Externally Validated Deep Learning Model for Multi-Disease Classification of Chest X-Rays

Weny Indah Kusumawati[1], Zendi Zakaria Raga Permana[2], Ira Puspasari[1*]

[1,3]*Computer Engineering Department, Universitas Dinamika*
Kedung Baruk No. 98 Surabaya, 60298, Indonesia

[2]*School of Electrical Engineering and Informatics, Institut Teknologi Bandung*
Ganesa No.10 Bandung, 40132, Indonesia

[*]*Corresponding author. Email: ira@dinamika.ac.id*

Abstract— Accurate classification of chest X-ray (CXR) images is vital for early detection of thoracic diseases such as COVID-19, Tuberculosis, and Pneumonia, particularly in regions with limited radiological expertise. While deep learning has shown promise in CXR interpretation, many existing models rely solely on internal datasets, risking overfitting and poor generalizability. Furthermore, inadequate tuning of network architectures may limit robustness across varied imaging conditions. This study presents an externally validated deep learning framework based on Convolutional Neural Networks (CNNs) for multi-disease CXR classification. This study compared a baseline CNN with two convolutional layers against a tuned architecture with three layers across multiple image resolutions (64×64, 112×112, 224×224). The proposed model employs transfer learning with a pre-trained CNN, fine-tuned for four-class classification using a softmax output layer. Training was performed with the Adam optimizer (learning rate: 0.0001, batch size: 32) and categorical cross-entropy loss, for up to 50 epochs with early stopping. Internal validation showed the tuned model outperformed the baseline, achieving 0.97 accuracy and an F1-score of 0.89. External validation confirmed superior generalizability, with the tuned model attaining an F1-score of 0.83 and an AUC of 0.97 at 112×112 resolution, compared to the baseline's F1-score of 0.79 and AUC of 0.94. These results highlight the potential of optimized CNN architectures as reliable, scalable tools for radiological decision support in resource-limited healthcare systems. Future work will incorporate explainable AI methods and real-world clinical validation to ensure safe, interpretable deployment.

Keywords— chest X-ray; deep learning; external validation; medical imaging; multi-disease classification

I. INTRODUCTION

The global COVID-19 pandemic has further strained healthcare systems and highlighted the urgent need for automated, rapid, and accurate diagnostic tools [1]. Tuberculosis in developing countries such as Indonesia, and pneumonia, which disproportionately affects both pediatric and elderly populations, continue to pose significant global public health challenges [2]. According to the Indonesian Ministry of Health, Indonesia ranks among the highest-burden countries for tuberculosis, with an estimated 969,000 new cases in 2022 [3], while pneumonia remains a leading cause of morbidity and mortality among children under five and older adults, especially in rural and underserved regions [4].

Given these circumstances, artificial intelligence (AI), particularly those based on deep learning (DL), has gained significant attention for its potential to enhance medical image interpretation. Medical image processing, when integrated with AI, enables the automatic extraction of relevant features from chest X-rays and CT scans, facilitating the classification of pulmonary diseases. Convolutional Neural Networks (CNNs) and transfer learning architectures such as VGG, ResNet, and DenseNet have demonstrated notable success in differentiating between normal and abnormal lung patterns [5]. A previous study proposes a robust, multi-step AI pipeline for TB screening. They begin by applying sophisticated segmentation networks to isolate lung regions, then use various CNN architectures to classify segmented CXRs, achieving a top accuracy of 99.1% [6]. These models can learn hierarchical representations of lung pathology, effectively distinguishing between viral infections like COVID-19, bacterial pneumonia, and mycobacterial infections such as TB.

Previous studies have explored binary or multi-class classification tasks focusing on subsets of these diseases [7], [8], [9]. Despite significant advances in medical image analysis, comprehensive deep learning approaches capable of simultaneously classifying normal lungs and accurately distinguishing among COVID-19, tuberculosis, and pneumonia remain scarce. Most existing models are disease-specific or limited to binary or ternary classifications, which restricts their clinical utility in real-world settings [8], [10]. A unified, multi-disease classification framework has the potential to substantially enhance diagnostic accuracy, streamline triage processes, and alleviate the burden on radiology services, particularly in low-resource settings or during periods of heightened demand, such as during pandemics [11].

Many models focus on individual diseases: e.g., ResNet-based systems for tuberculosis vs healthy, or VGG16 and DenseNet architectures for pneumonia vs normal, often evaluated in isolation and without multi-disease comparison [12], [13]. Although advanced deep learning architectures involving multiple stages have been proposed to enable thorough lung disease classification from chest radiographs, this approach presents several notable limitations. The use of

Received 9 July 2025, Revised 27 August 2025, Accepted 28 August 2025.
DOI: https://doi.org/10.15294/jte.v17i2.29892

sequential classification stages increases model complexity and significantly extends inference time, making it less practical for real-time or point-of-care deployment. The pipeline also demands extensive hyperparameter tuning for each stage, resulting in longer development cycles and reduced reproducibility [14].

Despite their potential, the use of deep features for detecting pulmonary abnormalities such as COVID-19, pneumonia, and tuberculosis from chest X-rays faces several limitations. Many datasets suffer from class imbalance, with significantly fewer COVID-19 or tuberculosis cases compared to normal and pneumonia, which may bias the model toward majority classes. In addition, most publicly available datasets are derived from limited sources and lack diversity in imaging conditions, patient demographics, and disease severity, thereby reducing the generalizability of the findings [15]. The extraction of a large number of deep features, such as those from VGG or AlexNet, can also introduce redundancy and limit interpretability, making it difficult to translate model decisions into clinical insights. Moreover, overlapping radiographic manifestations among these diseases often lead to clinically significant misclassifications, such as COVID-19 being classified as normal or tuberculosis as pneumonia. Another concern is the frequent reliance on internal validation and overall accuracy, without sufficient evaluation of sensitivity, specificity, and external validation, which are essential for assessing clinical applicability.

The joint diagnosis of pneumonia, COVID-19, and tuberculosis from chest X-ray images using deep learning faces several limitations [16]. First, the model was trained and evaluated using augmented versions of publicly available Kaggle datasets without independent external validation, raising concerns about overfitting and dataset bias. Second, relying solely on augmentation for variability may not adequately account for real-world differences in imaging conditions, equipment, settings, and patient demographics, limiting generalizability. Third, although the model reported very high accuracy (98.72%) and recall rates (99.66% for pneumonia, 98.10% for tuberculosis, and 96.27% for COVID-19), these results may be overoptimistic without rigorous testing on independent datasets. Finally, the exclusive dependence on homogeneous public datasets with limited clinical diversity in terms of populations, imaging devices, and geographic origins further undermines the model's applicability to broader and more diverse clinical settings.

A significant research gap remains in the development of deep learning models capable of simultaneously classifying normal, COVID-19, tuberculosis, and pneumonia from chest X-rays. Most existing models focus on one or two diseases and are limited to binary or ternary classification. Multi-class models often suffer from reduced accuracy and poor generalization as class complexity increases. Moreover, external validation is rarely performed, limiting real-world applicability; this highlights the need for a robust, generalizable, and externally validated deep learning framework that can accurately classify multiple respiratory diseases in diverse clinical settings. This study contributes to the field by proposing a unified deep learning framework designed to address current limitations in chest X-ray classification.

Specifically, the model is capable of simultaneously distinguishing among normal, COVID-19, tuberculosis, and pneumonia cases within a four-class classification setting, offering a more comprehensive diagnostic tool than existing binary or ternary models. To ensure robustness and generalizability, the framework is validated using external datasets and imaging conditions.

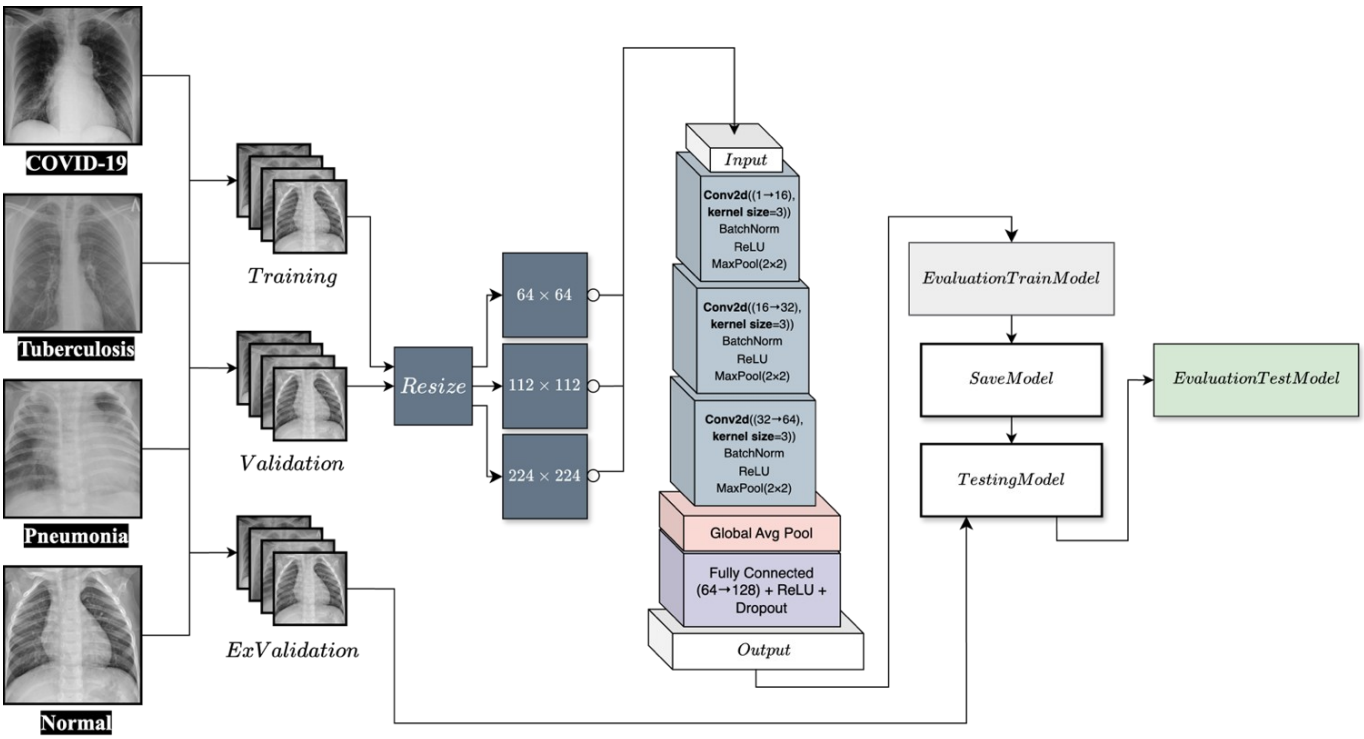

Figure 1. Research design of externally validated deep learning model for multi-disease classification of chest x-rays

TABLE I. DISTRIBUTION FOR TRAINING, TESTING, AND VALIDATION

Class	Training	Validation	Testing
COVID-19	460	10	106
Normal	1341	8	234
Pneumonia	3875	8	390
Tuberculosis	650	12	41

II. METHOD

The research workflow is depicted in Figure 1. The process begins with data collection from publicly available databases, followed by preprocessing to ensure consistency and quality of the input images. In this study, preprocessing consisted solely of image resizing to three target resolutions—64×64, 112×112, and 224×224 pixels—chosen to match the input requirements of the CNN architectures evaluated while preserving the original image characteristics. No cropping, rotation, contrast enhancement, or other image manipulation techniques were performed. Next, a CNN architecture specifically tailored for multi-disease classification is developed. The prepared dataset is then used to train the model, optimizing its parameters for accurate classification; this is followed by a comprehensive evaluation phase using standard performance metrics to quantify accuracy, precision, recall, and F1-score. Finally, external validation is performed on an independent dataset to rigorously assess the model's generalizability and robustness across unseen cases, thereby confirming its potential for deployment in real-world clinical scenarios.

A. Dataset Collection

The dataset used in this study was compiled from three publicly available sources [17],[18], [19]. Specifically, Dataset A contributed 5674 chest X-ray images, consisting of 326 COVID-19, 4273 pneumonia, 391 tuberculosis, and 684 normal cases [17]. Dataset B provided 762 images, with detailed class distribution as follows: 312 tuberculosis and 450 normal cases [18]. Finally, Dataset C contributed 699 images, including 250 COVID-19 and 449 normal cases [19]. Together, these sources ensured a balanced and diverse dataset for training, validation, and testing. The choice of public data was driven by the need for standardized, well-labeled images to enable reproducible model development and benchmarking. The focus on Indonesia is maintained by framing the study's relevance to local healthcare challenges—particularly the shortage of radiology specialists and the potential for AI-based tools to support diagnosis in resource-limited settings. A total of 7,135 images were utilized in this study, as detailed in Table I. Each image was adjusted to a fixed resolution of 64x64, 112x112, and 224×224 pixels, and normalized to standardize pixel intensity distributions.

B. Model Architecture

The proposed model leverages a transfer learning strategy built upon a pre-trained CNN architecture, a widely adopted approach in biomedical imaging [15], [20], [21], [22], [23], The network was fine-tuned on the training dataset using convolutional layers with a kernel size of 3, and a softmax activation function in the output layer to predict probabilities across the four target classes. To enhance generalization and mitigate overfitting, dropout and batch normalization layers were incorporated. Two architectures were evaluated: a baseline CNN comprising two convolutional layers and a tuned CNN extended to three convolutional layers. The input image resolutions explored were 64×64, 112×112, and 224×224 pixels. In the baseline model, the first convolutional layer utilized 16 filters (160 parameters), followed by a second layer with 32 filters (4,640 parameters). The models were trained using the Adam optimizer (initial learning rate = 0.0001, batch size = 32) with categorical cross-entropy loss. Training was conducted over 50 epochs with early stopping based on validation loss, ensuring convergence while preventing overfitting. This rigorous design and optimization provide a robust foundation for accurate multi-disease chest X-ray classification, balancing model complexity with generalization capability.

C. Evaluation Metrics

Classification performance was evaluated using several standard metrics—accuracy, precision, recall (sensitivity), F1-score, and the area under the receiver operating characteristic curve (AUC-ROC)—computed individually for each class. Confusion matrices were also generated to analyze misclassifications. Results were reported separately for internal validation and external testing to highlight generalizability. In classification analysis, True Positives (TP) represent instances where the model accurately identifies positive cases, while True Negatives (TN) denote correctly classified negative cases. Conversely, False Positives (FP) occur when the model incorrectly labels a negative instance as positive, and False Negatives (FN) arise when a positive instance is mistakenly classified as negative. These four metrics form the foundation for evaluating a model's performance, serving as the basis for key performance indicators such as accuracy, precision, recall, and F1-score, which are critical in assessing predictive reliability and generalizability across diverse datasets.

Accuracy (*Acc*) (1), measures the proportion of correctly classified samples among all predictions.

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

Precision (*Prec*) (2), evaluates the proportion of positive predictions that are correct.

$$Prec = \frac{TP}{TP + FP} \quad (2)$$

Recall (*Sens*) (3), measures the model's ability to correctly identify all positive instances.

$$Sens = \frac{TP}{TP + FN} \quad (3)$$

F1-Score (4) is the harmonic mean of precision and recall, providing a balanced metric in the presence of class imbalance.

$$F1 - Score = 2.\frac{Prec.x\ Sens}{Prec\ + Sens} \quad (4)$$

The AUC represents the degree of separability, where a value closer to 1.0 indicates better performance.

D. External Validation

To assess robustness, the trained model was tested on external datasets not seen during training or validation. Performance was evaluated using the same metrics to simulate deployment in real-world clinical scenarios. This step was essential to demonstrate the model's ability to generalize across diverse populations, image qualities, and disease distributions.

III. RESULTS AND DISCUSSION

To evaluate the effectiveness and generalizability of the proposed deep learning model for multi-disease classification of chest X-rays, conducted comprehensive internal training and external validation across four clinically significant categories: COVID-19, Tuberculosis, Pneumonia, and Normal. The model's performance was assessed using a range of metrics,

revealing its potential to assist in rapid and accurate diagnosis across diverse imaging datasets and clinical settings.

A. Classification Performance on Internal Dataset

The proposed deep learning model demonstrated robust classification performance on the internal dataset, effectively distinguishing between COVID-19, Tuberculosis, Pneumonia, and Normal cases, as reflected by consistently high accuracy, sensitivity, specificity, and F1-scores across all classes. The baseline CNN model with two convolutional layers was trained for 50 epochs using input image dimensions of 64 × 64, 112 × 112, and 224 × 224 pixels. Table II shows that across all configurations, the model consistently achieved a high training accuracy of 0.99, indicating effective feature learning. However, validation performance varied with input resolution. At 64 × 64, the validation accuracy reached only 0.78, suggesting limited generalization at lower resolutions. In contrast, 112 × 112 yielded the highest validation accuracy of 0.97, followed by 0.94 for 224 × 224. Notably, the lowest training loss of 0.68 was observed with the 224 × 224 input, highlighting its advantage in minimizing classification errors during training despite a slight drop in validation performance compared to 112 × 112.

In this study, the CNN model was further optimized by tuning its architecture to include three convolutional layers, aiming to enhance classification performance across different image resolutions [24]. The tuned model demonstrated notable improvements at each resolution. For a 64 × 64 input, the best performance was achieved at epoch 11, with a training accuracy of 0.97 and a validation accuracy of 0.84. For a 112 × 112, the optimal result occurred at epoch 15, yielding a training accuracy of 0.98 and validation accuracy of 0.89. At 224 × 224, the model reached its peak performance at epoch 14, with a training accuracy of 0.99 and an impressive validation accuracy of 0.97. Compared to the untuned model, these results represent an improvement of up to 3% in validation accuracy, underscoring the effectiveness of architectural tuning in enhancing generalization and predictive performance, as shown in Table III.

Table IV presents a comparison of training times between the baseline CNN and the tuned CNN models across varying image resolutions. The fastest training time was recorded for the tuned CNN using 64 × 64 images at 11 epochs, completing in 572.54 seconds. In contrast, the longest training time occurred with 224 × 224 images over 50 epochs, taking 5074.51 seconds. Notably, the optimal model in terms of both training and validation accuracy was achieved with 224 × 224 images at the 14th epoch, requiring 1332.13 seconds of training time, demonstrating a favorable balance between performance and computational efficiency.

TABLE II. THE RESULT OF TESTING THE 2 LAYER CNN MODEL ON THE INTERNAL DATASET

Dimensions (pixels)	Epoch	Accuracy		Loss
		Train	Validation	Train
64 x 64	50	0.99	0.78	0.94
112 x 112	50	0.99	0.97	1.30
224 x 224	50	**0.99**	**0.94**	**0.68**

TABLE III. THE RESULT OF TESTING THE TUNED CNN MODEL ON THE INTERNAL DATASET

Epoch	Accuracy		Loss Train
	Train	Validation	
11 "best epoch"	0.97	0.84	13.91
15 "best epoch"	0.98	0.89	7.37
14 "best epoch"	**0.99**	**0.97**	**3.22**

A comparative analysis of the Receiver Operating Characteristic (ROC) curves reveals a clear performance advantage of the tuned CNN model over the baseline CNN. The tuned model, which incorporates an additional convolutional layer and optimized training parameters, consistently achieved higher Area Under the Curve (AUC) scores across all image resolutions, indicating improved discriminative capability and reduced classification errors. In contrast, the untuned CNN showed lower ROC performance, particularly at lower image resolutions, suggesting limited generalization. These findings demonstrate the effectiveness of architectural tuning in enhancing the model's sensitivity and specificity, thereby improving overall diagnostic reliability, as shown in Figure 2 for the CNN model and Figure 3 for the tuned CNN model.

B. External Validation Performance

To assess the generalizability and real-world applicability of the proposed deep learning model, external validation was conducted using an independent chest X-ray dataset not seen during training. This evaluation serves as a critical benchmark for determining the model's robustness across diverse imaging sources and patient populations. The external validation results provide insight into the model's capacity to maintain high diagnostic performance when applied beyond the development dataset, thereby demonstrating its potential for clinical deployment in multi-disease classification tasks involving COVID-19, Tuberculosis, Pneumonia, and Normal cases.

Table V summarizes the external validation outcomes of the tuned CNN model. At an image resolution of 64 × 64, the model reached an accuracy of 0.89, an F1-score of 0.63, and an AUC of 0.94. When the resolution was increased to 112 × 112, the model delivered its best overall performance, achieving an accuracy of 0.90. At 224 × 224 resolution, the accuracy was 0.88.

Table VI further confirms these findings by presenting additional external validation results. At 64 × 64 resolution, the tuned CNN obtained an accuracy of 0.87, an F1-score of 0.69, and an AUC of 0.95. The 112 × 112 resolution yielded the most favorable outcomes, with an accuracy of 0.90, an F1-score of 0.83, and an AUC of 0.97—representing the highest performance across all tested configurations. At 224 × 224 resolution, the model attained an accuracy of 0.88, an F1-score of 0.77, and an AUC of 0.96. These results underscore the superior generalization and discriminative ability of the tuned CNN, particularly at the 112 × 112 resolution, highlighting its robustness for clinical application [25], [26].

A clear and systematic comparison between the baseline CNN and the tuned CNN models is now presented across the three evaluated input resolutions (64×64, 112×112, and 224×224 pixels) as shown in Figure 4 and Figure 5. The baseline CNN achieved an accuracy of 0.90, sensitivity of 0.81, specificity of 0.91, precision of 0.83, F1-score of 0.79, and an AUC of 0.94. In contrast, the tuned CNN model demonstrated superior performance with an accuracy of 0.90, sensitivity of 0.84, specificity of 0.92, precision of 0.85, F1-score of 0.83, and an AUC of 0.97.

While overall accuracy remained comparable, the tuned CNN exhibited notable gains in sensitivity, precision, F1-score, and especially AUC, reflecting a more balanced and clinically reliable classification performance. These improvements underscore not only the effectiveness of architectural optimization but also the robustness of the tuned model across multiple evaluation dimensions, thereby reinforcing the study's methodological and translational contributions.

TABLE IV. THE COMPARISON OF TRAINING TIME FOR CNN AND CNN TUNED MODELS

Models	Dimensions	Epoch	Time Training (s)
CNN	64 x 64	50	2595.26
	112 x 112	50	2777.02
	224 x 224	50	5074.51
CNN-Tuning	64 x 64	11	572.54
	112 x 112	15	834.30
	224 x 224	14	1332.13

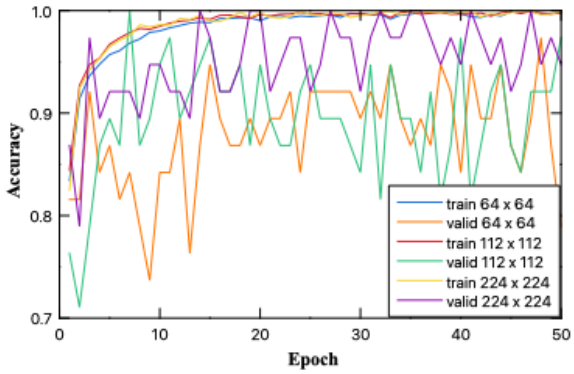

Figure 2. The ROC curve of the CNN model

The confusion matrix for the baseline CNN model under external data validation is presented in Figure 6. A single COVID-19 case was classified as Normal, reflecting an area where the model could be improved to ensure greater robustness in clinical application. Additionally, five Pneumonia cases were misclassified as Normal, potentially undermining clinical decision-making. These misclassifications highlight the limitations of the untuned model in accurately distinguishing between pathologies with overlapping radiographic features, underscoring the need for further optimization to enhance diagnostic reliability.

The confusion matrix for the tuned CNN model under external data validation is illustrated in Figure 7. Despite overall improved performance, the matrix reveals critical misclassification trends. Specifically, two COVID-19 cases were incorrectly classified as Normal, which could pose serious risks in real-world clinical settings where early detection is essential. Moreover, seven Pneumonia cases were also misclassified as Normal, indicating residual challenges in differentiating between subtle radiographic features of disease and healthy lung appearances. These findings emphasize that, while tuning improves overall metrics, further refinement is necessary to reduce false negatives in high-risk categories.

TABLE V. THE RESULT OF TESTING THE CNN MODEL ON THE EXTERNAL DATASET

Dimension	Class	Testing					
		Accuracy	Sensitivity	Specificity	F1-Score	Precision	AUC
64 x 64	Covid-19	0.96	0.89	0.97	0.86	0.87	0.99
	Normal	0.80	0.38	0.98	0.93	0.54	0.90
	Pneumonia	0.82	0.98	0.65	0.74	0.84	0.89
	TB	0.98	0.95	0.98	0.79	0.86	0.99
	Average	0.89	0.80	0.90	0.83	0.78	0.94
112 x 112	Covid-19	0.97	0.87	0.98	0.93	0.90	0.99
	Normal	0.82	0.44	0.98	0.94	0.60	0.88
	Pneumonia	0.84	0.98	0.69	0.76	0.86	0.91
	TB	0.97	0.95	0.97	0.67	0.78	0.99
	Average	**0.90**	**0.81**	**0.91**	**0.83**	**0.79**	**0.94**
224 x 224	Covid-19	0.97	0.84	0.99	0.98	0.91	0.99
	Normal	0.79	0.32	0.99	0.96	0.48	0.86
	Pneumonia	0.80	0.99	0.60	0.72	0.83	0.89
	TB	0.96	0.97	0.96	0.63	0.76	0.99
	Average	0.88	0.78	0.89	0.82	0.75	0.93

TABLE VI. THE RESULT OF TESTING THE TUNED CNN MODEL ON THE EXTERNAL DATASET

Dimension	Class	Testing					
		Accuracy	Sensitivity	Specificity	F1-Score	Precision	AUC
64 x 64	Covid-19	0.95	0.73	0.99	0.82	0.95	0.99
	Normal	0.78	0.32	0.98	0.47	0.89	0.89
	Pneumonia	0.81	0.98	0.64	0.84	0.74	0.94
	TB	0.93	0.95	0.93	0.61	0.45	0.98
	Average	0.87	0.75	0.89	0.69	0.76	0.95
112 x 112	Covid-19	0.97	0.89	0.98	0.90	0.90	0.99
	Normal	0.85	0.55	0.98	0.70	0.93	0.93
	Pneumonia	0.86	0.97	0.75	0.88	0.8	0.95
	TB	0.93	0.95	0.98	0.84	0.76	0.99
	Average	**0.90**	**0.84**	**0.92**	**0.83**	**0.85**	**0.97**
224 x 224	Covid-19	0.97	0.85	0.99	0.91	0.98	0.99
	Normal	0.79	0.33	0.99	0.49	0.96	0.91
	Pneumonia	0.79	0.99	0.58	0.83	0.71	0.95
	TB	0.98	0.97	0.98	0.86	0.76	0.99
	Average	0.88	0.79	0.89	0.77	0.85	0.96

C. Comparative Analysis

A direct comparison between the tuned and untuned CNN models highlights the impact of architectural optimization on classification performance. While both models demonstrated comparable accuracy during external validation—ranging from 0.87 to 0.90—the tuned CNN consistently outperformed the baseline in terms of F1-score and AUC, particularly at the 112 × 112 resolution, where it achieved an F1-score of 0.83 and an AUC of 0.97. When compared to prior works, the advantages of the tuned CNN become clearer. For example, previous studies utilizing AlexNet, VGG-16, and VGG-19 for pneumonia classification reported an accuracy of 94.1% [27].

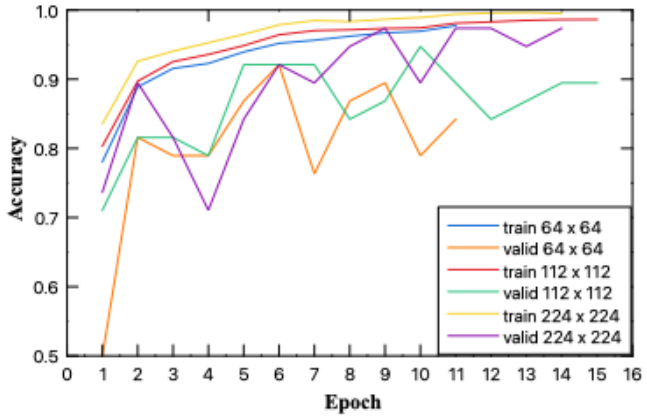

Figure 3. The ROC curve of the tuned CNN model

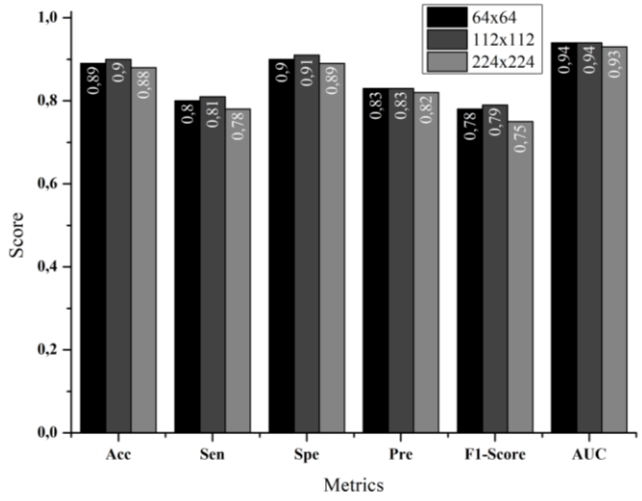

Figure 4. The comparation metrics of the CNN model

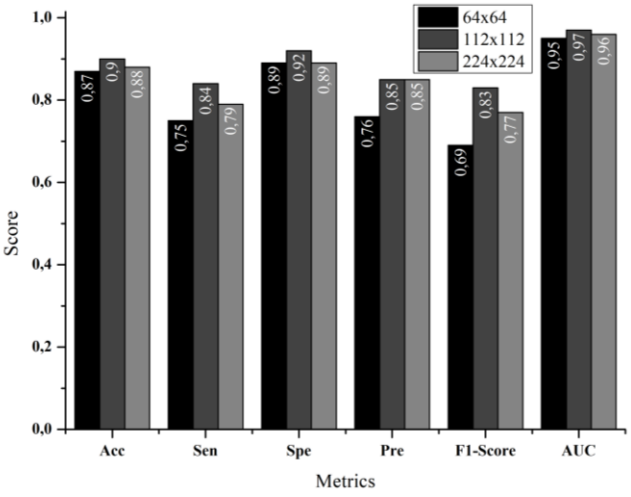

Figure 5. The comparation metrics of the tuned CNN model

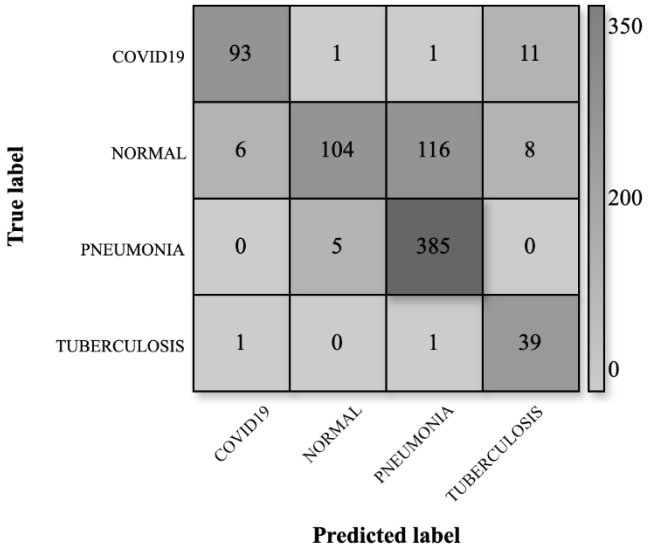

Figure 6. The confusion matrix test result of CNN model on the external dataset

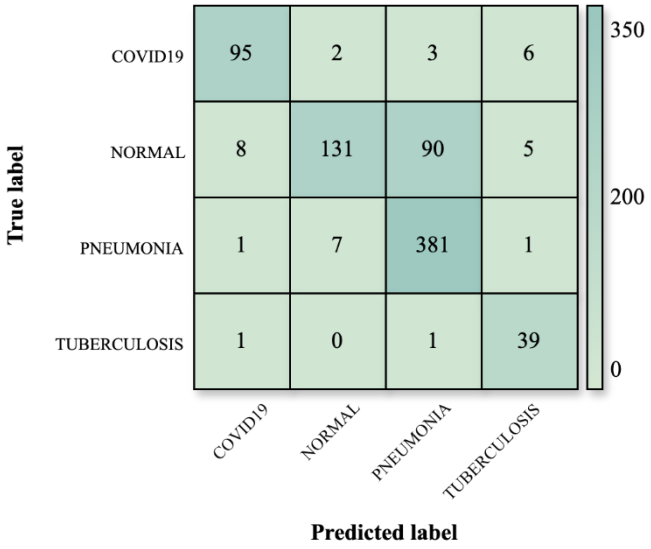

Figure 7. The confusion matrix test result of CNN tuned model on the external dataset

However, this performance was achieved through the combination of 300 deep features, which may have introduced redundancy and limited interpretability—an issue that this study streamlined architecture aims to overcome by reducing unnecessary feature overlap while maintaining high performance. Similarly, another study employing MobileNet V2 for multi-label classification of 14 lung diseases achieved an overall accuracy of 90% and an AUC of 0.810, but suffered from a markedly low sensitivity of 45.3% [28]. This low sensitivity underscores the challenge of effectively identifying positive cases, especially in clinically critical contexts. By contrast, the tuned CNN demonstrates a more balanced performance across metrics—including accuracy, sensitivity, specificity, precision, F1-score, and AUC—indicating not only stronger detection capability but also greater clinical reliability.

The findings of this study have important clinical implications, demonstrating that the tuned CNN model offers improved diagnostic performance for multi-disease classification of chest X-rays, particularly in detecting COVID-19, Tuberculosis, and Pneumonia, with potential application in resource-limited settings where radiological expertise is scarce.

Nevertheless, several limitations should be acknowledged. First, the external validation dataset may not fully capture the diversity of imaging conditions, patient demographics, and healthcare settings, which could affect the generalizability of the model. Second, class imbalance across categories, particularly with fewer samples for certain diseases, may have influenced model training and performance. Third, the risks associated with misclassification remain a concern, especially false negatives in high-risk categories such as COVID-19 and Tuberculosis, where delayed or missed diagnoses could have significant clinical consequences.

These challenges underscore the need for further optimization and validation. Future research should focus on addressing class imbalance through data augmentation or re-sampling techniques, enhancing model robustness via advanced architectures such as attention mechanisms or hybrid models, and expanding datasets to encompass broader populations and imaging conditions. Additionally, incorporating explainable AI will be critical for improving transparency and clinical trust, while prospective validation in real-world clinical workflows will be essential to ensure safe and effective deployment.

IV. CONCLUSION

In conclusion, this study presents a rigorously optimized deep learning model for accurate multi-disease classification of chest X-rays, capable of distinguishing COVID-19, Tuberculosis, Pneumonia, and Normal cases. Comparative analysis between the baseline (non-tuned) and tuned CNN models demonstrated that adding a third convolutional layer and optimizing training epochs significantly improved classification performance. In internal validation, the tuned model achieved higher accuracy (up to 0.97) and F1-scores compared to the baseline (0.94), particularly with 224×224 input resolution. External validation confirmed these improvements, with the tuned model attaining an F1-score of 0.83 and an AUC of 0.97 at 112×112 resolution, outperforming the baseline's 0.79 and 0.94, respectively. However, this study has several limitations. First, it relies on publicly available datasets, which may not fully represent the imaging variability and patient demographics of Indonesian clinical settings. Second, preprocessing was limited to image resizing, without enhancement or artifact correction, which may affect performance under suboptimal imaging conditions. Third, the absence of multi-center, real-world clinical validation limits the direct applicability of the findings. Future work will address these gaps by incorporating diverse datasets from Indonesian healthcare facilities, exploring advanced preprocessing and augmentation techniques, integrating explainable AI methods to improve interpretability, and conducting prospective clinical evaluations to ensure safe, reliable deployment in real-world environments.

ACKNOWLEDGEMENT

The authors wish to acknowledge Universitas Dinamika for its generous support and the resources that facilitated the successful execution and publication of this research work.

REFERENCES

[1] R. Filip, R. Gheorghita Puscaselu, L. Anchidin-Norocel, M. Dimian, and W. K. Savage, "Global Challenges to Public Health Care Systems during the COVID-19 Pandemic: A Review of Pandemic Measures and Problems," *J. Pers. Med.*, vol. 12, no. 8, 2022, doi: 10.3390/jpm12081295.

[2] A. Bhargava and M. Bhargava, "Tuberculosis deaths are predictable and preventable: Comprehensive assessment and clinical care is the key," *J. Clin. Tuberc. Other Mycobact. Dis.*, vol. 19, p. 100155, 2020, doi: 10.1016/j.jctube.2020.100155.

[3] I. Parwati *et al.*, "Evaluation of a real-time PCR assay performance to detect Mycobacterium tuberculosis, rifampicin, and isoniazid resistance in sputum specimens: a multicenter study in two major cities of Indonesia," *Front. Microbiol.*, vol. 15, no. May, pp. 1–8, 2024, doi: 10.3389/fmicb.2024.1372647.

[4] A. Keleb *et al.*, "Pneumonia remains a leading public health problem among under-five children in peri-urban areas of north-eastern Ethiopia," *PLoS One*, vol. 15, no. 9 September, pp. 1–15, 2020, doi: 10.1371/journal.pone.0235818.

[5] S. Sajed, A. Sanati, J. E. Garcia, H. Rostami, A. Keshavarz, and A. Teixeira, "The effectiveness of deep learning vs. traditional methods for lung disease diagnosis using chest X-ray images: A systematic review," *Appl. Soft Comput.*, vol. 147, p. 110817, 2023, doi: 10.1016/j.asoc.2023.110817.

[6] G. Nafisah, S. I., & Muhammad, "Tuberculosis detection in chest radiograph using convolutional neural network architecture and explainable artificial intelligence," *Neural Comput. Appl.*, vol. 36, no. 1, pp. 11–131, 2024.

[7] A. Makris, I. Kontopoulos, and K. Tserpes, "COVID-19 detection from chest X-ray images using deep learning and convolutional neural networks," *ACM Int. Conf. Proceeding Ser.*, pp. 60–66, 2020, doi: 10.1145/3411408.3411416.

[8] A. I. Khan, J. L. Shah, and M. M. Bhat, "CoroNet: A deep neural network for detection and diagnosis of COVID-19 from chest x-ray images," *Comput. Methods Programs Biomed.*, vol. 196, p. 105581, 2020, doi: 10.1016/j.cmpb.2020.105581.

[9] V. Pal, H. Pabari, S. Indoria, S. Patel, D. Krishnan, and V. Ravi, "Multifaceted Disease Diagnosis: Leveraging Transfer Learning with Deep Convolutional Neural Networks on Chest X-Rays for COVID-19, Pneumonia, and Tuberculosis," *Open Bioinforma. J.*, vol. 17, no. 1, pp. 1–18, 2024, doi: 10.2174/0118750362303182240516043224.

[10] M. M. A. Monshi, J. Poon, V. Chung, and F. M. Monshi, "CovidXrayNet: Optimizing data augmentation and CNN hyperparameters for improved COVID-19 detection from CXR," *Comput. Biol. Med.*, vol. 133, no. April, p. 104375, 2021, doi: 10.1016/j.compbiomed.2021.104375.

[11] P. Rajpurkar *et al.*, "Deep learning for chest radiograph diagnosis: A retrospective comparison of the CheXNeXt algorithm to practicing radiologists," *PLoS Med.*, vol. 15, no. 11, pp. 1–17, 2018, doi: 10.1371/journal.pmed.1002686.

[12] L. Venkataramana, D. V. V. Prasad, S. Saraswathi, C. M. Mithumary, R. Karthikeyan, and N. Monika, "Classification of COVID-19 from tuberculosis and pneumonia using deep learning techniques," *Med. Biol. Eng. Comput.*, vol. 60, no. 9, pp. 2681–2691, 2022, doi: 10.1007/s11517-022-02632-x.

[13] C. M. Huy, V. T. Q., & Lin, "An improved densenet deep neural network model for tuberculosis detection using chest x-ray images," *IEEE Access*, vol. 11, pp. 42839–42849, 2023.

[14] G. D. Deepak and S. K. Bhat, "A multi-stage deep learning approach for comprehensive lung disease classification from x-ray images," *Expert Syst. Appl.*, vol. 277, no. December 2024, p. 127220, 2025, doi: 10.1016/j.eswa.2025.127220.

[15] M. K. Mahbub, M. Biswas, L. Gaur, F. Alenezi, and K. C. Santosh, "Deep features to detect pulmonary abnormalities in chest X-rays due to infectious diseaseX: Covid-19, pneumonia, and tuberculosis," *Inf. Sci. (Ny).*, vol. 592, pp. 389–401, 2022, doi: 10.1016/j.ins.2022.01.062.

[16] M. S. Ahmed *et al.*, "Joint Diagnosis of Pneumonia, COVID-19, and Tuberculosis from Chest X-ray Images: A Deep Learning Approach," *Diagnostics*, vol. 13, no. 15, 2023, doi: 10.3390/diagnostics13152562.

[17] M. Kermany, Daniel; Zhang, Kang; Goldbaum, "Labeled Optical Coherence Tomography (OCT) and Chest X-Ray Images for Classification," 2018. https://data.mendeley.com/datasets/rscbjbr9sj/2

[18] M. E. Rahman, T., Khandakar, A., Kadir, M. A., Islam, K. R., Islam, K. F., Mazhar, R., ... & Chowdhury, "Reliable Tuberculosis Detection Using Chest X-Ray With Deep Learning, Segmentation and Visualization," *IEEE Access*, vol. 8, pp. 191586–191601, 2020.

[19] G. Maguolo and L. Nanni, "A critic evaluation of methods for COVID-19 automatic detection from X-ray images," *Inf. Fusion*, vol. 76, no. February, pp. 1–7, 2021, doi: 10.1016/j.inffus.2021.04.008.

[20] I. U. Yamani and B. Basari, "Leveraging Convolutional Neural Networks for Automated Detection and Grading of Diabetic Retinopathy from Fundus Images," *J. Tek. Elektro*, vol. 15, no. 2, pp. 68–73, 2024, doi: 10.15294/jte.v15i2.48769.

[21] B. N. Narayanan, V. S. P. Davuluru, and R. C. Hardie, "Two-stage deep learning architecture for pneumonia detection and its diagnosis in chest radiographs," no. March, p. 15, 2020, doi: 10.1117/12.2547635.

[22] T. Padma and C. Usha Kumari, "Deep Learning Based Chest X-Ray Image as a Diagnostic Tool for COVID-19," *Proc. - Int. Conf. Smart Electron. Commun. ICOSEC 2020*, no. October, pp. 589–592, 2020, doi:

10.1109/ICOSEC49089.2020.9215257.
[23] A. Iqbal, M. Usman, and Z. Ahmed, "Tuberculosis chest X-ray detection using CNN-based hybrid segmentation and classification approach," *Biomed. Signal Process. Control*, vol. 84, no. July 2022, p. 104667, 2023, doi: 10.1016/j.bspc.2023.104667.
[24] L. Alzubaidi *et al.*, *Review of deep learning: concepts, CNN architectures, challenges, applications, future directions*, vol. 8, no. 1. Springer International Publishing, 2021. doi: 10.1186/s40537-021-00444-8.
[25] D. Zheng, J., Lu, C., Hao, C., Chen, D., & Guo, "Improving the generalization ability of deep neural networks for cross-domain visual recognition," *IEEE Trans. Cogn. Dev. Syst.*, vol. 13, no. 1, pp. 607–620, 2020.
[26] P. Wang, Q., Xie, J., Zuo, W., Zhang, L., & Li, "Deep CNNs meet global covariance pooling: Better representation and generalization," *IEEE Trans. Cogn. Dev. Syst.*, vol. 43, no. 8, pp. 2582–2597, 2020.
[27] M. Toğaçar, B. Ergen, Z. Cömert, and F. Özyurt, "A Deep Feature Learning Model for Pneumonia Detection Applying a Combination of mRMR Feature Selection and Machine Learning Models," *Irbm*, vol. 41, no. 4, pp. 212–222, 2020, doi: 10.1016/j.irbm.2019.10.006.
[28] A. Souid, N. Sakli, and H. Sakli, "Classification and predictions of lung diseases from chest x- rays using mobilenet v2," *Appl. Sci.*, vol. 11, no. 6, 2021, doi: 10.3390/app11062751.