques such as K-Fold Cross Validation are used to split data into several subsets, or folds, and validate the model across these folds to minimize bias and variance [56][57]. In addition, performance metrics such as accuracy, precision, recall, and F1 score are used to holistically evaluate model performance, covering aspects of accuracy and consistency in predictions [58]. Moreover, performance metrics such as the Area Under the Receiver Operating Characteristic Curve (AUC-ROC) are critical for evaluating the diagnostic accuracy of the model. The AUC-ROC curve measures the balance between the true positive rate and the false positive rate, providing deep insight into the model’s ability to distinguish between different classes—in this case, the presence or absence of thoracic abnormalities [59].

Integrating lightweight deep learning models with advanced evaluation and visualization techniques aims to enhance the efficiency and accuracy of abnormality detection, especially in environments where traditional radiological resources are limited. By leveraging models such as MobileNet, EfficientNet, and ShuffleNet, this research seeks to develop powerful and mobile-friendly diagnostic tools that can operate effectively on devices with limited resources, thereby improving healthcare services and outcomes in underserved areas. This approach not only addresses the urgent need for scalable diagnostic solutions but also contributes to the broader goal of making advanced medical imaging technology more accessible and practical for widespread use in clinical practice. To achieve this objective, the study addresses the following research question: How can the implementation of lightweight and interpretable deep learning models (using Grad-CAM) enhance accuracy and efficiency in detecting thoracic abnormalities in chest X-ray (CXR) images within resource-constrained settings in Indonesia?

Table 1.Summary of the Significant Literature Studies

Focus	Author	Objective	Relevance
MobileNetV2, ShuffleNetV2	An et al. (2022) [50]	Develop a lightweight deep neural network (E-TBNet) for automatic detection of tuberculosis using X-ray DR imaging, optimized for devices with lower hardware levels.	MobileNetV2 achieved the highest accuracy (90%) among lightweight models, while ShuffleNetV2 excelled in size and efficiency. E-TBNet balanced accuracy (85%) and efficiency.
MobileNetV2	Velu (2023) [47]	Develop a fine-tuned MobileNetV2 model for accurate and rapid COVID-19 detection from chest X-rays.	Fine-tuned MobileNetV2 achieved superior performance (92.5% training accuracy, 93.75% validation accuracy), outperforming both scratch-trained CNN (81.4%) and fine-tuned ResNet50 (80.6%).
MobileNetV2	Akter et al. (2021) [46]	Develop a deep learning model for detecting COVID-19 from chest X-rays using CNN architectures.	Among the evaluated models (MobileNetV2, VGG16, ResNet50), MobileNetV2 achieved the highest accuracy (98%).
MobileNetV2	Kolonne et al. (2021) [45]	Develop a MobileNetV2-based model for classifying normal, pneumonia, and COVID-19 conditions from chest X-rays.	MobileNetV2 without transfer learning achieved the highest accuracy (98.65%) compared to the transfer learning approach (97.89%)
MobilenetV2, EfficientNetB0	Iqbal et al. (2024) [48]	Systematic review of AI methods for lung disease detection using chest X-rays.	MobileNetV2 achieved 94% accuracy on the CXR-14 dataset, while EfficientNetB0 achieved 98% accuracy in COVID-19 detection.
Transfer Learning Utilization,	Gu and Lee	Utilize deep transfer learning from general-purpose datasets (like	Lightweight models ShuffleNetV2 and MobileNetV2 showed significant performance

MobileNetV2, ShuffleNetV2	(2024) [49]	ImageNet) to classify pneumonia in X-rays.	improvements with transfer learning, achieving accuracies of 91.2% and 90.9%, respectively.
GradCAM Use	Gakhar et al. (2022) [60]	Develop a two-stage pipeline for thoracic abnormality detection and disease classification using fusion DCNNs.	The proposed fusion-based approach (ThoraciNet) achieved an AUC of 0.99 for CXR triaging and 0.79 for disease classification, outperforming single-stream DCNNs. GradCAM visualization was used to enhance interpretability.
EfficientNetB0	Kansal et al. (2024) [52]	Compare the performance of EfficientNetB0 and ResNet-50 for multi-centric lung abnormality classification.	EfficientNetB0 outperformed ResNet-50 with a testing accuracy of 99.62% on the Kaggle dataset and 99.78% on the Mendeley dataset.
EfficientNetB0, XAI Implementation	Sahin et al. (2024) [61]	Utilize CNN models with Grad-CAM++ to detect pneumonia from chest X-rays.	EfficientNetB0 achieved the highest accuracy (95.03%) and F-measure (96.12%), while VGG-19 and Inception-V3 also showed competitive performance.
EfficientNetB0	An et al. (2024) [53]	Develop a CNN model combining EfficientNetB0 and DenseNet121, enhanced by attention mechanisms for pneumonia detection.	Achieved high diagnostic accuracy (95.19%) with enhanced precision (98.38%) and F1 score (96.06%) due to integrating attention mechanisms.

2. Research Method

This section outlines the methodology employed in this study to develop and evaluate deep learning models for thoracic abnormality detection. First, we introduce the deep learning architectures used, including MobileNet, ShuffleNet, and EfficientNet, detailing their design principles and advantages. Next, we describe the dataset selection process, including dataset characteristics, preprocessing steps, and the strategy for creating a balanced subset. The research design is then presented, covering dataset partitioning, model training procedures, and evaluation strategies. Finally, we discuss the performance metrics used to assess model effectiveness, alongside interpretability techniques such as Grad-CAM, ensuring both accuracy and clinical reliability.

2.1 Deep Learning Models

2.1.1 MobileNet

MobileNetV2 is a lightweight deep learning architecture optimized for mobile and embedded applications. It improves upon the original MobileNet (as shown in Figure 1) by using depthwise separable convolutions to reduce computational complexity and inverted residuals with linear bottlenecks to enhance feature reuse and preserve important information [40][62]. These design choices make MobileNetV2 highly efficient, both in terms of speed and parameter count, without sacrificing classification performance.

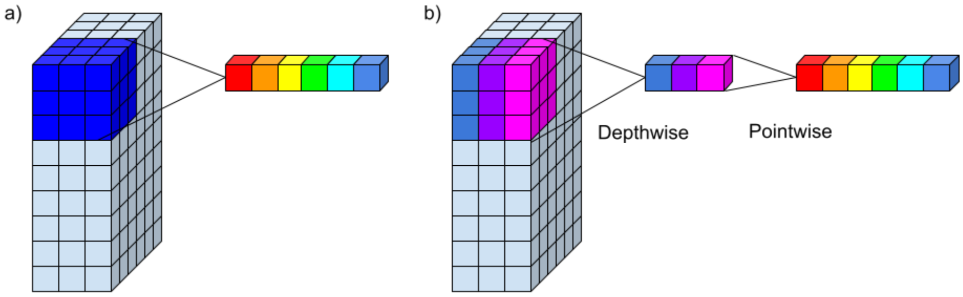
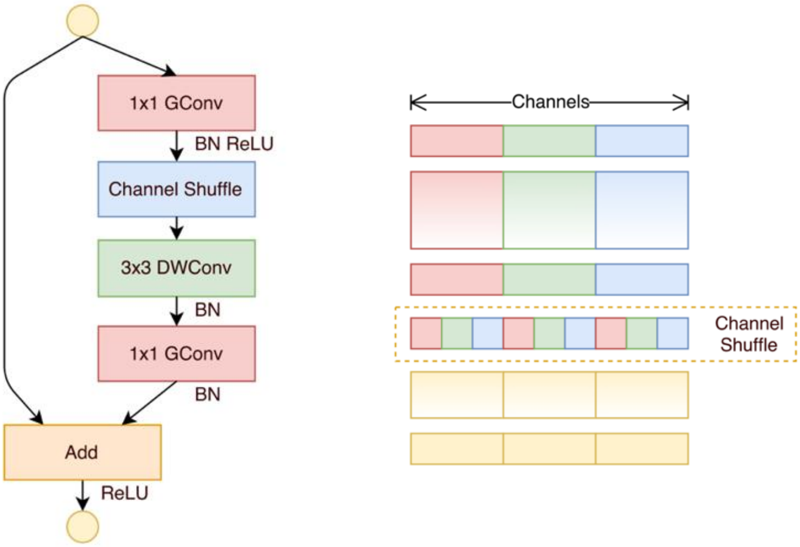


Figure 1. MobileNet Architecture: a) Normal Convolution Technique b) Depthwise Separable Convolution Technique (as base technique of MobileNet)[63]

Its suitability for medical imaging tasks has been demonstrated in multiple studies. For instance, MobileNetV2-based models have achieved over 90% accuracy in detecting COVID-19 and pneumonia from chest X-ray images [45][46][47]. These results support its use in this study for thoracic abnormality detection in resource-limited environments.

2.1.2 ShuffleNet

ShuffleNetV2 is a lightweight convolutional neural network optimized for real-time, low-power applications. It improves upon its predecessor by removing group convolutions (as seen in Figure 2a) to reduce memory access costs and introducing an equal channel width design and simplified operations for enhanced hardware efficiency [64]. A key strength of ShuffleNet architectures is the use of channel shuffling (as shown in Figure 2b), which improves information flow between feature groups after pointwise convolutions [64][65]. These architectural changes allow ShuffleNetV2 to maintain high speed and accuracy while minimizing computation—making it particularly suitable for embedded systems and mobile devices. In medical imaging, ShuffleNetV2 has demonstrated promising results. For example, Gu and Lee (2024) reported a pneumonia detection accuracy of 91.2% using transfer learning, outperforming MobileNetV2 and ResNet18 [49]. Similarly, An et al. (2022) found ShuffleNetV2 to be among the most efficient lightweight models for tuberculosis detection on embedded platforms, though with slightly lower recall compared to heavier networks like their proposed E-TBNet [50].



(a) (b)

Figure 2. ShuffleNet Architecture a) Shuffle Unit with Pointwise Group Convolution (GConv) and Channel Shuffle b) Detailed of Channel Shuffle [65]

2.1.3 EfficientNet

EfficientNet architecture, as shown in Figure 3, is a lightweight architecture widely recognized for its ability to strike an optimal balance between performance and efficiency. It introduces a new approach to model scaling known as compound scaling [41]. EfficientNet uses a systematic method of compound scaling to expand the model proportionally across three dimensions: network depth, network width, and input image resolution [41]. Unlike traditional scaling approaches that only adjust one dimension at a time, compound scaling adjusts all these dimensions simultaneously using a fixed scaling coefficient. This allows the model to improve performance efficiently without leading to overfitting or underfitting. EfficientNet simplifies the usually manual process of tuning model architecture by optimizing the scaling of the network for various tasks [41]. This approach produces a more balanced and efficient architecture compared to models that scale just one dimension at a time. EfficientNet is designed to deliver high performance while using fewer parameters and computational power [41]. This results in models that are not only smaller but also faster during inference. The EfficientNet family, ranging from variants B0 to B7, offers a variety of options depending on resource constraints and the desired level of accuracy. EfficientNetB0 is used in this study because it is a smaller variant, making it suitable for devices with limited power, which aligns with the need in Indonesia. In contrast, larger variants like B7 provide higher accuracy for more demanding applications. Additionally, EfficientNet demonstrates high energy efficiency, making it ideal for mobile device applications where battery usage is a critical factor. As such, EfficientNet is well-suited for tasks like medical image analysis on devices with limited computational power but requiring high accuracy results.

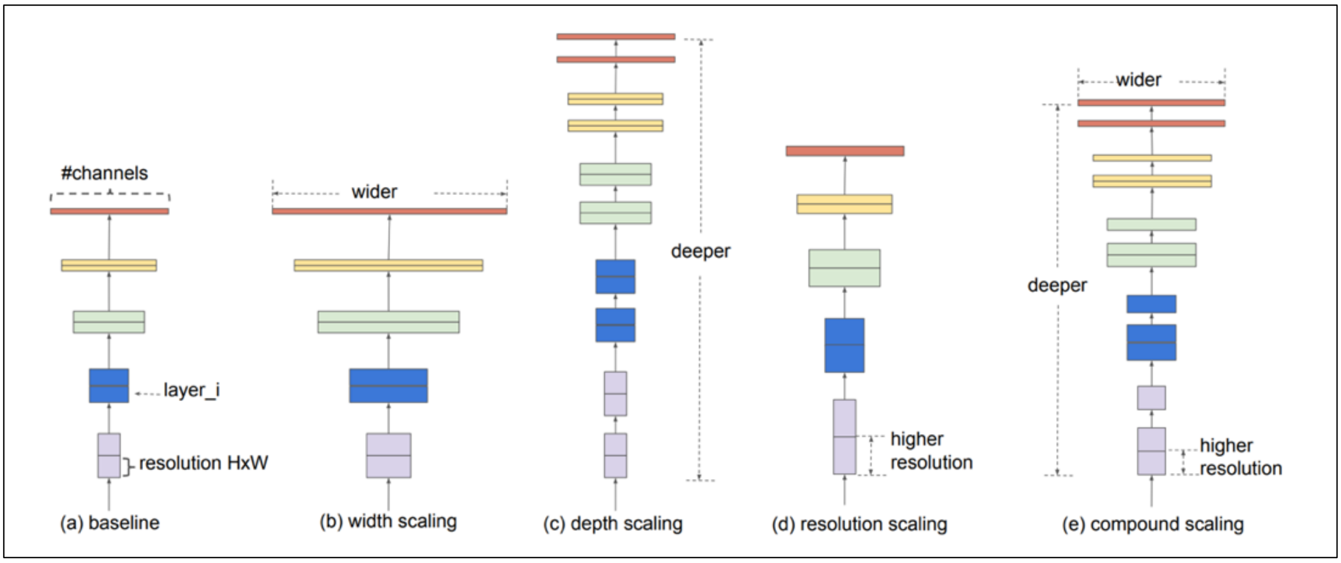


Figure 3. Scaling Model: (a) is the baseline network; (b)-(d) are conventional scaling methods that increase one dimension of width, depth, or resolution; (e) is the multiple scaling method of EfficientNet, which uniformly enlarges the three dimensions with a constant ratio [41]

2.2 Dataset

The dataset used in this research is ChestX-ray8, introduced by Wang et al. from the National Institutes of Health (NIH) [34]. It is one of the largest publicly available collections of chest X-ray images, consisting of 108,948 frontal-view X-ray images from 32,717 unique patients. Each image is labeled with one or more of eight common thoracic diseases: Atelectasis, Cardiomegaly, Effusion, Infiltration, Mass, Nodule, Pneumonia, and Pneumothorax. The labels were obtained through automated text mining from radiology reports using Natural Language Processing (NLP) tools (MetaMap and DNorm), with negation and uncertainty detection applied to exclude erroneous labels. The dataset is multi-label, meaning each image can exhibit multiple abnormalities.

The ChestX-ray8 dataset was chosen due to its clinical significance, particularly in the context of thoracic abnormality detection in Indonesia. The dataset covers a wide range of critical conditions such as pneumothorax, cardiomegaly, nodule/mass, consolidation, and infiltration, which are prevalent in the Indonesian population and represent significant diagnostic challenges (explained in Section 1: Introduction).

Each chest X-ray image in the ChestX-ray8 dataset is originally sized at 1024×1024 pixels. However, in this study, we resized all images to 224×224 pixels to match the input requirements of the pretrained ImageNet model, which serves as the backbone for our deep learning architecture. This resizing step ensures compatibility with the pretrained convolutional neural networks while preserving essential visual features needed for disease classification.

For this study, we focus on detecting five key abnormalities: Pneumothorax, Cardiomegaly, Nodule/Mass, Consolidation, and Infiltration, selected due to their clinical importance and representation in the dataset, as well as their relation to prevalent diseases in Indonesian. To create a balanced subset suitable for model training and evaluation, we selected 2,500 images per abnormality, totaling 12,500 images. This number was chosen based on the maximum available images for Cardiomegaly (2,776), which was capped at 2,500 for consistency across all classes. Although the dataset remains multi-label, this approach ensures equal representation across the five targeted abnormalities.

The number of images (presented in Table 2) for each selected abnormality before sampling was: Pneumothorax (2,534), Cardiomegaly (2,525), Nodule/Mass (3,494), Consolidation (2,953), and Infiltration (4,678). After balancing, the dataset comprises an equal number of images per abnormality, though multi-label overlaps remain, preserving the complexity of the classification task.

Table 2. Dataset Used with Five Selected Abnormalities

Abnormality	Number of Images
Pneumothorax	2534
Cardiomegaly	2525
Nodule/Mass	3494
Consolidation	2953
Infiltration	4678

The dataset was divided into training and testing sets with an 80:20 split. Table 3 presents the split dataset, where 80% (10,000 images – before k-fold) were used for model training and 20% (2,500 images) for testing. Additionally, to enhance model generalization and reduce overfitting, we employed 5-fold cross-validation (K=5) with a shuffled splitting method, ensuring that the training set was randomly reordered before being divided into five subsets. The model was trained on four subsets and validated on the fifth, iterating through all folds.

Table 3. Split Dataset for Training Purposes

Subset	Number of Images
Training	8000
Testing	2500
Validation (k-fold)	2000
Total	12500

This dataset configuration ensures a semi-balanced, multi-label, and clinically relevant input for training deep learning models, supporting the goal of developing efficient thoracic abnormality detection models for resource-limited healthcare settings.

2.3 Research Design

Figure 4 outlines the scenarios developed in this study, with the following details:

1. Dataset Repository: The chosen dataset was acquired from NIH repository, as explained in Section 2.2.
2. Dataset Reduction: Due to limited model training resources, the dataset is reduced to 12,500 images in total. These images contain five labels: pneumothorax, cardiomegaly, nodule/mass, consolidation, and infiltration.
3. Normalization: Before the model training process begins, the pixel values of the images need to be normalized. This helps make the training process more efficient and prevents gradient overflow, which can occur when irrelevant gradients are used during training.
4. Dataset Splitting: The dataset is split into a training set and a test set. The training set is used to train the model, while the test set is used to assess the final performance of the model.
5. K-Fold Cross Validation: After splitting the data into training and test sets, the training set is divided into training and validation sets. The K-fold cross-validation technique is applied, with $K = 5$, meaning the data is split into 80% for training and 20% for validation.
6. Model Training: Three models—MobileNet, ShuffleNet, and EfficientNet—are used in this study. These models are trained using the previously divided training dataset.
7. Test Data Prediction: To evaluate the performance of each model, the trained models are tested using the test dataset that was set aside earlier.
8. Model Saving: Once the models are trained, they are saved so they can be reused for future testing or deployment.
9. Final Evaluation: The models are evaluated based on the predictions made on the test data to determine which model provides the best performance. This evaluation offers insights into the effectiveness of each model.
10. Grad-CAM Visualization: Some images are predicted and shown along their abnormalities' localization. This could help radiologists to understand and interpret the abnormalities using the exhibited heatmap.

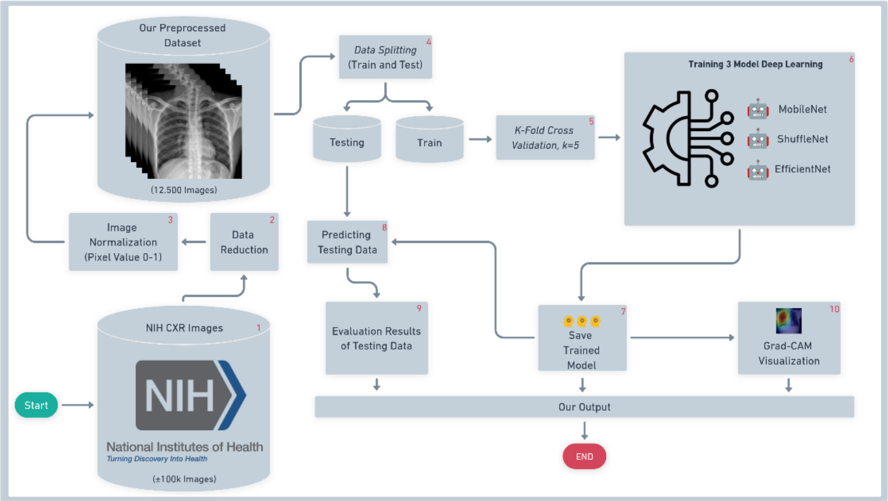

Our Preprocessed Dataset
(12.500 Images)
Data Splitting (Train and Test) 4
Testing
Train
K-Fold Cross Validation, k=5 5
Training 3 Model Deep Learning 6
MobileNet
ShuffleNet
EfficientNet
Image Normalization (Pixel Value 0-1) 3
Data Reduction 2
Predicting Testing Data 8
NIH CXR Images 1
NIH
National Institutes of Health
Turning Discovery Into Health
(±100k Images)
Start
Evaluation Results of Testing Data 9
Save Trained Model 7
Grad-CAM Visualization 10
Our Output
END

Figure 4. Research Workflow

Cite: A. Bauravindah, D. H. Fudholi, and R. T. Wahyuningrum, "Efficient Thoracic Abnormalities Detection Using Mobile Deep Learning Models", KINETIK, vol. 10, no. 3, Aug. 2025. https://doi.org/10.22219/kinetik.v10i3.2268

2.4 Evaluation Metrics and Interpretability

To ensure the robustness and generalization of the deep learning models for thoracic abnormality detection, comprehensive evaluation metrics and interpretability techniques are employed. **K-Fold Cross Validation** is implemented to mitigate overfitting and enhance model reliability by partitioning the dataset into *K* equal-sized folds, where each fold serves as a validation set once while the remaining *K* - 1 folds are used for training. This process is repeated *K* times, and the average performance across all folds is reported, ensuring every data point is utilized for both training and validation. Additionally, several performance metrics are used to assess diagnostic accuracy and consistency. **Accuracy**, which measures overall correctness, is calculated as the ratio of correctly predicted cases to total cases, as presented in Equation 1.

$$\mathrm{Accuracy} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \tag{1}$$

To evaluate classification quality, precision, recall, and F1 score are employed. **Precision** indicates the proportion of true abnormalities among all predicted abnormalities, which is calculated using Equation 2.

$$\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{2}$$

Recall measures the proportion of correctly identified abnormalities among all actual abnormalities, which is calculated using Equation 3.

$$\mathrm{Recall} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{3}$$

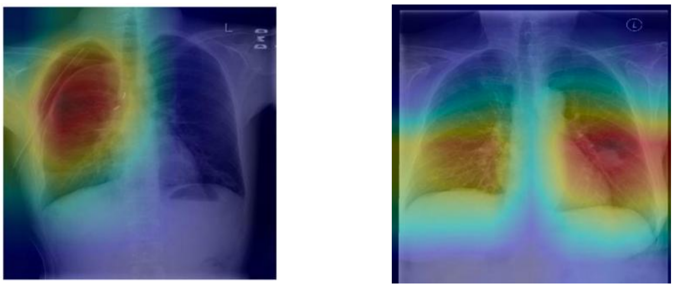
The **F1 score** provides a balance between precision and recall, which is particularly useful for imbalanced datasets, as calculated in Equation 4.

$$\mathrm{F1\ Score} = \frac{2 \times \mathrm{Precision} \times \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}} \tag{4}$$

Furthermore, **AUC-ROC** (Area Under the Receiver Operating Characteristic Curve) is used to evaluate the model's ability to differentiate between normal and abnormal cases, where a higher AUC signifies superior diagnostic performance. The ROC curve plots the true positive rate (sensitivity - TPR) against the false positive rate (FPR) across various classification thresholds, providing a comprehensive view of the model's discriminatory power; both are calculated using Equation 5.

$$\begin{gathered} \mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \\ \mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}} \end{gathered} \tag{5}$$

In addition to performance evaluation, interpretability is crucial for gaining clinician trust and ensuring practical applicability. Grad-CAM is employed to generate heatmaps that highlight the regions in chest X-rays contributing most to the model's predictions. As shown in Figure 5, Grad-CAM highlights different thoracic abnormalities: pneumothorax, nodule, and cardiomegaly. The heatmaps are overlaid on the original images, with the red areas indicating regions of highest model attention. For instance, the pneumothorax heatmap shows intense activation near the collapsed lung area, while the nodule detection highlights localized spots indicative of potential malignancies. In cardiomegaly detection, Grad-CAM emphasizes the enlarged heart region, aligning with radiological markers. By visualizing these activation maps, clinicians can interpret the model's focus areas, such as lung fields for tuberculosis or cardiomegaly, which shows strong activation around the heart area, consistent with an enlarged heart silhouette, a key indicator of heart failure. This not only enhances transparency but also facilitates clinical validation by ensuring the model's decision-making aligns with medical expertise. To further ensure reliability, expert feedback from radiologists is integrated into the evaluation process, allowing clinicians to assess whether the highlighted regions are consistent with their diagnostic reasoning. This combination of **quantitative evaluation** metrics and **qualitative interpretability** techniques ensures that the deep learning models are both accurate and clinically trustworthy, making them suitable for deployment in resource-limited settings, such as rural healthcare facilities in Indonesia.



Pneumothorax *Nodule* *Cardiomegaly*

Figure 5. Samples of Grad-CAM Implementation for Detecting 3 Different Abnormalities [66]

2.5 Hyperparmeter Configuration

To maintain consistency and fairness across all compared models (MobileNetV2, ShuffleNetV2, and EfficientNetB0), the Hyperparameter Configuration presented in Table 4 contains a standardized set of hyperparameters used during the training process.

Table 4. Hyperparameter Configuration

Hyperparameter	Value	Description
Input Image Size	(224, 224)	The size of the image after resizing (width × height), matching the requirements of the pretrained backbone.
Number of Channels	3	Images are in RGB format, consisting of three color channels.
Number of Classes	5	The model distinguishes between Pneumothorax, Cardiomegaly, Nodule/Mass, Consolidation, and Infiltration.
Batch Size	512	A large batch size is used to ensure training stability and memory efficiency.
Number of Epochs	50	A fixed number of training cycles to ensure fair evaluation of the models.
Initial Learning Rate	1.00E-04	The starting rate at which model weights are updated during optimization.
Optimizer	Adam	An adaptive optimizer chosen for its efficiency with relatively small datasets.
Loss Function	BCEWithLogitsLoss	Suitable for multi-label classification tasks.
Label Threshold	0.5	The probability threshold for predicting a positive label.
Backbone Freezing	No (False)	All layers of the model are fine-tuned, meaning no layers are frozen during training.

This standardized approach to hyperparameter selection ensures that the training process is conducted in a consistent and unbiased manner, allowing for a fair comparison between the performances of MobileNetV2, ShuffleNetV2, and EfficientNetB0. By employing the same training settings, differences in model performance can be attributed more accurately to the model architecture rather than variations in training methodology.

3. Results and Discussion

3.1 Evaluation of Model Performance on Validation Data

To determine the most effective deep learning model for thoracic abnormality detection, three lightweight architectures—MobileNetV2, ShuffleNetV2, and EfficientNetB0—were evaluated based on F1 score, accuracy, precision, recall, and AUC-ROC. The results, averaged across K-fold cross-validation, are presented in Table 5. The best model will continue to final evaluation in testing stage for comprehensive analysis.

Cite: A. Bauravindah, D. H. Fudholi, and R. T. Wahyuningrum, “Efficient Thoracic Abnormalities Detection Using Mobile Deep Learning Models”, KINETIK, vol. 10, no. 3, Aug. 2025. https://doi.org/10.22219/kinetik.v10i3.2268

Table 5. Model Performance on Validation Data (Average and Standar Deviation Across K-Folds)

Model	**F1 Score**	**Accuracy**	**Precision**	**Recall**	**AUC-ROC**
MobileNetV2	0.494 ± 0.008	0.732 ± 0.003	0.486 ± 0.009	0.505 ± 0.009	0.72 ± 0.005
ShuffleNetV2	0.463 ± 0.014	0.74 ± 0.008	0.512 ± 0.014	0.431 ± 0.016	0.697 ± 0.01
EfficientNetB0	**0.563 ± 0.009**	**0.773 ± 0.005**	**0.573 ± 0.007**	**0.558 ± 0.01**	**0.772 ± 0.005**

From Table 5, it can be seen that EfficientNetB0 consistently outperforms the other two models across all evaluation metrics, achieving the highest F1 score (0.563 ± 0.009), accuracy (0.773 ± 0.005), recall (0.558 ± 0.010), and AUC-ROC (0.772 ± 0.005). The low standard deviation across all metrics suggests stable performance across different validation folds. In contrast, MobileNetV2 performs moderately well, with an F1 score of 0.494 ± 0.008, but struggles with precision (0.486 ± 0.009), which indicates a higher false positive rate. ShuffleNetV2 performs the worst, with the lowest F1 score (0.463 ± 0.014) and recall (0.431 ± 0.016), indicating that it frequently fails to detect actual positive cases, a serious limitation in medical diagnostics.

MobileNetV2 demonstrates reasonable stability, particularly in accuracy (±0.003), but its higher variance in precision (±0.009) and recall (±0.009) suggests occasional inconsistencies in detecting abnormalities. In contrast, ShuffleNetV2 exhibits the highest performance fluctuations, with F1 score SD (±0.014) and recall SD (±0.016), making it the least reliable model due to its unpredictable false negative rates. This instability poses a significant risk in medical imaging, where missing abnormalities can be life-threatening.

While averaged results provide an overall performance assessment, selecting the best model requires identifying the highest F1 score achieved in any validation fold. Table 6 presents the per-fold F1 scores for each model.

Table 6. F1 Score per Fold for Each Model

Model	**Fold 1**	**Fold 2**	**Fold 3**	**Fold 4**	**Fold 5**	**Best**
MobileNetV2	0.488	0.496	0.501	**0.504**	0.483	Fold 4 (0.504)
ShuffleNetV2	0.454	0.453	0.448	0.473	**0.485**	Fold 5 (0.485)
EfficientNetB0	0.546	0.566	0.566	0.568	**0.569**	Fold 5 (0.569)

From this per-fold analysis, EfficientNetB0 achieves the highest F1 score of 0.569 in Fold 5, outperforming MobileNetV2 (0.504 in Fold 4) and ShuffleNetV2 (0.485 in Fold 5). Moreover, EfficientNetB0 consistently scores above 0.546 across all folds, whereas the other models exhibit greater variability in performance.

To finalize the model selection, Table 7 summarizes the highest F1 score achieved by each model and determines whether it qualifies as the best-performing model.

Table 7. Selecting the Best Model Based on Maximum F1 Score

Model	F1 Score	Selected as Best Model?
MobileNetV2	Fold 4 (0.504)	No
ShuffleNetV2	Fold 5 (0.485)	No
EfficientNetB0	**Fold 5 (0.569)**	**Yes**

With the highest single-fold F1 score (0.569 in Fold 5) and overall superior performance across all metrics, EfficientNetB0 is conclusively the best model for thoracic abnormality detection. Its higher recall (0.558 ± 0.010) is particularly valuable in medical imaging, as minimizing false negatives is crucial for avoiding missed diagnoses. Additionally, its AUC-ROC of 0.772 further supports its superior ability to distinguish between normal and abnormal cases. With EfficientNetB0 selected as the best model, the next step is to evaluate its performance on the test dataset.

3.2 Evaluation of Model Performance on Testing Data

To assess the real-world applicability of the models, we evaluated MobileNetV2, ShuffleNetV2, and EfficientNetB0 on the testing dataset, analyzing their ability to classify pneumothorax, cardiomegaly, nodule/mass, consolidation, and infiltration. Performance was measured using accuracy, precision, recall, F1 score, and AUC-ROC, ensuring a balanced evaluation of classification effectiveness. The analysis focuses on per-class performance, followed by a macro-average comparison to determine the most effective model for thoracic abnormality detection.

3.2.1 MobileNetV2 Testing Results Analysis

We evaluated the performance of MobileNetV2 in classifying five thoracic abnormalities—pneumothorax, cardiomegaly, nodule/mass, consolidation, and infiltration—using accuracy, precision, recall, F1 score, and AUC-ROC. Table 8 summarizes the results:

Table 8. MobileNetV2 Testing Results

Label	Accuracy	Precision	Recall	F1 Score	ROC-AUC
pneumothorax	0.835	0.609	0.590	0.599	0.842
cardiomegaly	0.710	0.398	0.415	0.406	0.663
nodule/mass	0.592	0.457	0.490	0.473	0.590
consolidation	0.802	0.480	0.528	0.503	0.786
infiltration	0.731	0.468	0.506	0.486	0.713
MacroAVG	0.734	0.483	0.506	0.494	0.719

The model demonstrated moderate performance, as indicated by a macro-average F1 score of 0.494 and AUC-ROC of 0.719. The recall (0.506) surpassing precision (0.483) suggests that MobileNetV2 detects more abnormal cases but at the cost of false positives. Pneumothorax detection was the strongest, achieving the highest accuracy (0.835), precision (0.609), and AUC-ROC (0.842), likely due to well-defined radiological features. Conversely, cardiomegaly exhibited the weakest performance with the lowest precision (0.398), recall (0.415), and F1 score (0.406), indicating difficulty in distinguishing enlarged hearts due to their subtle presentation. Consolidation detection showed a relatively strong recall (0.528) but lower precision (0.480), suggesting moderate performance. Both nodule/mass (AUC-ROC = 0.590) and infiltration (AUC-ROC = 0.713) had lower classification effectiveness, likely due to their variable and less distinct features.

The AUC-ROC curve on Figure 6 further illustrates MobileNetV2's ability to differentiate abnormalities. Pneumothorax had the highest AUC-ROC (0.842), reinforcing its strong classification capability. Consolidation followed with 0.786, indicating relatively good sensitivity. Infiltration scored 0.713, while cardiomegaly (0.663) and nodule/mass (0.590) performed the worst, with near-random classification ability. The model's limitations in detecting cardiomegaly and nodule/mass could be attributed to the subtle shape variations in the former and the small, diverse nature of lung nodules in the latter.

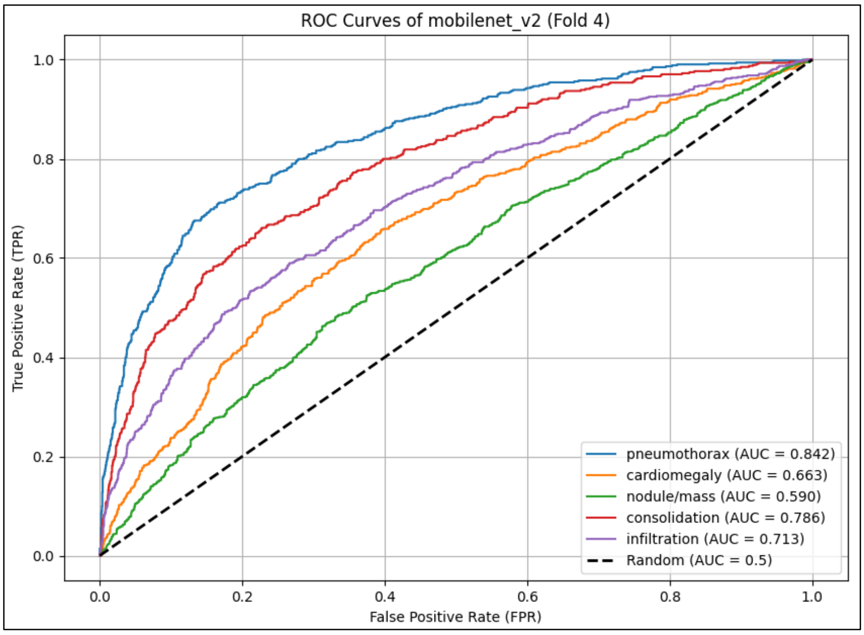


Figure 6. MobileNetV2 ROC-Curves

To improve performance, enhancements such as data augmentation (for better generalization on cardiomegaly cases) and advanced feature extraction (for detecting small lesions like nodules/masses) could be beneficial. While MobileNetV2 performs well for pneumothorax detection and may be useful for automated triage in resource-limited settings, it requires further optimization for detecting cardiomegaly and nodule/mass.

3.2.2 ShuffleNetV2 Testing Results Analysis

ShuffleNetV2's performance was consistently evaluated as in the previous section (**3.2.1**). The results are presented in Table 9.

Table 9. ShuffleNetV2 Testing Results

Label	Accuracy	Precision	Recall	F1 Score	ROC-AUC
pneumothorax	0.835	0.609	0.590	0.599	0.842
cardiomegaly	0.710	0.398	0.415	0.406	0.663
nodule/mass	0.592	0.457	0.490	0.473	0.590
consolidation	0.802	0.480	0.528	0.503	0.786
infiltration	0.731	0.468	0.506	0.486	0.713
MacroAVG	0.751	0.518	0.450	0.481	0.713

The macro-average F1 score of 0.481 indicates moderate classification performance, slightly lower than MobileNetV2 (0.494). ShuffleNetV2, however, achieves higher precision (0.518) but suffers from lower recall (0.450), leading to fewer false positives but more missed cases. Its best metrics are accuracy (0.751) and AUC-ROC (0.713), showing a fair ability to differentiate between normal and abnormal cases. However, recall (0.450) is the weakest, suggesting it misses a significant number of true positive cases.

ShuffleNetV2 excels in pneumothorax detection, achieving the highest accuracy (0.859) and AUC-ROC (0.827), with precision (0.692) being the best among all conditions. While it produces fewer false positives, its recall (0.588) is still suboptimal. Conversely, cardiomegaly detection is the model's weakest area, with the lowest recall (0.331), F1 score (0.372), and a poor AUC-ROC (0.662), indicating an inability to distinguish cardiomegaly from normal cases effectively.

Performance for nodule/mass (AUC-ROC = 0.614, F1 score = 0.461) and consolidation (AUC-ROC = 0.759, F1 score = 0.483) is moderate, with slight improvement over MobileNetV2 for nodule/mass but slightly worse results for consolidation. Infiltration detection (AUC-ROC = 0.704, F1 score = 0.455) also lags behind MobileNetV2, showing that ShuffleNetV2 struggles with diffuse lung conditions.

The AUC-ROC curve, as shown in Figure 7, further illustrates the model's classification ability across conditions. Pneumothorax (AUC = 0.827) is the best-performing class, with strong distinction from normal cases. Consolidation (AUC = 0.759) and infiltration (AUC = 0.704) show moderate classification ability, though with some misclassification issues. However, cardiomegaly (AUC = 0.662) and nodule/mass (AUC = 0.614) remain weak, struggling to differentiate from normal cases.

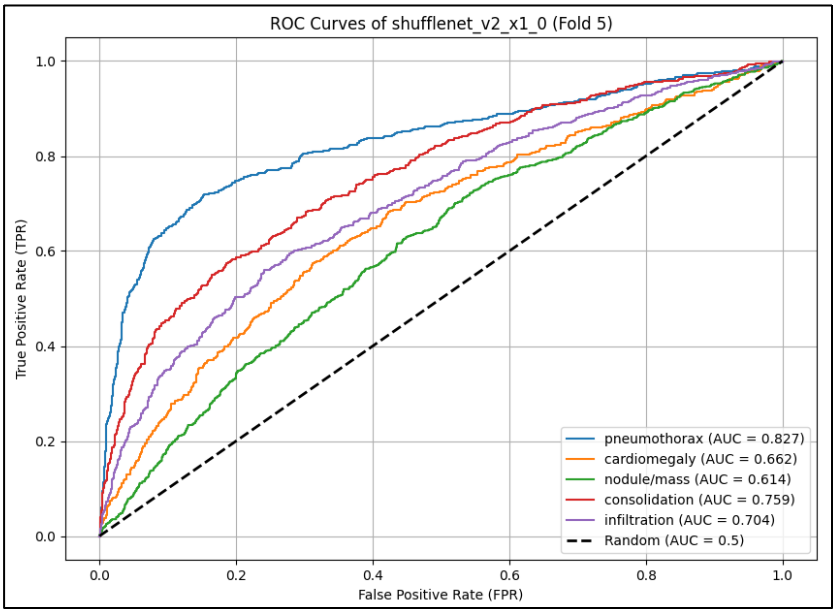


Figure 7. ShuffleNetV2 ROC-Curves

To enhance ShuffleNetV2's classification performance, optimizing recall for cardiomegaly is crucial, including weighted loss functions to improve sensitivity. Feature extraction enhancements, such as multi-scale convolutional filters and attention mechanisms, could help detect small abnormalities such as nodules.

3.2.3 EfficientNetB0 Testing Results Analysis

In this section, we analyze the test performance of EfficientNetB0, focusing on its ability, as shown in Table 10.

Table 10. EfficientNetB0 Testing Results

Label	Accuracy	Precision	Recall	F1 Score	ROC-AUC
pneumothorax	0.868	0.696	0.660	0.678	0.880
cardiomegaly	0.756	0.490	0.430	0.458	0.699
nodule/mass	0.618	0.488	0.519	0.503	0.646
consolidation	0.844	0.587	0.598	0.592	0.835
infiltration	0.772	0.547	0.548	0.548	0.764
MacroAVG	0.772	0.562	0.551	0.556	0.765

EfficientNetB0 outperforms MobileNetV2 and ShuffleNetV2 across all conditions, achieving the highest macro-average F1 score (0.556) and AUC-ROC (0.765), surpassing MobileNetV2 (0.494, 0.719) and ShuffleNetV2 (0.481, 0.713). It has the strongest recall (0.551), ensuring better identification of abnormal cases while minimizing false negatives.

The best performance is seen in pneumothorax detection, with the highest accuracy (0.868), precision (0.696), F1 score (0.678), and AUC-ROC (0.880). This suggests pneumothorax remains the easiest condition for CNN-based models to detect, with EfficientNetB0 significantly outperforming MobileNetV2 (AUC = 0.842) and ShuffleNetV2 (AUC = 0.827).

For nodule/mass detection, EfficientNetB0 demonstrates the greatest improvement, achieving the highest recall (0.519) and better AUC-ROC (0.646) than MobileNetV2 (0.590) and ShuffleNetV2 (0.614). This indicates improved sensitivity in detecting small lung abnormalities, although performance remains moderate and requires further optimization.

EfficientNetB0 also excels in consolidation detection, with the highest recall (0.598), precision (0.587), and AUC-ROC (0.835), making it highly effective in diagnosing pneumonia or lung infections. Similarly, infiltration detection shows notable improvement, with better recall (0.548), the highest F1 score (0.548), and an AUC-ROC of 0.764, outperforming MobileNetV2 (0.713) and ShuffleNetV2 (0.704).

While cardiomegaly detection remains challenging, EfficientNetB0 still achieves the highest AUC-ROC (0.699) and F1 score (0.458) compared to MobileNetV2 (0.663, 0.438) and ShuffleNetV2 (0.662, 0.431). Although classification remains difficult due to subtle heart enlargement, EfficientNetB0 shows measurable improvement.

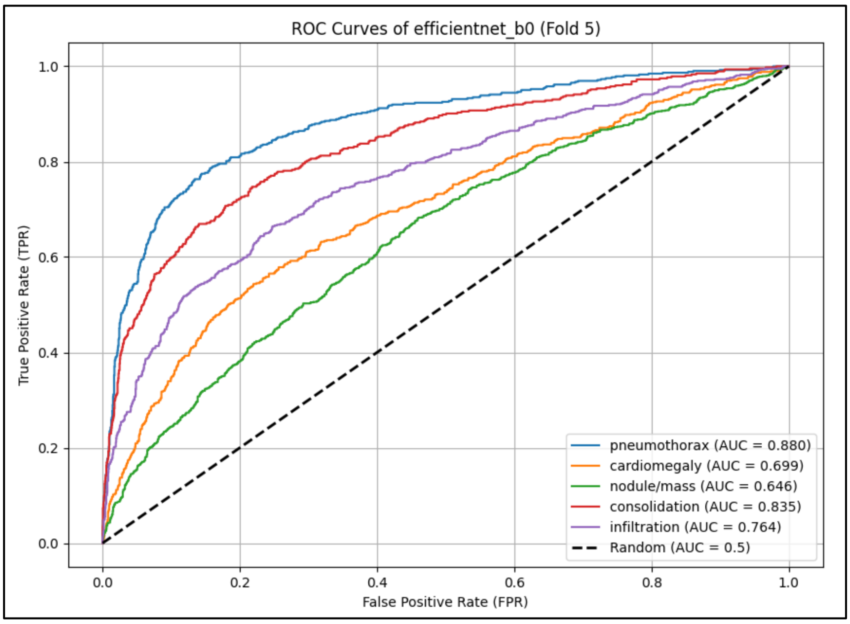


Figure 8. EfficientNetB0 ROC-Curves

EfficientNetB0 consistently achieves the highest AUC-ROC across all abnormalities, as shown in Figure 8 for its curve and Table 11 for a comparison across models, demonstrating superior classification ability. The most significant improvements are seen in pneumothorax, nodule/mass, and consolidation detection, while cardiomegaly remains the most difficult to classify.

Table 11. AUC-ROC Comparison of Models

Label	MobileNetV2	ShuffleNetV2	EfficientNetB0
pneumothorax	0.842	0.827	**0.880**
cardiomegaly	0.663	0.662	**0.699**
nodule/mass	0.590	0.614	**0.646**
consolidation	0.786	0.759	**0.835**
infiltration	0.713	0.704	**0.764**

To further optimize EfficientNetB0, cardiomegaly classification can be improved through segmentation-based pretraining and increasing labeled cases in the dataset. Nodule/mass detection can benefit from multi-scale feature extraction and attention mechanisms to enhance small lesion detection.

3.2.4 Macro-Average Metrics Analysis

The final comparative analysis evaluates the overall performance of MobileNetV2, ShuffleNetV2, and EfficientNetB0 based on macro-average accuracy, precision, recall, F1 score, and AUC-ROC. As seen in Table 12, EfficientNetB0 consistently outperforms the other models across all metrics, making it the most suitable for thoracic abnormality detection.

Table 12. Macro-Average Performance Comparison Across Models

Label	Accuracy	Precision	Recall	F1 Score	ROC-AUC
MobileNetV2	0.734	0.483	0.506	0.494	0.719
ShuffleNetV2	0.751	0.518	0.450	0.481	0.713
EfficientNetB0	**0.772**	**0.562**	**0.551**	**0.556**	**0.765**

EfficientNetB0 demonstrated superior classification balance in our study, achieving the highest accuracy (0.772), precision (0.562), recall (0.551), and F1 score (0.556). The high recall highlights its ability to minimize missed abnormal cases, which is crucial in medical imaging, while the high precision reduces false alarms. In contrast, MobileNetV2 (F1 score: 0.494, AUC-ROC: 0.719) and ShuffleNetV2 (F1 score: 0.481, AUC-ROC: 0.713) performed weaker, with ShuffleNetV2 notably struggling due to low recall (0.450), indicating an increase in false negatives. These findings align with the literature where MobileNetV2 has demonstrated notable accuracy in various contexts, such as 98.65% in COVID-19 detection without transfer learning (Kolonne et al. 2021) [45], 94% on the CXR-14 dataset (Iqbal et al. 2024) [48], and significant improvements with transfer learning, achieving 90.9% accuracy (Gu and Lee 2024) [49]. In lightweight model comparisons, MobileNetV2 excelled in accuracy (90%) while ShuffleNetV2 was more efficient in size (An et al. 2022) [50]. However, our study's relatively lower F1 scores for MobileNetV2 and ShuffleNetV2 also resonate with the limitations noted in previous studies, such as the lower F1 score of 0.435 in CheXNet, despite outperforming radiologists (Pranav Rajpurkar et al. 2017) [66]. This suggests that while these lightweight models are efficient, they may compromise on balanced classification, unlike EfficientNetB0, which shows more consistent performance across key metrics.

3.3 Comprehensive Performance Across Models

The performance evaluation of each model during the testing phase was based on three main factors presented in Table 13: inference time on 2,500 X-ray images, the number of trainable parameters, and final classification accuracy.

Table 13. Accuracy, Testing Time, and Model Complexity Across Models

Model	Accuracy	Testing Time (s)	Testing Time Per Image (ms)	Trainable Parameter
MobileNetV2	0.734	76.37	~30.5	2,230,277
ShuffleNetV2	0.751	44.04	~17.6	1,258,729
EfficientNetB0	0.772	82.53	~33.0	4,013,953

EfficientNetB0 achieved the highest accuracy (0.772), demonstrating strong generalization and feature representation. However, it also had the longest inference time (82.53 seconds) and the most parameters (over 4 million), making it less suitable for computationally limited environments. ShuffleNetV2 proved the most efficient, with the shortest inference time (44.04 seconds) and the fewest parameters (~1.26 million), while maintaining competitive accuracy (0.751), even surpassing MobileNetV2 (0.734), which had more parameters (~2.23 million) and a longer inference time (76.37 seconds). This shows that model efficiency does not always compromise performance when the architecture is well designed.

Regarding average inference time per image, ShuffleNetV2 took only 17.6 ms, significantly faster than EfficientNetB0 (33.0 ms) and MobileNetV2 (30.5 ms), making it ideal for rapid-response applications like mass screening or remote disease detection. A bubble chart (Figure 9) visualizes the trade-offs, with EfficientNetB0 in the upper right (high accuracy, high complexity), ShuffleNetV2 in the lower left (high efficiency, moderate accuracy), and MobileNetV2 in the middle (balanced performance).

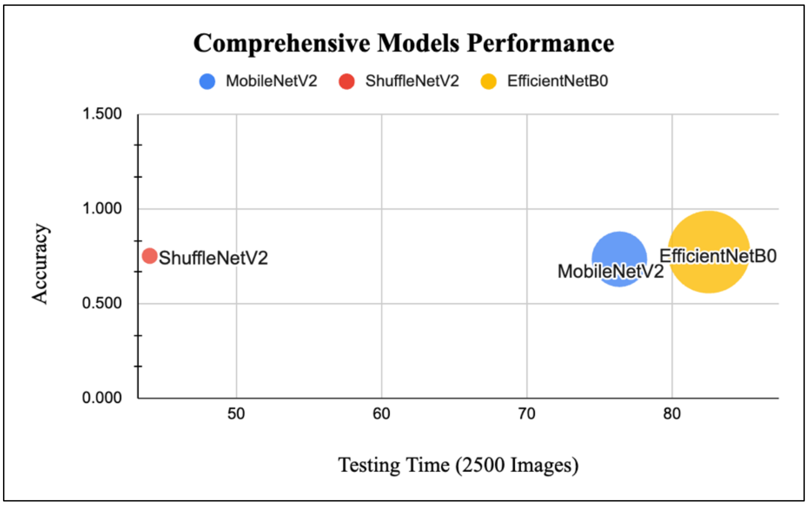
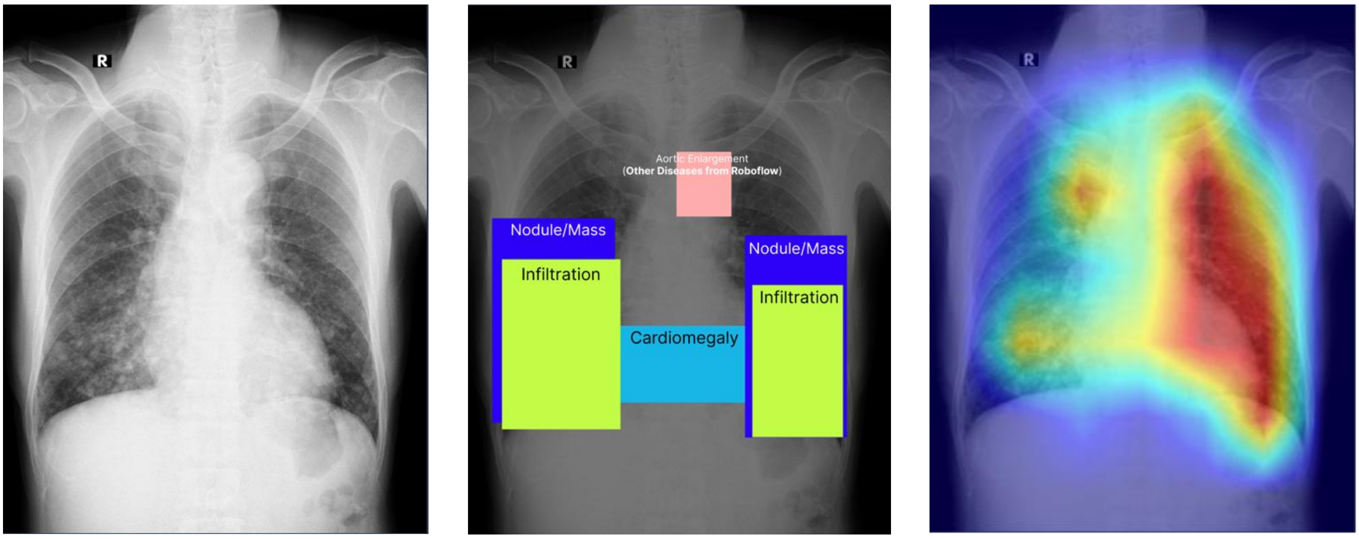


Figure 9. Comprehensive Performance Across Models in Bubble Chart

In summary, EfficientNetB0 offers the highest accuracy, ShuffleNetV2 excels in efficiency, and MobileNetV2 provides a balanced approach. These models are particularly relevant for mobile and edge applications where efficiency and portability are essential.

3.4 Grad-CAM Visualization Analysis

Based on Figure 10 and Table 14, the Grad-CAM visualization of EfficientNetB0 highlights both its strengths and limitations in thoracic abnormality detection. The true labels and ground-truth localization in this analysis are obtained from Roboflow's annotated dataset [67], which provides accurate bounding boxes for abnormalities. This dataset serves as the reference for evaluating the model's ability to correctly identify and localize thoracic diseases.



(a) (b) (c)

Figure 10. Lung Image with Grad-CAM Visualization: (a) Ground truth image, (b) true labels and localization taken from Roboflow [67], (c) Grad-CAM results predicted by EfficientNetB0

Table 14. Percentage of Prediction Result of 1 Image for Grad-CAM Visualization

Label	Actual Label	EfficientNetB0	is match?
pneumothorax	No	0.93 - Yes	No
cardiomegaly	Yes	0.06 - No	No
nodule/mass	Yes	0.01 - No	No
consolidation	No	0.00 - No	No
infiltration	Yes	0.70 - Yes	Yes

However, the model incorrectly predicts pneumothorax (0.93), likely confusing infiltration for a collapsed lung, while correctly identifying infiltration (0.70). Despite the presence of cardiomegaly and nodule/mass in the ground truth annotations, EfficientNetB0 assigns very low probabilities to these conditions (0.06 and 0.01, respectively), indicating poor sensitivity in their detection.

These findings align with previous research but also indicate some distinct challenges. Gakhar et al. (2022) achieved a high AUC for CXR triaging (0.99) and disease classification (0.79) [60]. Their use of GradCAM visualization enhanced interpretability, similar to the current study's approach. However, EfficientNetB0 in this study demonstrates a notable limitation in sensitivity for conditions like cardiomegaly and nodules. Additionally, Sahin et al. (2024) applied EfficientNetB0 with Grad-CAM++ for pneumonia detection, achieving the highest accuracy (95.03%) and F-measure (96.12%), indicating strong performance for specific conditions [61]. In contrast, the present analysis highlights the model's difficulty in detecting less distinct thoracic abnormalities, suggesting that EfficientNetB0's architecture may benefit from enhanced multi-scale feature extraction and tailored training for complex pathologies.

To enhance accuracy, segmentation-based pretraining for cardiomegaly, multi-scale feature extraction for nodule detection, and better-balanced training data are needed. While EfficientNetB0 demonstrates strong localization for lung opacities, it requires further refinement to minimize false positives and improve sensitivity to less distinct pathologies, particularly for cardiomegaly and nodules/masses.

3.5 Implications

The findings from this study have significant implications for healthcare in resource-limited settings, particularly in Indonesia, where there is a critical shortage of radiologists. The demonstrated effectiveness of EfficientNetB0 for detecting thoracic abnormalities, such as pneumothorax and consolidation, highlights the potential of AI-assisted diagnostic tools to reduce diagnostic delays and improve patient outcomes.

Integrating EfficientNetB0 into routine clinical practice could enhance the capacity of healthcare facilities to manage high patient volumes, especially in rural and underserved areas. The model's ability to localize abnormalities using Grad-CAM also supports its potential for aiding radiologists by visually identifying areas of concern, thereby speeding up the diagnostic process.

Furthermore, the relatively efficient performance of mobile-friendly models such as MobileNetV2 and ShuffleNetV2 suggests that lightweight AI solutions can be feasibly implemented in low-resource environments. These models, when integrated into mobile health applications, can support point-of-care diagnosis, which is crucial for early disease detection and timely intervention.

3.6 Limitation and Future Works

Although EfficientNetB0 demonstrated superior performance in detecting thoracic abnormalities, several limitations warrant attention. The model’s sensitivity for subtle conditions like cardiomegaly and nodules/masses remains limited, indicating challenges in distinguishing fine-grained features. Future research should enhance sensitivity through segmentation-based pretraining and multi-scale feature extraction.

The reliance on the ChestX-ray8 dataset may limit generalizability, as it might not fully capture the diversity of clinical cases encountered in real-world settings. Incorporating more diverse and locally sourced data would improve robustness. Additionally, while Grad-CAM provides visual interpretability, its outputs sometimes fail to align with clinical reasoning. Advanced interpretability techniques, such as Grad-CAM++ or attention-based methods, could enhance model transparency [53], [61].

EfficientNetB0's computational demands also pose challenges for deployment on low-power devices. Optimizing the model through techniques like pruning and quantization is essential for mobile implementation. Furthermore, prospective clinical validation is needed to confirm real-world applicability, as current evaluations are based solely on retrospective datasets.

Future work should focus on improving sensitivity, interpretability, mobile optimization, and clinical validation to enhance the model’s practical utility in resource-limited healthcare settings.

4. Conclusion

This study evaluated three mobile deep learning models—MobileNetV2, ShuffleNetV2, and EfficientNetB0—for automated thoracic abnormality detection in chest X-rays, addressing the critical shortage of radiologists in Indonesia. EfficientNetB0 emerged as the top-performing model, achieving a macro-average F1 score of 0.556 and an AUC-ROC of 0.765, outperforming both MobileNetV2 and ShuffleNetV2. This macro-average F1 score represents the model's balanced performance across multiple abnormalities, including pneumothorax and consolidation, where it demonstrated strong localization. These findings highlight EfficientNetB0's potential as a practical AI-assisted diagnostic tool, especially suitable for resource-limited healthcare settings in Indonesia. Despite its strengths, the model faced challenges in detecting cardiomegaly and nodules/masses, indicating the need for enhanced sensitivity through segmentation-based pretraining and multi-scale feature extraction. Future research should focus on improving model interpretability, reducing false positives, and integrating domain-specific fine-tuning to maximize clinical applicability. EfficientNetB0's high accuracy and efficiency make it a promising solution to support healthcare delivery in Indonesia, particularly in areas with limited medical expertise and resources.

References

[1] D. R. Wulan, "Kegiatan Puncak Hari Tuberkulosis Sedunia 2024: Gerakan Indonesia Akhiri Tuberkulosis," TBC Indonesia, 2024.

[2] S. Dayne, "Protecting children from the most-deadly infectious disease in Indonesia | UNICEF Indonesia.", 2024.

[3] Statista, "Indonesia: COPD projection 2017-2024," Statista., 2024.

[4] S. Andarini, E. Syahruddin, N. Aditya, J. Zaini, F. D. Kurniawan, S. Ermayanti, N. N. Soeroso, S. M. Munir, A. Infianto, A. Rima, U. A. Setyawan, L. Wulandari, H. Haryati, I. A. Jasminarti, and A. Santoso, "Indonesian Society of Respirology (ISR) Consensus Statement on Lung Cancer Screening and Early Detection in Indonesia," *J Respirol Indones*, vol. 43, no. 2, pp. 144–150, Apr. 2023. https://doi.org/10.36497/jri.v43i2.455

[5] F. R. Muharram, C. E. C. Z. Multazam, A. Mustofa, W. Socha, Andrianto, S. Martini, L. Aminde, and C. Yi-Li, "The 30 Years of Shifting in The Indonesian Cardiovascular Burden—Analysis of The Global Burden of Disease Study," *J Epidemiol Glob Health*, vol. 14, no. 1, pp. 193–212, Feb. 2024. https://doi.org/10.1007/s44197-024-00187-8

[6] R. Shah Gupta, A. Koteci, A. Morgan, P. M. George, and J. K. Quint, "Incidence and prevalence of interstitial lung diseases worldwide: a systematic literature review," *BMJ Open Resp Res*, vol. 10, no. 1, p. e001291, Jun. 2023. https://doi.org/10.1136/bmjresp-2022-001291

[7] nela, "Indonesia faces shortage of specialized doctors, hindering healthcare services," The Online Citizen, 2025.

[8] R. E. Yunus, "Radiology Loading and Coverage Hours in Indonesia," *Korean J Radiol*, vol. 25, no. 7, pp. 597–599, Jul. 2024. https://doi.org/10.3348/kjr.2024.0267

[9] S. Inui, W. Gonoi, R. Kurokawa, Y. Nakai, Y. Watanabe, K. Sakurai, M. Ishida, A. Fujikawa, and O. Abe, "The role of chest imaging in the diagnosis, management, and monitoring of coronavirus disease 2019 (COVID-19)," *Insights into Imaging*, vol. 12, no. 1, p. 155, Nov. 2021. https://doi.org/10.1186/s13244-021-01096-1

[10] S. H. Bradley, S. Abraham, M. E. Callister, A. Grice, W. T. Hamilton, R. R. Lopez, B. Shinkins, and R. D. Neal, "Sensitivity of chest X-ray for detecting lung cancer in people presenting with symptoms: a systematic review," *Br J Gen Pract*, vol. 69, no. 689, pp. e827–e835, Dec. 2019. https://doi.org/10.3399/bjgp19X706853

[11] T. Jewell, "Chest X-rays in Tuberculosis Diagnosis," Healthline, 2024.

[12] L. A. Eisen, J. S. Berger, A. Hegde, and R. F. Schneider, "Competency in Chest Radiography," *J Gen Intern Med*, vol. 21, no. 5, pp. 460–465, May 2006. https://doi.org/10.1111/j.1525-1497.2006.00427.x

[13] J. Rafferty, "Qualifications for interpreting/classifying chest roentenograms and maintenance of interpretation forms. | Occupational Safety and Health Administration.", 2025.

[14] J. B. Bomanji, N. Gupta, P. Gulati, and C. J. Das, "Imaging in TuberculosisImaging in Tuberculosis," *Cold Spring Harb Perspect Med*, vol. 5, no. 6, p. a017814, Jun. 2015. https://doi.org/10.1101/cshperspect.a017814

[15] Y. J. Jeong and K. S. Lee, "Pulmonary Tuberculosis: Up-to-Date Imaging and Management," *American Journal of Roentgenology*, vol. 191, no. 3, pp. 834–844, Sep. 2008. https://doi.org/10.2214/AJR.07.3896

[16] F. Gaillard, "Consolidation | Radiology Reference Article | Radiopaedia.org," Radiopaedia, 2025.

[17] K. S. Lee, J. Han, M. P. Chung, and Y. J. Jeong, "Consolidation," *Radiology Illustrated: Chest Radiology*, pp. 221–233, Aug. 2013. https://doi.org/10.1007/978-3-642-37096-0_22

[18] S. Pochepnia, E. M. Grabczak, E. Johnson, F. O. Eyuboglu, O. Akkerman, and H. Prosch, "Imaging in pulmonary infections of immunocompetent adult patients," *Breathe*, vol. 20, no. 1, Apr. 2024. https://doi.org/10.1183/20734735.0186-2023

[19] T. Cheng, H. Wan, Q. Cheng, Y. Guo, Y. Qian, L. Fan, Y. Feng, Y. Song, M. Zhou, Q. Li, G. Shi, and S. Huang, "Computed tomography manifestation of acute exacerbation of chronic obstructive pulmonary disease: A pilot study," *Experimental and Therapeutic Medicine*, vol. 11, no. 2, pp. 519–529, Feb. 2016. https://doi.org/10.3892/etm.2015.2930

[20] clevelandclinic, "Pneumothorax (Collapsed Lung)," Cleveland Clinic, 2025.

[21] P. T. King, "Inflammation in chronic obstructive pulmonary disease and its role in cardiovascular disease and lung cancer," *Clin Transl Med*, vol. 4, p. 26, Jul. 2015. https://doi.org/10.1186/s40169-015-0068-z

[22] C. L. McKnight and B. Burns, "Pneumothorax," in *StatPearls*, Treasure Island (FL): StatPearls Publishing, 2025.

[23] B. A. Rangelov, A. L. Young, J. Jacob, A. P. Cahn, S. Lee, F. J. Wilson, D. J. Hawkes, and J. R. Hurst, "Thoracic Imaging at Exacerbation of Chronic Obstructive Pulmonary Disease: A Systematic Review," *Int J Chron Obstruct Pulmon Dis*, vol. 15, pp. 1751–1787, Jul. 2020. https://doi.org/10.2147/COPD.S250746

[24] cancer.org, "Lung Cancer Early Detection, Diagnosis, and Staging."

[25] utswmed, "Lung Nodules | Condition | UT Southwestern Medical Center.", 2025.

[26] H. Amin and W. J. Siddiqui, "Cardiomegaly," in *StatPearls*, Treasure Island (FL): StatPearls Publishing, 2025.

[27] E. K. K. Brakohiapa, B. O. Botwe, B. D. Sarkodie, E. K. Ofori, and J. Coleman, "Radiographic determination of cardiomegaly using cardiothoracic ratio and transverse cardiac diameter: can one size fit all? Part one," *Pan Afr Med J*, vol. 27, p. 201, Jul. 2017. https://doi.org/10.11604/pamj.2017.27.201.12017

[28] clevelandclinic, "Enlarged Heart (Cardiomegaly): What It Is, Symptoms & Treatment," Cleveland Clinic, 2025.

[29] I. Cundrle, L. J. Olson, and B. D. Johnson, “Pulmonary Limitations in Heart Failure,” *Clin Chest Med*, vol. 40, no. 2, pp. 439–448, Jun. 2019. https://doi.org/10.1016/j.ccm.2019.02.010

[30] T. S. Metkus, “Pulmonary edema: MedlinePlus Medical Encyclopedia.”, 2025.

[31] L. Ridley, “Chest Radiograph Signs Suggestive of Pericardial Disease,” American College of Cardiology, 2025.

[32] J. G. Nam, M. Kim, J. Park, E. J. Hwang, J. H. Lee, J. H. Hong, J. M. Goo, and C. M. Park, “Development and validation of a deep learning algorithm detecting 10 common abnormalities on chest radiographs,” *Eur Respir J*, vol. 57, no. 5, p. 2003061, May 2021. https://doi.org/10.1183/13993003.03061-2020

[33] P. G. Anderson, H. Tarder-Stoll, M. Alpaslan, N. Keathley, D. L. Levin, S. Venkatesh, E. Bartel, S. Sicular, S. Howell, R. V. Lindsey, and R. M. Jones, “Deep learning improves physician accuracy in the comprehensive detection of abnormalities on chest X-rays,” *Sci Rep*, vol. 14, no. 1, p. 25151, Oct. 2024. https://doi.org/10.1038/s41598-024-76608-2

[34] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, “ChestX-ray8: Hospital-scale Chest X-ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases,” in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul. 2017, pp. 3462–3471. https://doi.org/10.1109/CVPR.2017.369

[35] L. Guo, C. Zhou, J. Xu, C. Huang, Y. Yu, and G. Lu, “Deep Learning for Chest X-ray Diagnosis: Competition Between Radiologists with or Without Artificial Intelligence Assistance,” *J Digit Imaging. Inform. med.*, vol. 37, no. 3, pp. 922–934, Feb. 2024. https://doi.org/10.1007/s10278-024-00990-6

[36] D. Kvak, A. Chromcová, M. Biroš, R. Hrubý, K. Kvaková, M. Pajdaković, and P. Ovesná, “Chest X-ray Abnormality Detection by Using Artificial Intelligence: A Single-Site Retrospective Study of Deep Learning Model Performance,” *BioMedInformatics*, vol. 3, no. 1, Art. no. 1, Mar. 2023. https://doi.org/10.3390/biomedinformatics3010006

[37] E. Blantz, “4 Key Challenges and Solutions to ICT Deployments for Rural Healthcare,” ICTworks, 2024.

[38] PricewaterhouseCoopers, “How can technology accelerate the digitisation of the Indonesian healthcare sector?,” PwC, 2024.

[39] S. Sarkar, “Overcoming Hospital Challenges with Enhanced Technology Solutions in Indonesia,” 2021.

[40] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, “MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications,” Apr. 16, 2017, ***arXiv***: arXiv:1704.04861, 2023. https://doi.org/10.48550/arXiv.1704.04861

[41] M. Tan and Q. V. Le, “EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks,” Sep. 11, 2020, ***arXiv***: arXiv:1905.11946. https://doi.org/10.48550/arXiv.1905.11946

[42] X. Zhang, X. Zhou, M. Lin, and J. Sun, “ShuffleNet: An Extremely Efficient Convolutional Neural Network for Mobile Devices,” Dec. 07, 2017, ***arXiv***: arXiv:1707.01083, 2024. https://doi.org/10.48550/arXiv.1707.01083

[43] K. G. van Leeuwen, M. de Rooij, S. Schalekamp, B. van Ginneken, and M. J. C. M. Rutten, “How does artificial intelligence in radiology improve efficiency and health outcomes?,” *Pediatr Radiol*, vol. 52, no. 11, pp. 2087–2093, Oct. 2022. https://doi.org/10.1007/s00247-021-05114-8

[44] M. Rybczak and K. Kozakiewicz, “Deep Machine Learning of MobileNet, Efficient, and Inception Models,” *Algorithms*, vol. 17, no. 3, Art. no. 3, Mar. 2024. https://doi.org/10.3390/a17030096

[45] S. Kolonne, C. Fernando, H. Kumarasinghe, and D. Meedeniya, “MobileNetV2 Based Chest X-Rays Classification,” in *2021 International Conference on Decision Aid Sciences and Application (DASA)*, Sakheer, Bahrain: IEEE, Dec. 2021, pp. 57–61. https://doi.org/10.1109/DASA53625.2021.9682248

[46] S. Akter, F. M. J. M. Shamrat, S. Chakraborty, A. Karim, and S. Azam, “COVID-19 Detection Using Deep Learning Algorithm on Chest X-ray Images,” *Biology (Basel)*, vol. 10, no. 11, p. 1174, Nov. 2021. https://doi.org/10.3390/biology10111174

[47] S. Velu, “An efficient, lightweight MobileNetV2-based fine-tuned model for COVID-19 detection using chest X-ray images,” *MBE*, vol. 20, no. 5, Art. no. mbe-20-05-368, 2023. https://doi.org/10.3934/mbe.2023368

[48] H. Iqbal, A. Khan, N. Nepal, F. Khan, and Y.-K. Moon, “Deep Learning Approaches for Chest Radiograph Interpretation: A Systematic Review,” *Electronics*, vol. 13, no. 23, Art. no. 23, Jan. 2024. https://doi.org/10.3390/electronics13234688

[49] C. Gu and M. Lee, “Deep Transfer Learning Using Real-World Image Features for Medical Image Classification, with a Case Study on Pneumonia X-ray Images,” *Bioengineering*, vol. 11, no. 4, p. 406, Apr. 2024. https://doi.org/10.3390/bioengineering11040406

[50] L. An, K. Peng, X. Yang, P. Huang, Y. Luo, P. Feng, and B. Wei, “E-TBNet: Light Deep Neural Network for Automatic Detection of Tuberculosis with X-ray DR Imaging,” *Sensors*, vol. 22, no. 3, p. 821, Jan. 2022. https://doi.org/10.3390/s22030821

[51] H. Mzoughi, I. Njeh, M. B. Slima, and A. BenHamida, “Deep efficient-nets with transfer learning assisted detection of COVID-19 using chest X-ray radiology imaging,” *Multimed Tools Appl*, vol. 82, no. 25, pp. 39303–39325, Oct. 2023. https://doi.org/10.1007/s11042-023-15097-3

[52] K. Kansal, T. B. Chandra, and A. Singh, “ResNet-50 vs. EfficientNet-B0: Multi-Centric Classification of Various Lung Abnormalities Using Deep Learning,” *Procedia Computer Science*, vol. 235, pp. 70–80, 2024. https://doi.org/10.1016/j.procs.2024.04.007

[53] Q. An, W. Chen, and W. Shao, “A Deep Convolutional Neural Network for Pneumonia Detection in X-ray Images with Attention Ensemble,” *Diagnostics (Basel)*, vol. 14, no. 4, p. 390, Feb. 2024. https://doi.org/10.3390/diagnostics14040390

[54] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, “Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization,” *Int J Comput Vis*, vol. 128, no. 2, pp. 336–359, Feb. 2020. https://doi.org/10.1007/s11263-019-01228-7

[55] S. Suara, A. Jha, P. Sinha, and A. A. Sekh, “Is Grad-CAM Explainable in Medical Images?,” vol. 2009, 2024, pp. 124–135. https://doi.org/10.1007/978-3-031-58181-6_11

[56] N. Le, B. Männel, M. Jarema, T. T. Luong, L. K. Bui, H. Q. Vy, and H. Schuh, “K-Fold Cross-Validation: An Effective Hyperparameter Tuning Technique in Machine Learning on GNSS Time Series for Movement Forecast,” in *Recent Research on Geotechnical Engineering, Remote Sensing, Geophysics and Earthquake Seismology*, A. Çiner, Z. A. Ergüler, M. Bezzeghoud, M. Ustuner, M. Eshagh, H. El-Askary, A. Biswas, L. Gasperini, K.-G. Hinzen, M. Karakus, C. Comina, A. Karrech, A. Polonia, and H. I. Chaminé, Eds., in Advances in Science, Technology & Innovation. , Cham: Springer Nature Switzerland, 2024, pp. 377–382. https://doi.org/10.1007/978-3-031-43218-7_88

[57] P. Refaeilzadeh, L. Tang, and H. Liu, “Cross-Validation,” in *Encyclopedia of Database Systems*, L. Liu and M. T. Özsu, Eds., Boston, MA: Springer US, 2009, pp. 532–538. https://doi.org/10.1007/978-0-387-39940-9_565

[58] G. Naidu, T. Zuva, and E. M. Sibanda, “A Review of Evaluation Metrics in Machine Learning Algorithms,” in *Artificial Intelligence Application in Networks and Systems*, R. Silhavy and P. Silhavy, Eds., Cham: Springer International Publishing, 2023, pp. 15–25. https://doi.org/10.1007/978-3-031-35314-7_2

[59] F. Melo, “Area under the ROC Curve,” in *Encyclopedia of Systems Biology*, W. Dubitzky, O. Wolkenhauer, K.-H. Cho, and H. Yokota, Eds., New York, NY: Springer, 2013, pp. 38–39. https://www.doi.org/10.1007/978-1-4419-9863-7_209

[60] M. Gakhar and A. Aggarwal, “ThoraciNet: thoracic abnormality detection and disease classification using fusion DCNNs,” *Phys Eng Sci Med*, vol. 45, no. 3, pp. 961–970, Sep. 2022. https://doi.org/10.1007/s13246-022-01137-z

[61] E. Sahin, A. Celikten, S. Demirel, A. Akpulat, K. Budak, and H. Karatas, “Pneumonia Detection in Chest X - Ray Images Using Deep Learning and Grad-CAM++ Visualization,” in *2024 Innovations in Intelligent Systems and Applications Conference (ASYU)*, Ankara, Turkiye: IEEE, Oct. 2024, pp. 1–5. https://doi.org/10.1109/ASYU62119.2024.10756959

[62] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, “MobileNetV2: Inverted Residuals and Linear Bottlenecks,” in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT: IEEE, Jun. 2018, pp. 4510–4520. https://doi.org/10.1109/CVPR.2018.00474

[63] M. Päpper, “Depthwise Separable Convolutions in PyTorch :: Päpper’s Machine Learning Blog — This blog features state of the art applications in machine learning with a lot of PyTorch samples and deep learning code. You will learn about neural network optimization and potential insights for artificial intelligence for example in the medical domain.,” Depthwise Separable Convolutions in PyTorch, 2024.

[64] N. Ma, X. Zhang, H.-T. Zheng, and J. Sun, “ShuffleNet V2: Practical Guidelines for Efficient CNN Architecture Design,” Jul. 30, 2018, *arXiv*: arXiv:1807.11164. https://doi.org/10.48550/arXiv.1807.11164

[65] X. Zhang, X. Zhou, M. Lin, and J. Sun, “ShuffleNet: An Extremely Efficient Convolutional Neural Network for Mobile Devices,” Dec. 07, 2017, *arXiv*: arXiv:1707.01083. https://doi.org/10.48550/arXiv.1707.01083

[66] P. Rajpurkar, J. Irvin, K. Zhu, B. Yang, H. Mehta, T. Duan, D. Ding, A. Bagul, C. Langlotz, K. Shpanskaya, M. P. Lungren, and A. Y. Ng, “CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning,” Dec. 25, 2017, *arXiv*: arXiv:1711.05225. https://doi.org/10.48550/arXiv.1711.05225

[67] Roboflow, “Chest abnormalities > Browse,” Roboflow.