JOURNAL OF
EMERGING INVESTIGATORS

Article

Lung region complexity affects Grand-CAM and Deep Taylor outcomes in pneumonia predictions

Anish Baghel[1,2], Viktoriia Liu[1]
[1] Aspiring Scholars Directed Research Program, Fremont, California
[2] Foothill High School, Pleasanton, California

SUMMARY

Accurate and timely pneumonia diagnosis is critical for patient care, and convolutional neural networks (CNNs) have shown significant effectiveness and high accuracy in automating this process. However, CNNs act as black boxes, which limits their interpretability in clinical settings. Explainable AI (XAI) techniques are designed to address this limitation by offering visual or quantitative explanations for model predictions. This study investigated two widely used XAI techniques—Gradient-weighted Class Activation Mapping (Grad-CAM) and Deep Taylor Decomposition (DTD)—for detecting pneumonia in chest X-ray images. Our analysis focuses on how different anatomical regions of the lungs affect the accuracy of explanations provided by these methods. We hypothesized that XAI methods will show statistically significant differences in activation and relevance across lung regions, with the lower regions exhibiting the highest importance values. Because pneumonia most commonly affects the lower lobes, we expect both methods to highlight these areas more strongly than the upper or central regions. Since CNNs learn increasingly complex features, subtle pneumonia patterns in the lower lobes or near the heart can be difficult to interpret. XAI techniques such as Grad-CAM and DTD help visualize these abstract features. Statistical analyses revealed consistent regional differences in how each method emphasized specific lung areas, with one region showing particularly strong relevance in both methods. These findings suggest that aligning interpretability techniques with clinically relevant anatomical regions can improve diagnostic support in medical imaging by helping identify subtle or complex abnormalities more effectively.

INTRODUCTION

Pneumonia is a major global health concern contributing to high rates of illness and death, especially among vulnerable groups such as children, the elderly, and the immunocompromised (1). As an infection that primarily affects the alveolar spaces of the lungs, pneumonia can escalate to severe illness if not accurately diagnosed and treated (1). Accurate and precise diagnosis is essential for effective treatment, yet traditional diagnostic methods like chest X-rays have significant limitations (1). Radiologists commonly use chest X-rays to detect pneumonia, identifying areas of opacity that indicate fluid buildup in the lungs (1). However, interpreting these images can be time-consuming, requires significant expertise, and often produces inconsistent results due to subjective differences among radiologists (2). Limited availability of skilled radiologists, particularly in resource-limited regions, highlighting the need for automated and dependable diagnostic approaches (2). To address these issues, deep learning models, particularly convolutional neural networks (CNNs), have emerged as powerful tools for automating the identification of pneumonia using chest X-ray images (2).

CNNs are designed to process and classify visual data, frequently achieving diagnostic accuracies above 90% when trained on large datasets (2). Their applications extend beyond pneumonia detection, including tasks such as tumor classification, retinal disease screening, and dermatological lesion analysis (2–3). Despite their strong performance, CNNs have been criticized for operating as opaque models, offering minimal transparency into the reasoning behind their predictions (2–3). To overcome this challenge, researchers have developed Explainable AI (XAI) methods that help interpret CNN decisions (4). In this study, we focus on two such XAI methods: Deep Taylor Decomposition (DTD) and Gradient-weighted Class Activation Mapping (Grad-CAM), which are effective in visualizing and explaining how neural networks arrive at their predictions (5).

We hypothesized that XAI methods, specifically Grad-CAM and DTD, will show statistically significant differences in activation and relevance across lung regions, with the lower regions exhibiting the highest importance values. Because pneumonia most commonly affects the lower lobes, we expect both methods to highlight these areas more strongly than the upper or central regions. We analyzed how differences between lung regions affect the clarity and usefulness of DTD and Grad-CAM explanations in detecting pneumonia. Specifically, we investigated which lung regions (upper, central, or lower) are most influential in model decision-making, and how each method visualizes these regions. We suggest that the complementary strengths of DTD's pixel-level precision and Grad-CAM's region-level feature localization may provide a more comprehensive and clinically relevant understanding of pneumonia indicators.

Statistical analysis supported our hypothesis, as it revealed that there are significant regional differences in relevance values with the lower lung region showing the highest importance in both methods. These results suggest that anatomical variation influences the effectiveness of XAI methods in highlighting areas relevant for diagnosis. Thus, aligning our explanation techniques with clinically significant regions could enhance interpretability, help radiologists with subtle finds, and improve the diagnostic accuracy in medical imaging.

https://doi.org/10.59720/25-039

RESULTS

We evaluated the performance of our trained CNN model in detecting pneumonia from chest X-rays. We trained the model using a dataset of 5,863 pediatric chest X-rays (patients ages 1–5) sourced from Kaggle, labeled by medical professionals as either healthy, labeled as normal or pneumonia-affected (6, 16). To maximize training data while ensuring unbiased evaluation, we split the dataset into 89.1% training, 10.7% testing, and 0.3% validation, with a class distribution of 73% pneumonia and 27% healthy. We used the larger training set to enable the model to learn robust features, the testing set to independently assess performance, and the small validation set primarily for occasional hyper-parameter tuning and early stopping without significantly reducing training data. The model achieved 95% training accuracy based on the training images. To evaluate the model's effectiveness on unseen data, we visualized test results with a confusion matrix (**Figure 1**). We also calculated key performance metrics: 94% accuracy, 96.1% recall, 94.4% precision, and a 91.9% F1 score. These metrics indicate strong predictive performance, particularly in identifying pneumonia cases, supporting the model's potential utility in clinical settings.

We next applied two XAI methods, DTD and Grad-CAM, to analyze chest X-rays for pneumonia detection. Both methods were used to create heatmaps that highlight the regions of the lungs most influential in CNN's pneumonia prediction (**Figures 2-3**). DTD provided pixel-level relevance, emphasizing the most critical regions for decision-making, while Grad-CAM highlighted broader regions of importance based on the network's deeper feature maps.

We overlaid the original chest X-ray with heatmaps generated by Grad-CAM and DTD, which highlighted regions indicative of pneumonia and observed general characteristics of pneumonia including hazy opacities (commonly linked to pneumonia-related infiltrates or effusion) in the lower and upper corners of the lung (7) (**Figure 2A**). The DTD overlay emphasized high-relevance regions for the model's prediction, notably assigning significant weight to the central area around the heart and highlighting scattered areas in the lower right lung with bright red regions, indicating strong relevance scores according to the DTD visualization (**Figure 2B**) (8). In contrast, the Grad-CAM heatmap highlighted broader and more defined areas, such as the top left and bottom right regions of the lung (**Figure 2C**). These regions often align with radiologist-identified zones for pneumonia detection, particularly when examining for lobar or multilobar pneumonia (9).

We applied a similar analysis to another chest X-ray image (**Figure 3**). The original X-ray displayed more significant opacity, particularly in the upper right region, which could have indicated pneumonia infiltrate or effusion (**Figure 3A**). The DTD overlay highlighted similar relevance patterns to those observed in the first example, concentrating on the central and lower portions of the lungs (**Figure 3A-3B**). These locations correspond to the fluid accumulations due to pneumonia (10). DTD notably emphasized areas near the diaphragm and along the lower edges, which could correlate with fluid accumulation. The Grad-CAM heatmap similarly focused on the outer edges of the lung, particularly near the diaphragm, and assigned significant weight to the areas surrounding the heart (**Figure 3C**).

The results of the statistical analysis demonstrated significant differences in how Grad-CAM and DTD methods highlighted various anatomical regions of the lungs in pneumonia detection. To facilitate the analysis, the lung length was divided into three equal sections, creating three distinct regions for examination: Region #1 (upper area of lungs), Region #2 (central area of lungs), and Region #3 (lower area of lungs) since the prevalence of pneumonia identifiers varies greatly based on these regions (**Figure 4**) (10–11). These divisions were made to ensure that each region was proportionally representative of the entire lung and that the analysis could capture any variation in the relevance of different lung areas in pneumonia detection. The boundaries of these regions were drawn based on the measured lung length, ensuring consistency across all images included in the study. This approach allowed for a systematic comparison of how different parts of the lung contributed to the model's decision-making process using Grad-CAM and DTD methods. We calculated mean intensity values and normalized intensity values in these three regions in the Grad-CAM and DTD heatmaps of ten randomly selected pneumonia X-ray images (**Figures 5-6**). The mean pixel intensity values for Region #1, Region #2, and Region #3 were 94.98, 99.47, and 111.70 for Grad-CAM and 36.24, 36.95, and 42.50 for DTD respectively. A one-way ANOVA indicated significant differences among regions for both Grad-CAM (p = 0.0387 < 0.05) and DTD (p = 0.0152 < 0.05). Post-hoc Tukey HSD tests showed that Region #3 had significantly higher intensity than Region #1 for both methods (Grad-CAM: p = 0.0378); DTD: p = 0.0201), and Region #3 was also higher than Region #2 for DTD (p = 0.0416). Normalized mean intensities showed a similar trend, with Grad-CAM values of 0.542, 0.605, and 1.000 and DTD values of 0.045, 0.051, and 0.079 for Regions #1, #2, and #3 respectively, indicating the strongest activation in Region #3. Paired t-tests further showed that Grad-CAM intensities

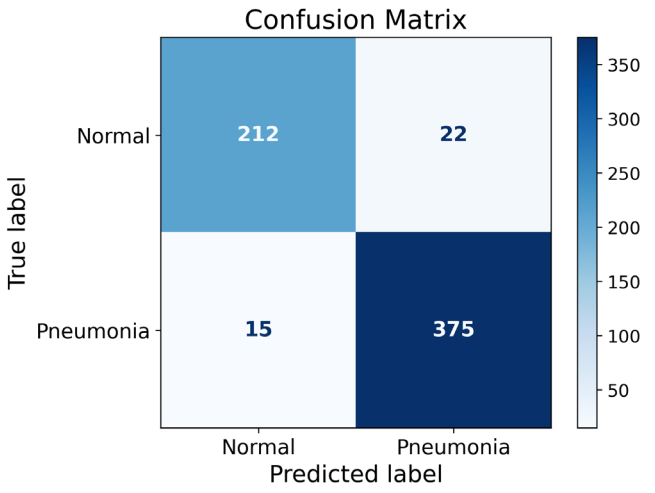

Figure 1: Confusion matrix for the testing dataset. The confusion matrix shows the performance of the model on the test data. True labels are represented on the y-axis, while predicted labels are on the x-axis. The model correctly classified 212 normal images and 375 pneumonia images and misclassified 22 normal images as pneumonia and 15 pneumonia images as normal.

https://doi.org/10.59720/25-039

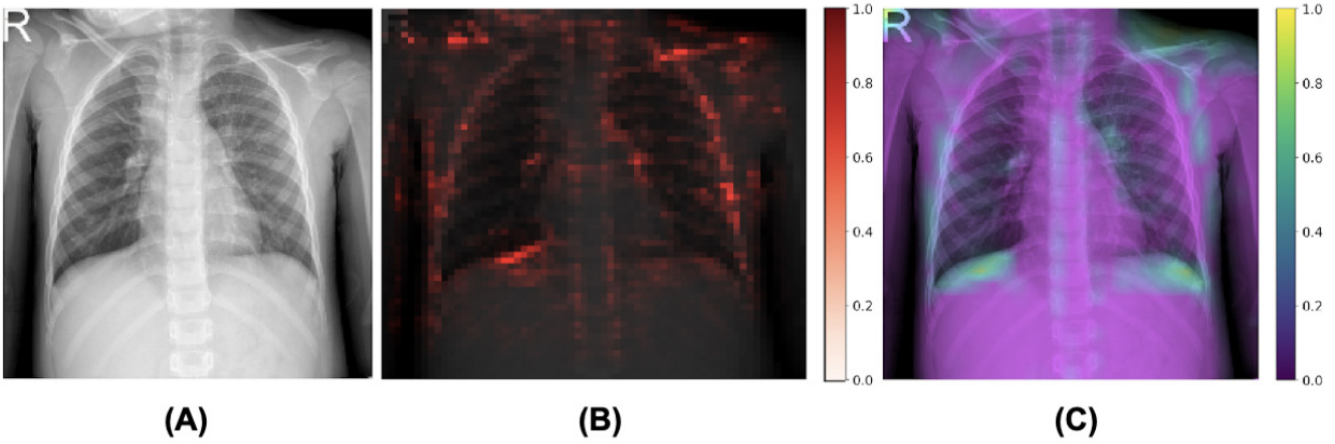

Figure 2: Comparison of Deep Taylor Decomposition (DTD) and Gradient-weighted Class Activation Mapping (Grad-CAM) heatmaps for a pneumonia case (Example 1). A) Original chest X-ray image of a child (ages 1–5) diagnosed with pneumonia, B) the DTD heatmap, and C) and the Grad-CAM heatmap. All images were sourced from the publicly available pediatric chest X-ray dataset on Kaggle, labeled by medical professionals (6). Both heatmaps highlight regions the model used for its prediction. B) In the DTD map, bright red areas indicate higher relevance scores, while in the C) Grad-CAM map yellow regions represent stronger model activation.

were significantly higher than DTD across all regions (Region #1: t = -9.45, p < 0.00001; Region #2: t = -7.22, p < 0.00003; Region #3: t = -7.61, p < 0.000014). These findings supported the hypothesis that different lung regions contributed variably to the effectiveness of XAI explanations and highlighted the importance of Region #3 as the most influential area in pneumonia detection across both explainability techniques.

We further analyzed the heatmaps and found that while both Grad-CAM and DTD identified Region #3 as the most critical for pneumonia detection, the two methods highlighted different features within the same anatomical areas.

DISCUSSION

We hypothesized that XAI methods, specifically Grad-CAM and DTD, would show statistically significant differences in activation and relevance across lung regions, with the lower regions exhibiting the highest importance values. Because pneumonia most commonly affects the lower lobes, we expected both methods to highlight these areas more strongly than the upper or central regions.

To test this hypothesis, we trained a CNN to distinguish pneumonia from healthy cases. Our CNN model achieved 94% accuracy on the test dataset, demonstrating strong performance in identifying key pneumonia-related features. We then applied XAI techniques, specifically Grad-CAM and DTD, to interpret CNN's predictions by highlighting image regions that influenced the model's decision. This integrated approach not only confirmed the model's accuracy but also provided visual insights into the features driving its diagnostic predictions. Although the CNN had strong performance, the relatively small size of the training and validation datasets highlights an important limitation. As only 627 images were used for testing and 14 for validation, which may have weakened the CNN's ability to generalize and could have contributed to fluctuations in accuracy and interpretability metrics. Specifically, the small validation set could have created some instability in the consistency of the Grad-Cam and DTD relevance maps. So future studies using larger validation and testing datasets are needed for giving a more reliable evaluation of the model and the comparative performance of the XAI methods.

Our comparative analysis of Grad-CAM and DTD in identifying pneumonia-affected regions in chest X-rays demonstrated that both techniques are valuable tools for model interpretability, though each exhibited distinct strengths and limitations. Grad-CAM tended to focus on the more prominent and widely recognized areas of pneumonia, particularly in the upper lung regions. It effectively highlighted broader areas, including regions near the heart and upper chest, which are traditionally associated with pneumonia (11). However, Grad-CAM's reliance on the final convolutional layers of the neural network resulted in lower precision when identifying smaller, subtler features. This limitation is significant in medical imaging, where minor details often hold clinical importance. The focus on generalized patterns by the deeper layers of the network limited Grad-CAM's sensitivity to subtle abnormalities, making it necessary to complement this method with a more granular approach like DTD (11-12).

In contrast, DTD produced more detailed and spatially nuanced attributions by highlighting smaller, less prominent regions that may be clinically relevant. The DTD heatmap demonstrated finer granularity—particularly around the heart and diaphragm—capturing subtle features potentially associated with pneumonia that Grad-CAM overlooked (**Figure 2**). This impression of higher spatial precision is supported by entropy and pixel intensity distribution metrics, which indicate that DTD assigns relevance more diffusely across the region. However, DTD's output is sensitive to initial conditions, such as the selection of the root point and specific activation paths, which can influence the final heatmap (12-13). This inherent variability can be mitigated by complementing DTD with additional explainability techniques like Grad-CAM to cross-validate region importance and reduce interpretability uncertainty.

By combining DTD and Grad-CAM, the interpretability and reliability of pneumonia detection in chest X-rays were significantly improved compared to using either method alone. DTD's pixel-level relevance attribution complemented

https://doi.org/10.59720/25-039

Grad-CAM's broader localization of feature regions, resulting in a more comprehensive and accurate visualization of pneumonia indicators. The enhanced accuracy of these visualizations is supported by the closer alignment of DTD's highlighted regions with those identified by medical experts. Furthermore, this integration supports the understanding that the complexity and nature of features within specific lung regions can significantly influence the effectiveness of XAI techniques. Although each method possessed inherent strengths and limitations, their combined application provided a more complete insight into how the model arrives at its predictions.

The significant differences observed between Grad-CAM and DTD values across all lung regions indicate that the two explainable AI approaches interpret model decisions in different ways. Grad-CAM generally produced higher importance values and showed greater variability between images, reflected by its larger standard deviations and F-statistics (14). This suggests that Grad-CAM is more sensitive to image-specific patterns and may highlight particular regions more strongly depending on the image. In contrast, DTD generated lower-intensity values with reduced variability, indicating a more consistent distribution of relevance across images. These statistically significant differences suggest that each method emphasizes different aspects of the model's decision process. When used together, Grad-CAM's ability to localize important regions and DTD's pixel-level attribution can provide a more complete and interpretable explanation of pneumonia-related patterns in chest X-rays. Therefore, selecting an appropriate explainability method may depend on the desired balance between visual interpretability and consistency in medical image analysis: Grad-CAM offers more intuitive visual explanations, while DTD provides more stable estimates of feature relevance.

Despite these differences, both Grad-CAM and DTD consistently identified the lower lung area, Region #3, as the most important region for pneumonia detection, suggesting that this region likely contains important diagnostic features that have a greater influence on model decision making. The statistically significant post hoc differences between the upper lung region, Region #1, and the lower lung region, Region #3, for both methods further support this observation, highlighting the unique role that Region #3 plays in pneumonia detection. The observation is consistent with known clinical patterns of pneumonia, which typically affects the lower lung regions, which are more susceptible to infection due to gravity (15). This can often be observed in cases where fluid accumulation and inflammation occur in the alveoli, leading to consolidation, which is a characteristic feature of pneumonia (15). Despite the general agreement regarding the importance of Region #3, the weak Pearson correlation coefficients indicate that Grad-CAM and DTD focus on different features within the same regions, suggesting that these methods may produce different results in terms of explainability (**Figures 5-6**).

Additional analysis of variability revealed, that Grad-CAM exhibits greater variation in assigning importance scores across different images compared to DTD (**Figure 5**). The higher standard deviations seen in Grad-CAM in all three lung regions implies a lack of consistency, which could create challenges in clinical settings as stable and reliable explanations are necessary in these settings for decision-making. Conversely, the lower standard deviations and more stable F-statistics associated with DTD suggest that it produces more consistent relevance estimates, potentially making it a more reliable method for highlighting critical regions in pneumonia detection.

Overall, the differences in how Grad-CAM and DTD assign importance demonstrate that different XAI methods can reveal distinct aspects of model behavior. Grad-CAM's stronger but more variable activations may provide more dynamic insights, yet can also introduce uncertainty, whereas DTD's stability brings greater confidence in the model's behavior. The findings support the idea that different lung regions contribute differently to model explanations and highlight the particularly important role of Region #3 in guiding accurate pneumonia predictions

Future work could expand this analysis by incorporating radiologist reader studies to evaluate how XAI heatmaps

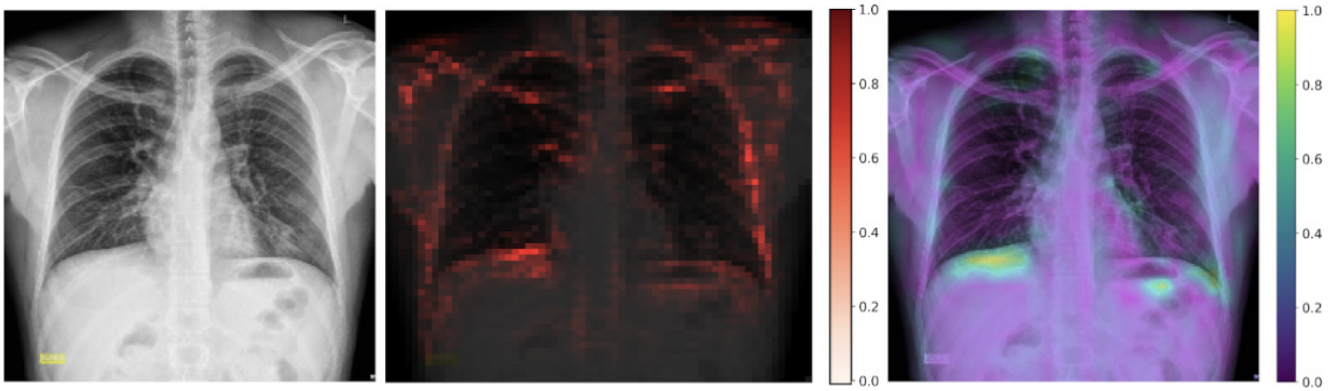

Figure 3: Comparison of Deep Taylor Decomposition (DTD) and Gradient-weighted Class Activation Mapping (Grad-CAM) heatmaps for a pneumonia case (Example 2). A) Original chest X-ray image of a child (ages 1–5) diagnosed with pneumonia, B) the DTD heatmap, and C) and the Grad-CAM heatmap. All images were sourced from the publicly available pediatric chest X-ray dataset on Kaggle, labeled by medical professionals (6). Both heatmaps highlight regions the model used for its prediction. B) In the DTD map, bright red areas indicate higher relevance scores, while in the C) Grad-CAM map yellow regions represent stronger model activation.

https://doi.org/10.59720/25-039

influence diagnostic confidence and accuracy. Region-specific perturbation analysis could also help verify whether the areas identified as important by Grad-CAM and DTD truly affect model prediction. Additional metrics, such as entropy or sparsity, may further quantify how attention is distributed across images. Finally, involving clinical experts in evaluating explanation maps would help determine how useful these explainability techniques are in real-world diagnostic settings and strengthen the practical relevance of AI-assisted pneumonia detection.

MATERIALS AND METHODS

Study dataset

The model was trained on a dataset of 5,863 chest X-ray images of children aged 1 to 5 years, sourced from Kaggle (6, 16). Each image was pre-labeled by medical professionals as either pneumonia or normal. Of the total images, 4,279 depicted pneumonia cases and 1,584 showed healthy lungs, representing 73% and 27% of the dataset, respectively. The dataset was split into three subsets: 5,222 images were used for training, 627 for testing, and 14 for validation.

Data preprocessing

All images were resized to 224×224 pixels with three color channels to ensure compatibility with the CNN input requirements. Pixel values were normalized to a 0–1 range. To increase data variability and reduce overfitting, augmentation techniques were applied, including horizontal flips, ±15° rotations, and zooming.

Model architecture and training

A custom sequential CNN was developed for binary classification. It consisted of five convolutional layers with ReLU activation functions. The number of filters increased from 32 to 128 across layers, then decreased back to 32. Each convolutional layer was followed by a 2×2 max-pooling layer. After feature extraction, the network included a flattening layer, a dense layer with 128 neurons and ReLU activation, and a final softmax output layer with two units representing the pneumonia and normal classes. The model was trained using the Adam optimizer and binary cross-entropy loss. Training was performed over 25 epochs, with early stopping based on validation loss to prevent overfitting.

XAI integration

To enhance the interpretability of the model's predictions, two XAI techniques were applied: Grad-CAM and DTD. DTD, a variant of layer-wise relevance propagation, assigns a relevance score to each pixel in the image to quantify its contribution to the model's prediction (17–18). For these relevance scores to be valid and interpretable, they must be non-negative, as negative relevance is not meaningful in the context of feature importance. Additionally, the total relevance across all pixels must sum to the overall relevance detected by the model, ensuring no overestimation of any particular region's contribution. These relevance scores were then mapped to a heatmap, which can be superimposed on the original image, which offered a visual representation of the areas that were the most important to the model's prediction (19).

DTD interprets the predictions of a neural network by using a Taylor series approximation to estimate the relevance of input features. This approach backpropagated through the network model to estimate the significance of each pixel in the input image (20). The method requires choosing a root point which serves as a reference for calculating pixel relevancy. The choice of the root point is critical in this approach, with different rules guiding the selection. In this study, we used the z^+ rule for convolution layers and the ε rule for fully connected layers, following the methodology of Kohlbrenner *et al.* (2020) (21). We chose the z^+ rule for this study because it is particularly effective in image classification tasks where we aim to identify the most relevant and positively contributing regions. Our goal was to focus on areas in chest X-rays that actively support pneumonia classification, making the explanations more clinically intuitive and visually interpretable.

The DTD approach approximates the neural network's behavior using the first term of a Taylor series expansion, centered around a root point, $\tilde{x}$. This approximation defines a relevance function, R, which is calculated as the gradient of the network function at the root point, multiplied by the difference between the pixel value, x, and the root point using the following equation:

$$R(x) = \left(\frac{\partial f(x)}{\partial x}\Big|_{x=\tilde{x}}\right) \cdot (x - \tilde{x}) \quad \text{(Equation 1)}$$

This local linear approximation simplifies the neural network's complex feature space, enabling relevance attribution at each layer. By applying this process across all layers, DTD generates a comprehensive heat map that

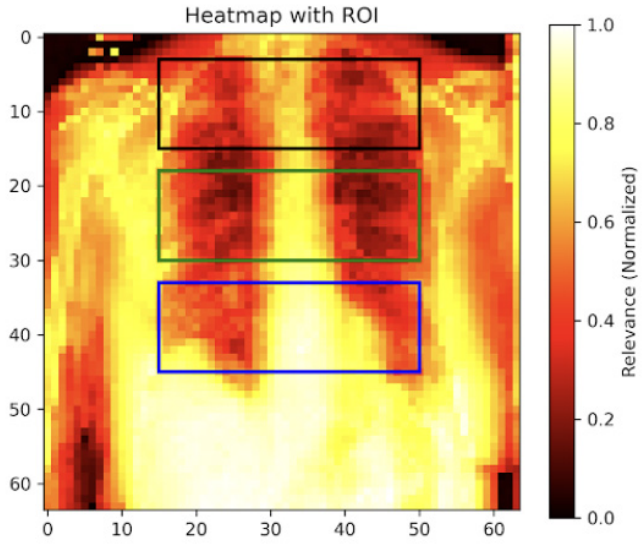

Figure 4: Delineation of lung regions for analysis. The chest X-ray heatmap was divided into three anatomical regions based on proportional lung length to enable region-specific evaluation of Gradient-weighted Class Activation Mapping (Grad-CAM) and Deep Taylor Decomposition (DTD) outputs. The total lung length was measured from the apex (top) to the base (bottom) of the lungs and divided into three equal sections, represented by rectangular bounding boxes: Region #1 (upper lung, black), Region #2 (middle lung, green), and Region #3 (lower lung, blue). The overlaid normalized relevance map illustrates how the regions were standardized across images for consistent comparison of XAI-derived activation and relevance values.

https://doi.org/10.59720/25-039

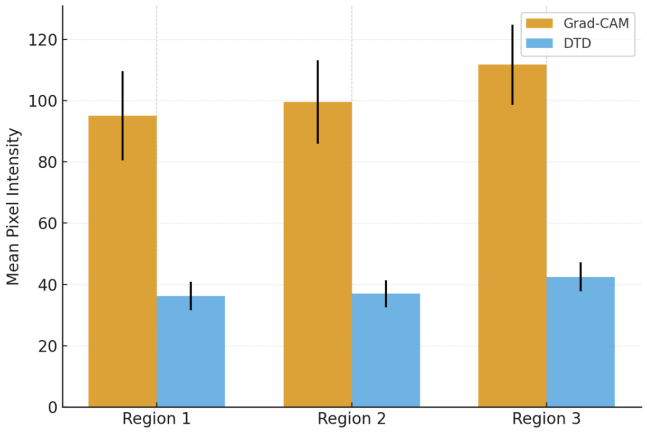

Figure 5: Mean pixel intensity across lung regions for Grad-CAM and DTD heatmaps. The average pixel intensity values for Regions #1, #2, and #3 in the Gradient-weighted Class Activation Mapping (Grad-CAM, orange) and Deep Taylor Decomposition (DTD, blue) heatmaps. The grouped bars compare how each explainability method highlights important regions of the lungs. ANOVA revealed statistically significant differences in pixel intensity across regions for both Grad-CAM (p = 0.0387) and DTD (p = 0.0152). Region #3 exhibited the highest mean intensity (Grad-CAM = 111.70; DTD = 42.50), indicating its greater contribution to model decision-making.

highlights the specific regions that contributed most to the model's output.

Grad-CAM was used to create class-specific heatmaps by analyzing the influence of feature maps on the model's predictions (22-23). Each feature map was assigned a weight to indicate the contribution of different image regions to these feature maps to highlight image regions contributing most to the final decision. Feature maps were generated for each of the convolutional layers by creating a model identical to the original CNN except for the target layer. This was done by replicating the original model for that target layer. The model's prediction for a giving image was then recorded, and backpropagation was applied to compute gradients that indicated any changes in the output score relative to the variations found in the feature maps (24). These gradients were then used to weight the feature maps, which were then subsequently combined. A ReLU activation function was applied to retain only positive contributions, and the resulting heatmaps were resized to match the original image, to visually represent the regions that were the most important to the model's decision.

To judge the interpretability of pneumonia detection in chest X-rays, Grad-CAM and DTD were applied to generate pixel-level relevance maps for ten randomly selected images labeled as pneumonia. Each X-ray was analyzed with both methods to identify the regions most associated with the model's prediction. To enable regional analysis, the total lung length was measured from the apex to the base of the lungs. This length was divided into three equal sections to define three anatomical regions: Region #1 (upper area of lungs), Region #2 (central area of lungs), and Region #3 (lower area of lungs). These divisions between lung regions were then standardized across all images to help ensure proportional representation and consistent region boundaries. Relevance maps from Grad-CAM and DTD were then generated and overlaid on the original images. Then the pixel intensity values for each of the tree lung regions were extracted. This procedure enabled statistical comparisons of how each XAI method attributed importance across the different lung areas.

Statistical analysis

To evaluate the differences in how each method highlighted the lung regions, pixel intensity values (Grad-CAM) and relevance scores (DTD) were analyzed using one-way ANOVA across the three anatomical regions. Post hoc Tukey HSD tests were employed to identify specific differences between regions. Paired *t*-tests were conducted to compare Grad-CAM and DTD values within each region. Pearson correlation coefficients were calculated to assess the relationship between Grad-CAM and DTD values for each region. Variance analysis, including the calculation of standard deviations and F-statistics, was performed to determine the consistency and variability of both methods across the lung regions. The average difference in importance values assigned by Grad-CAM and DTD was analyzed to highlight the relative contributions of each method. A significance threshold of $p < 0.05$ was used for all statistical tests. Each analysis was performed using 10 pneumonia chest X-ray images as biological replicates.

Packages

All analyses were conducted using Python 3.10 within the Anaconda distribution, running in a Jupyter Notebook environment to facilitate code development and interactive data exploration (25-27). The deep learning models were implemented using TensorFlow v2.12.0 with Keras serving as the high-level API for model architecture and training (28). To support explainability, the iNNvestigate toolbox

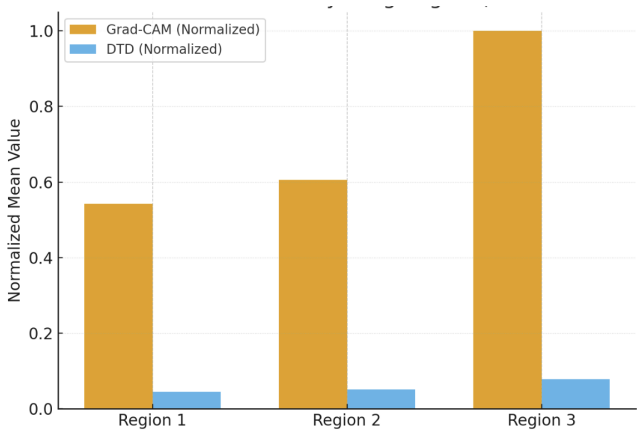

Figure 6: Normalized mean relevance comparison of Grad-CAM and DTD across lung regions. Statistical comparison of Grad-CAM (orange) and DTD (blue) relevance across the three lung regions in pneumonia X-rays. The normalized mean values illustrate the consistency and variability of both methods. Pearson correlation coefficients between Grad-CAM and DTD values were weak (r = 0.22–0.34), suggesting that the two methods emphasize different features within the same regions. Variance analysis (coefficient of variation = 0.12–0.16) showed that Grad-CAM exhibited greater variability than DTD, reflecting differences in how each method visualizes regional importance.

https://doi.org/10.59720/25-039

v1.0.9 was used to generate DTD relevance maps. Grad-CAM was implemented using Grad-CAM v1.4.8 and custom Keras-compatible functions validated against established interpretability benchmarks. Statistical analyses were performed using scikit-learn v1.3.0 for preprocessing, correlation, and performance metrics (29-31). Advanced statistical testing, including ANOVA and Tukey HSD, was carried out using SciPy v1.11.1 and StatsModels v0.14.0 (32-33). All plots and visualizations, including Grad-CAM and DTD heatmaps as well as statistical graphs, were generated using Matplotlib v3.7.2 and Seaborn v0.12.2 to ensure clarity and consistency in visual representation (34-35).

Received: January 28, 2025
Accepted: June 23, 2025
Published: July 29, 2026

REFERENCES

1. Lim, Wei Shen. "Pneumonia—Overview." *Encyclopedia of Respiratory Medicine,* 2nd ed., Academic Press, 2022, pp.185-97. https://doi.org/10.1016/B978-0-12-801238-3.11636-8.
2. Khan, Wasif, *et al*. "Intelligent Pneumonia Identification from Chest X-Rays: A Systematic Literature Review." *IEEE Access*, vol. 9, 30 March 2021, pp. 51747−71, https://doi.org/10.1109/access.2021.3069937.
3. Becker, Judith, *et al*. "Artificial Intelligence-Based Detection of Pneumonia in Chest Radiographs." *Diagnostics*, vol. 12, no. 6, 14 June 2022, p. 1465, https://doi.org/10.3390/diagnostics12061465.
4. Yang, Yuting, *et al*. "Explainable Deep Learning Models on the Diagnosis of Pneumonia." *2021 IEEE/ACM Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE)*, Washington, DC, USA, 2021, pp. 134−8, https://doi.org/10.1109/CHASE52844.2021.00032.
5. Alam, Mahbub Ul, *et al*. "Exploring LRP and Grad-CAM Visualization to Interpret Multi-Label-Multi-Class Pathology Prediction Using Chest Radiography." *2022 IEEE 35th International Symposium on Computer-Based Medical Systems (CBMS)*, Shenzen, China, 2022, pp. 258−63, https://doi.org/10.1109/CBMS55023.2022.00052.
6. Mooney, Paul Timothy. "Chest X-Ray Images (Pneumonia)." *Kaggle*, 19 Mar. 2018. https://www.kaggle.com/datasets/paultimothymooney/chest-xray-pneumonia. Accessed 20 Nov. 2024.
7. Reissig, Angelika *et al*. "Transthoracic sonography of diffuse parenchymal lung disease: the role of comet tail artifacts." *J Ultrasound Med*, *vol.* 22(2), Feb. 2003, pp. 173−80, https://doi.org/10.7863/jum.2003.22.2.173.
8. Franquet, T. "Imagine of pneumonia: trends and algorithms." *Eur Respir J.,* vol. 18, Jul. 2001, pp. 196−08, https://doi.org/10.1183/09031936.01.00213501.
9. Marchiori, Edson *et al*. "High-resolution computed tomography findings from adult patients with Influenza A (H1N1) virus-associated pneumonia." *Eur J Radiol.,* vol. 74, Apr. 2010, pp. 93−8, https://doi.org/10.1016/j.ejrad.2009.11.005.
10. Hansell, David M *et al*. "Fleischner Society: glossary of terms for thoracic imaging." *Radiology*, vol. 246, Mar. 2008, pp. 697−22, https://doi.org/10.1148/radiol.2462070712.
11. Ahamed, Khabir Uddin, *et al*. "DTLCx: An Improved ResNet Architecture to Classify Normal and Conventional Pneumonia Cases from COVID-19 Instances with Grad-CAM-Based Superimposed Visualization Utilizing Chest X-ray Images." *Diagnostics*, vol. 13, no. 3, 2 Feb. 2023, p. 551, https://doi.org/10.3390/diagnostics13030551.
12. Du, Jiajie, and Yang, Jiaqi. "Classification of Chest X-ray Images (Pneumonia) Based on ResNet and Grad-CAM." *Proceedings of the 2023 6th International Conference on Signal Processing and Machine Learning (SPML '23)*, Association for Computing Machinery, New York, 17 Oct. 2023, pp. 156–64, https://doi.org/10.1145/3614008.3614031.
13. Jan, Chia-Wei, *et al*. "Optical Based Gradient-Weighted Class Activation Mapping and Transfer Learning Integrated Pneumonia Prediction Model." *Tech Science Press*, vol. 47, no. 3, 9 Nov. 2023, pp.2989−10, https://doi.org/10.32604/csse.2023.042078.
14. Koyyada, and Singh, Thipendra P. "An Explainable Artificial Intelligence Model for Identifying Local Indicators and Detecting Lung Disease from Chest X-ray Images." *Healthcare Analytics*, vol. 4, Dec. 2023, p. 1002−06, https://doi.org/10.1016/j.health.2023.100206.
15. Naz, Zubaira, *et al*. "An Explainable AI-Enabled Framework for Interpreting Pulmonary Diseases from Chest Radiographs." *Cancers*, vol. 15, no. 1, 3 Jan. 2023, p. 314, https://doi.org/10.3390/cancers15010314.
16. Kermany, Daniel S. et.al. "Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning." *Cell*, vol 172, 5, 22 Feb 2018, pp. 1122-31, https://doi.org/10.1016/j.cell.2018.02.010.
17. Hassan, Mohammad Mehedi *et al*. "Explaining COVID-19 Diagnosis with Taylor Decompositions." *Neural Comput & Applic,* vol. 35, 17 Nov. 2023, pp. 22087–100, https://doi.org/10.1007/s00521-022-08021-7.
18. Kauffmann, Jacob, *et al*. "Towards Explaining Anomalies: A Deep Taylor Decomposition of One-Class Models." *Pattern Recognition*, vol. 101,1 May 2020, pp. 107−98, https://doi.org/10.1016/j.patcog.2020.107198.
19. Montavon, Grégoire, *et al*. "Explaining Nonlinear Classification Decisions with Deep Taylor Decomposition." *Pattern Recognition*, vol. 65, 1 May 2017, pp. 211–22, https://doi.org/10.1016/j.patcog.2016.11.008.
20. Sixt, Leon, and Landgraf, Tim. "A Rigorous Study of the Deep Taylor Decomposition." *Transactions on Machine Learning Research*, 14 Nov. 2020, https://doi.org/10.48550/arXiv.2211.08425.
21. Kohlbrenner, Maximilian, *et al*. "Towards Best Practice in Explaining Neural Network Decisions with LRP." *arXiv,* 13 Jul. 2020., https://doi.org/10.48550/arXiv.1910.09840.
22. Selvaraju, Ramprasaath R., *et al*. "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization." *2017 IEEE International Conference on Computer Vision (ICCV)*, Venice, Italy, 2017, pp. 618−26, https://doi.org/10.1109/ICCV.2017.74.
23. Pitroda, Vidhi, *et al*. "An Explainable AI Model for Interpretable Lung Disease Classification." *2021 IEEE International Conference on Internet of Things and Intelligence Systems (IoTaIS)*, Bandung, Indonesia,

https://doi.org/10.59720/25-039

24 Nov. 2021, pp. 98-103,https://doi.org/10.1109/IoTaIS53735.2021.9628573 .
24. Rajaraman, Sivaramakrishnan, *et al*. "Visualization and Interpretation of Convolutional Neural Network Predictions in Detecting Pneumonia in Pediatric Chest Radiographs." *Applied Sciences*, vol. 8, no. 10, 2018, p. 1715, https://doi.org/10.3390/app8101715.
25. Python Software Foundation. *Python*: Version 3.10. Python Software Foundation, 2021, https://www.python.org/
26. Anaconda, Inc. *Anaconda Distribution*. Version 2023, Anaconda, Inc., https://www.anaconda.com/
27. Kluyver, Thomas, *et al*. *Jupyter Notebook.* Project Jupyter, 2016, https://jupyter.org/
28. Abadi, Martín, *et al*. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. 2015, https://www.tensorflow.org/
29. Alber, Maximilian, *et al*. iNNvestigate Neural Networks: Version 1.0.9. 2018, https://github.com/albermax/innvestigate
30. Chattopadhay, Aditya, *et al*. Grad-CAM: Version 1.4.8. 2020, https://github.com/jacobgil/keras-grad-cam
31. Pedregosa, Fabian, *et al*. Scikit-Learn: Machine Learning in Python. Version 1.3.0, 2011, https://scikit-learn.org/
32. Virtanen, Pauli, *et al*. SciPy: Version 1.11.1, *SciPy Community*, 2023, https://scipy.org/
33. Seabold, Skipper, and Josef Perktold. StatsModels: Version 0.14.0. 2010, https://www.statsmodels.org/
34. Hunter, John D., *et al*. Matplotlib: Version 3.7.2. Matplotlib Development Team, 2023, https://matplotlib.org/
35. Waskom, Michael L. Seaborn: Version 0.12.2. 2023, https://seaborn.pydata.org/