Journal of Applied Informatics and Computing (JAIC)
Vol.9, No.6, December 2025, pp. 3383~3393
e-ISSN: 2548-6861

Vision Transformer for Pneumonia Classification with Grad-CAM Explainability

Immanuel Julius Darmawan [1], Catur Supriyanto [2]*
Department of Informatics Engineering, Universitas Dian Nuswantoro, Semarang, Indonesia
111202213983@mhs.dinus.ac.id [1], catur.supriyanto@dsn.dinus.ac.id [2]

Article Info

Article history:

Received 2025-10-21
Revised 2025-11-17
Accepted 2025-11-26

Keyword:

Chest X-Ray,
Grad-CAM,
Pneumonia Classification,
Vision Transformer.

ABSTRACT

Pneumonia is still one of the main causes of death around the world, especially in kids and older people. To lower the death rate, early and accurate diagnosis is very important. Chest X-ray (CXR) imaging is widely used for this purpose, but manual reading of CXR images can be time-consuming and may lead to differences in interpretation between observers. To address this problem, this study presents a pneumonia classification model based on the Vision Transformer (ViT) architecture combined with Gradient-weighted Class Activation Mapping (Grad-CAM) to make the model's decisions more interpretable. The model was trained on a publicly available CXR dataset with 5,863 images that were split into Normal and Pneumonia classes, using a 70:15:15 split for training, validation, and testing. The ViT model achieves an accuracy of 96.41% on the test set and a high recall for pneumonia cases, while class weighted loss helps to maintain more balanced predictions between the two classes. The Area Under the Curve (AUC) of 0.975 indicates strong discrimination between pneumonia-positive and normal samples. Grad-CAM visualizations, supported by a randomization test and occlusion analysis, provide an initial qualitative view of the lung regions that influence the model's predictions and often overlap with radiologically plausible areas. However, the heatmaps have not been formally evaluated by radiologists, and the correspondence between highlighted regions and pneumonia consolidation patterns has not yet been quantitatively validated. Therefore, the proposed ViT Grad-CAM framework should be regarded as an exploratory step toward explainable pneumonia classification on chest X-rays rather than a system that is ready for clinical deployment.

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

I. INTRODUCTION

Pneumonia remains a major global health burden, particularly in developing countries, and continues to be among the leading causes of illness and death in children and the elderly [1]. Early detection and timely treatment are essential to reduce pneumonia related illness and death. Among available imaging modalities, chest X-ray remains the most common and cost effective diagnostic tool in clinical practice [2]. However, manual CXR interpretation requires experienced radiologists and is prone to inter-observer variability, diagnostic fatigue, and limited specialist availability in many healthcare facilities [3]. So it is important to build an automated decision support system based on artificial intelligence to make pneumonia diagnosis more consistent and efficient.

Deep learning particularly convolutional neural networks (CNN) has significantly advanced CXR-based pneumonia detection. Yen and Tsao proposed a lightweight CNN for multi-class classification on a large CXR dataset and reported test accuracies of 97.10% for pneumonia and 97.86% for normal cases, showing that carefully designed CNN can be both accurate and computationally efficient [4]. Manickam et al. [5] evaluated several transfer learning CNN architectures, such as Inception-V3 and ResNet variants, and demonstrated that these models can achieve high diagnostic performance for pneumonia classification on CXR images. Other studies

have reported more moderate performance. Mardianto et al. [6] obtained an overall accuracy of 91% for binary classification of normal versus pneumonia using a CNN model. Usman et al. [7] reported a lower accuracy of about 79% when the model was trained without extensive data augmentation, illustrating that performance can drop substantially on more challenging and imbalanced data. Lestari et al. [8] investigated ResNet50V2 and InceptionV3-based CNN pipelines for pneumonia detection and reported validation accuracies in the range of 0.90–0.94 depending on the architecture and training configuration. These results indicate that CNN based methods form a strong yet heterogeneous baseline, with performance heavily influenced by dataset characteristics, class balance, and training strategy.

A key limitation of CNNs is their reliance on local receptive fields, which restricts their ability to capture long-range spatial relationships across the entire lung field. Pneumonia often presents with diffuse or subtle patterns distributed across multiple lung regions, making global context particularly important for accurate classification [5]. Moreover, CNN-based systems are frequently trained on imbalanced datasets where pneumonia cases substantially outnumber normal cases. Chang and Huang [9] showed that the classification performance for lung opacities in CXR could vary substantially F1-scores between 0.60 and 0.78 depending on the balance of positive and negative samples, highlighting that CNN models are sensitive to skewed data distributions.

The Vision Transformer (ViT) offers a promising alternative to purely convolutional approaches by representing images as patch sequences and applying self-attention to capture global relationships across patches. This architecture allows ViT models to reason over long-range context without being restricted by local kernels. Manzari et al. [10] introduced MedVit, demonstrating that ViT-based architectures can achieve robust performance across several medical imaging tasks. Tyagi et al. [11] showed that a ViT model outperformed CNN and VGG16 baselines on CXR pneumonia detection, achieving training accuracy around 96%, although validation accuracy remained lower around 86.38%, indicating potential challenges in generalization. Singh et al. [12] reported ViT-based framework for pneumonia detection can achieve accuracy in the high-ninety range on a curated CXR datasets. Ko et al. [13] demonstrated that ViT variants can achieve competitive performance when optimized properly. These findings suggest that ViT can surpass CNN baselines in certain conditions, especially when capturing global structures is important.

However, most ViT-based studies in this domain focus primarily on maximizing classification accuracy, with limited emphasis on model explainability. In many publications, heatmaps are shown only as qualitative illustrations, without systematic evaluation of whether the highlighted regions truly correspond to radiologically meaningful lesions. This lack of rigorous explanation analysis raises concerns regarding the trustworthiness and clinical adoption of AI-based diagnostic systems. In medicine, high accuracy alone is not sufficient, clinicians also need to understand why a model produces a particular decision before they can rely on it in practice.

Explainable Artificial Intelligence (XAI) techniques have been introduced to address this issue by providing visual or textual explanations for model predictions. Gradient-weighted Class Activation Mapping (Grad-CAM) is one of the most widely used XAI methods and can generate heatmaps indicating the image regions that contribute most strongly to a model's prediction [14]. Nonetheless, several studies have shown that Grad-CAM heatmaps can sometimes emphasize irrelevant structures or artifacts, so careful validation is required to avoid misleading interpretations. Suara et al. [15] explicitly questioned whether Grad-CAM is reliably explainable for medical images and emphasized the need to check the alignment between heatmaps and true pathological regions on the image. Purwono et al. [16] highlighted that many XAI studies in medical imaging do not include rigorous quantitative criteria to assess whether the visual explanations faithfully reflect the model's internal behavior. Chen et al. [17] further argued that explainable medical-imaging AI systems must be designed in a human-centered manner and should provide explanations that deliver concrete benefits for clinicians in real clinical workflows.

Based on these gaps, this study proposes a Vision Transformer based pneumonia classification model equipped with Grad-CAM for visual explanation. The model is developed to classify CXR images into Normal and Pneumonia categories while generating interpretable heatmaps to highlight regions most influential in the diagnostic decision. To address concerns related to explainability, this study incorporates additional validation procedures, including parameter randomization sanity checks and occlusion sensitivity analysis, to assess whether the generated heatmaps truly depend on learned model parameters rather than input artifacts. These procedures are used to verify that the model's explanations remain focused on relevant lung regions. This study aims to develop a ViT-based pneumonia classifier with reliable Grad-CAM visualizations and competitive performance compared to CNN baselines. Through the integration of ViT and systematic XAI evaluation, this work contributes to more interpretable and methodologically robust approaches for CXR image analysis.

II. METHOD

A. Dataset

The dataset used in this study originates from a publicly available collection of chest X-ray (CXR) images on the Kaggle platform titled Chest X-Ray Images (Pneumonia). This dataset was developed by the research team at Guangzhou Women and Children's Medical Center and was first published by Kermany et al. s first published by [18] as part of their study on medical image based lung disease detection. It has been widely used in pneumonia classification

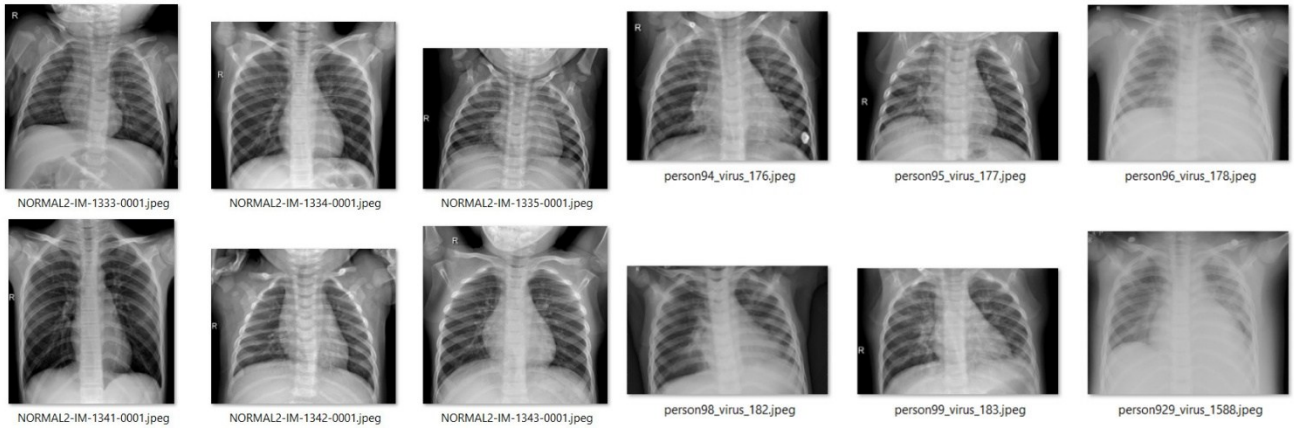

Figure 1. Normal and Pneumonia Chest X-Ray

TABLE I
DATASET DISTRIBUTION

Split	Normal	Pneumonia
Train	1104	2973
Val	228	656
Test	247	616
Overall	1579	4245

research due to its high quality, large sample size, and medically validated labeling [18].

In this study, a total of 5,824 images are used, the detailed dataset distribution is summarized in Table I. As shown in Table I, the number of Pneumonia images is considerably higher than the number of Normal images overall ratio 1:2.69, indicating a clear class imbalance that must be addressed during model training. Example images from each class are presented in Figure 1, illustrating visual differences between normal lungs and pneumonia lungs.

B. Proposed Method

This study aims to develop a classification for chest X-ray images using a Vision Transformer (ViT) model, and to complement the predictions with interpretable visual explanations based on Gradient-weighted Class Activation Mapping (Grad-CAM). The flowchart of the proposed method is illustrated in Figure 2 and consists of the following main stages:

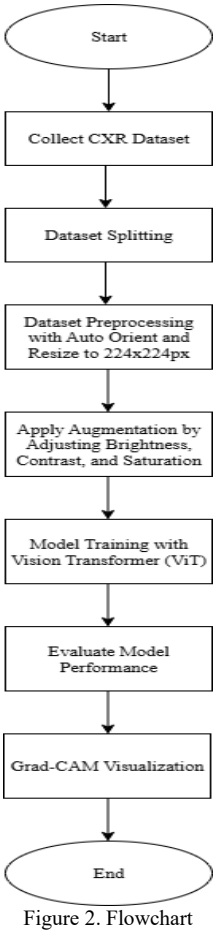

Figure 2. Flowchart

1. Data Collection: The dataset used is publicly available on Kaggle and consists of chest X-ray images of pneumonia. This dataset contains chest X-ray images from patients with normal conditions and pneumonia, which are used to train and test the model.

2. Dataset Splitting: The full dataset is divided using a simple stratified hold-out split with 70% of the images for training, 15% for validation, and 15% for testing, as reported in Table I. No k-fold cross-validation is applied in this study. This choice is motivated by two considerations to keeping a fixed and completely independent test set for the final evaluation of the model, and avoiding the substantial computational cost of repeatedly training a Vision Transformer with Grad-CAM analyses on limited GPU resources. To partially compensate for the absence of cross-validation, early stopping based on the validation set is used to reduce overfitting, and the class distribution is preserved across all splits.

3. Dataset Preprocessing: All images are automatically re-oriented (auto orientation) to ensure a

consistent view and then resized to 224×224 pixels to match the input size required by the Vision Transformer architecture [19]. This step is important because it ensures that all data is the same size, allowing the model to consistently find features.

4. Dataset Augmentation: To increase visual diversity and reduce overfitting while preserving anatomical structures, a lightweight color-based augmentation is applied to training images brightness ±10%, contrast ±10%, and saturation ±5%. No aggressive geometric transformations are used to avoid unrealistic deformation of the thoracic anatomy.

5. Training the model with Vision Transformer (ViT): This stage is the most important part of the research. The ViT model is used because it can learn global spatial relationships between parts of an image using self attention. ViT, on the other hand, divides images into patches of the same size and then converts them into vector tokens that are processed by a transformer encoder in several layers [20]. The platform that used to trained the model is in Google Colab with a NVIDIA T4 GPU.

6. Analyzing the Model: After training was complete, the model's performance was evaluated using several metrics, namely accuracy, precision, recall, F1-score, and Area Under Curve (AUC). The confusion matrix also shows how many predictions for each class were correct and how many were incorrect. This test helps determine how well the model can consistently recognize patterns of pneumonia.

7. Grad-CAM Illustration: The final step is to use Gradientweighted Class Activation Mapping (Grad-CAM) is applied to the ViT model to generate heatmaps that highlight regions contributing most strongly to NORMAL or PNEUMONIA predictions [21]. To verify the reliability of these explanations, two sanity checks are performed model-parameter randomization, where Grad-CAM from the trained ViT is compared with Grad-CAM from a randomly re-initialized ViT of the same architecture, and occlusion sensitivity, where small patches of the input image are systematically masked to measure the drop in prediction probability. These analyses help ensure that the explanations truly depend on the learned parameters and clinically meaningful lung regions rather than on spurious artifacts.

C. Design of Experiment

This section describes the experimental setup, computing environment, training configuration, and strategy for handling class imbalance. The main objective is to ensure that the experiment is reproducible and that the results are accurately measurable.

1. Experimental Environment: All experiments are conducted on Google Colab using an NVIDIA Tesla T4 GPU (16 GB VRAM). Training and inference are executed on the GPU, while data loading and augmentation are performed on the CPU using a PyTorch DataLoader with num_workers=4 and pin_memory=True to maintain an efficient data pipeline.

TABLE II
EXPERIMENTAL ENVIRONMENT

Component	Specification
Platform	Google Collab
GPU	NVIDIA Tesla T4 (16GB VRAM)
Framework	PyTorch 2.0
Programming Language	Python 3.10
Main Libraries	timm 1.0.9, TorchVision, Albumentations 1.4.15, OpenCV 4.10, scikit-learn 1.5.2, torchmetrics 1.4.0, einops 0.8.0, pytorch-grad-cam

The implementation uses Python 3.10 and PyTorch 2.0, along with the timm library for the Vision Transformer and additional packages for visualization and explainability. The main hardware and software specifications are summarized in Table II.

2. Model Architecture and Training Configuration: The classification model used in this study is the vit_small_patch16_224 variant of the Vision Transformer provided by the timm library. The Vision Transformer (ViT) was chosen because it can use a self-attention mechanism to capture the global context between patches. This has been shown to work well for classifying medical images [19]. In this architecture, each CXR image is divided into 16x16 non-overlaping patches that are embedded into a token sequence. A learnable class toen is prepended to be sequece, which is then processed by stacked Transformer encoder blocks consisting of multihead self attention and feed forward layers with residual connections and layer normalization. The final class token representation is fed into a fully connected layer to produce the Normal or Pneumonia class. The ViT backbone is initialized with ImageNet-1k pretrained weights and fine-tuned on the pneumonia dataset used in this study. Model training uses a batch size of 32 and the AdamW optimizer with an initial learning rate of 2×10^{-4} and weight decay of 1×10^{-4}. A short warm-up phase is followed by cosine decay, and training runs for up to 30 epochs with early stopping (patience = 6) to limit overfitting. The key hyperparameters are summarized in Table III.

3. Handling class imbalance: As shown in Table I, with normal class substantially fewer than pneumonia class. Ignoring this imbalance may bias the model toward the majority class and degrade performance on normal cases. To mitigate this issue, a class weighted cross entropy loss is used. With training counts of 1,104 normal and 2,973 pneumonia samples, the computed class weights are 1.846 and 0.686, respectively Table III, providing a stronger gradient contribution for the minority class. A label-smoothing factor of 0.05 is also applied to improve calibration. Alternative balancing strategies such as oversampling, SMOTE, or class specific augmentation were not employed because they risk

producing unrealistic or duplicated CXR patterns, especially in medical datasets. Instead, this study adopts a conservative approach by preserving the original data distribution, correcting imbalance at the loss level through class-weighting, and applying only mild symmetric augmentation to both classes.

TABLE III
TRAINING PARAMETERS

Parameter	Value
Architecture	ViT (vit_small_patch16_224), timm
Input Size	224 x 224 pixels
Pretrained Weights	ImageNet-1k
Batch Size	32
Epochs (max)	30
Early Stopping	Patience = 6
Optimizer	AdamW
Learning Rate	2e-4
Weight Decay	1e-4
Loss Function	Class Weighted Cross Entropy, label smoothing 0.05
Augmentation	Brightness 10%, Contrast 10%, Saturation 5%
DataLoader	num_workers = 4 , pin_memory = True

TABLE IV
MODEL PERFORMANCE EVALUATION RESULTS

	Precision	Recall	F1-Score
Normal	0.9655	0.9069	0.9353
Pneumonia	0.9636	0.9870	0.9751
Accuracy			0.9641
Macro Avg	0.9645	0.9469	0.9552
Weighted Avg	0.9641	0.9641	0.9637

D. Performance Metrics

This study evaluated the model's performance using both quantitative and qualitative methods. Quantitative evaluation checks how well the model sorts data, while qualitative evaluation uses visual aids to show how easy it is to understand what the model says.The quantitative evaluation employed several established metrics, including Accuracy, Precision, Recall, F1-Score, and Area Under the Curve (AUC), all derived from the Confusion Matrix. The math formulas for these metrics are below:

1. *Accuracy:* Accuracy measures the proportion of correctly classified samples compared to the total number of samples. It represents the overall effectiveness of the model in distinguishing between the two classes. The formula is shown in Equation (1)

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

2. *Precision:* Precision, also known as Positive Predictive Value, indicates how many of the samples predicted as positive (pneumonia) are actually correct. A high precision means fewer false positives. The formula is shown in Equation (2)

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

3. *Recall:* Recall measures the ability of the model to correctly identify positive samples out of all actual positives. It reflects the completeness of the model's positive predictions. The formula is shown in Equation (3)

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

4. *F1-Score:* The F1-Score is the harmonic mean of Precision and Recall, balancing both metrics into a single measure of performance. It is particularly useful when the dataset has class imbalance. The formula is shown in Equation (4):

$$F1 = \frac{Precission\ x\ Recall}{Precission + Recall} \quad (4)$$

5. *Area Under Curve (AUC):* AUC evaluates how well the model distinguishes between positive and negative classes at various threshold settings. It is derived from the Receiver Operating Characteristic (ROC) curve, which plots True Positive Rate (TPR) versus False Positive Rate (FPR). A higher AUC indicates better discriminative capability.

The Confusion Matrix was also used to show how predicted and real labels were related. This helped find mistakes that were different for each class. The researchers used Gradient-weighted Class Activation Mapping (Grad-CAM) as a qualitative way to understand the results, along with the numerical evaluation. Grad-CAM makes heatmaps that show which parts of the chest X-ray image were most important to the model when it made its prediction. This allows them to check whether the model is focuses on clinically relevant parts of the lungs [22].By using both quantitative and qualitative evaluations, researchers can evaluate the proposed Vision Transformer model not only on how accurate its predictions are, but also on how easy it is to understand, making sure that both performance and clinical transparency are maintained.

III. RESULTS AND DISCUSSION

In this section, the researchers will describe and analyze all results obtained from the pneumonia classification experiment using the Vision Transformer (ViT) model and the Grad-CAM visualization explainability.

A. Model Evaluation Performance

This subsection shows the results from the Vision Transformer (ViT) model that was trained to tell the difference between normal and pneumonia using chest X-ray images. The researchers used Accuracy, Precision, Recall,

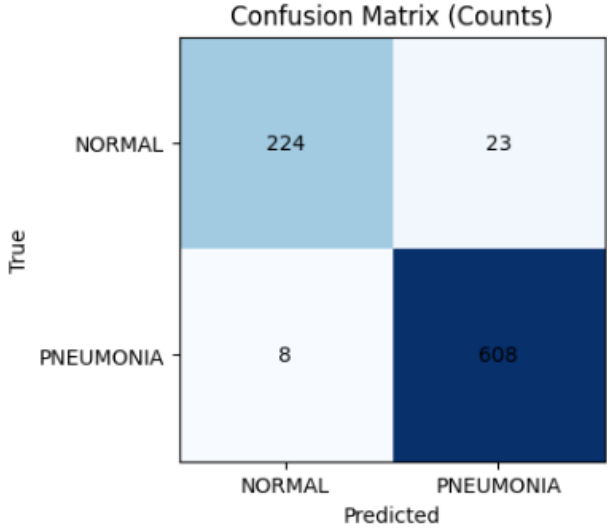

Figure 3. Confusion Matrix

F1-Score, and Area Under the Curve (AUC) to see how well the model worked. The evaluation was performed on the unseen test set to evaluate the generalization ability of the trained model. The evaluation results are in the Table IV.

Although the dataset is imbalanced, the model maintained consistent performance across both classes. This stability was achieved through the use of a class weighting strategy during training, which assigns greater importance to the minority class to counteract the imbalance. Specifically, the computed class weights were 1.846 for Normal and 0.686 for Pneumonia, ensuring that misclassifications in the underrepresented Normal class were penalized more heavily.

By integrating these weights into the cross entropy loss function, the Vision Transformer (ViT) model was guided to learn balanced decision boundaries, thus preventing bias toward the majority class [23]. This strategy is consistent with recent findings in medical imaging research, where weighted loss functions are widely applied to mitigate skewed class distributions and maintain diagnostic reliability.

As a result, the recall value of 98.70% demonstrates that the model effectively identified pneumonia positive cases while maintaining strong precision on normal samples confirming the success of the class balancing mechanism.

The Confusion Matrix shown in Figure 3 further illustrates the prediction distribution between the two classes. The ViT model correctly classified the vast majority of samples, with only a few minor misclassifications between Normal and Pneumonia, emphasizing its robustness in distinguishing pathological features in chest X-rays.

The confusion matrix in Figure 3 shows the distribution of predictions for both the Normal and Pneumonia classes. Out of 247 Normal test images, 224 were correctly classified, while 23 were misclassified as Pneumonia. On the other hand, the model correctly identified 608 out of 616 Pneumonia images, with only 8 misclassified as Normal.

These results demonstrate that the Vision Transformer (ViT) model performs reliably across both classes despite the

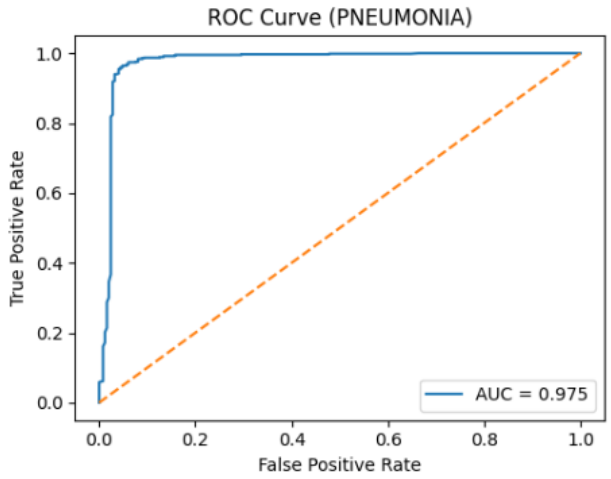

Figure 4. Receiver Operating Characteristic (ROC) Curve

class imbalance in the dataset. The relatively small number of misclassifications indicates that the model effectively distinguishes the visual patterns of healthy and infected lungs, especially considering the subtle opacity differences often present in chest X-ray images.

Furthermore, the confusion matrix highlights the model's strong sensitivity to pneumonia cases an essential aspect for medical diagnosis where false negatives can have serious clinical implications. The high true positive rate for the Pneumonia class aligns with the recall value shown in Table IV, confirming that the model prioritizes correct identification of disease cases without losing much precision.

To further evaluate the model's discriminative capability, the Receiver Operating Characteristic (ROC) curve and Area Under the Curve (AUC) were analyzed. The ROC curve provides a graphical representation of the trade off between sensitivity and specificity across different classification thresholds, while the AUC value quantifies the overall ability of the model to separate positive and negative classes.

The ROC curve shown in Figure 4 illustrates the discriminative performance of the Vision Transformer (ViT) model on the Pneumonia class. The model achieved an AUC score of 0.975, indicating excellent separability between pneumonia positive and normal chest X-ray samples. The curve demonstrates that the ViT maintained a consistently high true positive rate even at low false positive rates, reflecting strong sensitivity and specificity.

This performance confirms that the model is capable of distinguishing pneumonia lungs with high reliability across varying decision thresholds. A high AUC value above 0.9 is generally considered outstanding in medical image analysis, as it signifies the model's ability to make accurate predictions even under threshold variations [24]. Therefore, the obtained AUC of 0.975 strengthens the conclusion that the ViT model provides robust and discriminative classification results suitable for clinical decision support systems.

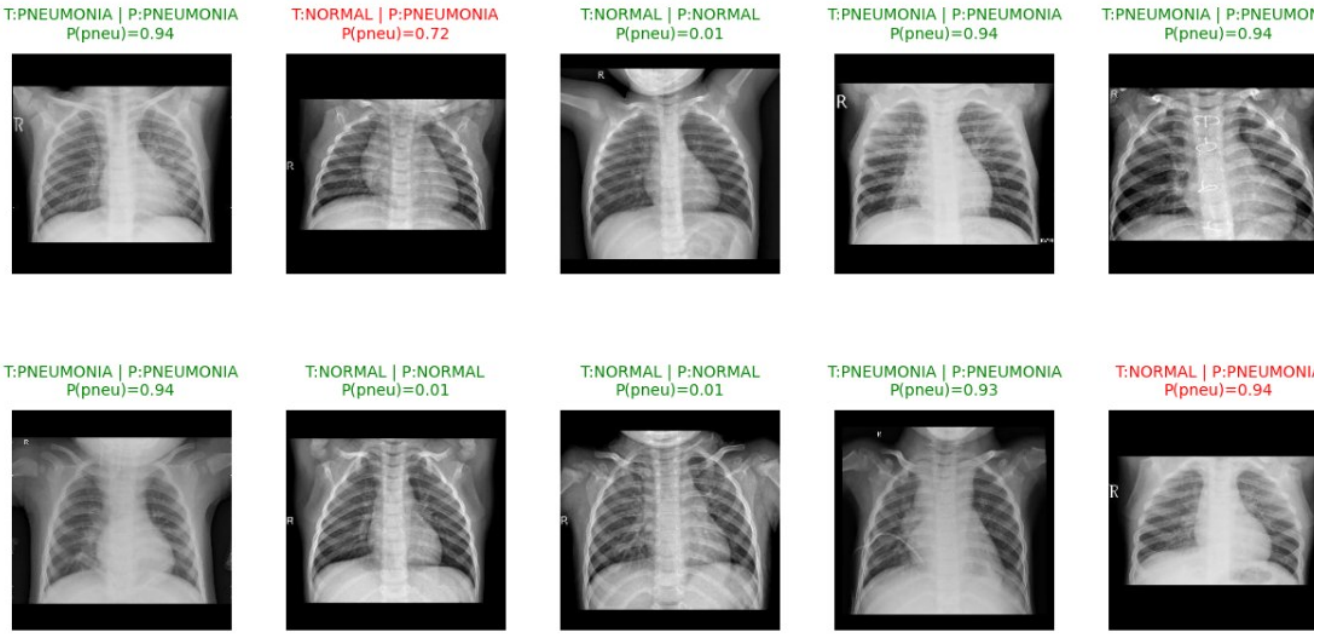

Figure 1. Predicting on Test Set

B. Classification Test and Prediction Results

This section describes how well the Vision Transformer (ViT) model performed on the test dataset. The model was trained to classify chest X-ray images into two categories Normal and Pneumonia based on its prediction outcomes and confidence scores. Each image in the test set was processed using the trained ViT model, which produced both the predicted label and a probability score representing the model's confidence. These scores are highly valuable in medical classification tasks, as they indicate the reliability of each prediction and can assist in making informed diagnostic decisions.

The results show that the ViT model achieved consistent and confident predictions throughout the test set. Confidence scores were generally above 0.90, and the majority of samples were correctly classified. For instance, normal lung images without visible opacities were accurately recognized with an high confidence. Similarly, pneumonia cases showing dense or patchy opacities were detected with strong confidence, indicating that the model effectively learned the visual characteristics of infected lungs.

However, there are some misclassifications occurred. Some normal chest X-rays were mistakenly predicted as pneumonia with moderate confidence (e.g., P(pneu) ≈ 0.72). Upon closer examination, these images contained faint or low contrast opacities near the lower lobes, which may have caused the model to misinterpret them as potential infection. Conversely, a few pneumonia images were predicted as normal when the infection appeared mild or partially obscured by other anatomical structures. These errors highlight the sensitivity of the Vision Transformer's attention mechanism to subtle texture and contrast variations a challenge that even trained radiologists can face when interpreting chest X-rays.

The model also displayed a well balanced confidence distribution. When the model was correct, its confidence was consistently high, while incorrect predictions were typically accompanied by moderate confidence. This indicates that the ViT model exhibits good self awareness in its decision making process it becomes cautious when uncertain and confident when sure.

From a clinical perspective, this behavior is desirable because the model prioritizes sensitivity detecting pneumonia whenever there is doubt. In medical diagnosis, false negatives are more critical than false positives, as missing a true pneumonia case can delay proper treatment. In contrast, false positives can be clarified through follow up tests.

Figure 2. Grad-CAM Visualization Results for Normal Samples

Therefore, the model's slight bias toward predicting pneumonia demonstrates a clinically appropriate balance between recall and precision. Figure 5 presents several representative examples from the test set, showing the true and predicted labels along with their corresponding confidence values. Green texts indicate correctly classified samples, while red texts denote misclassified ones. These examples illustrate that the ViT model remains highly confident in clear case predictions but exhibits caution in

more ambiguous instances, reflecting sound clinical interpretability.

C. Grad-CAM Visualization and Explainability Analysis

This subsection analyzes the interpretability of the Vision Transformer (ViT) model using Gradient-weighted Class Activation Mapping (Grad-CAM). Grad-CAM is employed to highlight image regions that contribute most to the model's prediction, providing a qualitative view of whether the network attends to lung areas that are visually plausible for normal or pneumonia category. The implementation is based on the PyTorch Grad-CAM library, using the final normalization block of the ViT (model.blocks[-1].norm1) as the target layer. Each Grad-CAM map is overlaid on the original chest X-ray, where warmer colors (red/yellow) indicate higher contribution to the decision and cooler colors (green/blue) indicate lower contribution.

Figure 6 presents Grad-CAM visualizations for three normal test images that were correctly classified. In these examples, the attention patterns are generally distributed across the thoracic cavity, with noticeable activations around the mediastinum, clavicles, and rib structures, and more diffuse responses within the lung fields. This suggests that, for clearly normal cases, the model does not concentrate on a single focal opacity but rather integrates information from broader anatomical regions and texture patterns that are compatible with a normal appearance. Some hotspots also appear along bony structures and at the upper chest boundary, which indicates that the model partially relies on contextual and structural cues outside the central lung regions.

Figure 7 shows Grad-CAM maps for three correctly classified pneumonia images. Compared with the normal examples, the activation tends to be more localized and intense in parts of the lung fields where radiographic opacities are present. In particular, several cases exhibit high-activation areas in the middle and lower lung zones, overlapping with regions that visually appear denser or blurred, which is consistent with pneumonia-related infiltrates. At the same time, some activation is still observed close to the diaphragm and chest wall, reflecting the tendency of transformer-based models to also incorporate contrast transitions and surrounding anatomical structures into their decision process.

Overall, these Grad-CAM examples provide an initial qualitative indication that the ViT model often attends to radiologically meaningful regions when distinguishing normal and pneumonia class, although the attention is not restricted exclusively to pathological areas. The analysis in this subsection is descriptive and based on a limited number of test images. It does not constitute a formal clinical validation, and the heatmaps have not been systematically assessed by radiologists. Therefore, the visual explanations should be interpreted as a supportive tool for understanding the model's behavior rather than as evidence that the system is ready for routine clinical use. Further studies involving expert review and larger-scale evaluation would be required before considering any clinical deployment.

Figure 3. Grad-CAM Visualization Results for Pneumonia Samples

D. Sanity Checks Randomization and Occlusion

To verify that the Grad-CAM visualizations depend on the parameters learned during training and not only on the model architecture or image structure, two additional sanity checks

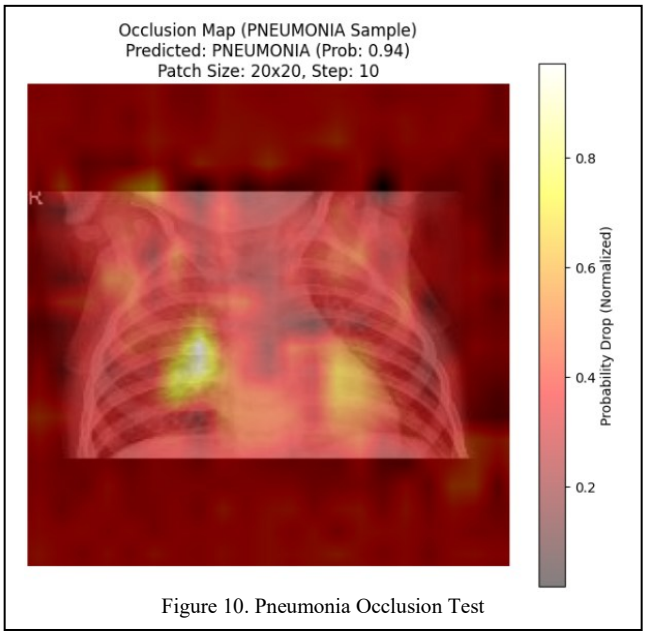

Figure 10. Pneumonia Occlusion Test

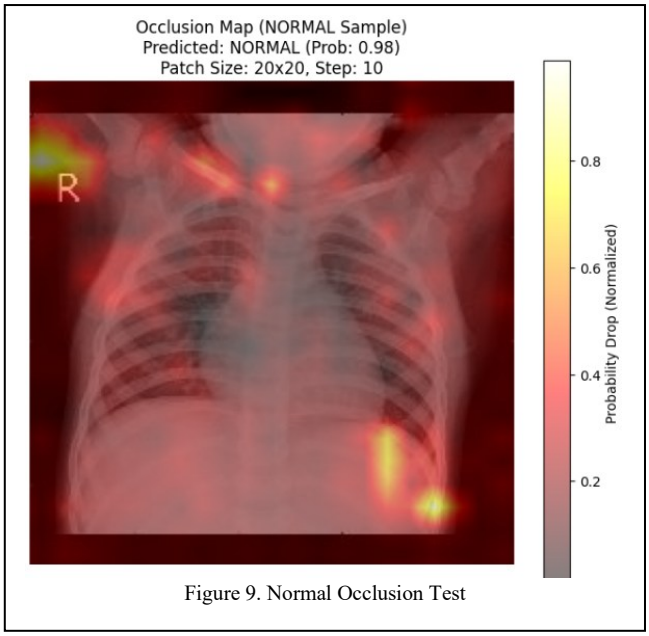

Figure 9. Normal Occlusion Test

model parameter randomization test and occlusion sensitivity analysis were performed. For the randomization test, a second Vision Transformer model was created with the same architecture (vit_small_patch16_224 with two output classes) but with randomly initialized weights by setting pretrained=False and not performing any training on the pneumonia dataset. This "random model" therefore has the same structure as the trained ViT but does not contain any task-specific knowledge.

A random image from the independent test set was then selected using a fixed random seed for reproducibility, and Grad-CAM maps were generated for both models using the same target layer, namely the last normalization block of the encoder (blocks[-1].norm1). For each model, the predicted class and its softmax probability were computed, and Grad-CAM was run with respect to the predicted class.

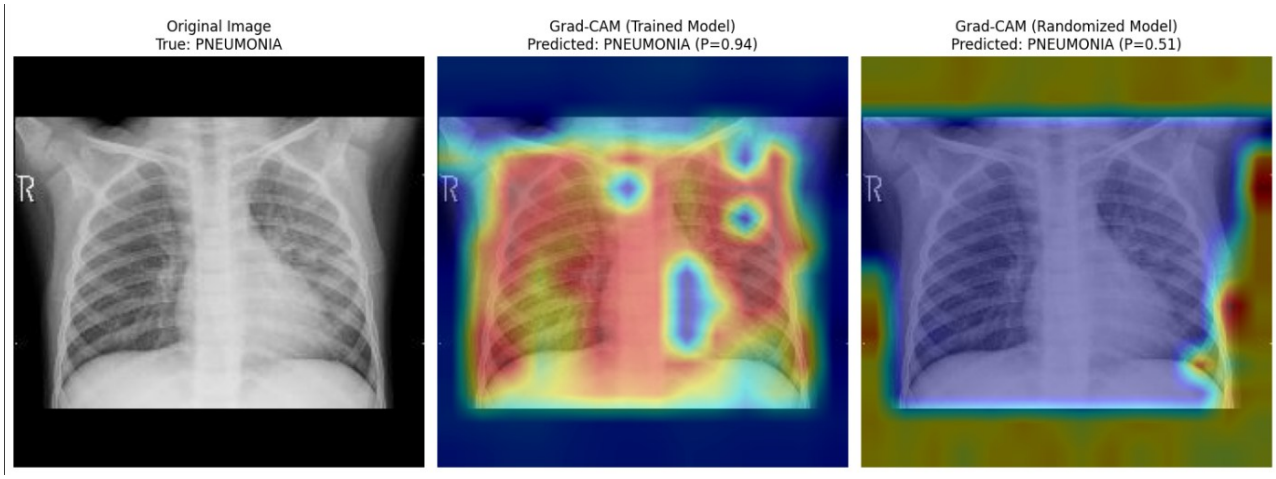

Figure 8. Randomization Sanity Check

Figure 8 demonstrate the randomization sanity check. The trained model predicts pneumonia with a probability of 0.94, and its Grad-CAM heatmap shows relatively structured activations over the lung fields, with higher intensity around regions where opacities are visible on the radiograph. In contrast, the random model produces a noisier prediction (probability 0.51) and a Grad-CAM heatmap that is largely diffuse and shifted toward the image borders, with much weaker focus on the parenchymal lung regions.

This qualitative difference suggests that the Grad-CAM maps of the trained model are influenced by the learned parameters and not solely by the ViT architecture or input statistics. Nevertheless, this analysis is based on a limited number of examples and should be regarded as an initial sanity check rather than a formal quantitative evaluation.

The second sanity check investigates how sensitive the model's prediction is to local perturbations in different regions of the image. For this purpose, an occlusion sensitivity map was computed for one correctly classified normal image and one correctly classified pneumonia image from the test set.

The procedure is implemented as follows. For a given test image, the model is first evaluated to obtain the predicted class and its softmax probability. Then a 20×20 pixel gray patch (RGB value (127, 127, 127), corresponding to a normalized intensity of 0.5) is slid across the image with a step size of 10 pixels in both horizontal and vertical directions. At each patch location, the patched image is re-evaluated by the model (using the same preprocessing as in validation), and the drop in probability for the original predicted class is recorded. These probability differences are stored in an occlusion map, which is later normalized to the [0, 1] range and resized to the original image resolution for visualization. Brighter colors on the occlusion map indicate regions where masking causes a larger reduction in confidence.

The resulting occlusion maps are shown in Figure 9 and figure 10. For the normal sample, the probability drops are relatively moderate and distributed over several thoracic regions, without a single dominant hotspot. This pattern indicates that the model's normal prediction is based on a combination of global structural cues and overall lung clarity rather than on one specific small area.

For the pneumonia sample, a stronger probability drop is observed when the occlusion patch covers a region in the lower lung field where a radiographic opacity is present, whereas masking more peripheral regions results in smaller changes in probability. This suggests that the model relies more heavily on that opacity region when deciding that the image represents pneumonia, in line with the corresponding Grad-CAM heatmap.

Taken together, the randomization and occlusion experiments provide complementary qualitative evidence that the Grad-CAM explanations are linked to the trained ViT parameters and that at least part of the model's decision is influenced by radiologically plausible lung regions. However, these sanity checks were performed on a small number of images and have not been systematically reviewed by

radiologists. Consequently, they should be interpreted as preliminary interpretability analyses rather than definitive proof of clinical readiness.

IV. Conclusion and Future Work

This study proposed a pneumonia classification model based on the Vision Transformer (ViT) architecture combined with Gradient-weighted Class Activation Mapping (Grad-CAM) for visual explainability. The model was trained and evaluated on a publicly available pediatric chest X-ray dataset consisting of 5,824 images divided into NORMAL and PNEUMONIA classes. The main objective was to investigate whether a ViT-based classifier, equipped with explainability tools, can provide reliable predictions and qualitatively interpretable decision patterns for pneumonia detection on CXR images.

The experimental results show that the ViT model achieved high Accuracy, Precision, Recall, F1-Score, and an AUC of 0.975 on the held-out test set, despite the imbalanced class distribution. The use of class-weighted cross-entropy loss helped to compensate for the dominance of the pneumonia class and encouraged more balanced decision boundaries between normal and pneumonia. The confusion matrix analysis confirmed that the number of false negatives for pneumonia was relatively low, which is important from a clinical safety perspective.

Grad-CAM visualizations, together with the randomization test and occlusion sensitivity analysis, provided an initial qualitative view of the model's behavior. In many correctly classified normal cases, the attention maps were broadly distributed over the thoracic region without focusing on a single focal opacity, whereas in pneumonia cases, higher activations often overlapped with regions showing visible opacities. The randomization experiment indicated that Grad-CAM maps from the trained ViT are more structured than those from an untrained random model, and the occlusion maps showed that masking opacity regions causes a larger drop in pneumonia prediction confidence. These findings suggest that the model frequently attends to radiologically plausible lung regions, although the attention is not restricted exclusively to pathological areas.

Several limitations of this work should be acknowledged. First, the evaluation was conducted on a single public dataset with pediatric CXR images and an imbalanced class distribution, which may limit the generalizability of the model to other age groups, institutions, or imaging protocols. Second, a single stratified train validation test split was used instead of full k-fold cross-validation or external test sets, so the reported metrics may still be sensitive to the chosen split. Third, the Grad-CAM and sanity check analyses were performed on a limited number of samples and have not been systematically assessed by radiologists. Therefore, the interpretability results should be considered preliminary and not as a substitute for expert visual inspection.

Future work will focus on addressing these limitations by evaluating the model on multi-institutional and multi-center CXR datasets, exploring more robust validation strategies such as cross-validation and external test cohorts, and involving radiologists in structured reader studies to assess the clinical usefulness of the generated explanations. In addition, extensions such as integrating lung segmentation, comparing different transformer variants, and combining visual explanations with uncertainty estimation may further improve the robustness and interpretability of ViT-based systems for pneumonia detection in real world clinical settings.

Bibliography

[1] S. Safiri *et al.*, "Global burden of lower respiratory infections during the last three decades," Jan. 2023. doi: https://doi.org/10.3389/fpubh.2022.1028525.

[2] F. Khan *et al.*, "AI-assisted detection for chest X-rays (AID-CXR): a multi-reader multi-case study protocol," *BMJ Open*, vol. 14, no. 12, Dec. 2024, doi: 10.1136/bmjopen-2023-080554.

[3] J. Becker *et al.*, "Artificial Intelligence-Based Detection of Pneumonia in Chest Radiographs," *Diagnostics*, vol. 12, no. 6, Jun. 2022, doi: 10.3390/diagnostics12061465.

[4] C. T. Yen and C. Y. Tsao, "Lightweight convolutional neural network for chest X-ray images classification," *Sci Rep*, vol. 14, no. 1, Dec. 2024, doi: 10.1038/s41598-024-80826-z.

[5] A. Manickam, J. Jiang, Y. Zhou, A. Sagar, R. Soundrapandiyan, and R. Dinesh Jackson Samuel, "Automated pneumonia detection on chest X-ray images: A deep learning approach with different optimizers and transfer learning architectures," *Measurement (Lond)*, vol. 184, Nov. 2021, doi: 10.1016/j.measurement.2021.109953.

[6] M. F. F. Mardianto, A. Yoani, S. Soewignjo, I. K. P. K. A. Putra, and D. A. Dewi, "Classification of Pneumonia from Chest X-ray images using Support Vector Machine and Convolutional Neural Network," 2024. [Online]. Available: www.ijacsa.thesai.org

[7] C. Usman, S. U. Rehman, A. Ali, A. M. Khan, and B. Ahmad, "Pneumonia Disease Detection Using Chest X-Rays and Machine Learning," *Algorithms*, vol. 18, no. 2, Feb. 2025, doi: 10.3390/a18020082.

[8] D. Lestari, A. Mulya, A. Tatamara, R. R. Haiban, and H. D. Khalifah, "Deep Learning for Pneumonia Detection in Chest X-Rays using Different Algorithms and Transfer Learning Architectures," *Public Research Journal of Engineering, Data Technology and Computer Science*, vol. 3, no. 1, pp. 1–9, Jul. 2025, doi: 10.57152/predatecs.v3i1.1656.

[9] I. Y. Chang and T. Y. Huang, "Deep learning-based classification for lung opacities in chest x-ray radiographs through batch control and sensitivity regulation," *Sci Rep*, vol. 12, no. 1, Dec. 2022, doi: 10.1038/s41598-022-22506-4.

[10] O. N. Manzari, H. Ahmadabadi, H. Kashiani, S. B. Shokouhi, and A. Ayatollahi, "MedViT: A Robust Vision Transformer for Generalized Medical Image Classification," Feb. 2023, doi: 10.1016/j.compbiomed.2023.106791.

[11] K. Tyagi, G. Pathak, R. Nijhawan, and A. Mittal, "Detecting Pneumonia using Vision Transformer and comparing with other techniques," in *2021 5th International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, IEEE, Dec. 2021, pp. 12–16. doi: 10.1109/ICECA52323.2021.9676146.

[12] S. Singh, M. Kumar, A. Kumar, B. K. Verma, K. Abhishek, and S. Selvarajan, "Efficient pneumonia detection using Vision Transformers on chest X-rays," *Sci Rep*, vol. 14, no. 1, Dec. 2024, doi: 10.1038/s41598-024-52703-2.

[13] J. Ko, S. Park, and H. G. Woo, "Optimization of vision transformer-based detection of lung diseases from chest X-ray

images," *BMC Med Inform Decis Mak*, vol. 24, no. 1, Dec. 2024, doi: 10.1186/s12911-024-02591-3.

[14] A. Alqutayfi *et al.*, "Explainable Disease Classification: Exploring Grad-CAM Analysis of CNNs and ViTs," *Journal of Advances in Information Technology*, vol. 16, no. 2, pp. 264–273, 2025, doi: 10.12720/jait.16.2.264-273.

[15] S. Suara, A. Jha, P. Sinha, and A. A. Sekh, "Is Grad-CAM Explainable in Medical Images?," Jul. 2023, doi: 10.1007/978-3-031-58181-6_11.

[16] P. Purwono, A. Nabila, E. Wulandari, and K. Nisa, "Explainable Artificial Intelligence (XAI) in Medical Imaging: Techniques, Applications, Challenges, and Future Directions," *Review*, vol. 1, no. 1, pp. 52–66, Jun. 2025, doi: 10.53623/amms.v1i1.692.

[17] 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," Dec. 01, 2022, *Nature Research*. doi: 10.1038/s41746-022-00699-2.

[18] D. Kermany, K. Zhang, and M. Goldbaum, "Labeled Optical Coherence Tomography (OCT) and Chest X-Ray Images for Classification," 2018, doi: 10.17632/rscbjbr9sj.2.

[19] A. Dosovitskiy *et al.*, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," Jun. 2021, [Online]. Available: http://arxiv.org/abs/2010.11929

[20] O. N. Manzari, H. Ahmadabadi, H. Kashiani, S. B. Shokouhi, and A. Ayatollahi, "MedViT: A robust vision transformer for generalized medical image classification," *Comput Biol Med*, vol. 157, p. 106791, May 2023, doi: 10.1016/j.compbiomed.2023.106791.

[21] 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, doi: 10.1007/s11263-019-01228-7.

[22] Y. Yang, G. Mei, and F. Piccialli, "A Deep Learning Approach Considering Image Background for Pneumonia Identification Using Explainable AI (XAI)," *IEEE/ACM Trans Comput Biol Bioinform*, vol. 21, no. 4, pp. 857–868, Jul. 2024, doi: 10.1109/TCBB.2022.3190265.

[23] E. Chamseddine, N. Mansouri, M. Soui, and M. Abed, "Handling class imbalance in COVID-19 chest X-ray images classification: Using SMOTE and weighted loss," *Appl Soft Comput*, vol. 129, Nov. 2022, doi: 10.1016/j.asoc.2022.109588.

[24] D. Park, "A Comprehensive Review of Performance Metrics for Computer-Aided Detection Systems," Nov. 01, 2024, *Multidisciplinary Digital Publishing Institute (MDPI)*. doi: 10.3390/bioengineering11111165.