 ISSN 2579-7298

Evaluation Of A Feature-Concatenated Model For Multiclass Diagnosis Of Pulmonary Diseases on An Imbalanced Dataset

Wahyu Ajitomo [1], Dyah Aruming Tyas [2,*], Agus Harjoko [3]

Department of Computer Science and Electronics, Faculty of Mathematics and Natural Sciences, Universitas Gadjah Mada, Sekip Utara, Yogyakarta 55281, Indonesia
[1] wahyuajitomo2001@mail.ugm.ac.id; [2] dyah.aruming.t@ugm.ac.id *; [3] aharjoko@ugm.ac.id
* corresponding author

ARTICLE INFO

Article history
Received
Revised
Accepted

Keywords
Imbalanced dataset
Chest X-ray classification
Lung disease diagnosis
CNN concatenation
Multiclass Focal Loss
Grad-CAM interpretability

ABSTRACT

Lung diseases such as pneumonia, tuberculosis, and COVID-19 pose serious global health challenges, particularly in X-ray image classification where class distribution is often imbalanced. To address this issue, this study proposes a hybrid model based on concatenated CNN architectures and applies class weighting using Multiclass Focal Loss. The dataset consists of 7,135 X-ray images divided into four main classes: pneumonia, tuberculosis, COVID-19, and normal. Focal loss with a gamma parameter of 2.0 is employed to enhance the model's focus on minority classes. Evaluation results show that combined models such as DenseNet121 + VGG16 and VGG16 + ResNet50 achieve F1-scores of up to 0.87, outperforming single models. Grad-CAM visualizations also indicate that the combined models can recognize pathological areas more comprehensively and accurately. This approach proved effective in improving the accuracy and sensitivity of AI-based diagnostic systems.

1. Introduction

Respiratory diseases such as pneumonia, tuberculosis (TB), and COVID-19 remain major global health threats. Pneumonia is one of the leading causes of death in children under five [1], TB continues to report more than 5 million new cases annually [2], and the COVID-19 pandemic has caused unprecedented mortality worldwide [3]. Chest X-ray imaging is a widely accessible and essential tool for pulmonary disease diagnosis [4], yet manual interpretation is time-consuming and subject to inter-observer variability, motivating the adoption of deep learning for automated diagnosis [5].

Early studies often emphasized binary classification tasks such as COVID-19 vs. normal, achieving very high accuracy [6], [7], but performance generally declined when extended to multiclass classification [8], [9]. A persistent barrier is class imbalance, where minority diseases such as TB or COVID-19 are underrepresented. Preprocessing approaches including SMOTE [10], undersampling [11], ADASYN [12], and hybrid resampling [13], improved recall in some cases but may not reflect real-world data distributions. By contrast, algorithm-level strategies such as cost-

sensitive learning [14], Focal Loss [15], Class-Balanced Loss [16], and Batch-Balanced Focal Loss [17] have demonstrated more robust improvements for imbalanced clinical datasets.

Architectural advances further shaped pulmonary disease detection. Hybrid deep learning models such as CNN–LSTM [18] and CNN–ELM [19] achieved strong accuracy in multiclass classification, while ensemble CNNs [20] and DenseNet–ResNet hybrids (e.g., DenResCov-19) reported up to 95% accuracy [9]. Multimodal concatenation approaches have also been explored. For example, combined CT and X-ray modalities to classify two classes (Normal vs. COVID-19), achieving 99.87% accuracy using cross-entropy loss [21]. Similarly, concatenated Xception and ResNet50V2 for COVID-19 vs. pneumonia classification [22]. However, these datasets were balanced and binary, which do not capture the complexity of real-world multiclass diagnosis.

Interpretability is also a critical requirement for clinical application. Explainable AI (XAI) techniques such as Grad-CAM [23], Score-CAM [24], and Guided Grad-CAM [25] provide visual justifications for model predictions, enhancing clinician trust. Broader surveys confirm that XAI remains underdeveloped in multiclass medical imaging tasks [26], [27].

Despite these advances, several research gaps remain. First, most works emphasize binary classification, whereas clinical practice demands multiclass diagnosis [6], [9]. Second, reliance on preprocessing-based balancing may distort clinical data distributions [10], [12]. Third, while multiclass frameworks exist [9], [19], [21], they often fail to achieve balanced sensitivity across minority classes. Finally, the integration of interpretability into multiclass deep learning remains limited [23], [26].

To address these gaps, this study proposes a feature-concatenated CNN architecture (VGG19 + DenseNet121) optimized with Multiclass Focal Loss. Unlike prior works that relied on single architectures or binary tasks, the proposed model (1) tackles multiclass classification of Normal, Pneumonia, TB, and COVID-19 under severe imbalance, (2) applies loss-function-level optimization to enhance minority sensitivity, and (3) integrates Grad-CAM visualizations for interpretability consistent with radiologic findings. VGG19 is well-suited for fine-grained pulmonary texture detection, while DenseNet121 improves feature reuse and gradient propagation; concatenation strengthens representation learning. Multiclass Focal Loss further emphasizes hard-to-classify minority cases, aligning with real-world data distributions [15], [16], [28]. By combining these strategies, this study contributes one of the first frameworks that jointly addresses multiclass imbalance, accuracy, and clinical interpretability in chest X-ray diagnosis.

The main contributions of this study are threefold: (1) the design of a feature-concatenated dual-backbone architecture that leverages complementary strengths of DenseNet121 and VGG19/ResNet50 to enrich representation learning, (2) the application of Multiclass Focal Loss ($\gamma = 2$, α computed automatically from class distribution) to mitigate severe class imbalance and improve sensitivity to minority classes, and (3) a comprehensive evaluation in a realistic multiclass setting, supported by Grad-CAM visualizations to verify that model attention aligns with clinically relevant lung regions.

2. Method

This study adopts a structured workflow for multiclass classification of pulmonary diseases using chest X-ray images. The process begins with dataset acquisition and partitioning into training, validation, and testing subsets. Prior to modeling, all images are preprocessed through resizing to 224 × 224 pixels and intensity normalization to [0,1], ensuring consistency across inputs. The core architecture integrates a feature-concatenated network of DenseNet121 and VGG19, selected for their complementary feature extraction capabilities. To address the inherent class imbalance, the model is trained with Multiclass Focal Loss, which dynamically emphasizes

minority classes. Model performance is assessed using accuracy, precision, recall, F1-score, and AUC-ROC, providing a comprehensive evaluation of classification effectiveness. Finally, to enhance interpretability and clinical trust, Grad-CAM visualizations are generated, highlighting the lung regions most influential in the model's decision-making process.

Fig 1. Research Flow Diagram

2.1. Data Acquisition

The dataset used in this study consists of chest X-ray images categorized into four main classes: pneumonia, tuberculosis, COVID-19, and normal. The dataset was obtained from Kaggle under the name Chest X-Ray (Pneumonia, COVID-19, Tuberculosis)[29]. It comprises a total of approximately 7,135 labeled images, annotated by expert radiologists, this imbalance was deliberately preserved to reflect real-world clinical data distribution. as shown in Table 1.

Table 1. Distribution Dataset

Class	*Frequency*	*Percentage*
Pneumonia	4273	59.90%
Tuberculosis	703	9.85%
COVID-19	576	8.07%
Normal	1583	22.19%

2.2. Data Preprocessing

Data preprocessing is a critical stage to ensure that the input images are structured, standardized, and suitable for deep learning model training and evaluation. In this study, a total of 7,135 chest X-ray images comprising Pneumonia, COVID-19, Tuberculosis, and Normal cases were obtained from publicly available Kaggle repositories. Preprocessing was implemented using the ImageDataGenerator module in Keras, which normalized pixel intensities to the [0,1] range and facilitated dataset partitioning into training, validation, and testing subsets. The flow_from_directory method automated the loading process by reading folder structures that represented each class and converting the labels into numerical form through one-hot encoding.

All images were resized to 224 × 224 pixels to match the input requirements of the CNN architectures employed in this study, ensuring uniform input dimensions across all samples and enabling consistent training. The parameter class_mode='categorical' was specified to activate the multiclass label format, while shuffle=True was applied to the training subset to improve generalization and reduce overfitting. For validation and test subsets, shuffle=False was used to preserve the sequence of data during evaluation.

This preprocessing strategy standardized the dataset, reduced variability across inputs, and promoted robust model generalization to unseen samples, thereby supporting effective and reliable multiclass classification of pulmonary diseases.

2.3. Concatenated CNN

The proposed concatenated CNN architecture is designed to exploit the complementary strengths of DenseNet121 and VGG19 for multiclass classification of chest X-ray images. Input images are resized to 224 × 224 pixels with three channels (RGB) to align with the input requirements of both backbones.

In the first stage, both DenseNet121 and VGG19 perform parallel feature extraction**.** DenseNet121 employs dense connectivity, where each layer receives feature maps from all preceding layers, facilitating efficient information flow and reducing parameter redundancy. In contrast, VGG19 applies a series of 3×3 convolutional filters in a structured deep architecture, enabling the extraction of fine-grained and complex visual patterns in medical images.

Following feature extraction, Global Average Pooling (GAP) is applied to transform feature maps into one-dimensional vectors that summarize filter responses. Batch normalization is then used to stabilize and accelerate convergence, while dropout regularization mitigates overfitting. Each branch is followed by fully connected (dense) layers that refine the extracted representations.

The feature vectors from DenseNet121 and VGG19 are subsequently merged through a concatenation layer, forming a richer and more representative combined feature space. This joint representation is passed to a final dense layer with a softmax activation function, producing class probabilities for the four diagnostic categories: Pneumonia, Tuberculosis, COVID-19, and Normal**.**

By integrating DenseNet121's efficient information reuse with VGG19's stable hierarchical feature extraction, the concatenated architecture enhances representational capacity and is expected to improve classification performance, particularly for underrepresented classes. The workflow of the proposed model is illustrated in Figure 2**.**

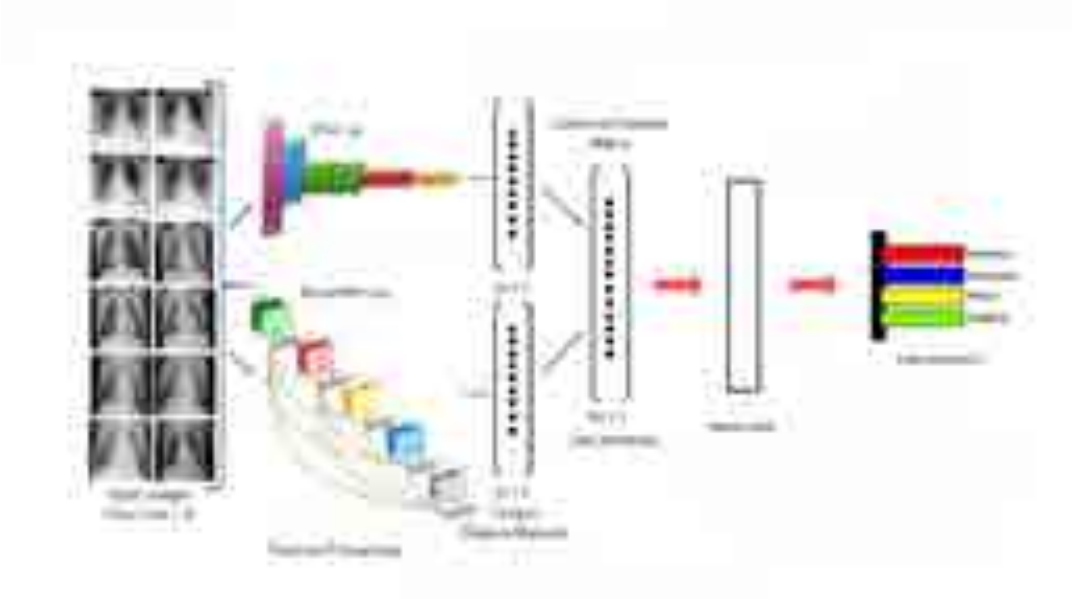


Fig 2. Workflow of the Concatenated CNN Architecture.

An essential strategy employed in the proposed architecture is class weighting, which addresses the problem of dataset imbalance. This approach assigns higher weights to minority classes during training while reducing the weights for majority classes, ensuring that errors on underrepresented categories have a greater impact on the optimization process. As a result, the model is encouraged to focus more on difficult and minority cases, thereby improving sensitivity across all classes.

In this study, class weighting is realized through the use of Focal Loss, a loss function originally introduced for binary classification of dense object detection. Focal Loss modifies the standard cross-entropy loss by introducing a modulating factor that reduces the relative loss contribution of well-classified examples, allowing the model to concentrate on hard-to-classify samples. To extend its applicability to this work, Focal Loss is adapted into a multiclass formulation by integrating the categorical cross-entropy function. This Multiclass Focal Loss dynamically adjusts the contribution of each class based on both prediction difficulty and class imbalance, providing a more effective optimization strategy for real-world multiclass medical imaging tasks.

2.4. Evaluation Model

The method used to measure the performance of a classification model, which includes actual and predicted outcomes, is the confusion matrix. A confusion matrix is a matrix used to evaluate a classifier's performance that can be used to evaluate the performance of a predictor[31].

Accuracy is the most commonly used metric for evaluating classification performance. It estimates the probability that the predicted class label matches the actual class label[32]. The formula for calculating accuracy is shown in Equation (1).

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \times 100\% \quad (1)$$

Precision measures how accurately a model predicts the positive class. The formula for precision is presented in Equation (2).

$$Precision = \frac{TP}{TP+FP} \times 100\% \quad (2)$$

Recall or sensitivity measures the completeness or accuracy of positive information retrieved by the system compared to the total amount of actual positive information[32]. The calculation of sensitivity is shown in Equation (3).

$$Recall = \frac{TP}{TP+FN} \times 100\% \quad (3)$$

F1 score indicates perfect recall and precision, whereas a lower F1 score suggests a lack of recall or precision. The F1 score is evaluated as shown in Equation (4).

$$F_1\ Score = \frac{Precision \times sensitivity}{Precision + sensitivity} \times 100\% \quad (4)$$

AUC-ROC is an evaluation metric used to measure a model's ability to distinguish between positive and negative classes in binary or multiclass classification tasks (using a one-vs-all approach).
The ROC Curve is a plot that illustrates the relationship between the True Positive Rate (TPR) (Recall) on the Y-axis, as defined in Equation (5), and the False Positive Rate (FPR) on the X-axis, as shown in Equation (6).

TPR (*Recall*):

$$TPR_x = \frac{TP_x}{TP_x+FN_x} \quad (5)$$

FPR:

$$FPR_x = \frac{FP_x}{FP_x+TN_x} \quad (6)$$

This study did not employ questionnaires or expert assessments; all model decisions and evaluations were based solely on quantitative measures, including classification metrics and Grad-CAM analyses.

2.5. Grad-CAM Visualization

Gradient-weighted Class Activation Mapping (Grad-CAM) is employed in this study to enhance the interpretability of the proposed model. Grad-CAM generates a class-discriminative heatmap that highlights the regions of an image most influential to the model's prediction, thereby providing insight into the decision-making process of the CNN. The technique works by computing the gradients of the target class score with respect to the feature maps in the final convolutional layer. These gradients are used as weights to combine the feature maps, resulting in an activation map that emphasizes class-relevant regions. The activation map is then superimposed on the original chest X-ray, producing a heatmap visualization that indicates areas most critical for classification.

This method provides important benefits in medical imaging. It enables clinicians to verify whether the model attends to clinically relevant lung regions rather than artifacts, while also improving the transparency and reliability of predictions. By linking model performance with

interpretability, Grad-CAM supports greater trust and acceptance of AI-assisted diagnosis. The visualization workflow is illustrated in Figure 3.

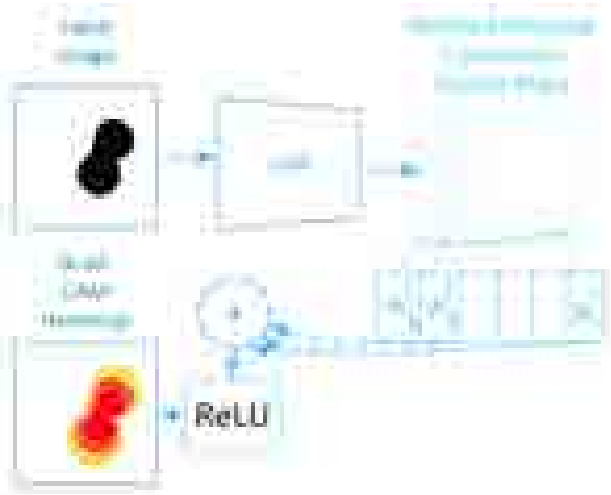


Fig 3. Grad-CAM Visualization Workflow[30]

3. Results and Discussion

3.1. Data Preprocessing

The preprocessing stage consisted of two main steps: resizing and normalization. First, all chest X-ray images were resized to 224 × 224 pixels to match the input requirements of the CNN architectures employed in this study. Second, pixel intensity values were normalized from the original range of [0, 255] to [0, 1] using the parameter rescale=1./255 in the Keras ImageDataGenerator. This normalization ensured a uniform distribution of pixel values, enabling the model to learn more stably and efficiently. An example of the preprocessing results is illustrated in Figure 4.

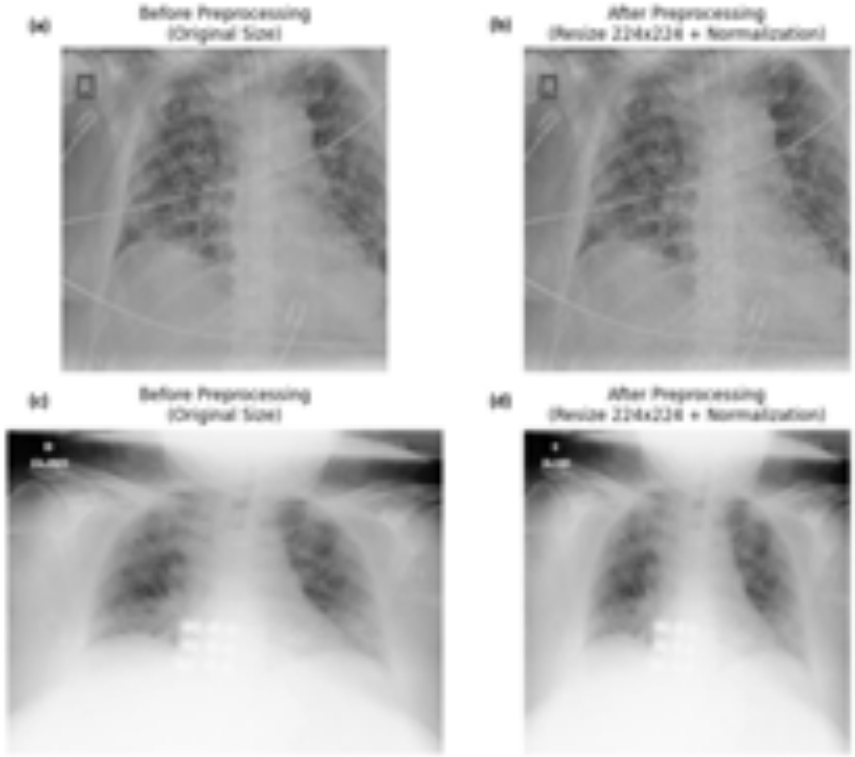


Fig 4. Chest X-ray preprocessing results: (a) and (c) before preprocessing, (b) and (d) after preprocessing (resize 224×224 and normalization).

Visually, the left column of Figure 4 presents the original chest X-ray images prior to preprocessing, where image dimensions varied and pixel values were unnormalized. The right column shows the processed images with standardized dimensions and normalized intensity values. Despite resizing and scaling, the essential diagnostic features remain intact, indicating that the applied preprocessing steps effectively prepare the data for model training without compromising clinically relevant details.

3.2. Impact of Focal Loss on Model Performance

Focal Loss is a loss function specifically designed to address the issue of class imbalance, which was a notable challenge in this study. The distribution of training samples was skewed, with Tuberculosis and COVID-19 classes containing substantially fewer images than Pneumonia and Normal. To mitigate this limitation, a Multiclass Focal Loss was employed with a focusing parameter $\gamma = 2.0$ and class weights (α) automatically computed using the compute_alpha() function. Mechanistically, Focal Loss extends the standard cross-entropy by incorporating a focusing factor that reduces the relative contribution of well-classified (easy) samples while amplifying the penalty for hard-to-classify samples. The compute_alpha() function assigns weights inversely proportional to class frequencies, ensuring that minority classes (e.g., Tuberculosis and COVID-19) receive greater emphasis. These weights are normalized so that the total gradient contribution remains stable, thereby directing updates toward underrepresented classes without destabilizing the optimization process.

1) Baseline without Focal Loss

When trained with standard cross-entropy loss, the DenseNet121 model achieved an overall accuracy of 0.83 and a macro recall of 0.82. Class-specific results revealed the underlying imbalance: Pneumonia achieved high recall (0.95) but lower precision (0.82) due to false positives, while COVID-19 and Normal showed poor recall (0.67 and 0.66, respectively). Tuberculosis reached perfect recall (1.00) but very low precision (0.61), indicating frequent misclassifications. These outcomes highlight the tendency of cross-entropy loss to favor majority classes, reducing sensitivity for minority categories. Detailed performance metrics are provided in Table 2.

Table 2. Baseline without Focal Loss

Class	*Precision*	*Recall*	*F1-score*
COVID-19	0.95	0.67	0.78
Normal	0.88	0.66	0.76
Pneumonia	0.82	0.95	0.88
Tuberculosis	0.61	1.00	0.76
Accuracy			**0.83**
Macro Avg	0.81	0.82	0.79
Weighted Avg	0.84	0.83	0.82

2) Baseline with Focal Loss

Applying Multiclass Focal Loss to the DenseNet121 model led to a clear improvement in overall performance, particularly in clinically critical metrics. The model's accuracy increased from 0.83 to 0.87, while macro recall rose from 0.82 to 0.88. Recall for COVID-19 showed the most substantial gain, rising from 0.67 to 0.92, and Normal also improved from 0.66 to 0.79,

significantly reducing false negatives in these vital categories. Pneumonia achieved a more balanced trade-off, with precision at 0.90 and recall at 0.89, while Tuberculosis retained high recall (0.93) alongside improved precision (0.72). These findings confirm that Multiclass Focal Loss effectively mitigates class imbalance by enhancing sensitivity for minority classes without sacrificing overall performance. The detailed improvements are presented in Table 3.

Table 3. Baseline with Focal Loss

Class	*Precision*	*Recall*	*F1-score*
COVID-19	0.82	0.92	0.87
Normal	0.86	0.79	0.83
Pneumonia	0.90	0.89	0.90
Tuberculosis	0.72	0.93	0.81
Accuracy			**0.87**
Macro Avg	0.83	0.88	0.85
Weighted Avg	0.87	0.87	0.87

3) Class-wise Delta Improvement

A direct comparison between the two training scenarios demonstrates how Multiclass Focal Loss shifts the model toward higher sensitivity for minority and clinically significant classes. COVID-19 recall increased by 0.25, and Normal improved by +0.13, substantially reducing false negatives in these categories. Tuberculosis showed a slight decrease in recall (−0.07) but gained significantly in precision (+0.11), while Pneumonia experienced a minor drop in recall (−0.06) accompanied by improved precision. Overall, these adjustments resulted in a more balanced performance profile across all four classes. Detailed class-wise precision, recall, and F1-scores are provided in Table 4.

Table 4. Class-wise Delta Improvement between Cross-Entropy and Multiclass Focal Loss

Class	*Support*	*Recall (CE)*	*Recall (Focal)*	*Δ Recall*	*F1 (CE)*	*F1 (Focal)*	*Δ F1*
COVID-19	106	0.67	0.92	**+0.25**	0.78	0.87	**+0.09**
Normal	234	0.66	0.79	**+0.13**	0.76	0.83	**+0.07**
Pneumonia	390	0.95	0.89	**−0.06**	0.88	0.90	**+0.02**
Tuberculosis	41	1.00	0.93	**−0.07**	0.76	0.81	**+0.05**

Overall, the integration of Multiclass Focal Loss with class distribution–based α weighting substantially improved the balance between precision and recall across all categories, resulting in an increase of +0.06 in macro recall and +0.04 in accuracy. This outcome indicates that the model became less biased toward majority classes and more effective in recognizing minority cases such as COVID-19 and Tuberculosis. From a clinical perspective, this improvement is particularly important, as it reduces false negatives in critical disease categories and thereby enhances the reliability and applicability of automated diagnosis in real-world medical settings.

3.3. Model Evaluation Results

Testing was performed on fourteen model configurations, comprising both single models and concatenated models. The single models employed a single pretrained CNN backbone, whereas the

concatenated models combined two distinct architectures through feature-level concatenation to enhance representational capacity. The complete set of evaluated model combinations is presented in Table 5.

Table 5. Performance comparison of single and concatenated CNN models

Model	*Accuracy*	*Precision*	*Recall*	*F1-Score*	*AUC-ROC*	*Time*	*Time without FL*
DenseNet121	**84%**	**85%**	**84%**	**84%**	**96%**	18:30.54	19:44.84
VGG19	82%	82%	82%	81%	95%	22:12.79	25:27.23
VGG16	84%	85%	84%	84%	96%	19:42.28	24:19.73
ResNet50	78%	80%	78%	78%	94%	18:21.11	26:14.75
InceptionV3	79%	80%	79%	79%	94%	19:31.16	27:35.79
Xception	83%	84%	83%	83%	95%	17:22.14	20:34.16
DenseNet121 + VGG19	86%	87%	86%	86%	97%	36:27.96	37:54.27
DenseNet121 + VGG16	**87%**	**87%**	**87%**	**87%**	**97%**	34:55.69	35:57.81
DenseNet121 + InceptionV3	85%	85%	85%	85%	96%	26:32.16	29:01.56
DenseNet121 + Xception	86%	87%	86%	86%	97%	26:45.70	26:01.41
VGG19 + ResNet50	84%	84%	84%	83%	96%	32:34.17	33:38.43
VGG19 + InceptionV3	83%	84%	83%	82%	96%	34:55.03	33:35.72
VGG19 + Xception	85%	86%	85%	84%	96%	35:41.77	34:28.94
VGG16 + ResNet50	**87%**	**88%**	**87%**	**87%**	**97%**	39:31.56	33:51.29

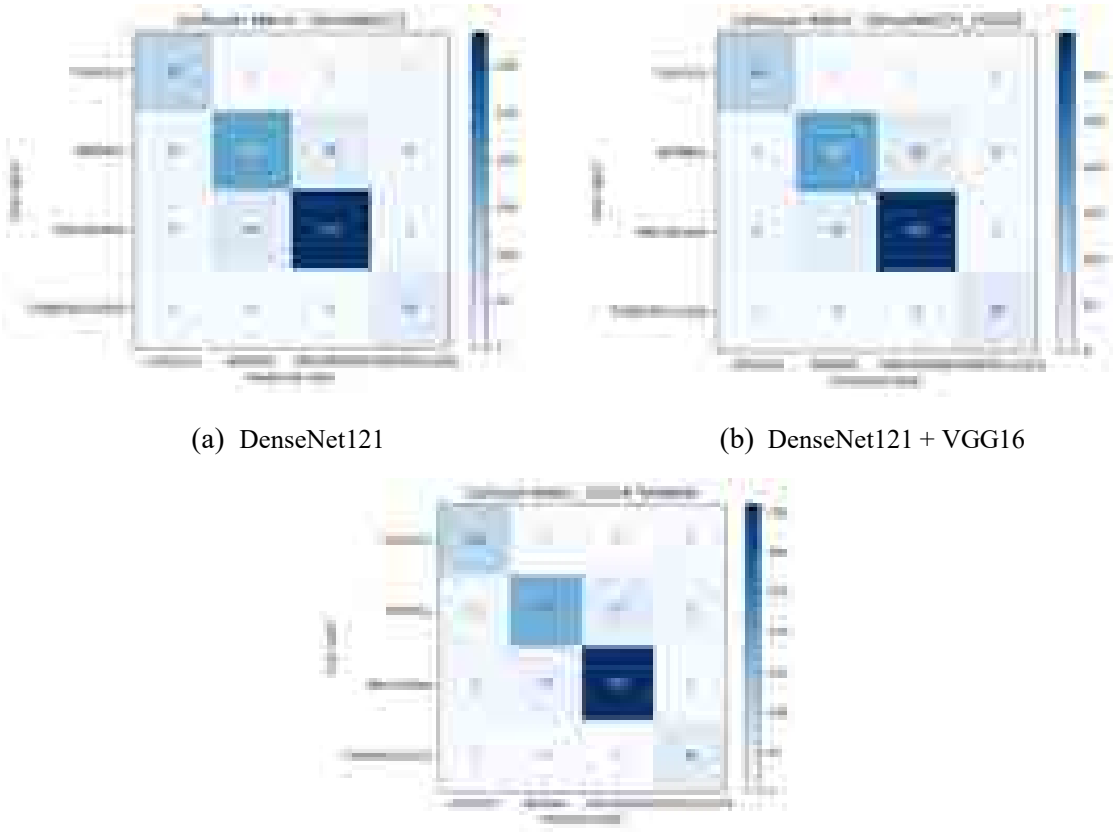
Based on the experimental results, the concatenated models generally outperformed the single backbones, confirming the advantage of combining complementary feature extractors. The VGG16 + ResNet50 and DenseNet121 + VGG16 architectures achieved the highest accuracy of 87%, with precision (88%), recall (87%), and F1-score (87%) showing a well-balanced performance profile. Both models also reached an AUC-ROC of 97%, indicating excellent capability in distinguishing between disease categories. These findings suggest that feature-level concatenation enriches image representation, allowing the model to capture both fine-grained textures and deep contextual patterns, which is particularly valuable in multiclass settings.

Nevertheless, some single models also demonstrated competitive performance. For instance, DenseNet121 and VGG16 achieved 84% accuracy, precision and recall in the 84–85% range, and an AUC-ROC of 96%. Although slightly lower than concatenated models, these results highlight that single models remain viable, especially when computational efficiency is prioritized, since they require fewer parameters and shorter training times.

Overall, the consistent performance gains observed in concatenated models underline the importance of leveraging complementary architectures in multiclass classification of imbalanced chest X-ray datasets. Unlike binary studies that often report near-perfect accuracy under balanced conditions, the present results demonstrate that hybrid models are more robust in complex multiclass scenarios, achieving both high accuracy and balanced sensitivity. From a clinical standpoint, this improvement is meaningful, as it reduces the risk of misdiagnosis across minority classes such as COVID-19 and Tuberculosis, thereby enhancing the reliability of AI-assisted diagnostic support.

To complement the quantitative performance metrics, AUC-ROC curves and confusion matrices were generated to provide a more detailed evaluation of class-wise performance. These visualizations help illustrate the distribution of prediction errors and reveal the sensitivity of the models to each disease category. The analysis focused on the three best-performing models DenseNet121, DenseNet121 + VGG16, and VGG16 + ResNet50, which consistently achieved optimal results across accuracy, precision, recall, F1-score, and AUC-ROC.

The purpose of this analysis is to assess how effectively each model distinguishes between classes and to identify patterns of recurring false positives and false negatives. The confusion matrices for the three models are presented in Figure 5, offering a granular view of classification outcomes and class-specific diagnostic strengths and limitations.



(a) DenseNet121 (b) DenseNet121 + VGG16

(c) VGG16 + ResNet50

Fig 5. Confusion matrices of the evaluated models: (a) DenseNet121 single model, (b) DenseNet121 + VGG16 concatenated model, and (c) VGG16 + ResNet50 concatenated model.

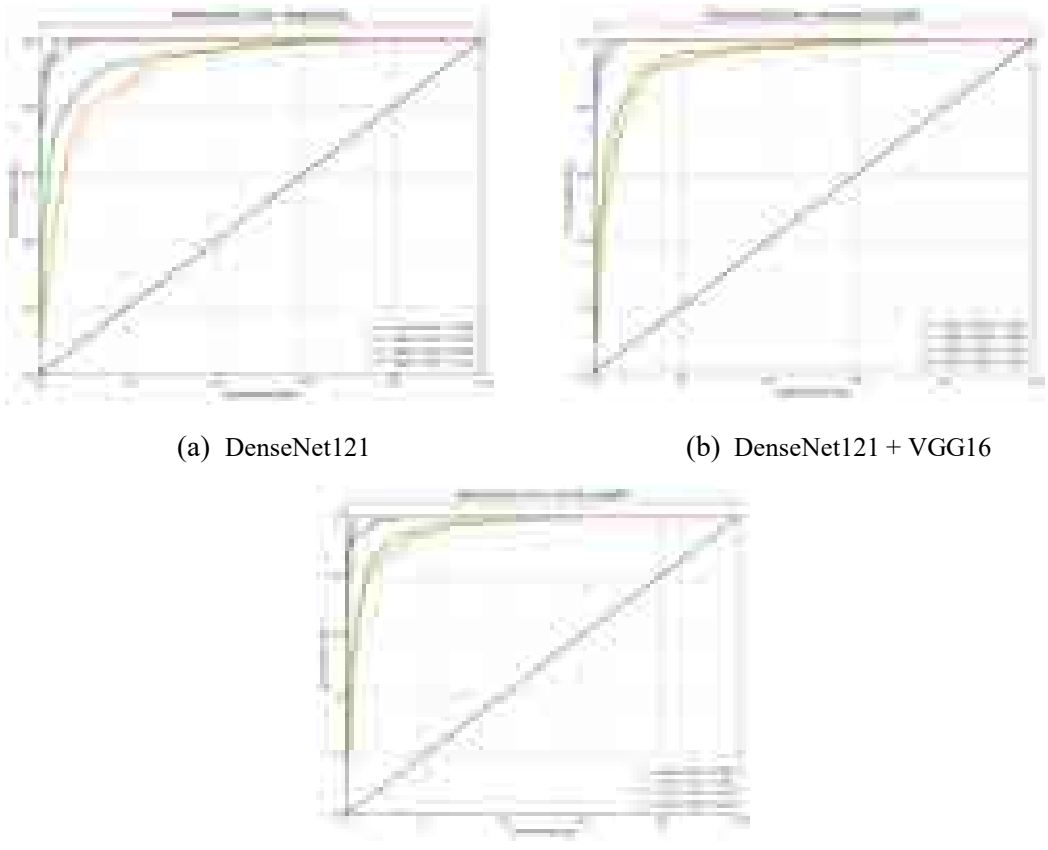
Figure 5 illustrates the progressive improvement in classification performance from single to concatenated models. In Figure 5(a), the single DenseNet121 model achieved satisfactory overall accuracy but exhibited a tendency to overpredict the Pneumonia class while misclassifying minority categories such as Tuberculosis. This reflects the bias of single backbones toward majority classes, leading to clinically concerning errors.

In Figure 5(b), the concatenated DenseNet121 + VGG16 model produced a more balanced prediction distribution, with reduced misclassifications and improved sensitivity to minority classes. This improvement stems from the complementary strengths of the two backbones: VGG16 effectively captures fine-grained textures, while DenseNet121 provides deeper feature representation. Consequently, recall for COVID-19 and Tuberculosis increased, while overprediction of Pneumonia decreased.

Further enhancement is shown in Figure 5(c) with the VGG16 + ResNet50 model, which demonstrated greater stability and consistency across all classes. This model achieved a more balanced trade-off between precision and recall and yielded lower overall misclassification rates. Importantly, false negatives for COVID-19 and Tuberculosis, two clinically critical classes were significantly reduced.

Overall, these results confirm that the concatenation strategy not only improves global evaluation metrics (recall and AUC-ROC) but also distributes errors more evenly across classes, thereby enhancing the robustness and reliability of deep learning–based clinical decision support systems.

In addition to categorical evaluation metrics, Receiver Operating Characteristic (ROC) curves and Area Under the Curve (AUC) values were generated to assess the sensitivity of each model to individual classes. The ROC curve illustrates the trade-off between the true positive rate (TPR) and the false positive rate (FPR) across varying decision thresholds, while the AUC provides a quantitative measure of overall discriminative ability. A higher AUC indicates stronger performance in distinguishing between disease categories. The ROC-AUC results for the evaluated models are presented in Figure 6.



(a) DenseNet121 (b) DenseNet121 + VGG16

(c) VGG16 + ResNet50

Fig 6. Multiclass ROC Curves: (a) DenseNet121, (b) DenseNet121 + VGG16, (c) VGG16 + ResNet50

Figure 6 presents the ROC curves for both single and concatenated models. In Figure 6(a), the single DenseNet121 model achieved strong AUC values between 0.92 and 0.99, although slight variations across classes reflected residual imbalance in sensitivity. In Figure 6(b), the concatenated DenseNet121 + VGG16 model demonstrated more dominant ROC curves, with AUC values consistently above 0.94 and peaking at 1.00, indicating improved stability and generalization.

Similarly, Figure 6(c) shows that the VGG16 + ResNet50 model maintained stable performance, with AUC values ranging from 0.95 to 1.00 across all categories.

These results confirm that concatenated models consistently outperform single architectures by reducing inter-class variability and enhancing sensitivity to minority classes. This improvement underscores that concatenation not only raises accuracy but also strengthens the model's generalization ability across multiple pulmonary disease categories. While the proposed dual-backbone (DenseNet121 + VGG19) was the main focus, other concatenated models such as VGG16 + ResNet50 also achieved comparable performance, underscoring the general effectiveness of the concatenation strategy.

To evaluate the stability of the proposed approach, three independent training runs with different random seeds were performed for the best single and concatenated models. The results are reported as mean ± standard deviation for accuracy, macro-F1, and macro-AUC. As shown in Table 6, the variance across runs was minimal, confirming that the models produced consistent results and that the improvements were not due to chance.

Table 6. Performance across three repeated runs (mean ± SD) for selected models

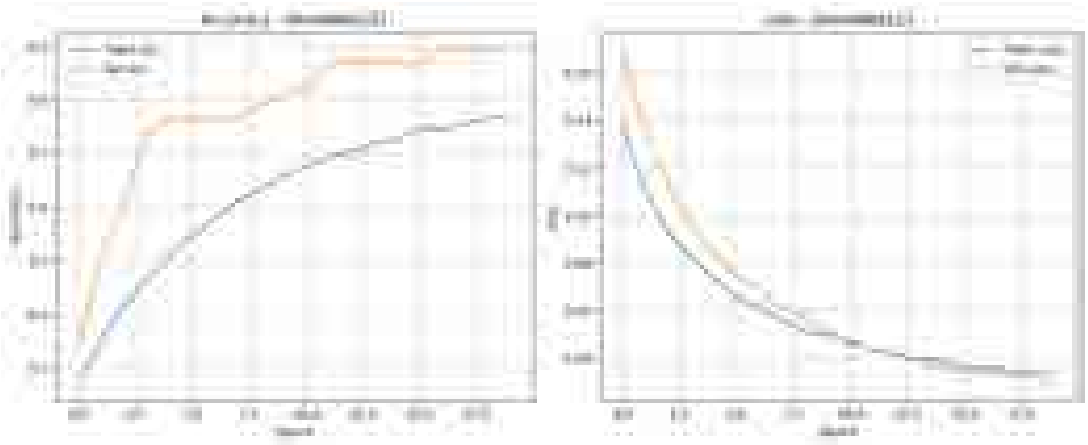
Models	*Accuracy*	*Macro-F1*	*Macro-Recall*	*Macro-AUC*
DenseNet121 (single)	0.83 ± 0.01	0.82 ± 0.02	0.82 ± 0.02	0.96 ± 0.01
DenseNet121 + VGG16	0.87 ± 0.01	0.87 ± 0.01	0.88 ± 0.01	0.97 ± 0.01
VGG16 + ResNet50	0.87 ± 0.01	0.87 ± 0.02	0.87 ± 0.01	0.97 ± 0.01

The low variance across runs confirms the robustness of the models. Importantly, the concatenated architectures achieved higher macro-recall (0.87–0.88) than the single DenseNet121 model (0.82), highlighting their superior ability to detect minority classes such as COVID-19 and Tuberculosis.

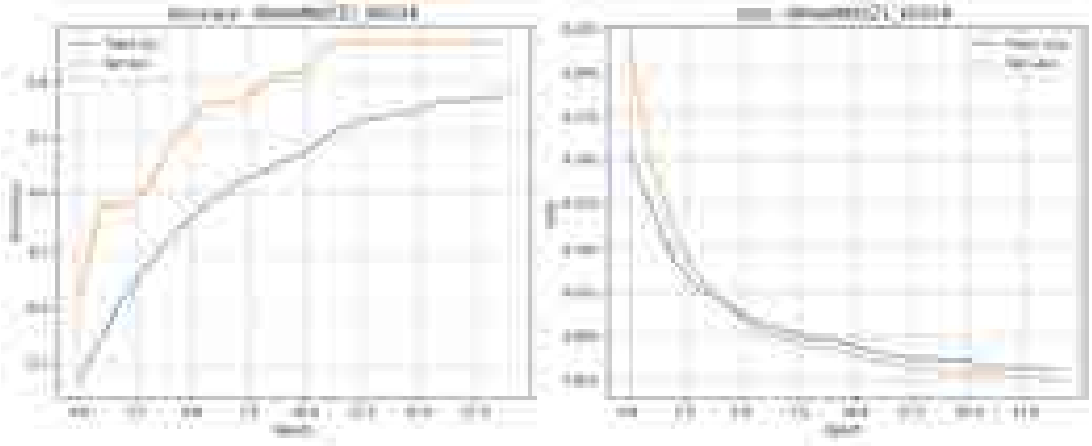
Based on the overall experimental evaluation, several key findings can be summarized:

1) Multiclass Focal Loss substantially improved recall for minority classes (COVID-19 and Normal), reducing false negatives while maintaining overall accuracy.
2) Concatenated architectures (DenseNet121 + VGG16, VGG16 + ResNet50) consistently achieved higher macro-recall (0.87–0.88) compared to single models (0.82), balancing sensitivity across all classes.
3) Concatenation also boosted macro-F1 and AUC, confirming that complementary feature extraction enriches representation and improves classification robustness.
4) Grad-CAM visualizations revealed consistent attention to clinically relevant lung regions, thereby enhancing interpretability and clinical trust.
5) Repeated training runs produced stable results with low variance, proving that the improvements were consistent and not dependent on a single experimental setup.

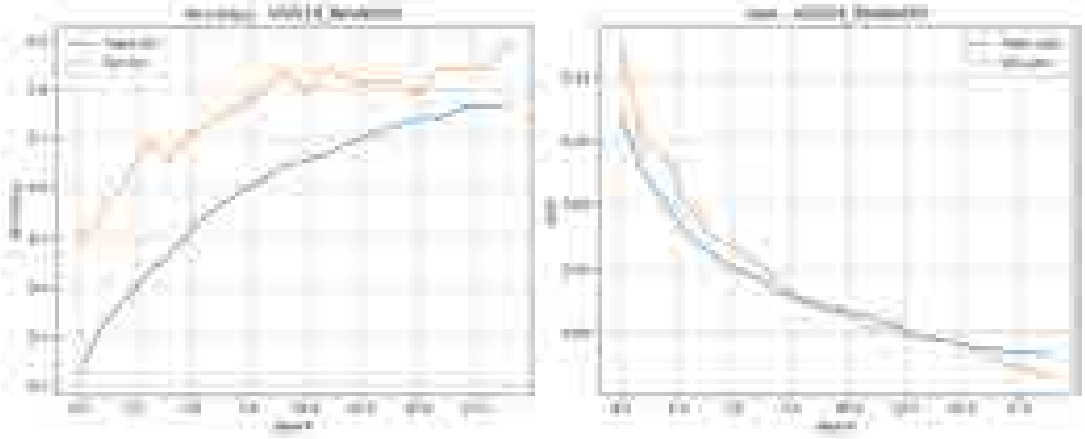
Following these findings, we further analyzed the training dynamics, convergence stability, and generalization ability of the models using accuracy and loss curves across training epochs. These visualizations provide additional evidence on hyperparameter suitability, overfitting risks, and the effect of concatenation on convergence speed and generalization. The complete results are presented in Figure 7.



(a) DenseNet121



(b) DenseNet121 + VGG16



(c) VGG16 + ResNet50

Fig 7. Training and Validation Performance: (a) DenseNet121, (b) DenseNet121 + VGG16, (c) VGG16 + ResNet50

Figure 7(a) shows that the single model DenseNet121 quickly reached stable validation accuracy at approximately 0.87. The parallel decline of training and validation loss with only a small gap suggests stable convergence and the absence of overfitting. However, the test recall (84%) and AUC (0.96) indicate that, despite good generalization, the model still struggled with minority classes, reflecting its bias toward majority categories. In Figure 7(b), the concatenated

DenseNet121 + VGG16 model converged faster and maintained validation accuracy at 0.87, with validation loss consistently below training loss. The narrow training–validation gap demonstrates strong generalization capacity. This is reinforced by higher test recall (87%) and AUC (0.97), showing that concatenation reduced false negatives in clinically critical classes such as COVID-19 and Tuberculosis. Figure 7(c) illustrates the VGG16 + ResNet50 model, which achieved gradual improvement in validation accuracy approaching 0.90, while validation loss steadily decreased with only minor mid-epoch fluctuations. The close alignment of training and validation curves indicates reliable convergence without overfitting. On the test set, this model also reached a recall of 87% and an AUC of 0.97, confirming its robustness and balanced performance across classes.

Taken together, these results highlight that while single backbones can achieve stable convergence, concatenated models provide faster learning, improved generalization, and superior sensitivity to minority classes. This directly addresses the imbalance challenge and enhances the clinical reliability of automated chest X-ray classification. Overall, the accuracy–loss curves demonstrate that all evaluated models converged stably, maintained small training–validation gaps, and avoided overfitting. Importantly, concatenated models not only delivered higher recall and AUC but also preserved robust generalization throughout training, confirming their superiority over single backbones in multiclass chest X-ray classification.

As a final step, Grad-CAM visualizations were generated to interpret the spatial focus of the models during chest X-ray classification. These heatmaps not only provide insight into the regions most influential for prediction but also allow evaluation of whether the models rely on clinically relevant features.

For this analysis, the three best-performing models DenseNet121, DenseNet121 + VGG16, and VGG16 + ResNet50, were selected based on their consistent performance in accuracy, F1-score, and AUC-ROC. The visualization results are presented in Figure 8.

(a) DenseNet121 (b) DenseNet121 + VGG16

(c) VGG16 + ResNet50

Fig 8. Grad-CAM visualizations of the best-performing models: (a) DenseNet121, (b) DenseNet121 + VGG16, (c) VGG16 + ResNet50. The first row shows correct predictions, while the second row shows misclassified cases

Figure 8 presents Grad-CAM visualizations from the three best-performing models applied to chest X-ray images labeled as Normal. In Figure 8(a), the single DenseNet121 model primarily highlighted the central and upper lung fields, but several activations extended outside the lungs, such as the chest wall and shoulders, which led to misclassifications such as tuberculosis for a normal case, indicating that the single backbone is more prone to attending irrelevant regions.

In Figure 8(b), the VGG16 + ResNet50 model displayed a more centralized and symmetrical activation pattern, with stronger focus on the lung parenchyma, particularly posterior regions; this pattern was relatively stable and clinically consistent, although a misclassification as COVID-19 occurred. In Figure 8(c), the DenseNet121 + VGG16 model exhibited the most evenly distributed activations, concentrating on the hilar and lower lung zones that are clinically relevant, and despite one misclassification as COVID-19, the model consistently prioritized lung-related areas rather than peripheral artifacts. Overall, these findings demonstrate that concatenated models not only achieved higher accuracy but also produced more clinically meaningful attention compared to single backbones, thereby reducing reliance on irrelevant regions and reinforcing their potential as reliable decision-support tools in medical imaging.

4. Conclusion

Through a series of repeated experiments, the proposed feature-concatenated dual-backbone combined with Multiclass Focal Loss consistently improved macro-F1 and macro-AUC while reducing false negatives in minority classes. These gains were stable across multiple runs and further validated by Grad-CAM, which confirmed that the models focused on clinically relevant lung regions.

This study has demonstrated that concatenating CNN architectures (DenseNet121, VGG16, and ResNet50) with Multiclass Focal Loss provides an effective solution for multiclass lung

disease classification under imbalanced conditions. The approach improved sensitivity and AUC, particularly for minority classes such as Tuberculosis and COVID-19, thereby reducing clinically critical false negatives. Focal Loss emphasized underrepresented categories without destabilizing training, while concatenated architectures accelerated convergence and enhanced generalization. Grad-CAM visualizations confirmed that the models attended to clinically relevant lung regions, strengthening both interpretability and clinical trust.

Despite these promising results, the study is limited by the use of a relatively small, publicly available dataset, restriction to X-ray modality, and increased computational demands of concatenated models. Future studies should validate performance on larger and more diverse datasets, apply domain adaptation to improve generalizability, and evaluate integration into clinical workflows with radiologists. Extending interpretability with quantitative uncertainty measures and improving efficiency through pruning, compression, or knowledge distillation will be important for real-world deployment. Strengthening these aspects will further support the use of deep learning as a trustworthy clinical decision-support tool for early and accurate lung disease diagnosis.

Acknowledgment

This work was supported by the Department of Computer Science and Electronics, Universitas Gadjah Mada under the Publication Funding Year 2025.

Data and Software Availability Statements

The data supporting the findings of this study are openly available on Kaggle at https://www.kaggle.com/datasets/jtiptj/chest-xray-pneumoniacovid19tuberculosis.

References

[1] WHO, “Pneumonia.” Accessed: Sep. 02, 2024. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/pneumonia

[2] WHO, *Global tuberculosis report 2023*. 2023.

[3] WHO, “World Health Organization 2023 data.who.int.” Accessed: Sep. 02, 2024. [Online]. Available: https://data.who.int/dashboards/covid19/more-resources

[4] G. D. Rubin *et al.*, “The role of chest imaging in patient management during the covid-19 pandemic: A multinational consensus statement from the fleischner society,” *Radiology*, vol. 296, no. 1, pp. 172–180, 2020, doi: 10.1148/radiol.2020201365.

[5] 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.

[6] I. D. Apostolopoulos and T. A. Mpesiana, “Covid-19: automatic detection from X-ray images utilizing transfer learning with convolutional neural networks,” *Phys. Eng. Sci. Med.*, vol. 43, no. 2, pp. 635–640, 2020, doi: 10.1007/s13246-020-00865-4.

[7] 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.

[8] A. Abbas, M. M. Abdelsamea, and M. M. Gaber, “Classification of COVID-19 in chest X-ray images using DeTraC deep convolutional neural network,” *Appl. Intell.*, vol. 51, no. 2, pp. 854–864, 2021, doi: 10.1007/s10489-020-01829-7.

[9] M. Mamalakis *et al.*, “DenResCov-19: A deep transfer learning network for robust automatic classification of COVID-19, pneumonia, and tuberculosis from X-rays,” *Comput. Med. Imaging Graph.*, vol. 94, 2021, doi: 10.1016/j.compmedimag.2021.102008.

[10] N. V Chawla, K. W. Bowyer, L. O. Hall, and W. S. Philip Kegelmeyer, “synthetic minority over-sampling Technique,” *J Artif Intell Res*, vol. 16, p. 16, 2018.

[11] M. Buda, A. Maki, and M. A. Mazurowski, "A systematic study of the class imbalance problem in convolutional neural networks," *Neural Networks*, vol. 106, pp. 249–259, 2018, doi: https://doi.org/10.1016/j.neunet.2018.07.011.

[12] H. He, Y. Bai, E. A. Garcia, and S. Li, "ADASYN: Adaptive synthetic sampling approach for imbalanced learning," in *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, Jun. 2008, pp. 1322–1328. doi: 10.1109/IJCNN.2008.4633969.

[13] A. Fernández, S. García, F. Herrera, and N. V Chawla, "SMOTE for Learning from Imbalanced Data: Progress and Challenges," *J. Artif. Intell. Res.*, vol. 61, pp. 863–905, 2018.

[14] B. Zadrozny, "Learning and evaluating classifiers under sample selection bias," in *Proceedings of the Twenty-First International Conference on Machine Learning*, in ICML '04. New York, NY, USA: Association for Computing Machinery, 2004, p. 114. doi: 10.1145/1015330.1015425.

[15] T. Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal Loss for Dense Object Detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 2, pp. 318–327, 2020, doi: 10.1109/TPAMI.2018.2858826.

[16] Y. Cui, M. Jia, T. Y. Lin, Y. Song, and S. Belongie, "Class-balanced loss based on effective number of samples," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2019-June, pp. 9260–9269, 2019, doi: 10.1109/CVPR.2019.00949.

[17] J. Singh *et al.*, "Batch-balanced focal loss: a hybrid solution to class imbalance in deep learning.," *J. Med. Imaging*, vol. 10, no. 5, pp. 051809-, Jun. 2023, doi: 10.1117/1.jmi.10.5.051809.

[18] A. Hamza *et al.*, "COVID-19 classification using chest X-ray images: A framework of CNN-LSTM and improved max value moth flame optimization," *Front. Public Heal.*, vol. 10, 2022, doi: 10.3389/fpubh.2022.948205.

[19] M. Nahiduzzaman *et al.*, "Parallel CNN-ELM: A multiclass classification of chest X-ray images to identify seventeen lung diseases including COVID-19," *Expert Syst. Appl.*, vol. 229, no. PA, Nov. 2023, doi: 10.1016/j.eswa.2023.120528.

[20] D. Kuzinkovas and S. Clement, "The Detection of COVID-19 in Chest X-rays Using Ensemble CNN Techniques," *Inf.*, vol. 14, no. 7, 2023, doi: 10.3390/info14070370.

[21] N. Hilmizen, A. Bustamam, and D. Sarwinda, "The Multimodal Deep Learning for Diagnosing COVID-19 Pneumonia from Chest CT-Scan and X-Ray Images," *2020 3rd Int. Semin. Res. Inf. Technol. Intell. Syst. ISRITI 2020*, pp. 26–31, 2020, doi: 10.1109/ISRITI51436.2020.9315478.

[22] M. Rahimzadeh and A. Attar, "A modified deep convolutional neural network for detecting COVID-19 and pneumonia from chest X-ray images based on the concatenation of Xception and ResNet50V2," *Informatics Med. Unlocked*, vol. 19, p. 100360, 2020, doi: 10.1016/j.imu.2020.100360.

[23] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM," *Rev. do Hosp. das Clnicas*, vol. 17, pp. 331–336, 2016, [Online]. Available: http://arxiv.org/abs/1610.02391

[24] H. Wang *et al.*, "Score-CAM: Score-weighted visual explanations for convolutional neural networks," *IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Work.*, vol. 2020-June, pp. 111–119, 2020, doi: 10.1109/CVPRW50498.2020.00020.

[25] A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, "Grad-CAM++: Generalized Gradient-Based Visual Explanations for Deep Convolutional Networks," in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, Mar. 2018, pp. 839–847. doi: 10.1109/WACV.2018.00097.

[26] B. H. M. van der Velden, H. J. Kuijf, K. G. A. Gilhuijs, and M. A. Viergever, "Explainable artificial intelligence (XAI) in deep learning-based medical image analysis," *Med. Image Anal.*, vol. 79, p. 102470, 2022, doi: 10.1016/j.media.2022.102470.

[27] H. Chen, C. Gomez, C. M. Huang, and M. Unberath, "Explainable medical imaging AI needs human-centered design: guidelines and evidence from a systematic review," *npj Digit. Med.*, vol. 5, no. 1, 2022, doi: 10.1038/s41746-022-00699-2.

[28] L. Wang, C. Wang, Z. Sun, S. Cheng, and L. Guo, "Class Balanced Loss for Image Classification," *IEEE Access*, vol. 8, pp. 81142–81153, 2020, doi: 10.1109/ACCESS.2020.2991237.

[29] JtiptJ, "Chest X-Ray (Pneumonia,Covid-19,Tuberculosis)," Kaggle. Accessed: Apr. 03, 2024. [Online]. Available: https://www.kaggle.com/datasets/jtiptj/chest-xray-pneumoniacovid19tuberculosis

[30] J. C. Chien, J. Der Lee, C. S. Hu, and C. T. Wu, "The Usefulness of Gradient-Weighted CAM in Assisting Medical Diagnoses," *Appl. Sci.*, vol. 12, no. 15, 2022, doi: 10.3390/app12157748.

[31] M. Kuhn and K. Johnson, *Applied predictive modeling*, 1st ed. Springer New York, NY, 2013. doi: 10.1007/978-1-4614-6849-3.

[32] M. Bekkar, H. K. Djemaa, and T. A. Alitouche, "Evaluation Measures for Models Assessment over Imbalanced Data Sets," *J. Inf. Eng. Appl.*, vol. 3, no. 10, pp. 27–38, 2013, [Online]. Available: http://www.iiste.org/Journals/index.php/JIEA/article/view/7633