

CoSegXAI: A Pipeline for Context vs. Segmented ROI Images in Lung Disease Classification through Saliency Map Insights

CoSegXAI: uma Pipeline para Imagens Contextuais vs. ROI Segmentadas na Classificação de Doenças Pulmonares Através de Percepções do Mapa de Saliência

Graciella dos Santos Favoreto^{1*}, Mariana Aya Suzuki Uchida¹, Pedro Augusto Luiz¹, Erikson Julio de Aguiar¹, Marcel Koenigkam Santos², Agma Juci Machado Traina¹

Abstract: This research proposed CoSegXAI, an experimental pipeline to investigate the impact of lung segmentation on the explainability and performance of deep neural networks applied to multiclass chest X-ray classification. We employed the **COVID-19 Radiography Database** and trained ResNet50, DenseNet201, and VGG19 models on both contextual (original) and ROI-based (segmented) images across four diagnostic classes: COVID-19, lung opacity, viral pneumonia, and normal cases. To evaluate model decisions, we generate the saliency maps using Grad-CAM and assess them through visual analysis and quantitative metrics, including Insertion Correlation (IC), Deletion Correlation (DC), and Sparsity. Our findings show that models trained on contextual images achieved superior classification performance. DenseNet201 achieves an accuracy of 92.80% and 88.08% trained on contextual and ROI-based images, respectively. However, explainability analysis reveals that models trained on ROI-based (segmented important regions) images produce more focused attention patterns with better calibration, achieving superior IC, DC, and Sparsity scores. These results suggest that models are more reliable and provide clinicians with greater explainability. The source code is available at GitHub repository.

Keywords: saliency maps — chest x-ray — deep learning — explainable AI

Resumo: Esta pesquisa propõe o CoSegXAI, um pipeline experimental para investigar o impacto da segmentação pulmonar na explicabilidade e no desempenho de redes neurais profundas aplicadas à classificação multiclasse de radiografias torácicas. Utilizamos as imagens do **COVID-19 Radiography Database** disponibilizado publicamente e treinamos as redes ResNet50, DenseNet201 e VGG19 tanto em imagens contextuais (originais) quanto baseadas em ROI (segmentadas) em quatro classes diagnósticas: COVID-19, opacidade pulmonar, pneumonia viral e casos normais. Para avaliar as decisões dos modelos, geramos mapas de saliência usando Grad-CAM e os avaliamos através de análise visual e métricas quantitativas, incluindo Correlação de Inserção (IC), Correlação de Deleção (DC) e Esparsidade. Nossos resultados mostram que modelos treinados em imagens contextuais alcançaram desempenho de classificação superior. A rede DenseNet201 atinge uma acurácia de 92,80% e 88,08% quando treinada em imagens contextuais e baseadas em ROI, respectivamente. No entanto, a análise de explicabilidade revela que modelos treinados em imagens baseadas em ROI revelam padrões de atenção mais focalizados com melhor calibração, alcançando métricas de IC, DC e Esparsidade superiores. Esses resultados sugerem que os modelos são mais confiáveis e fornecem aos radiologistas maior explicabilidade. O código-fonte está disponível no repositório do GitHub.

Palavras-Chave: mapas de saliência — radiografia torácica — aprendizado profundo — explicabilidade

¹ Instituto de Matemática e Ciência da Computação, Universidade de São Paulo (USP), Brazil

² Departamento de Imagens Médicas, Hematologia e Oncologia Clínica, Universidade de São Paulo (USP), Brazil

*Corresponding author: graciellafavoreto@usp.br

DOI: <http://dx.doi.org/10.22456/2175-2745.150955> • Received: 15/10/2025 • Accepted: 03/11/2025

CC BY-NC-ND 4.0 - This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.

1. Introduction

Artificial Intelligence (AI) has been transforming the traditional Radiological techniques and image diagnoses, providing advances in analyzing complex data and diseases [1]. Although radiologists have a high ability to evaluate these images, they must deal with a large data volume and a shortage of professionals. Also, radiologists are prone to a longer, exhaustive, and subjective diagnostic process [2, 3]. Tools based on AI can help mitigate these challenges and evolve the analysis dynamics.

Medical image classification plays a crucial role in clinical treatment, particularly in pulmonary cases, where physicians utilize images in the diagnosis process and quantify the severity of diseases [4]. Deep Learning (DL) methods are promising for automated classification tasks, automatically extracting and selecting features to classify them, providing radiologists with faster and accurate guidance to aid in analysis [5, 4]. Extracting useful information from these images is relevant for respiratory diseases, which remain among the leading causes of global morbidity and mortality [6].

Deep Learning is a black-box method and usually employs explainability to generate visual explanations of the model's decisions, such as saliency maps from Gradient-weighted Class Activation Mapping (Grad-CAM) [7]. These post hoc techniques highlight regions on the image at the pixel level that the model deemed most relevant during classification [8]. Explainability is a valuable strategy for medical image classifiers, as they handle sensitive data and necessitate a more precise model.

Beyond classification, the segmentation task is a valuable step for medical image analysis. Segmentation removes the background and focuses on the Region of Interest (ROI) or essential image areas. In developing DL solutions for disease detection, it is important to guard them against features that could severely affect reliable decision-making. Segmentation can serve as a preprocessing technique to prevent a classifier from learning unnecessary pixel information when categorizing an image [9, 10].

This research aims to evaluate the impact of segmentation as a preprocessing technique in a lung disease classification task, and to analyze saliency maps generated by Grad-CAM in detail. We propose a comparative pipeline to investigate deep learning multiclass classification on pulmonary images, focusing on the trade-off between region-of-interest (ROI) images and contextual images in terms of model performance and explainability. We apply the COVID-19 Radiography Database [11, 12] to investigate both original (called contextual or entire images) and ROI-based (Lung region segmented) images. To further analyze the model's performance, we explore the model's decisions using Grad-CAM saliency maps and conduct both visual inspections and quantitative evaluations applying established metrics. This paper's contributions are as follows:

1. Analyzing and quantifying explainability over contex-

tual and ROI-based images to investigate the impact of specific regions of interest in identifying diseases.

2. A Comparative analysis of Deep Learning models' performance trained on contextual and ROI-based chest X-ray images.
3. CoSegXAI, an experimental pipeline to quantify the saliency maps' behavior on chest X-Ray images and specific ROIs employing quantitative and visual aspects.

Section 2 presents the main concepts and related work. Section 3 shows the proposed methodology. Section 4 presents results and discussions. Section 5 concludes the paper.

2. Background and Related Works

This section describes the fundamental concepts and related work on medical image segmentation, classification, and explainability methods.

Segmentation. Image segmentation can be defined as the partition of the contextual image into a set of anatomical structures or regions of interest. In a medical scenario, segmentation is a crucial step that aims to clarify anatomical or pathological structure changes, thereby improving diagnosis efficiency and accuracy [13, 14]. Medical image segmentation can be performed through different approaches: manual segmentation, where experts manually delineate regions of interest; semi-automatic segmentation, which combines manual input with computational assistance; and fully automatic segmentation, which relies entirely on algorithmic approaches without human intervention [15]. [16] compared two CNNs: TernaNet-16 and Mask R-CNN to segment skin lesions in dermoscopic images, revealing that Mask R-CNN outperforms TernaNet-16 in accurately delineating lesions and attribute boundaries, with the best Mask R-CNN model achieving a Dice Similarity Coefficient of 84.76% for lesion segmentation [16]. Segmentation can also be used as a tool in classification tasks, serving as an essential preprocessing technique that shields a classifier from unnecessary information. Proper segmentation can increase the overall accuracy of medical image classifiers [9]. To demonstrate the use of iteratively pruned deep learning models for COVID-19 detection, [10] used a segmentation-classification pipeline to guard the training process against irrelevant features that could affect reliable decision-making. A U-Net model was trained to generate lung masks to be superimposed on Chest X-ray images, allowing for cropping as a bounding box that contains the lung pixels. The best-performing models, including VGG-16, VGG-19, and Inception-V3, were iteratively pruned to reduce complexity and improve memory efficiency, resulting in an accuracy of 99.01% with model ensemble.

Classification. Image classification is an essential step in clinical treatment, helping clinicians to diagnose and quantify the severity of diseases in the human body. Deep Learning techniques, such as Convolutional Neural Networks (CNNs),

significantly improve performance in medical image classification by extracting and selecting features from these samples to classify them [4]. Addressing the bias in automatic skin tone classification, [17] evaluated different preprocessing methods to ensure balanced performance despite the scarcity of annotated samples. The authors concluded that combining the best approaches for each skin color enhances accuracy across diverse skin types. Rodrigues et al. [18] compared four CNN architectures in a COVID-19 classification task, finding that shallow fine-tuning and data augmentation can assist the classification performance of AlexNet, VGG-11, SqueezeNet, and DenseNet-12 models with a small number of training samples.

Explainability. Proposed by [7], Grad-CAM produces class-discriminative localization maps applicable to any CNN-based differentiable model without re-training, and allows the visualization of the learned behavior of deep learning models [7, 10], aiding interpretability and bias detection via saliency maps. Saliency maps are visual explainability techniques that highlight the most relevant regions of an input concerning a model's prediction. These techniques assign importance scores to pixels or regions, enabling the interpretation of which parts of the input have the most significant influence on the network's decision. Several studies aim to validate whether the regions highlighted by saliency maps are relevant to the decision process, such as regions that a human expert would consider important for the model's decision, rather than relying on irrelevant features [19, 20, 21]. Rajaraman et al. [10] used Grad-CAM to verify the learned behavior of the pruned models, and it was observed that pruned models precisely localize the salient ROI when a visual comparison with ground truth annotations from specialists is applied. However, a quantitative analysis of the Grad-CAM's saliency maps wasn't applied. Kuiry et al. [22] point out that, while explainability methods contribute to model understanding, they also face important limitations in terms of stability, consistency, and computational efficiency.

3. CoSegXAI: Proposed Pipeline

This study proposed CoSegXAI, a four-phase experimental pipeline, to examine the trade-offs between classification performance and explainability in chest X-ray analysis. Figure 1 shows CoSegXAI phases: (i) **phase 1** is image processing, including dataset collecting and segmentation; (ii) **phase 2** is train and test Deep Learning models using hyperparameters; (iii) **phase 3** is the model performance evaluation using state-of-the-art metrics; and (iv) **phase 4** is visual and quantitative analysis of explainability using visual assessment and metrics to summarize the saliency maps.

3.1 Dataset

We employed the publicly available **COVID-19 Radiography Database** from Kaggle [11, 12] in **phase 1**, which provides an ideal foundation for comparing contextual versus ROI-based training. The dataset comprises 21,165 chest X-ray

images with a resolution of 299×299 pixels, paired with corresponding ground-truth lung segmentation masks. The dataset includes four diagnostic classes: 3,616 COVID-19 positive cases, 10,192 normal cases, 6,012 cases of lung opacity (non-COVID lung infections), and 1,345 other viral pneumonia cases.

This dataset was selected explicitly for three critical reasons that directly support our experimental objectives: (i) the availability of segmentation masks for all four classes enables reliable ROI extraction, (ii) the representation across respiratory conditions allows for robust multiclass analysis, and (iii) the substantial dataset size provides sufficient samples for both training and comprehensive explainability evaluation. Our comparative analysis encompasses the ground-truth masks, enabling the creation of two distinct experimental scenarios: contextual images (original radiographs) and ROI-based images (lung-segmented regions).

3.2 Data Preprocessing and Preparation

The **Phase 1** of CoSegXAI establishes the preprocessing flow that ensures consistent model training and evaluation. The dataset is randomly partitioned into training (70%), test (15%), and validation (15%) sets using a hold-out strategy. We organized identical splits across both contextual and ROI-based experimental scenarios, thereby avoiding overfitting and ensuring a fair performance comparison. The images are standardized to 224×224 pixels to match the input requirements of pre-trained CNN architectures. Table 1 demonstrates the sample size across the class on each subset.

Table 1. Dataset distribution by class and split.

Class	Train	Validation	Test
COVID	2531	542	543
Normal	7134	1528	1530
Viral Pneumonia	941	201	203
Lung Opacity	4208	901	903
Total	14814	3172	3179

3.3 Model Training and Architecture Settings

The **Phase 2** of CoSegXAI defines the training settings employing ResNet50 [23], DenseNet201 [24], and VGG19 [25] CNN architectures. We trained the architecture on contextual and ROI-based scenarios, resulting in six distinct model variants for comparative analysis.

We added a simple data augmentation strategy aimed at balancing the enhancement of generalization with the maintenance of evaluation consistency. Training employs a multi-stage pipeline: initial resize to 256×256 pixels, followed by random resized cropping to 224×224 (scale 0.8-1.0) to introduce spatial variability, random horizontal flip (probability 0.5), random rotation ($\pm 15^\circ$), and random erasing (probability 0.1, scale 0.02-0.2). These data augmentation variations enable models to observe different spatial regions during training, enhancing generalization capabilities. Conversely, validation and testing use only resizing to 224×224 pixels,

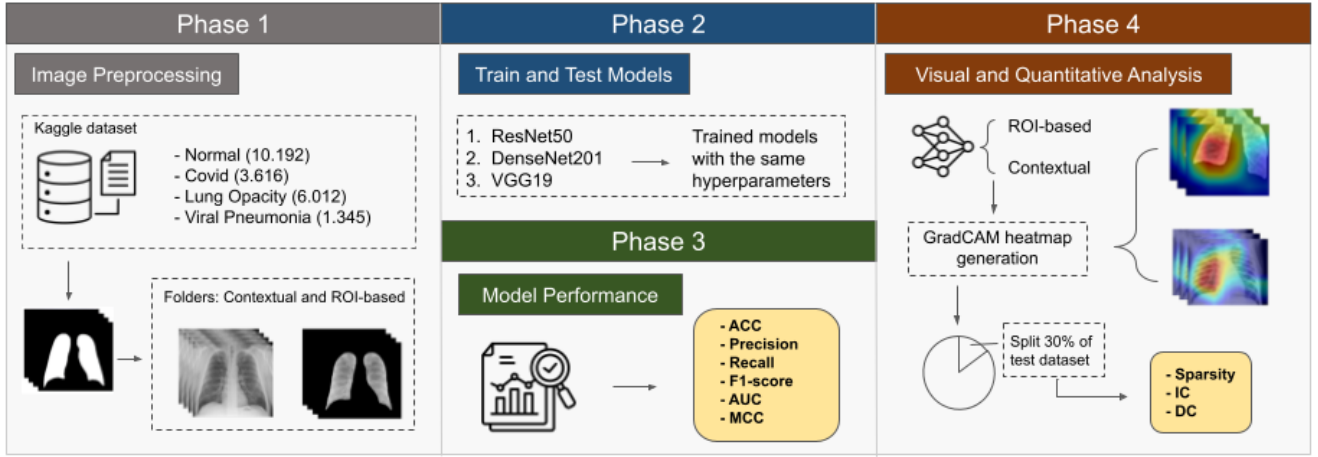


Figure 1. Overview of the pipeline CoSegXAI with the four phases proposed

ensuring consistent evaluation conditions across all model-dataset combinations.

3.4 Classification Performance Analysis

After training DL models as presented in the previous phase, the actual **Phase 3** aims to analyze the performance of multi-class classification. Each trained model was evaluated using state-of-the-art classification metrics: accuracy, precision, recall, F1-score, Area Under the Curve (AUC), and Matthew's Correlation Coefficient (MCC). This standardized evaluation protocol enables direct comparison between contextual and ROI-based training across different architectural choices, providing the performance baseline necessary for understanding the explainability trade-offs investigated in Phase 4.

3.5 Explainability Assessment

The **Phase 4** of CoSegXAI analyzes the model explainability through visual inspection and objective metrics to quantify saliency maps, such as Sparsity - Eq. (4), Deletion Correlation (DC) - Eq. (1), and Insertion Correlation (IC) - Eq. (2) [26]. Such metrics provide insight into how well the model's attention aligns with decision-making processes across both contextual and ROI-based training. Note that the formalization and application of these metrics in such a scenario is one of our contributions.

For measuring the saliency metrics, we consider a 30% split of the test dataset. This was chosen because computing these metrics is expensive; the subset selected was chosen randomly while preserving the class labels.

Deletion and Insertion Correlations address saliency map calibration by measuring the linear correlation between saliency scores and predicted confidence variations, rather than computing area under curves.

Formally, a Gaussian blur is used to degrade image regions while maintaining natural characteristics. The process sorts saliency map pixels from highest to lowest values, creating a priority queue for evaluation. At each iteration, mask regions are updated according to the saliency importance of each area

to ensure precise measurement of its contribution to the final prediction.

Given an input image $I \in \mathbb{R}^{H \times W \times 3}$ and saliency map $S \in \mathbb{R}^{H' \times W'}$, we sort S from the highest to the lowest element. At each $S_{i'j'}$, we create a blurred counterpart I_{blur} using Gaussian convolution with $\sigma = 8$. This process operates through $k = 4$ steps, where each step processes pixels ranked by saliency importance. The model is assessed on perturbed images to obtain confidence scores for the initially predicted class.

Deletion Correlation (DC) quantifies how well saliency scores predict confidence drops when replacing salient pixels with blurred versions:

$$DC = \text{corr}(v_k, s_k) = \text{corr}(c_k - c_{k+1}, s_k) \quad (1)$$

Insertion Correlation (IC) measures how well saliency scores predict confidence improvements when revealing salient pixels in a blurred image:

$$IC = \text{corr}(v_k, s_k) = \text{corr}(c_{k+1} - c_k, s_k) \quad (2)$$

where s_k represents the saliency score of the area masked at step k , and c_k is the model's confidence for the initially predicted class after progressively masking the most salient regions. High DC or IC values indicate well-calibrated saliency maps, implying that confidence score variations correspond to high saliency scores.

Sparsity complements these measures by quantifying attention concentration, with the highest values indicating focused activations within specific regions. In contrast, the lowest values suggest the most dispersed attention patterns, which may be less clinically interpretable. Formally, **Sparsity** evaluates the visual focus of a saliency map, indicating whether it concentrates on particular regions.

$$S' = \frac{S - S_{\min}}{S_{\max} - S_{\min}} \quad (3)$$

where $S_{\max}$ and $S_{\min}$ are the maximum and minimum scores of S . In this normalized form, $S'_{\max} = 1$, allowing sparsity to be expressed as:

$$\text{Sparsity} = \frac{1}{S'_{\text{mean}}} \quad (4)$$

4. Results and Discussion

This section presents the experimental results on the performance of models in contextual and ROI-based scenarios. Furthermore, we evaluate the explainability via saliency maps visualization and extract explainability scores using Sparsity, Insertion Correlation (IC), and Deletion Correlation (DC).

Experimental Setup: We conducted the experiments on a server equipped with an NVIDIA A100-SXM4-40GB GPU using Python version 3.13.5 and PyTorch [27] version 2.8.0 with CUDA 12.8 support. Furthermore, we set up the following hyperparameters across all scenarios: SGD optimizer, learning rate of 1×10^{-3} , momentum of 0.9, and weight decay of 1×10^{-4} . Additionally, we incorporate a dropout rate of 0.3 and gradient clipping with a maximum norm of 1.0 to prevent overfitting. Finally, the models are trained for 20 epochs using a batch size of 32.

4.1 Models Performance Evaluation

We evaluate the performance of models trained on contextual and ROI-based images. Table 2 presents performance scores across different architectures, ranging from contextual and ROI-based scenarios. Note that these data assess how different data representations affect classification metrics from DL models.

Table 2. Models' performance on contextual and ROI-based images

Model	ACC $\uparrow$	Pr $\uparrow$	Rc $\uparrow$	F1 $\uparrow$	AUC $\uparrow$	MCC $\uparrow$
<i>Contextual</i>						
ResNet50	90.81	91.45	90.81	90.71	98.30	0.860
DenseNet201	92.80	93.23	92.80	92.73	98.62	0.891
VGG19	88.17	89.44	88.17	87.92	98.12	0.821
<i>ROI-based</i>						
ResNet50	85.75	87.31	85.75	85.25	97.17	0.783
DenseNet201	88.08	88.96	88.08	87.79	97.90	0.818
VGG19	82.70	85.15	82.70	81.90	96.90	0.737

Note: Bold indicates the best value. All metrics (except MCC) are expressed in (%). An upward arrow ($\uparrow$) indicates that higher values represent better performance. **ACC:** Accuracy, **Pr:** Precision, **Rc:** Recall, **F1:** F1-score, **AUC:** Area Under Curve, **MCC:** Matthew's Correlation Coefficient.

The performance metrics presented in Table 2 consistently demonstrate DenseNet201's superiority across all metric values, achieving the highest performance regardless of whether trained on contextual or ROI-based samples. This superiority can be attributed to the network's architecture, since DenseNet201 is structured to reuse features and reduce parameter count. Notably, all three architectures exhibit superior performance when trained on contextual images compared

to their ROI-based versions. The worst model trained on contextual images, (VGG19 at 88.17% accuracy), is even slightly better than the best model trained on ROI-based images (DenseNet201 at 88.08% accuracy).

4.2 Explainability Analysis

This section explores the visual and quantitative analysis of explainability. Grad-CAM helps visualize the learning process of models trained with contextual or ROI-based samples. According to [9], this visualization can help ensure that the model captures the correct features. The following subsections detail these analyses.

4.3 Visual Assessment

Figure 2 presents examples of Grad-CAM saliency maps produced by our ResNet50, DenseNet201, and VGG19 alongside the original classification sample and its label for both dataset types. The red parts of the saliency maps indicate the primary section of the image that the network focused on during the classification. The orange/yellow regions indicate areas of medium importance, while green/blue areas represent the least important areas to classify a sample. Further quantitative analysis of visual representations provides a deeper understanding of the relationship between saliency maps and their corresponding quantitative results. Figure 2 corroborates these quantitative findings, showing that ROI-based models produce saliency maps with the highest Sparsity, which means more localized activation regions. In contrast, contextual models often exhibit diffuse attention patterns that can extend beyond the lung boundaries.

4.4 Quantitative Analysis

We evaluated the saliency maps from Grad-CAM, using sparsity, insertion correlation (IC), and deletion correlation (DC). Figure 3 presents the means for sparsity, IC, and DC regarding training scenarios and models. The x -axis displays the various models and dataset types used for training, while the y -axis represents the mean performance of each experiment. For the three metrics, VGG19 achieved the highest mean values for the ROI-based dataset scenario, whereas models trained within the contextual scenario presented the lowest metrics for sparsity and DC. Note that the lowest sparsity represents less focused features on saliency maps.

Moreover, models trained on contextual images may also focus on areas outside the lung regions, suggesting that they could rely on spurious background information for classification. Conversely, saliency maps from ROI-based models demonstrate more targeted attention within lung regions, achieving superior saliency metrics that indicate both more focused activations and stronger correlation between selected pixels and classification decisions. Note that the Deletion Correlation (DC) metric assesses how well saliency scores affect model confidence when salient regions are gradually blurred. ROI-based models showed significantly better performance on DC, with improvements ranging from 0.389 points for DenseNet201 (from 0.072 to 0.761) to 0.645 points

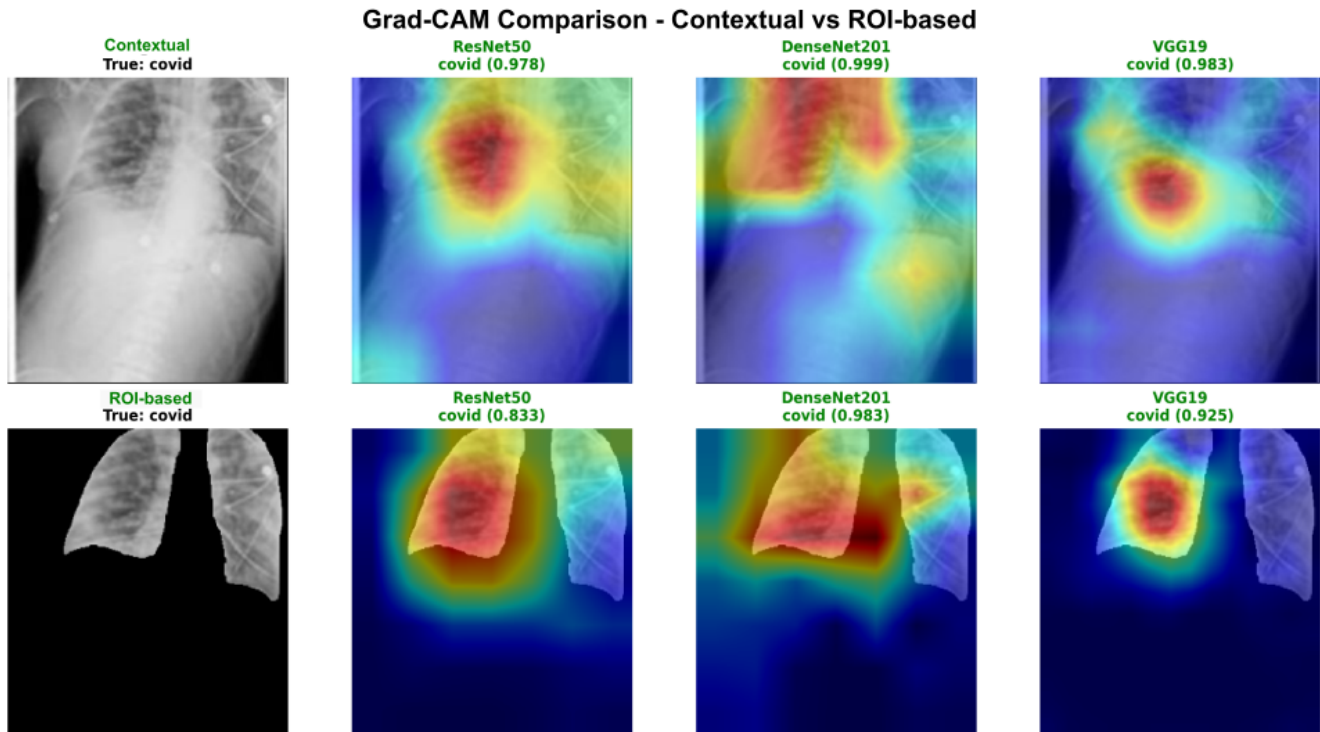


Figure 2. Visual comparisons of Grad-CAM generated by each model in both dataset types. The red areas of the saliency maps highlight the primary sections of the image that the network concentrated on during classification. The orange and yellow regions indicate areas of medium importance, while the green and blue areas represent the least important sections for classifying a sample.

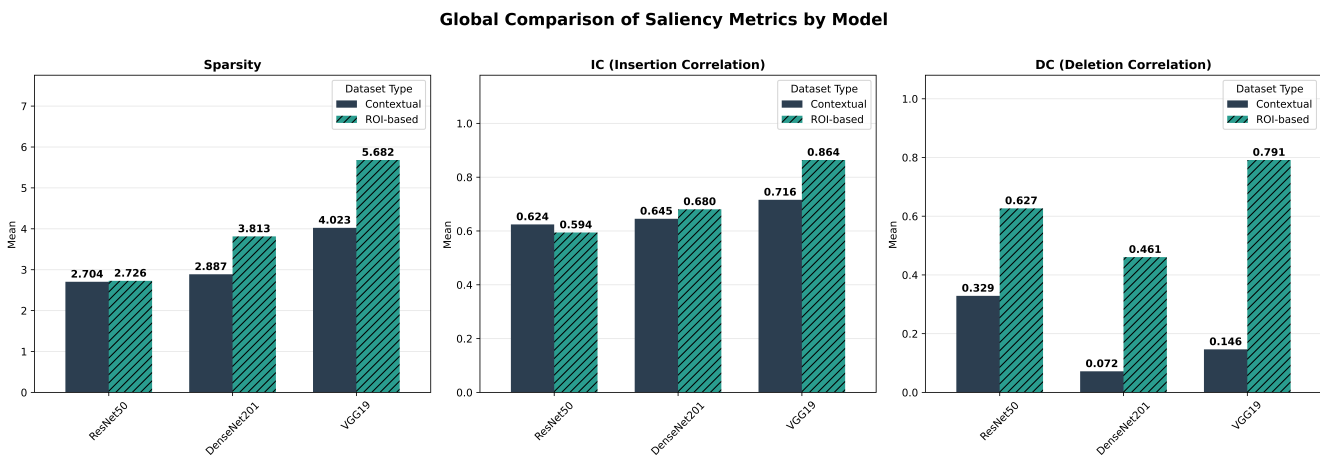


Figure 3. Quantitative comparison of saliency map (Sparsity, IC, and DC) obtained from ResNet50, DenseNet201, and VGG19 in contextual and ROI-based scenarios.

for VGG19 (from 0.146 to 0.791). This suggests that ROI-based models exhibit significant connections between the highlighted regions of saliency maps and the model's actual decision-making process.

4.5 Discussion

Our findings indicate a trade-off between classification performance and explainability in chest X-ray analysis. Models trained on contextual images outperformed models trained on

ROI-based images in quantitative metrics for all architectures. However, ROI-based images provided more interpretable saliency maps, focusing more consistently on the pathological regions, not relying on features that might not be related to the diagnosis. This suggests that contextual images might contain spurious patterns (e.g., image artifacts, external markings) that could mislead the model. DenseNet201 achieves the best performance, presenting accuracy that outperforms ROI-based models by 4.82%. However, the explainability

analysis reveals that models trained on ROI-based images produce more focused outputs, rather than models trained on contextual images, which achieve higher performance. Note that performance analysis alone cannot be enough to imply that a model is trustworthy.

Furthermore, our insights show that VGG19 exhibits the most disparity between performance and explainability. VGG19 achieves the lowest classification accuracy among contextual models (88.17%). However, VGG19 trained on ROI-based images demonstrated significant improvements in explainability, where Sparsity increased from 4.023 to 5.682, IC improved from 0.71 to 0.86, and DC improved from 0.146 to 0.791. These metrics in VGG19 assess the potential, even considering a simpler model, of this capability to be more explainable, despite the other model, more complex, like DenseNet201, having more expressive numbers.

Finally, the Deletion Correlation (DC), which assesses how well saliency scores affect model confidence when salient regions are gradually blurred. ROI-based models showed better DC, improving from 0.389 points for DenseNet201 (from 0.072 to 0.761) to 0.645 points for VGG19 (from 0.146 to 0.791). This suggests that ROI-based models exhibit significant connections between the highlighted regions of saliency maps and the model's actual decision-making process.

5. Conclusions

The key findings demonstrate that models trained on contextual (complete) images achieve superior classification performance. However, explainability analysis using Grad-CAM saliency maps and quantitative metrics (Insertion Correlation, Deletion Correlation, and Sparsity) consistently favors ROI-based models, which produce more focused attention patterns.

The superior performance of contextual models may be attributed to the utilization of non-pulmonary features, such as medical devices (wires and cables), extrapulmonary alterations (pleural effusion, cardiomegaly), and skeletal changes that correlate with patient age and clinical status. While these contextual cues provide valuable prognostic information in clinical practice, their incorporation into automated classification systems raises concerns about whether models are truly learning to detect lung opacities versus relying on indirect clinical markers that correlate with lung disease risk.

This observed pattern aligns with recent literature highlighting the complexity of medical image analysis, where optimal preprocessing strategies may vary depending on the primary objective. Our findings suggest that the choice between a contextual and an ROI-based pipeline should be informed by the specific clinical application, risk tolerance, and interpretability requirements rather than solely by classification performance metrics.

As future work, we intend to expand the analytical framework to incorporate additional explainability methods beyond Grad-CAM, including Grad-CAM++ and SHAP, thereby providing different views on the model's decision-making processes. These enhancements could support advancing the field

toward more robust and interpretable medical image analysis systems.

Acknowledgments

This study was financed in part by the São Paulo Research Foundation (FAPESP – grants 2021/08982-3, 2023/18026-8, 2024/13328-9, 2024/18940-4 and 2025/15353-3), the National Research Council (CNPq), and the Coordination for Higher Education Personnel Improvement (CAPES – Finance Codes 001, 88887.914087/2023-00, 88887.008174/2024-00)

Author Contributions

Graciella S. Favoreto: Conceptualization, model implementation, result analysis and drafting of the manuscript; Mariana Aya S. Uchida: writing - original draft, editing, and discussion of results; Pedro A. Luiz: writing - original draft, code, editing, and discussion of results; Erikson J. Aguiar: writing, editing, discussion of results and review; Marcel K. Santos: supervision and review; Agma Juci M. Traina: supervision and review. All authors read and approved the final version of the text.

References

- [1] BRIN, D. et al. Assessing gpt-4 multimodal performance in radiological image analysis. *European Radiology*, Springer, v. 35, n. 4, p. 1959–1965, 2025.
- [2] WEI, J. et al. Deep learning model for diagnosing and classifying subtypes of chronic pulmonary aspergillosis in chest ct. *Mycoses*, Wiley Online Library, v. 68, n. 4, p. e70061, 2025.
- [3] FU, X. et al. Explainable hybrid transformer for multi-classification of lung disease using chest x-rays. *Scientific Reports*, Nature Publishing Group UK London, v. 15, n. 1, p. 6650, 2025.
- [4] CHEN, C.; ISA, N. A. M.; LIU, X. A review of convolutional neural network based methods for medical image classification. *Computers in biology and medicine*, Elsevier, v. 185, p. 109507, 2025.
- [5] PLESNER, L. L. et al. Commercially available chest radiograph ai tools for detecting airspace disease, pneumothorax, and pleural effusion. *Radiology*, Radiological Society of North America, v. 308, n. 3, p. e231236, 2023.
- [6] PESSOA, D. et al. Ensemble deep learning model for dimensionless respiratory airflow estimation using respiratory sound. *Biomedical Signal Processing and Control*, Elsevier, v. 87, p. 105451, 2024.
- [7] SELVARAJU, R. R. et al. Grad-cam: Visual explanations from deep networks via gradient-based localization. In: *2017 IEEE International Conference on Computer Vision (ICCV)*. [S.l.: s.n.], 2017. p. 618–626.

- [8] VENKATESH, K. et al. Gradient-based saliency maps are not trustworthy visual explanations of automated ai musculoskeletal diagnoses. *Journal of Imaging Informatics in Medicine*, Springer, v. 37, n. 5, p. 2490–2499, 2024.
- [9] HERTEL, R.; BENLAMRI, R. A deep learning segmentation-classification pipeline for x-ray-based covid-19 diagnosis. *Biomedical Engineering Advances*, Elsevier, v. 3, p. 100041, 2022.
- [10] RAJARAMAN, S. et al. Iteratively pruned deep learning ensembles for covid-19 detection in chest x-rays. *IEEE Access*, IEEE, v. 8, p. 115041–115050, 2020.
- [11] RAHMAN, T. et al. Exploring the effect of image enhancement techniques on covid-19 detection using chest x-ray images. *Computers in biology and medicine*, Elsevier, v. 132, p. 104319, 2021.
- [12] CHOWDHURY, M. E. H. et al. Can ai help in screening viral and covid-19 pneumonia? *IEEE Access*, v. 8, p. 132665–132676, 2020. ISSN 2169-3536.
- [13] AZAD, R. et al. Medical image segmentation review: The success of u-net. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 46, n. 12, p. 10076–10095, 2024. ISSN 1939-3539.
- [14] WANG, R. et al. Medical image segmentation using deep learning: A survey. *IET Image Processing*, v. 16, n. 5, p. 1243–1267, 2022. ISSN 1751-9667.
- [15] KUIRY, S. et al. Ga-rise: Posthoc model agnostic explanations of black-box classifiers using genetic algorithm-based optimized masks—a case study on chest x-ray images. *International Journal on Artificial Intelligence Tools*, World Scientific, v. 34, n. 01, p. 2550006, 2025.
- [16] RODRIGUES, A. V. d. M.; OLIVEIRA, R. B. Segmentation of skin lesions and their attributes in dermatoscopic images based on convolutional neural networks. v. 32, p. 99–106, 2025. ISSN 2175-2745.
- [17] MANOEL, L. A. V.; PONTI, M. A. Leveraging optimal methods for diverse skin types classification in images for reduced bias. *Revista de Informática Teórica e Aplicada*, v. 32, n. 1, p. 250–256, 2025. ISSN 2175-2745.
- [18] RODRIGUES, L. et al. Evaluating convolutional neural networks for covid-19 classification in chest x-ray images. In: *XVI Workshop de Visão Computacional (WVC)*. SBC, 2020. p. 52–57. ISSN 0000-0000. Disponível em: <https://sol.sbc.org.br/index.php/wvc/article/view/13480>.
- [19] ARUN, N. et al. Assessing the trustworthiness of saliency maps for localizing abnormalities in medical imaging. *Radiology: Artificial Intelligence*, v. 3, n. 6, p. e200267, 2021. Disponível em: <https://doi.org/10.1148/ryai.2021200267>.
- [20] SAPORTA, A. et al. Benchmarking saliency methods for chest x-ray interpretation. *Nature Machine Intelligence*, v. 4, n. 10, p. 867–878, 2022. ISSN 2522-5839. Disponível em: <https://doi.org/10.1038/s42256-022-00536-x>.
- [21] ZENG, Y. et al. Inconsistency between human observation and deep learning models: Assessing validity of postmortem computed tomography diagnosis of drowning. *Journal of Imaging Informatics in Medicine*, v. 37, n. 3, p. 1–10, 2024. ISSN 2948-2933.
- [22] KUIRY, S. et al. Ga-rise: Posthoc model agnostic explanations of black-box classifiers using genetic algorithm-based optimized masks — a case study on chest x-ray images. *International Journal on Artificial Intelligence Tools*, World Scientific Publishing Company, 2025. Disponível em: <https://www.worldscientific.com/worldscinet/ijait>.
- [23] HE, K. et al. Deep residual learning for image recognition. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. [S.l.: s.n.], 2016. p. 770–778.
- [24] HUANG, G. et al. Densely connected convolutional networks. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. [S.l.: s.n.], 2017. p. 4700–4708.
- [25] SIMONYAN, K.; ZISSERMAN, A. *Very Deep Convolutional Networks for Large-Scale Image Recognition*. 2015. Disponível em: <https://arxiv.org/abs/1409.1556>.
- [26] GOMEZ, T.; FRÉOUR, T.; MOUCHÈRE, H. Metrics for Saliency Map Evaluation of Deep Learning Explanation Methods. *Emerging Topics in Pattern Recognition and Artificial Intelligence*, p. 101–121, 2022.
- [27] PASZKE, A. et al. *PyTorch: An Imperative Style, High-Performance Deep Learning Library*. 2019. Disponível em: <https://arxiv.org/abs/1912.01703>.