**Engineering and Technology Journal e-ISSN: 2456-3358**

**Volume 11 Issue 02 February-2026, Page No.-8927-8934**

**DOI: 10.47191/etj/v11i02.17, I.F. – 8.482**



# Explainable Deep Learning for Lung Disease Detection from Chest X-rays via Layer-wise Relevance Propagation

**Benny Sukma Negara[1], Muhammad Irsyad[2*], Laila Nurul Fauziyyah[3]**

[1,2,3]Informatics Engineering Study Program, Faculty of Science and Technology, Universitas Islam Negeri Sultan Syarif Kasim, Pekanbaru-Riau, Indonesia.

**ABSTRACT:** This study proposes an approach to classify lung diseases based on X-ray images using the VGG16 architecture equipped with the Layer-wise Relevance Propagation (LRP) interpretability method. The dataset consists of three classes: COVID-19, pneumonia, and normal, which are processed through augmentation and normalization. The model is trained with a data ratio of 70:30, a learning rate of 0.001, a batch size of 32, and an Adam optimizer. The training results show high accuracy of 96.78% with balanced precision, recall, and F1-score values. The LRP method was used to highlight important areas in the image that contributed to the model's prediction, thereby increasing decision transparency. The main contribution of this research is the integration of VGG16 with LRP in multi-class X-ray image classification, which provides accurate results along with visual interpretations that support confidence in medical applications.

**KEYWORDS:** Image Classification, Chest X-Ray, VGG16, Interpretability, Layer-wise Relevance Propagation.

## I. INTRODUCTION

The leading cause of death worldwide is lung disease [1]. Various diseases affect human respiratory function and quality of life, including Coronavirus Disease 2019 (COVID-19) and Pneumonia. COVID-19 is a disease that can be transmitted through direct contact with an infected person when they cough, sneeze, or exhale. A person can become infected through breathing if they are in close proximity to someone with COVID-19 [2]. Meanwhile, pneumonia is the leading cause of death in Indonesia in the *post-neonatal* group (aged 29 days to 11 months), accounting for 14% of deaths, up from 9.8% in 2020 [3]. The high risk of COVID-19 transmission and the high mortality rate due to pneumonia highlight the importance of accurate diagnosis of lung disease. Diagnosis of lung disease can be made through medical history, physical examination, sputum or phlegm analysis, and additional laboratory tests. One method of additional laboratory testing is using X-rays. *Chest X-rays* are a practical choice for initial *screening* because the equipment is readily available [4]. *Chest X-rays* (CXR) visualize the structure of the lungs and are used to detect abnormalities such as nodules, masses, and fluid accumulation, which are signs of lung disease [5].

On the other hand, advances in technology, particularly *deep learning*, have encouraged the development of effective models for detecting diseases based on medical images [6]. The most widely used architecture is *the convolutional neural network* (CNN), which is capable of recognizing complex patterns in images, including *X-ray* images that are often undetectable to the human eye. A number of studies have shown that CNNs perform reliably in clinical practice for medical classification tasks and have been widely used. The use of CNNs has achieved an accuracy rate of up to 99.90% in disease classification tasks [7]. One of the most widely used CNN architectures is VGG16, which is known to be effective in visual feature extraction. Study [8] reported that the application of *transfer learning* using VGG16 for brain tumor detection in MRI images successfully achieved an accuracy of up to 100%, demonstrating the great potential of this approach in clinical applications. The implementation of CNN models in disease detection using images has the potential to improve the quality of medical diagnosis.

The main weakness of *deep learning* is its *black-box* behavior, making it difficult to understand how the model makes decisions [9]. This has led to the development of *explainable artificial intelligence* (XAI), which aims to explain and provide a logical understanding behind AI decisions and optimize algorithms for better performance. The use of XAI in the field of health is important to provide interpretability of model decision results, thereby increasing confidence in the use of the system in diagnosis [10]. One XAI method that can be used is *layer- wise relevance propagation* (LRP).

*Layer-wise Relevance Propagation* (LRP) is a technique designed to explain *deep learning* model decisions by remapping the relevance of model predictions to specific parts of *the input*, such as specific areas on *X-ray* images [11]. This method allows for a clearer understanding of

which features or areas most influence the results provided by the model [12]. In a medical context, LRP helps identify the parts of an image that contribute most to the model's decision, making the prediction results more transparent and easier for medical personnel to understand [13]. One application of LRP that has been proven effective is shown in previous research that integrated LRP into the VGG16 model for the task of classifying kidney-ureter-bladder (KUB) X-ray images. The study successfully identified the presence of kidney stones with high accuracy, while also providing visualization of the image areas that were the focus of the model's attention through an *explainable artificial intelligence* (XAI) approach [14]. The combination of *transfer learning* and LRP has been proven to produce strong interpretability and increase user confidence in the prediction results.

However, the scope of this research is still limited to binary classification (kidney stones vs. normal) and is only applied to one type of medical image, namely KUB *X-rays*. There has not been much exploration into the application of LRP in the classification of other, more complex medical images, such as chest *X-rays,* which cover a variety of disease conditions with overlapping visual characteristics. In addition, challenges in multi-class classification and the need for high interpretability in medical decision-making remain gaps that have not been widely addressed.

This study aims to address these limitations by applying a similar approach in a broader context, namely the classification of chest *X-ray* images into three categories of lung disease: COVID-19, normal lungs, and pneumonia. In this context, model interpretability is not only important for understanding algorithmic decisions, but also crucial in increasing clinical confidence in deep learning systems. Therefore, this study aims not only to achieve high classification accuracy through the VGG16 model, but also to produce clear and meaningful visual interpretations using the LRP method, thereby contributing to the development of a more transparent and reliable *deep learning*-based diagnosis system.

## II. METHOD

The method proposed in this study consists of two major stages. In the first stage, chest *X-rays* are classified to detect diseases in the categories of COVID-19, normal lungs, and pneumonia. In the second stage, *explainable artificial intelligence* (XAI) is implemented using the *layer-wise relevance propagation* (LRP) method to interpret the model's performance results.

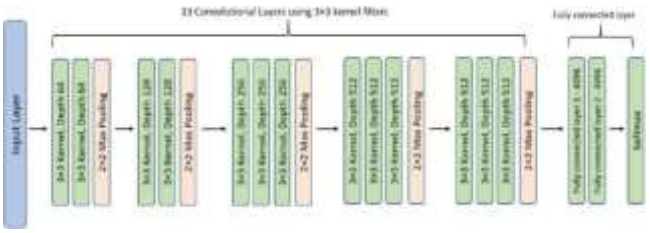


Figure 3. VGG16 architecture model

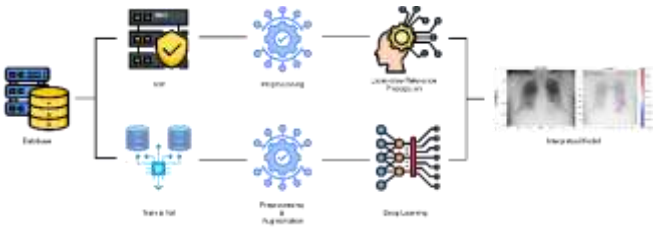


**Figure 1. Research Stages**

**Dataset**

The dataset used in this study consists of chest X-ray images divided into three categories, namely COVID-19, normal, and pneumonia. All images were obtained from the Mendeley Data website, which provides open access to scientific datasets. For the deep learning model training and evaluation stages, 1,000 images were used for each category, resulting in a total of 3,000 X-ray images. The images were organized into separate folders based on their categories to facilitate data labeled.

Furthermore, in the implementation stage of the interpretability technique using layer-wise relevance propagation (LRP), one image from each category was used, taken from the same dataset source but not included in the previous training or validation process. This was done to test the model's ability to provide interpretations of predictions on new data that had never been seen before. Thus, the total data used was 3003 images.

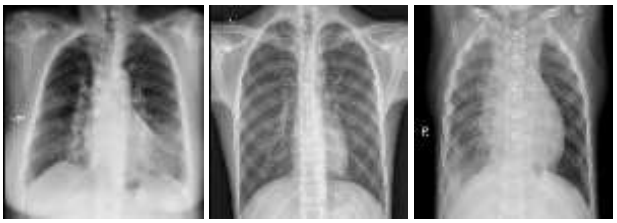


**Figure 2. X-ray images COVID-19, Normal, Pneumonia.**

### *A. Preprocessing*

In this study, each chest X-ray image first undergoes a preprocessing stage before being used in the model training process. This stage includes resizing the image to 224×224 pixels to match the input format of the VGG16 architecture, converting the image into tensor format, and normalizing it based on the mean and standard deviation of the ImageNet dataset. These steps aim to ensure that the data has a format and scale that is consistent with the pre-

trained model, thereby accelerating convergence and improving training stability [15].

### B. Deep Learning Process

*1. Visual Geometric Group 16*

VGG16 is a CNN architecture consisting of 16 layers with trainable weights, namely 13 convolutional layers and 3 fully connected layers. All convolutional layers in VGG16 use small 3x3 kernels with padding and stride 1. In addition, every few convolutional layers, a *max pooling* operation with a size of 2x2 and a stride of 2 is used to reduce the spatial dimensions of the features. There are also two *fully connected* layers, each with 4096 units, before finally passing to the last *fully connected* layer that functions as *a classifier* [16]. *The* final *output* is generated by the *softmax* function, which converts the final values into probabilities of each target class.

2. Model Training Configuration

This study uses the VGG16 architecture with pre-trained weights from ImageNet, where the *feature extractor* layer is frozen to maintain general feature extraction capabilities, and the final classification layer is modified to match the number of target classes. This strategy follows the *transfer learning* approach, which has been proven effective in medical image classification because it can utilize knowledge from source domains such as ImageNet and reduce training time and large data requirements [17]. To increase the diversity of the training data and prevent *overfitting*, data augmentation was performed using *horizontal flipping* and *random rotation* [18]. This augmentation has been shown to be effective in improving the accuracy of *X-ray* image classification, as demonstrated by a study in [19], which shows that *random rotation* and *horizontal flipping* can help models learn more diverse features. The model was trained using *CrossEntropyLoss*, which is a standard loss function for multi- class classification, and *the Adam optimizer,* which is known to be efficient in accelerating the convergence process and adaptive to gradient scales [20].

Freezing convolutional parameters and retraining only the final layer enables training efficiency without losing performance in the target domain. In this study, the model was trained using a *batch* size of 32 and a number *of epochs* set to 30. *The batch size* of 32 was chosen based on considerations of balance between training stability and computational efficiency. A smaller *batch* size tends to result in more frequent gradient updates and can help the model achieve more stable convergence, while a *batch* size that is too large can reduce the model's generalization ability. A total of 30 *epochs* was chosen to provide sufficient training time for the model to learn patterns from the data without *overfitting*. To overcome *overfitting*, only the best model, meaning that during the training phase, if *the epoch* validation accuracy is higher than the highest accuracy, then that model is saved [21].

*Hyperparameter* exploration was conducted using several variations *of learning rates* and dataset splits as presented in the model training parameter table.

**Table 1. Parameters**

| Parameter | Value |
|---|---|
| Train & val data ratio | 90:10, 80:20, 70:30 |
| Optimizer | Adam |
| Activation function | ReLU |
| Epoch | 30 |
| Batch Size | 32 |
| Learning rate | 0.01, 0.001, 0.0001, 0.00001 |

3. Model Evaluation

A common way to evaluate the classification performance of a model is to use *a confusion matrix*. *A confusion matrix* works by representing the model's

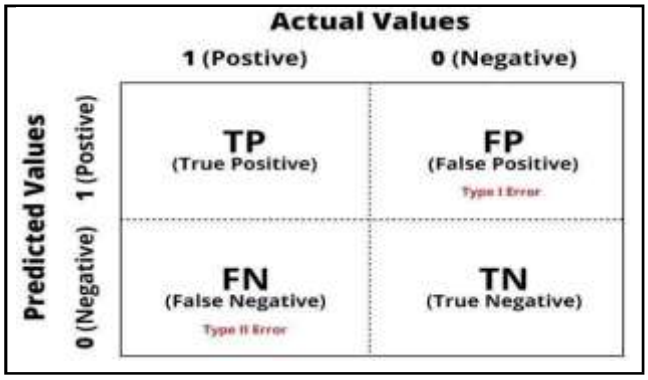



**Figure 4. Confusion Matrix**

prediction results in a table, showing four different combinations of predicted and actual values [22]. This table includes:

The explanation of each term is as follows: *True Positive* (TP) is the number of data with a positive class that has been correctly classified by the system. *True Negative* (TN) indicates the number of data with a negative class that has also been classified correctly. *False Negative* (FN) is the number of data that actually belongs to the positive class, but is incorrectly classified as negative by the system. Conversely, *False Positive* (FP) refers to the number of data that should belong to the negative class but are incorrectly classified as positive by the system. There are several equations that use the above values, as follows:

a) Accuracy

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

b) F1-Score

$$F1\ Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (2)$$

c) Recall

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

d) Precision

$$Precision = \frac{TP}{TP + FP} \quad (4)$$

### *C. Implementation of XAI using LRP*

1. Explainable Artificial Intelligence

Explainable Artificial Intelligence (XAI) refers to methodologies designed to make the results of AI models more transparent and understandable to users. This concept is particularly important in applications related to health, finance, and other critical decision-making, where users need to trust and understand how AI reaches certain decisions [23]. XAI aims to convert "black box" models, where decisions are unclear, into more transparent "white box" models, allowing users to understand and evaluate the results [24].

2. Layer-wise Relevance Propagation

*Layer-wise Relevance Propagation* (LRP) is a technique for interpreting *the output* of deep learning models, particularly for image classification. This method works by calculating the contribution of each *neuron* in the neural network to the final decision of the model. Using LRP, each pixel of an image can be evaluated to determine how much it contributes to the resulting classification. This provides deep insight into the model's decision-making process and makes it easier for users to understand the reasoning behind each prediction generated by the model by redistributing the relevance values associated with *neurons* in the upper layers to the *input* layer [25].

The operation of LRP is as follows:

$$R_j = \sum_k \frac{Z_{jk}}{\sum_j Z_{jk}} \quad (5)$$

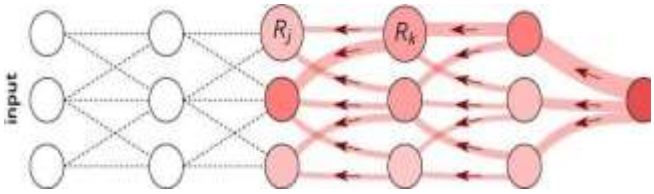



**Figure 5. LRP Workflow**

The LRP technique works by redistributing the relevance values associated with *neurons* in the *output* layer to the lower *input* layers. This procedure is subject to conservation properties, which means that the relevance received by a particular layer must be redistributed to the lower layers in the same amount. The distribution of the relevance score *Rk* in a particular layer to *the* generic *neuron* j of a lower layer can be obtained by applying a rule where *Rj* corresponds to the relevance score of *neuron j* due to *neuron k*. The quantity *zjk* models how much *neuron j* has contributed to making *neuron k* relevant, while the denominator enforces the conservation property. The method of calculating *zjk* is denoted as a rule and differs depending on the type of layer contained in the network.

## III.RESULT AND DISCUSSION

In this study, the entire implementation process was carried out using the Google Colab platform with the Python programming language and the PyTorch *deep learning* library as the main framework. This environment was chosen based on its ease of access, availability of GPUs for intensive computing, and good compatibility with various supporting libraries required in deep learning modeling and interpretability.

### *A. Hyperparameter Testing Results*

To obtain the best model configuration, a series of experiments were conducted by varying the training and validation data ratios and learning rate values. The table below presents the results of twelve experiments, each using a combination of data ratios (90:10, 80:20, and 70:30) and learning rate values (0.01, 0.001, 0.0001, and 0.00001) with a batch size of 32 and 30 epochs.

**Table 2. Test Results**

| Expt. No. | Data Ratio | LR | Accc | Val Acc | Loss | Val Loss |
|---|---|---|---|---|---|---|
| 1 | 90:10 | 0.01 | 87.11 | 94.67 | 0.6525 | 0.1904 |
| 2 | 90:10 | 0.001 | 94.56% | 96.33 | 0.1551 | 0.0900 |
| 3 | 90:10 | 0.0001 | 94.52 | 96.33% | 0.1620 | 0.1304 |
| 4 | 90:10 | 0.00001 | 90.44 | 92.33 | 0.3122 | 0.2778 |
| 5 | 80:20 | 0.01 | 90.21% | 96.00 | 0.5045 | 0.2379 |
| 6 | 80:20 | 0.001 | 95.33% | 96.50 | 0.1400 | 0.1082 |
| 7 | 80:20 | 0.0001 | 95.04% | 96.00% | 0.1563 | 0.1394 |
| 8 | 80:20 | 0.00001 | 90.88 | 93.33 | 0.3218 | 0.2877 |
| 9 | 70:30 | 0.01 | 90.90% | 95.44 | 0.4815 | 0.2102 |
| 10 | **70:30** | **0.001** | **95.57%** | **96.78** | **0.1292** | **0.1117** |
| 11 | 70:30 | 0.0001 | 90.00 | 92.89 | 0.3528 | 0.3113 |
| 12 | 70:30 | 0.00001 | 95.38 | 96.50 | 0.1400 | 0.1082 |

Based on the test results, it is known that the model with a data ratio of 70:30 and a learning rate of 0.001 (Experiment 10) produced the best performance with the highest validation accuracy of 96.78% and the lowest val loss of 0.1117.

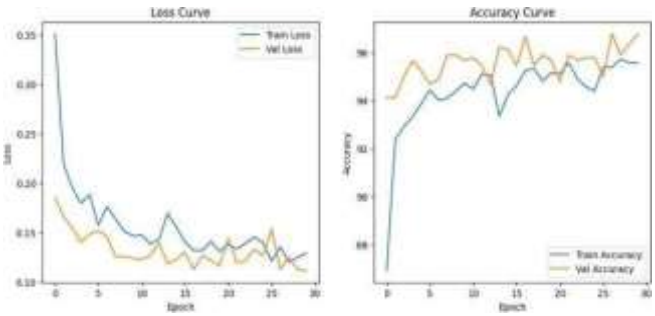



**Figure 6. Loss and accuracy graph of the best model**

*The visualization of the loss curve during training and accuracy with the best performance in the 10th training shows that the training process for 30 epochs provides stable and fairly optimal model performance. On the loss curve, it can be seen that the training loss value decreases sharply from the start of training (~0.35) to a value close to 0.11 at the end of the epoch. This decrease indicates that the model has successfully learned patterns from the training data gradually. Meanwhile, the validation loss shows a value that tends to be lower than the training loss and decreases steadily without extreme fluctuations, indicating that there is no significant overfitting.*

The accuracy curve in the 10th training shows a similar positive trend. Training *accuracy* gradually increases from around 87% to nearly 95%, while *validation* accuracy consistently remains high from the start of training and reaches over 96% at the end of the process. Interestingly, the validation accuracy was higher than the training accuracy, especially at the beginning and middle *of the epoch*. This could be due to regularization techniques, data augmentation, or a validation data distribution that was easier to classify. Overall, there were no significant differences between the model's performance on the training and validation data, so it can be concluded that the model has good generalization capabilities and does not experience significant *overfitting* or *underfitting*.

Based on the combination of *accuracy*, *validation accuracy*, and loss values, the 10th experiment with a ratio of 70:30 and a learning rate of 0.01 was selected as the best model configuration to be used in the final training process and interpretability analysis using the LRP method.

### *B. Classification Evaluation*

To evaluate the performance of the VGG16 classification model in detecting lung conditions based on X- ray images, a confusion matrix was used to describe the number of correct and incorrect predictions of the model for each class: COVID-19, normal, and pneumonia. The following confusion matrix was generated from the inference process on all test data:

*The confusion matrix* in Figure 6 shows that the model is able to recognize the COVID-19 class very well, with 300 correct predictions out of 300 data points, while there are several errors in the classification of the normal and pneumonia classes**.** *The confusion matrix* can calculate the performance of each class in more detail using metric analysis on *precision*, *recall,* and *F1-score*.

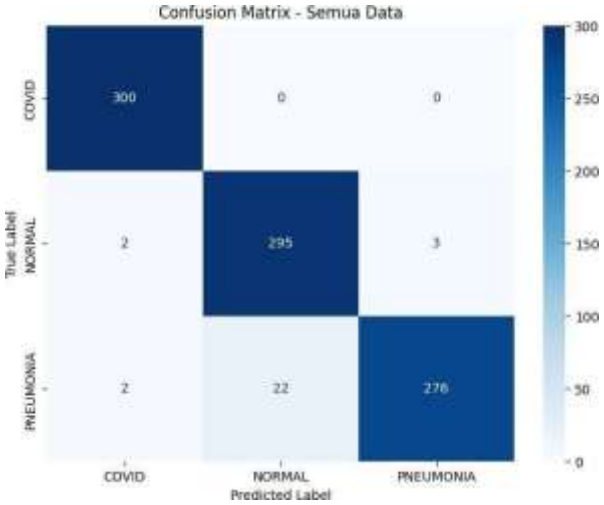



**Figure 7. Confusion Matrix**

**Table 3. Classification Report**

| | Precision | Recall | F1-Score | Support |
|---|---|---|---|---|
| Covid | 0.99 | 1.00 | 0.99 | 300 |
| Normal | 0.93 | 0.98 | 0.96 | 300 |
| Pneumonia | 0.99 | 0.92 | 0.95 | 300 |
| Accuracy | | | 0.97 | 900 |

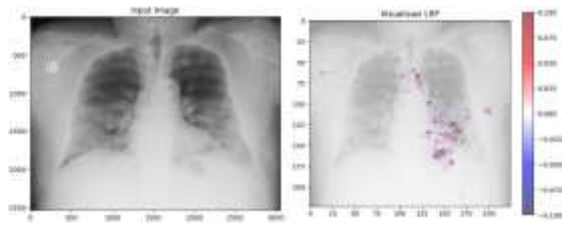
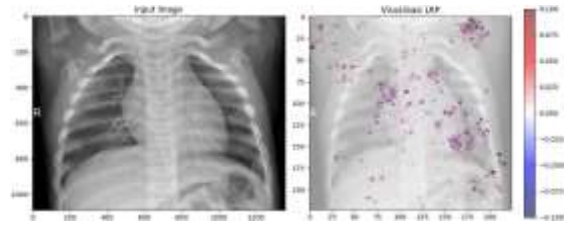
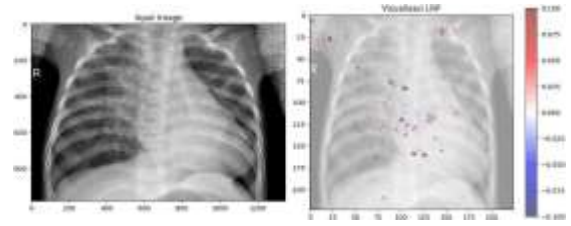
The evaluation results shown in Table 3 indicate that the model performs best in the COVID-19 class with a *precision* value of 0.99 and *a* perfect *recall* of 1.00, meaning that the model is able to recognize all COVID-19 images without detection errors. For the Normal class, *the precision* and *recall* are 0.93 and 0.98, respectively, indicating high performance but still some false predictions. The Pneumonia class obtained *a precision* of 0.99 and *a recall* of 0.92, showing that although positive predictions are very accurate, some Pneumonia cases are still missed by the model. Overall, the model achieved an accuracy of 97%, with an average *F1-score* of 0.97. This shows that the model has excellent overall classification performance in distinguishing the three disease categories.

### *C. LRP Implementation*

After the VGG16 *deep learning* model was successfully trained to classify *X-ray* images into three categories: COVID-19, normal, and pneumonia, the next step was to interpret the model's decisions using the *layer-wise relevance propagation* (LRP) method. LRP was implemented to obtain a visual representation of the parts of the image that the model considered relevant in making its predictions. In this process, the input was preprocessed to adjust the prediction model. The trained model was then reloaded with the best parameters obtained during training, followed by backward propagation from *the output* to the *input* layer to

calculate the contribution of each pixel to the resulting prediction.

**Table 4. LRP Implementation Results**

| Visualization of input and output | |
|---|---|
| COVID-19 |  |
| NORMAL |  |
| PNEUMONIA |  |

Above table shows the results of LRP implementation in a *deep learning* model for chest *X-ray* image classification. This visualization consists of two parts: the image on the left (image *input*) is the original *X-ray* image that is input into the VGG16 model, while the image on the right (LRP visualization) is the result of LRP, which shows the areas in the image that are considered relevant to the model in making decisions. The dark red to pink area (0.1 to 0.0) shows a positive contribution supporting the decision made by the model. Meanwhile, the dark blue to light blue area (0.0 to -0.1) shows a negative contribution arising from features that contradict the model's decision (contribution to other classes). Gray or uncolored areas indicate pixels that do not influence the model's prediction. The range of values in the legend is from -0.1 to +0.1, which is the range of *relevance scores* calculated from LRP propagation.

The blue and red areas shown are often adjacent to each other, which can be confusing. This occurs because *the convolution* and *pooling layers* in CNNs can cause neighboring pixels to have different influences on active features. Thus, the adjacent contributions of red and blue indicate areas that are significantly active, either supporting or opposing the model's decision. Furthermore, non-medical artifacts such as the identification letters D or R located at the edge of the image also receive positive relevance scores. The inclusion of artifacts in the relevance map indicates that the model learns *shortcut features*, which are visual patterns that are statistically correlated with the label but are not directly related to the target pathology.

Additionally, in the normal class image, it can be seen that the left shoulder area obtained a fairly high positive relevance score. This indicates that the model considers this area to be relevant in making decisions regarding the normal label. This pattern likely emerged because in the training data, the left shoulder area consistently appeared clean of abnormalities and had uniform lighting, so it was learned as a non-pathological visual feature. In other words, the model not only recognizes the presence of pathological features, but also utilizes the absence of abnormalities as a supporting indicator. In the context of LRP, the positive contribution of this clean area reinforces the interpretation that the model learns based on data distribution statistics, not solely on medical meaning. This finding also underlines the importance of a comprehensive evaluation of LRP visualization to identify potential biases or shortcut learning in the model.

## CONCLUSIONS

Based on the results of the discussion, the classification model developed shows high validation accuracy, reaching 96.78%. However, high accuracy does not guarantee that the predictions produced are always semantically and clinically correct. The results of interpretation using the *layer-wise relevance propagation* (LRP) method reveal that there are a number of cases where the model provides predictions that are correct in terms of labels, but not based on relevant features that logically support the decision. This phenomenon shows that the model has the potential for *shortcut learning*, which is the tendency to rely on patterns or visual artifacts in images that have no medical relevance, but happen to appear frequently in certain labels.

To address this issue, it is recommended that future training of classification models on medical images use datasets that have been cleaned of non-pathological artifacts, such as writing, markers, or medical equipment in the images. In this way, the model can focus on learning significant pathological characteristics, resulting in fairer, more accurate, and more reliable decisions in real clinical contexts. Furthermore, to improve the clinical validity of the model's interpretation results, it is highly recommended that the relevance visualizations generated by LRP be evaluated by a medical team, particularly radiologists or pulmonologists. This evaluation aims to assess the correspondence between the areas highlighted by the model and the pathological areas that are clinically

considered significant. By involving medical expert review, model interpretation will not only be more accurate but also more accountable in actual clinical practice.

**REFERENCES**


1. D. Septhya *et al.*, "Implementation of Decision Tree and Support Vector Machine Algorithms for Lung Cancer Classification," *Malcom Indonesian Journal of Machine Learning and Computer Science*, vol. 3, no. 1, pp. 15–19, 2023, doi: 10.57152/malcom.v3i1.591.
2. N. Jain, A. Choudhury, J. Sharma, V. Kumar, D. De, and R. Tiwari, "A review of novel coronavirus infection (Coronavirus Disease-19)," *Global Journal of Transfusion Medicine*, vol. 5, no. 1, pp. 22–26, 2020, doi: 10.4103/GJTM.GJTM_24_20.
3. M. drg. Oscar Primadi and M. dr. Anas Ma'ruf, "PROFILE OF HEALTH IN INDONESIA 2020," 2020.
4. F. N. Azizah and D. Juniati, "Analysis of Lung Disease Types Based on Chest X-Ray Using the Fuzzy C-Means Method," Mathunesa Jurnal Ilmiah Matematika, vol. 9, no. 2, pp. 322–331, 2021.-Ray Using the Fuzzy C-Means Method," *Mathunesa Scientific Journal of Mathematics*, vol. 9, no. 2, pp. 322–331, 2021, doi: 10.26740/mathunesa.v9n2.p322-331.
5. S. Saturi and S. Banda, "Advanced Lung Disease Detection and Classification Using Ge-U-Net-ODL with Gabor Filters and Entropy- Based Feature Selection," *Journal of Sensors, IoT & Health Sciences*, vol. 2, no. 2, pp. 69–86, Jun. 2024, doi: 10.69996/jsihs.2024011.
6. M. Kranthi, S. Sailaja, and E. Jyothi, "Deep Learning Approaches for Medical Image Processing in the Big Data Era," vol. 01, no. 01,
7. pp. 24–31, 2024, doi: 10.58599/ijsmcse.2024.1108. M. E. Karar, E. E. Hemdan, and M. A. Shouman, "Cascaded Deep Learning Classifiers for Computer-Aided Diagnosis of COVID-19 and Pneumonia Diseases in X-Ray Scans," *Complex & Intelligent Systems*, vol. 7, no. 1, pp. 235–247, 2020, doi: 10.1007/s40747-020- 00199-4.
8. C. Karima, N. Bourbia, K. Messaoudi, and E.-B. Bourennane, "Analysis and Classification of Medical Images Using Deep Learning Algorithms," *Recent Advances in Computer Science and Communications*, vol. 18, Oct. 2024, doi: 10.2174/0126662558327739240925073925.
9. M. M. Rahman, K. Matsuo, S. Matsuzaki, and S. Purushotham, "DeepPseudo: Pseudo Value Based Deep Learning Models for Competing Risk Analysis," *Proceedings of the Aaai Conference on Artificial Intelligence*, vol. 35, no. 1, pp. 479–487, 2021, doi: 10.1609/aaai.v35i1.16125.
10. P. Gort, C. Claessens, F. van der Sommen, and P. H. N. de With, "Evaluating the Interpretability of Prototype Networks for Medical Image Analysis," p. 54, 2025, doi: 10.1117/12.3046678.
11. D. A. Safitri, Y. P. M. Mamesah, and J. F. J. Timban, "Chest X-ray Findings in Patients with Pulmonary Tuberculosis and Chronic Kidney Disease at Prof. Dr. R. D. Kandou General Hospital, January–June 2022," *Medical Scope Journal*, vol. 4, no. 1, pp. 93–98, 2023, doi: 10.35790/msj.v4i1.44722.
12. D. Kim, J. Lee, J. Moon, and T. Moon, "Interpretable Deep Learning-based Hippocampal Sclerosis Classification," *Epilepsia Open*, vol. 7, no. 4, pp. 747–757, 2022, doi: 10.1002/epi4.12655.
13. S. Gupta *et al.*, "Four Transformer-Based Deep Learning Classifiers Embedded with an Attention U-Net-Based Lung Segmenter and Layer-Wise Relevance Propagation-Based Heatmaps for COVID-19 X-ray Scans," *Diagnostics*, vol. 14, no. 14, p. 1534, Jul. 2024, doi: 10.3390/diagnostics14141534.
14. F. Ahmed *et al.*, "Identification of kidney stones in KUB X-ray images using VGG16 empowered with explainable artificial intelligence," *Sci Rep*, vol. 14, no. 1, p. 6173, Mar. 2024, doi: 10.1038/s41598-024-56478-4.
15. L. Sağır, E. Kaba, M. H. Yiğit, F. Taşçı, and H. Uzun, "Predicting Semen Analysis Parameters From Testicular Ultrasonography Images Using Deep Learning Algorithms: An Innovative Approach to Male Infertility Diagnosis," *J Clin Med*, vol. 14, no. 2, p. 516, 2025, doi: 10.3390/jcm14020516.
16. D. Shen, G. Wu, and H.-I. Suk, "Deep Learning in Medical Image Analysis.," *Annu Rev Biomed Eng*, vol. 19, pp. 221–248, Jun. 2017, doi: 10.1146/annurev-bioeng-071516-044442.
17. H. E. Kim, A. Cosa-Linan, N. Santhanam, M. Jannesari, M. E. Maros, and T. Ganslandt, "Transfer learning for medical image classification: a literature review," *BMC Med Imaging*, vol. 22, no. 1, p. 69, Dec. 2022, doi: 10.1186/s12880-022-00793-7.
18. R. Hao, K. Namdar, L. Liu, M. A. Haider, and F. Khalvati, "A Comprehensive Study of Data Augmentation Strategies for Prostate Cancer Detection in Diffusion-Weighted MRI Using

Convolutional Neural Networks," *Journal of Digital Imaging*, vol. 34, no. 4,

19. pp. 862–876, Aug. 2021, doi: 10.1007/s10278-021-00478-7. B. Liu, C. Tan, S. Li, J. He, and W. Hong-yan, "A Data Augmentation Method Based on Generative Adversarial Networks for Grape Leaf Disease Identification," *IEEE Access*, vol. 8, pp. 102188–102198, 2020, doi: 10.1109/access.2020.2998839.
20. H. Iiduka, "Appropriate Learning Rates of Adaptive Learning Rate Optimization Algorithms for Training Deep Neural Networks," 2020, doi: 10.48550/arxiv.2002.09647.
21. G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely Connected Convolutional Networks," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Jul. 2017, pp. 2261–2269. doi: 10.1109/CVPR.2017.243.
22. S.-C. Kim and Y. Cho, "Implementation of a Predictive System to Improve the Accuracy of Urine Self-Diagnosis Using Smartphones: Application of a Confusion Matrix-Based Learning Model Through RGB Semiquantitative Analysis," *Sensors*, vol. 22, no. 14,
23. p. 5445, 2022, doi: 10.3390/s22145445. A. K. M. Nor, S. R. Pedapati, M. Muhammad, and V. Leiva, "Overview of Explainable Artificial Intelligence for Prognostic and Health Management of Industrial Assets Based on Preferred Reporting Items for Systematic Reviews and Meta-Analyses," *Sensors*, vol. 21, no. 23, p. 8020, 2021, doi: 10.3390/s21238020.
24. S. Lebovitz, H. Lifshitz-Assaf, and N. Levina, "To Engage or Not to Engage With AI for Critical Judgments: How Professionals Deal With Opacity When Using AI for Medical Diagnosis," *Organization Science*, vol. 33, no. 1, pp. 126–148, 2022, doi: 10.1287/orsc.2021.1549.
25. P. R. A. S. Bassi, S. S. J. Dertkigil, and A. Cavalli, "Improving deep neural network generalization and robustness to background bias via layer-wise relevance propagation optimization," *Nat Commun*, vol. 15, no. 1, p. 291, 2024, doi: 10.1038/s41467-023-44371-z.