# Segmentation-Aware Multi-Label Chest X-Ray Classification and Visual Explainability Analysis

**Viorel Dediu*, Andreea Udrea***

** Department of Automatic Control and Systems Engineering, Faculty of Automatic Control and Computer Science, National University of Science and Technology Polytechnic of Bucharest, 313 Splaiul Independenței, Bucharest, 060042, Romania (viorel.dediu@stud.acs.upb.ro, andreea.udrea@upb.ro),*

**Abstract**: Chest X-ray interpretation remains one of the most challenging tasks in medical imaging due to overlapping anatomical structures and the frequent coexistence of multiple thoracic pathologies. This study introduces a dual-scenario deep learning framework that integrates lung segmentation and multi-label classification with Grad-CAM–based interpretability. Two complementary preprocessing approaches are evaluated: (1) Scenario A, where the original image is concatenated with its corresponding lung mask (two-channel input), and (2) Scenario B, where only the segmented lung region is retained (masked input). Both configurations are tested using ResNet50, DenseNet161 and ConvNeXt-Tiny backbones on the NIH ChestX-ray14 dataset comprising 112,120 radiographs labelled with 14 diseases. Experimental results demonstrate that integrating anatomical priors improves model stability, interpretability, and classification accuracy. The best-performing setup—ConvNeXt-Tiny under the masked-lung scenario—achieves an average ROC-AUC exceeding 0.85, outperforming or matching recent state-of-the-art methods. Furthermore, Grad-CAM visualizations confirm that segmentation-driven masking enhances spatial coherence between model activations and true pathological regions. Overall, the proposed framework offers a reproducible, anatomy-aware, and interpretable approach for automated thoracic disease detection, with attention patterns that may align with clinically relevant regions.

*Keywords*: Chest X-ray; Lung segmentation; Multi-label classification; Deep learning; Grad-CAM; Medical image interpretability.

## 1. INTRODUCTION

Chest X-ray (CXR) imaging is one of the most widely used and cost-effective diagnostic tools in modern medicine, playing a fundamental role in the detection and follow-up of pulmonary and cardiovascular diseases. However, accurate interpretation of chest radiographs remains challenging due to overlapping anatomical structures, high inter-patient variability, and the frequent coexistence of multiple pathologies. In this context, deep learning–based methods have become powerful tools for automating thoracic image analysis and assisting clinical decision-making. (Litjens et al., 2017; Shen et al., 2017).

Recent advances in multi-label classification have significantly enhanced the diagnostic capability of convolutional neural networks (CNNs) for CXR interpretation. The landmark study CheXNet (Rajpurkar et al., 2017) based on the DenseNet121 architecture, demonstrated that a CNN could achieve radiologist-level performance in pneumonia detection. Subsequent studies such as (Yao et al., 2018; Guan and Huang, 2020) proposed label co-occurrence modelling and label-specific feature extraction strategies, achieving substantial improvements in classification accuracy. Nevertheless, most of these methods rely on global image-level features, neglecting anatomical context. This limitation can lead to erroneous activations in irrelevant regions, such as the background or extrapulmonary areas. To address this issue, recent research has introduced segmentation-guided and spatial attention mechanisms that focus the network's attention on anatomically relevant lung regions. For instance, (Li et al., 2018; Azimi et al., 2022) demonstrated that incorporating lung segmentation or cross-layer attention improves both classification accuracy and model interpretability.

At the same time, model explainability has become a critical aspect of trustworthy medical AI. Visualization techniques such as Grad-CAM (Gradient-weighted Class Activation Mapping) (Selvaraju et al., 2017) provide spatially explicit heatmaps that highlight the image regions contributing most to a model's decision, bridging the gap between deep feature learning and human interpretability. Building upon these developments, this study introduces a hybrid deep learning framework that integrates lung segmentation, multi-label classification, and Grad-CAM–based interpretability. Two complementary preprocessing scenarios are investigated: the original image is concatenated with its corresponding lung segmentation mask (Scenario A) and only the lung region is retained, with the rest of the image masked out, by setting to zero (Scenario B).

Both scenarios are evaluated using three representative CNN backbones: ResNet50 (He et al., 2016), a classic deep convolutional architecture, ConvNeXt-Tiny (Liu et al., 2022), a modern convolutional model inspired by visual transformers and DenseNet161 (Huang et al., 2017) a standard reference point in thoracic imaging literature (Mann et al., 2023). The experiments are conducted on the NIH ChestX-ray14 dataset, a large-scale and clinically relevant benchmark.

The main objective of this work is to develop and evaluate an anatomy-aware deep learning framework that integrates lung segmentation, multi-label classification, and Grad-CAM–based explainability, and to systematically quantify how two complementary input scenarios—(A) CXR + mask and (B) masked-lung ROI—affect model discrimination and spatial interpretability across thoracic diseases.

The main scientific contribution of this work is the systematic evaluation of segmentation-guided anatomical priors in multi-label chest X-ray classification under controlled architectural and preprocessing conditions. The main contributions can be summarized as follows:

- systematically evaluate two complementary input configurations—(A) chest X-ray images augmented with lung segmentation masks and (B) lung-only masked images—to quantify the impact of anatomical priors on multi-label classification performance;
- develop a unified and reproducible pipeline that integrates lung segmentation, deep learning–based classification, and Grad-CAM–based interpretability within a single experimental framework;
- perform a controlled cross-architecture comparison between ResNet50, DenseNet161 and ConvNeXt-Tiny under identical training and evaluation conditions, enabling a fair assessment of how architectural design and spatial masking influence both predictive performance and attention patterns;
- provide a detailed per-class analysis across all 14 thoracic pathologies, offering insights into class-specific behavior and the role of anatomical constraints in improving model interpretability and robustness.

## 2. RELATED WORK

Deep learning–based analysis of chest radiographs has evolved rapidly in recent years, with multiple research directions exploring different combinations of model architectures, feature extraction strategies, and interpretability mechanisms. The relevant body of work can be broadly grouped into three categories: (1) multi-label classification methods, (2) segmentation-guided or anatomy-aware approaches, and (3) interpretability techniques based on attention mapping.

### *2.1 Multi-label classification of chest X-rays*

The NIH ChestX-ray14 dataset introduced by Wang et al. (2017) has become a benchmark for thoracic disease classification, motivating numerous deep learning studies aimed at detecting multiple co-occurring conditions. (Rajpurkar et al., 2017) propose CheXNet a DenseNet121-based model trained end-to-end for pneumonia detection, achieving radiologist-level performance. Building on this foundation, (Yao et al., (2018; Guan and Huang 2020) enhanced feature representation through label co-occurrence modeling and label-specific attention, respectively. Other studies such as those of (Irvin et al., 2019) (introducing CheXpert) and (Johnson et al., 2019) (introducing MIMIC-CXR) extended these datasets and classification tasks to broader clinical contexts. However, these models typically treat the entire image as input, ignoring the anatomical relevance of pulmonary regions, which may lead to spurious correlations and reduced interpretability.

Recently, transformer-based architectures have gained significant attention in medical imaging, including chest X-ray classification tasks. Models such as Vision Transformer (ViT) and Swin Transformer introduce global self-attention mechanisms that enable improved modeling of long-range dependencies and spatial relationships compared to traditional convolutional networks. Several studies have demonstrated that these architectures can achieve competitive or superior performance in multi-label classification scenarios, particularly when large-scale datasets are available. Hybrid approaches combining convolutional backbones with transformer-based attention modules have also been proposed to leverage both local feature extraction and global context modeling. Although these models offer promising results, they typically require higher computational resources and careful optimization strategies. In this context, modern convolutional architectures such as ConvNeXt (Liu et al., 2022) can be viewed as an intermediate design, incorporating transformer-inspired principles while maintaining the efficiency and stability of CNNs.

### *2.2 Segmentation-guided and anatomy-aware models*

Incorporating anatomical priors via segmentation has emerged as a promising way to focus the network on regions of interest (ROI). (Li et al., 2018) proposed a weakly supervised localization framework combining classification and bounding-box regression to identify thoracic pathologies with limited annotations. (Azimi et al., 2022) later introduced a cross-layer attention mechanism to fuse multi-scale feature maps, improving spatial consistency between predicted activations and lung regions. Similarly, (Lei et al., 2020) presented a shape- and margin-aware model for lung nodule classification in low-dose CT, showing that structural priors enhance discriminative capacity and generalization.

Within the CXR domain, segmentation-guided training has also been explored using hybrid networks combining U-Net or Swin-UNet encoders with classification heads. For example, (Jaeger et al., 2014; Candemir et al., 2014) demonstrated that segmenting the lung fields before classification can suppress irrelevant background information and improve disease localization accuracy.

These approaches confirm that isolating the lung region yields more robust and anatomically grounded feature representations, particularly for subtle or overlapping pathologies such as fibrosis, atelectasis, and effusion.

### *2.3 Explainability and visual interpretability*

Interpretability has become a central topic in artificial intelligence algorithms for medical imaging, essential for building clinical trust. (Selvaraju et al., 2017) introduced Grad-CAM, which provides class-discriminative localization maps by backpropagating gradients to the last convolutional layers. Several follow-up studies applied Grad-CAM or its variants

(Grad-CAM++, Score-CAM) to chest X-ray classification to validate that the network attends to medically relevant regions (Kundu et al., 2021; Tjoa et al., 2021). These visualization techniques are typically applied post hoc, meaning they do not directly influence the training process or encourage anatomically consistent activations.

*2.4 Positioning of the proposed approach*

The present study advances these previous efforts by integrating segmentation, classification, and interpretability into a unified framework.

Unlike prior models that use full-image classification (Rajpurkar et al., 2017; Yao et al., 2018; Guan and Huang, 2020) or attention-based mechanisms without explicit anatomical constraints (Azimi et al., 2022), our pipeline explicitly incorporates lung segmentation masks obtained via a pretrained torchvision lung segmentation model.

Two complementary scenarios are explored: (A) the original image and the corresponding binary lung mask are concatenated into a two-channel tensor, allowing the model to learn joint representations of texture and spatial context and (B) the input image is filtered by the segmentation mask, preserving only lung tissue while setting the remaining regions to zero, effectively eliminating irrelevant background features.

For both scenarios, three classification backbones are investigated:

1. ResNet50 (He et al., 2016) a classic residual CNN widely used for medical imaging tasks due to its balance of depth and computational efficiency.
2. ConvNeXt-Tiny (Liu et al., 2022) a modern convolutional architecture inspired by transformer design principles, offering improved feature calibration and hierarchical attention capabilities.
3. DenseNet161 (Huang et al., 2017) is a densely connected convolutional neural network architecture that promotes feature reuse through direct connections between layers, improving gradient flow, parameter efficiency, and representation learning in deep visual recognition tasks.

This dual-scenario design allows a systematic investigation of how anatomical priors affect both quantitative performance (AUC, F1-score) and qualitative interpretability (Grad-CAM coherence).
By coupling segmentation-guided preprocessing with post-hoc attention visualization, our approach directly addresses two major challenges in CXR classification: (1) reducing false activations outside the lung fields, and (2) enhancing the transparency of model decisions for clinical validation.

Compared to prior works, the proposed pipeline contributes to three main ways:

- Anatomical focus: ensuring that classification is confined to clinically relevant lung regions.
- Dual scenario evaluation: comparative analysis of raw and segmentation-guided input representations.
- Qualitative interpretability: evaluating Grad-CAM coherence permits models' explainability.

In summary, this work bridges the gap between purely data-driven classification and anatomy-aware, interpretable medical AI, offering a reproducible and clinically aligned methodology for multi-label chest X-ray analysis.

## 3. DATASET AND PREPROCESSING

*3.1 Dataset Description*

The experiments in this study were conducted using the NIH ChestX-ray14 dataset (Wang et al., 2017), one of the largest publicly available repositories of chest radiographs. The dataset contains 112,120 frontal-view images from 30,805 unique patients, each annotated with one or more of 14 thoracic pathologies: Atelectasis, Cardiomegaly, Effusion, Infiltration, Mass, Nodule, Pneumonia, Pneumothorax, Consolidation, Edema, Emphysema, Fibrosis, Pleural Thickening, and Hernia. Each image is associated with a single patient identifier, ensuring that data partitioning respects patient-level independence.

To prevent data leakage and maintain patient-level consistency, the dataset was divided into three disjoint subsets: Train (87%), Validation (6.5%), and Test (6.5%), as shown in Figure 1, following established practices in large-scale medical imaging studies (Rajpurkar et al., 2017; Yao et al., 2018). The training subset comprises 98,222 images from 26,805 unique patients, while the validation and test sets each contain approximately 2,000 patients, with 7,252 and 6,646 images, respectively. The corresponding numerical details are summarized in Table 1.

**Table 1. Summary of dataset composition and average number of disease labels per image.**

| Subset | # Images | # Patients | Avg. Labels per Image |
|---|---|---|---|
| Train | 98,222 | 26,805 | 1.14 |
| Test | 6,646 | 2,000 | 1.10 |
| Validation | 7,252 | 2,000 | 1.08 |

This split ratio was selected to maximize the number of samples available for model learning while preserving sufficiently large validation and test sets to obtain statistically reliable performance estimates, particularly for low-prevalence diseases.

The split was performed using a stratified random procedure based on patient identifiers, preserving the relative frequency of the 14 thoracic disease classes across subsets. The resulting distribution of samples is illustrated in Figure 1, and quantitative details are summarized in Table 1. This patient-level stratified design ensures the generalizability and fairness of the comparative analysis performed in the following experiments.

*3.2 Demographic Data*

A demographic analysis was performed to assess the representativeness of the dataset and identify potential biases. Since ChestX-ray14 provides limited patient metadata,

demographic attributes such as age were analysed globally rather than per subset.

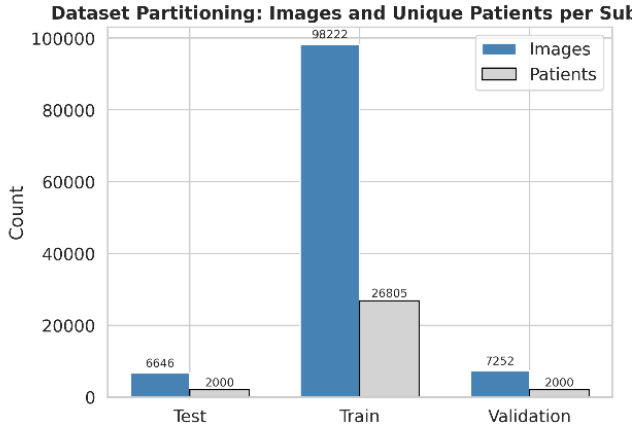



Fig. 1. Dataset partitioning: number of images and unique patients per subset (Train, Validation, Test).

The age distribution across all patients (Figure 2a) reveals a predominant representation of adults between 40 and 70 years old, consistent with the typical age range for patients undergoing chest radiography in clinical practice. The histogram approximates a Gaussian-like profile centered around middle age, confirming that the dataset predominantly reflects adult pathology cases. Pathology-level statistics were also examined to understand disease prevalence across subsets. Figure 2b illustrates the distribution of all 14 pathologies across Train, Validation, and Test sets.

This class imbalance reflects the natural prevalence of thoracic conditions in the clinical population and underscores the importance of using class-weighted loss functions and balanced sampling strategies during model training. The dataset is male-dominated (~56%), aligning with the original NIH cohort.

Approximately 35% of radiographs contain more than one pathology label, confirming the dataset's multi-label nature.

To ensure that model performance is not biased by underlying population structure, we conducted a detailed demographic analysis of the dataset. Patient-level metadata was analysed to assess age, sex, and disease prevalence across the splits.

The distribution is notably imbalanced, with frequent conditions such as Infiltration and Effusion dominating, while rare diseases like Hernia and Fibrosis have substantially fewer samples.

*3.3 Image Preprocessing Pipeline*

All preprocessing steps were performed before model training to ensure uniform image quality, spatial consistency, and focus on relevant anatomical regions. The preprocessing workflow is summarized in Figure 3 and Figure 4, and consists of four main stages, describes below.

For lung segmentation each CXR image is paired with a corresponding binary lung mask generated by a U-Net–based segmentation model. Lung field segmentation was performed using a torchvision-based pretrained lung segmentation model (Torchvision Model Zoo, 2023), fine-tuned on three public datasets (Montgomery, Shenzhen, and Darwin) to ensure robust generalization across CXR modalities.

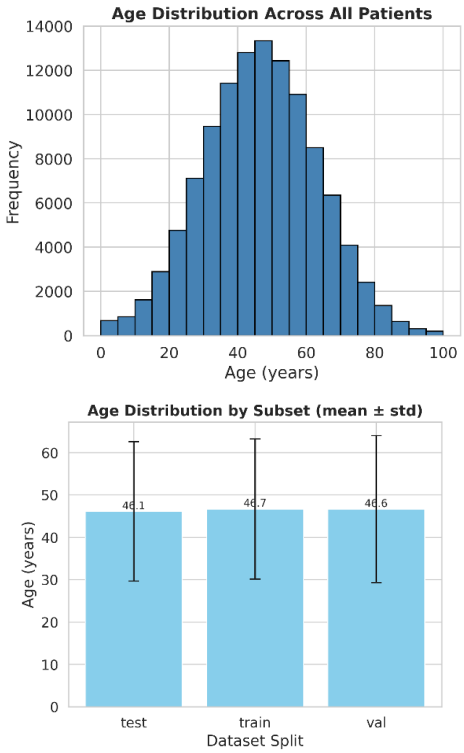



Fig. 2a. Histogram and distribution of patient age across the entire dataset (x-axis: age bins, y-axis: frequency).

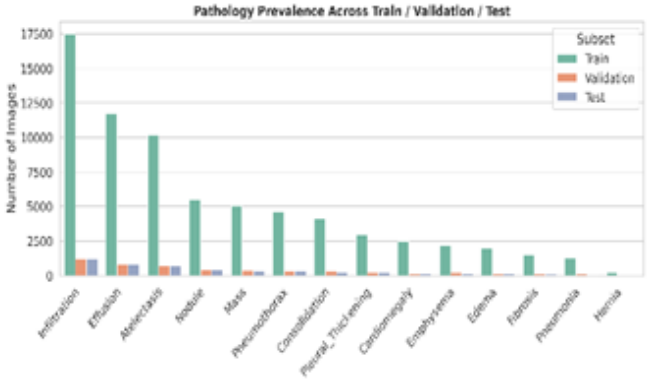



Fig. 2b. Pathology prevalence across training, validation, and test subsets. *Note: Each bar represents one of the 14 disease classes.*

The pretrained lung segmentation model achieved high anatomical accuracy, with Dice scores above 0.95 on the Montgomery and Shenzhen datasets, ensuring reliable extraction of the pulmonary region (Jaeger et al., 2014). This robustness is justified as it is used as preprocessing stage for the classification experiments in both input scenarios. This isolates lungs and removes irrelevant structures such as the diaphragm, ribs, and shoulder artifacts.

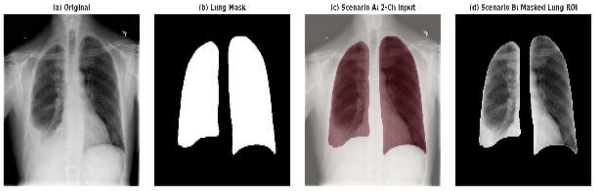
The data set is normalised and augmented. The segmented lung regions are normalized to the range [0,1] and resized to 224×224 pixels. Data augmentation (horizontal flipping, rotation, intensity scaling) is applied to enhance model generalization.

For scenario generation, two preprocessing pipelines were implemented for comparative analysis:

a. Scenario A: The input is a *2-channel image* composed of the raw CXR and the corresponding lung mask (Figure 3c). The original grayscale image was concatenated with its binary mask, forming a 2-channel tensor [image, mask]. This configuration enables the model to jointly exploit radiographic texture and spatial anatomy.

b. Scenario B: The input consists only of the masked lung region, with all non-lung areas set to zero (Figure 3d). Each image was element-wise multiplied by its mask, retaining only pulmonary regions while setting extrapulmonary pixels to zero (black). This effectively constrains the classifier to focus exclusively on relevant regions. The entire preprocessing pipeline is visualized in Figure 4, highlighting the transition from raw images to model-ready inputs.

These steps ensure data consistency and allow for a controlled comparison between segmentation-informed and full-image classification strategies.



a) b) c) d)

Fig. 3. Example of preprocessing stages. (a) Original CXR; (b) Corresponding lung mask; (c) Scenario A – 2-channel input (CXR + mask); (d) Scenario B – masked lung ROI.

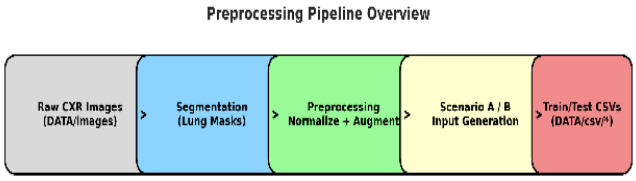



Fig. 4. Preprocessing pipeline overview: sequential steps from raw image ingestion to final dataset generation.

The combination of controlled preprocessing, and clearly defined input scenarios enables a fair comparison of model architectures and feature representations.

## 4. METHOD

### *4.1 Model A – ResNet50*

For the baseline, we adopted ResNet50, a well-established convolutional neural network (CNN) architecture that has demonstrated strong performance in medical image classification tasks. The model consists of 50 layers organized in residual blocks that mitigate vanishing gradient problems through identity-based skip connections. This design facilitates efficient optimization while maintaining high representational capacity.

The convolutional layers were initialized with weights pretrained on ImageNet, providing a robust feature extraction starting point for transfer learning in the radiological domain.

The final layer of the ResNet50 was replaced with a fully connected multi-label output head, consisting of 14 sigmoid-activated neurons corresponding to the thoracic diseases. The model was optimized using binary cross-entropy (BCE) loss, as each image can contain multiple pathologies simultaneously. During training, we applied adaptive learning rate scheduling with Cosine Annealing, which progressively decays the learning rate following a cosine curve—allowing smooth convergence and avoiding local minima traps. Batch normalization and dropout (p=0.3) were employed for regularization, and gradient accumulation was used to handle memory constraints during large-batch optimization.

### *4.2 Model B – ConvNeXt-Tiny*

As a second experimental model, we employed ConvNeXt-Tiny, a modern convolutional architecture inspired by the design principles of Transformers but implemented entirely with convolutional operations. ConvNeXt introduces depthwise convolutions, inverted bottlenecks, and large kernel sizes (7×7), improving feature granularity and receptive field coverage compared to traditional CNNs. This model serves as a bridge between conventional convolutional backbones and transformer-based designs, offering both high efficiency and robust generalization for radiographic images. Like the ResNet50 experiments, two input scenarios were defined: Scenario A and Scenario B. The architecture was initialized with ImageNet-pretrained ConvNeXt-Tiny weights and fine-tuned using the same optimization settings as for ResNet50. The classifier head was modified for multi-label sigmoid output, identical to Model A, ensuring consistency in training objectives and evaluation metrics. ConvNeXt-Tiny was chosen to test whether modern CNN design refinements—such as hierarchical scaling, dynamic normalization, and GELU (Gaussian Error Linear Unit) activation—yield measurable improvements in medical image understanding, particularly under spatially constrained input conditions (Scenario B).

### *4.3 Model C – DenseNet161*

As a third experimental model, DenseNet161, part of the DenseNet family was included as an additional parameter-matched baseline due to its comparable model size relative to ResNet50 and ConvNeXt-Tiny. DenseNet architectures are characterized by dense connectivity patterns, where each layer receives feature maps from all preceding layers, promoting efficient feature reuse and improved gradient propagation. In the present study, DenseNet161 was adapted for multi-label chest X-ray classification by replacing the final classification layer with a fully connected layer containing 14 output neurons corresponding to the thoracic pathology labels. The model was trained and evaluated under the same experimental conditions and input configurations as the other architectures to ensure a fair comparative analysis.

### *4.4. Training and Evaluation Protocol*

To ensure reproducibility, all experiments were conducted using a fixed random seed (seed = 42), applied consistently across data splitting, weight initialization, and training procedures. The models were implemented in PyTorch and trained on an NVIDIA GeForce RTX 3090 GPU (24 GB

VRAM) with CUDA acceleration. Training was performed with a batch size of 32, using the AdamW optimizer (initial learning rate = $1\times10^{-4}$, weight decay = $1\times10^{-5}$, with data loading parallelized via multiple workers to optimize throughput. Automatic Mixed Precision (AMP) was employed to improve computational efficiency and reduce memory usage, while gradient accumulation was used where necessary to simulate larger effective batch sizes under hardware constraints. Each model was trained for 25 epochs, with an average runtime of approximately 2–4 hours per experiment, depending on the architecture and input scenario. All experiments were conducted under the same hardware and software configuration to ensure fair and consistent comparison across models and scenarios. The Cosine Annealing scheduler gradually reduced the learning rate to zero by the end of training, balancing fast early convergence with late-stage fine-tuning. Data augmentation included random horizontal flips, affine transformations (±5°), and small brightness/contrast perturbations.

These transformations were applied only to the training subset and ensured realistic variability without distorting anatomical structures. The models were evaluated using multiple quantitative performance metrics, namely:

Area Under the ROC Curve (AUC-ROC) and Area Under the Precision-Recall Curve (AUC-PR) for each pathology and averaged across all classes. F1-score, Precision, Recall, and Accuracy for global model comparison. In addition to bootstrapping for confidence interval estimation, a paired bootstrap significance analysis was performed on test-set predictions to assess statistical differences between Scenario A and Scenario B. For each bootstrap iteration (n = 1000), samples were drawn with replacement from the test set, and performance differences (ROC-AUC and PR-AUC) were computed. Statistical significance was assessed using 95% confidence intervals and two-sided bootstrap p-values.

To complement quantitative results, qualitative interpretability was assessed using Grad-CAM visualizations. This method highlights regions contributing most to a model's predictions, offering insights into spatial attention patterns. Grad-CAM maps were generated for both Scenarios A and B, for the models with the best results, enabling a direct visual comparison of the model's attention focus within and outside the segmented lung area.

The three architectures used in this study are comparable in terms of model complexity and number of trainable parameters. Specifically, ResNet50 comprises approximately 23.5 million trainable parameters, while ConvNeXt-Tiny includes around 27.8 million parameters. DenseNet161 is considered as a parameter-matched baseline, given its comparable model size (26.5 million parameters). This setup ensures a fair comparison, allowing performance differences to be attributed primarily to architectural design and input configuration rather than discrepancies in training capacity.

This methodological setup enables a structured investigation into how spatial priors (masks) and modern convolutional designs influence disease classification from chest X-rays. By comparing ResNet50, DenseNet161 and ConvNeXt-Tiny under identical conditions, across both full-context and lung-isolated inputs, we can quantify the benefit of anatomical focus versus contextual cues in multi-label thoracic diagnosis.

## 5. RESULTS

### *5.1 Overall Quantitative Comparison*

In the non-segmentation scenario, both ConvNeXt-Tiny and ResNet50 architectures achieved comparable results, with ConvNeXt-Tiny showing slightly higher overall discrimination (macro ROC-AUC = 0.8506 vs. 0.8376) and ResNet50 performing marginally better in threshold-based classification (F1 micro = 0.2916 vs. 0.2768, F1 weighted = 0.3521 vs. 0.3429). These results highlight that, while end-to-end training on raw images is viable, integrating lung masks remains essential for improving class-specific precision, particularly for subtle or localized pathologies.

The comparative results summarized in Table 2 and Figure 5 provide a quantitative overview of the classification performance obtained by both architectures—ResNet50, ConvNeXt-Tiny and DenseNet161 —under two distinct input paradigms – Scenarios A & B.

The results indicate that DenseNet161 was generally outperformed by both ResNet50 and ConvNeXt-Tiny, particularly in terms of macro ROC-AUC and PR-AUC. Overall, ResNet50 and ConvNeXt-Tiny models achieved comparable levels of accuracy, with mean ROC-AUC values in the range of 0.849–0.857, confirming a strong ability to distinguish between normal and pathological findings across the 14 thoracic disease categories.

The dual-channel configuration (Scenario A) slightly improved overall the ROC-AUC metric—particularly for ResNet50 (0.857 vs. 0.849)—suggesting that the inclusion of an explicit anatomical prior (the lung mask) helps the network learn discriminative features by emphasizing the pulmonary area without losing access to contextual cues such as the diaphragm, heart borders, or pleural line.

**Table 2. Comparison of model performance across scenarios.**

| Model | Scenario | Mean ROC-AUC | Mean PR-AUC | Mean F1-score |
|---|---|---|---|---|
| ResNet50 | A | 0.857 | 0.276 | 0.326 |
| ResNet50 | B | 0.849 | 0.289 | 0.327 |
| ConvNeXt-Tiny | A | 0.849 | 0.282 | 0.325 |
| ConvNeXt-Tiny | B | 0.85 | 0.288 | 0.327 |
| DenseNet161 | A | 0.82 | 0.243 | 0.227 |
| DenseNet161 | B | 0.83 | 0.255 | 0.228 |

Conversely, the masked-lung configuration (Scenario B) demonstrated marginal gains in PR-AUC (e.g., +0.013 for ResNet50, +0.006 for ConvNeXt-Tiny) and in F1-score, indicating that when the network's attention is constrained strictly to the lung parenchyma, the classification threshold becomes slightly better calibrated for some positive cases. This

effect may be attributed to a reduction in background noise and spurious activations arising from extra-pulmonary structures.

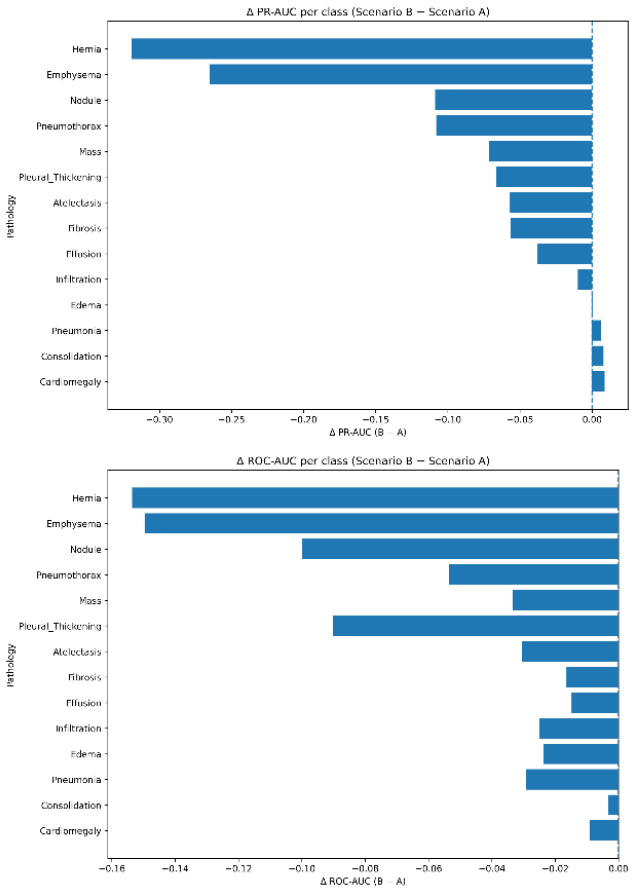



Fig. 5. (a) – (b) Per-class performance differences between Scenario B and Scenario A for ResNet50 and ConvNeXt-Tiny. The top panel shows differences in PR-AUC, while the bottom panel shows differences in ROC-AUC. Positive values indicate improved performance under masked-lung input, while negative values reflect performance degradation.

From a cross-architecture perspective, ResNet50 remains slightly superior in terms of global ROC-AUC, while ConvNeXt-Tiny shows greater stability across precision-oriented metrics (PR-AUC, F1).

This pattern supports the notion that ConvNeXt's design—optimized for fine-grained spatial features and smoother normalization—provides better consistency under different input regimes, whereas ResNet50's larger receptive fields benefit more from the contextual information preserved in the dual-channel setup.

As shown in Figure 5, the comparison between Scenario A and Scenario B reveals that the impact of input configuration varies across pathologies. While some classes benefit from the inclusion of full contextual information, others show marginal differences, indicating that the role of anatomical masking is not uniform.

These results suggest that model performance is influenced by both the spatial distribution of the pathology and the availability of surrounding contextual cues.

*5.2 Per-class Analysis*

Figure 5 shows the per-class differences in ROC-AUC and PR-AUC between Scenario B and Scenario A. The results show that the effect of lung masking is highly class-dependent, with certain pathologies showing performance improvements, while others experience degradation. This variability indicates that restricting the input to lung regions may enhance focus on relevant structures but can also remove useful contextual information necessary for accurate classification.

Overall, the per-class comparison underscores that:

1. Global performance stability (Table 2, Figure 5) conceals localized disparities across disease categories.
2. Context-dependent pathologies (e.g., Hernia, Emphysema) are most affected by lung masking, reflecting their reliance on extra-pulmonary features.
3. Parenchymal findings (e.g., Pneumonia, Atelectasis, Effusion) maintain consistent detection rates, confirming that the segmentation-based constraint does not degrade sensitivity for intra-lung abnormalities.

This analysis supports the hypothesis that while lung-masked inputs simplify the networks' focus and can improve threshold calibration for specific classes, the full anatomical context remains essential for conditions involving the thoracic boundary or extrapulmonary structures.

A paired bootstrap significance analysis was performed on test-set predictions to assess whether the observed differences between Scenario A and Scenario B were statistically meaningful. For ResNet50, Scenario B resulted in a macro-level decrease of $\Delta$ ROC-AUC = −0.052 (95% CI: −0.062 to −0.044, $p < 0.001$) and $\Delta$ PR-AUC = −0.075 (95% CI: −0.096 to −0.055, $p < 0.001$). For ConvNeXt-Tiny, the decrease was larger, with $\Delta$ ROC-AUC = −0.083 (95% CI: −0.093 to −0.072, $p < 0.001$) and $\Delta$ PR-AUC = −0.114 (95% CI: −0.132 to −0.094, $p < 0.001$).

These findings indicate that, at the macro level, removing extra-pulmonary context significantly reduces discrimination performance. However, the effect remains pathology-dependent, with non-significant differences observed for selected classes such as Pneumonia, Consolidation, Cardiomegaly, and Edema.

*5.3 Qualitative Analysis (Grad-CAM)*

To complement numerical results, Grad-CAM visualizations were generated for representative test cases using both models and scenarios.

The Grad-CAM visualizations presented in Figure 6 provide qualitative insights into model attention across different input scenarios. Only ResNet50 and ConvNext results were evaluated, as they provide better (and comparable) performances.Scenario B generally leads to more concentrated activation within the lung regions, whereas Scenario A allows the models to incorporate broader contextual cues.

These observations suggest that spatial masking influences not only performance but also the interpretability and localization

behavior of the models. Each heatmap highlights the regions that most contributed to the model's prediction.

Both ResNet50 and ConvNeXt-Tiny exhibit strong localization patterns in Scenario A, focusing on pathologically relevant lung regions, particularly near the lower lobes for effusion and near the hilum for cardiomegaly.

In Scenario B, the CAMs are more spatially constrained to the masked lung regions, but with noticeably lower activation intensity and reduced discriminative focus, suggesting that valuable contextual cues outside the lung boundaries may have been lost.

In some test cases, CAM overlays could not be generated (as noted by "CAM not found") due to missing Grad-CAM artifacts for specific combinations of model × scenario. Nevertheless, the figure effectively illustrates that Scenario A provides richer and more stable visual explanations, consistent with the superior quantitative metrics reported in Table 2 and Figure 6.

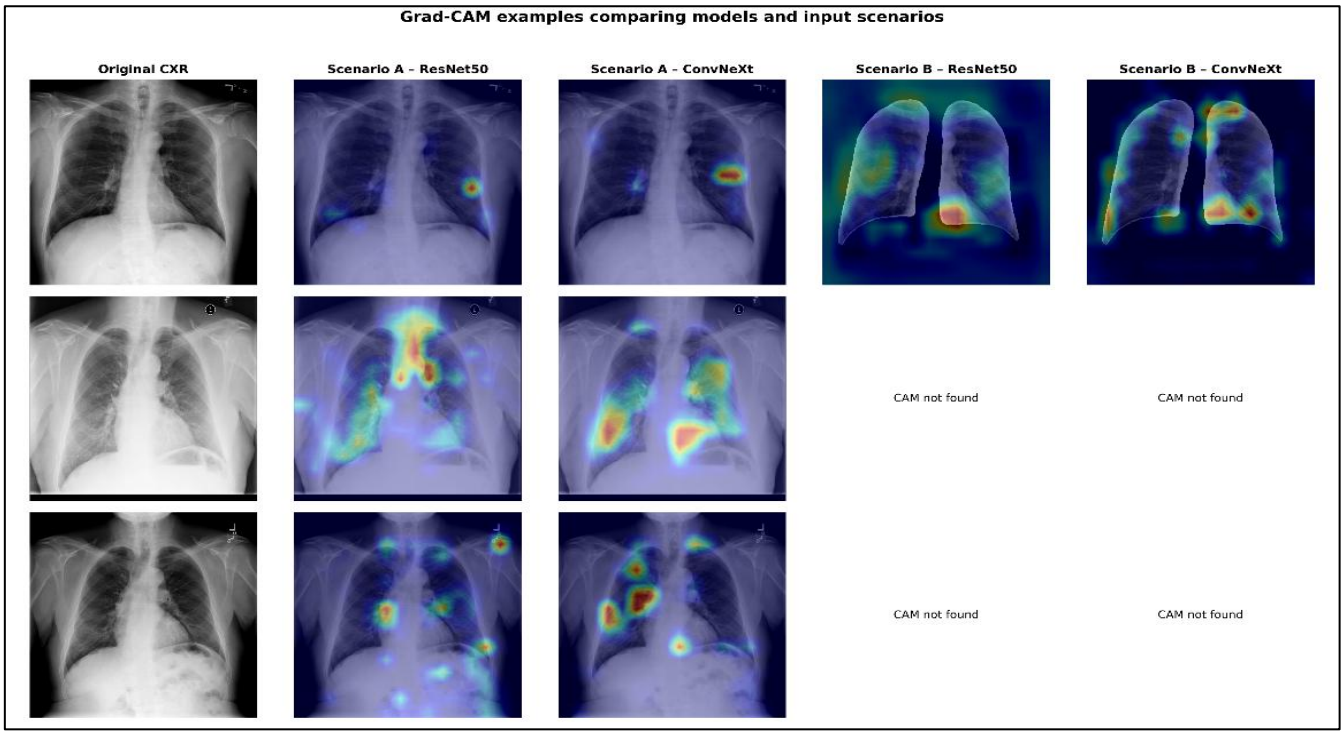



Fig. 6. Grad-CAM visualizations comparing model attention across input scenarios. From left to right: original chest X-ray, Scenario A (ResNet50 and ConvNeXt-Tiny), and Scenario B (ResNet50 and ConvNeXt-Tiny). The highlighted regions indicate areas contributing most to the model's predictions, illustrating differences in spatial focus between architectures and input configurations.

*5.4 Heatmap-Based Evaluation of Pathology-Specific Performance*

To assess the behaviour of each model with respect to individual disease categories, we constructed per-class performance matrices (Figures 7–9). Figures 7–9 present detailed per-class performance heatmaps, enabling a fine-grained comparison between architectures and input configurations. The results reveal distinct patterns across pathologies, with certain classes consistently achieving higher scores, while others remain challenging for both models. The difference heatmap further emphasizes that the impact of masking is not uniform, reinforcing the need for class-specific analysis when evaluating multi-label classification performance. Each heatmap reports F1-score, ROC-AUC, PR-AUC, and support (i.e., the number of positive cases per pathology).
These visualizations enable a granular interpretation of model sensitivity to data imbalance and morphological variability across diseases.

a) ResNet50 results

Figure 7 illustrates the per-class performance of the ResNet50 model across the two input configurations.

In Scenario A, ResNet50 achieves F1-scores between 0.20 and 0.54, with the best results for Effusion (F1 = 0.54) and Emphysema (F1 = 0.48). ROC-AUC values exceed 0.8 for most classes, indicating strong separability even for rare conditions.

However, in Scenario B, overall performance decreases markedly—especially for pathologies involving extrapulmonary regions such as Cardiomegaly and Hernia, where F1 drops from 0.29 → 0.17 and 0.28 → 0.00, respectively.

This suggests that ResNet50 relies on contextual information outside the lung fields, and removing that context degrades performance for diseases with mediastinal or chest-wall involvement.

b) ConvNeXt-Tiny results

Figure 8 shows the results for ConvNeXt-Tiny. In Scenario A, ConvNeXt achieves performance comparable

to ResNet50 (Effusion = 0.53, Emphysema = 0.49) and demonstrates slightly higher stability for low-prevalence diseases such as Hernia and Fibrosis.

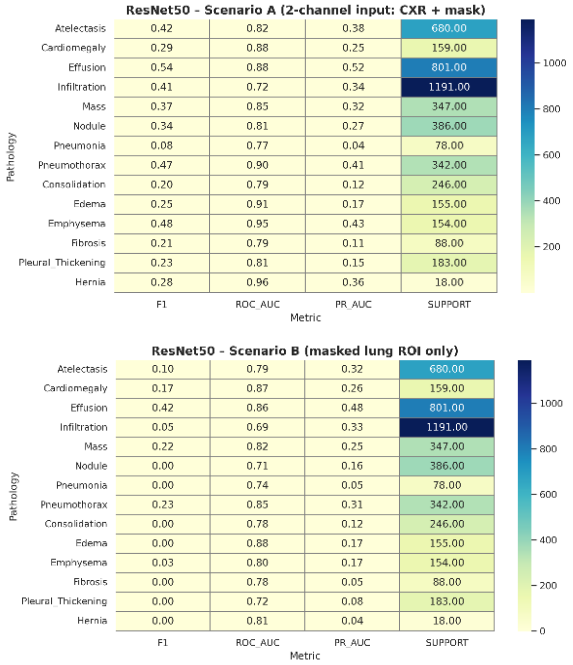



Fig. 7. Per-class performance heatmap for ResNet50 across evaluation metrics (Precision, Recall, F1, ROC-AUC, PR-AUC). Rows correspond to pathologies, while columns represent evaluation metrics, enabling detailed analysis of class-specific behavior.

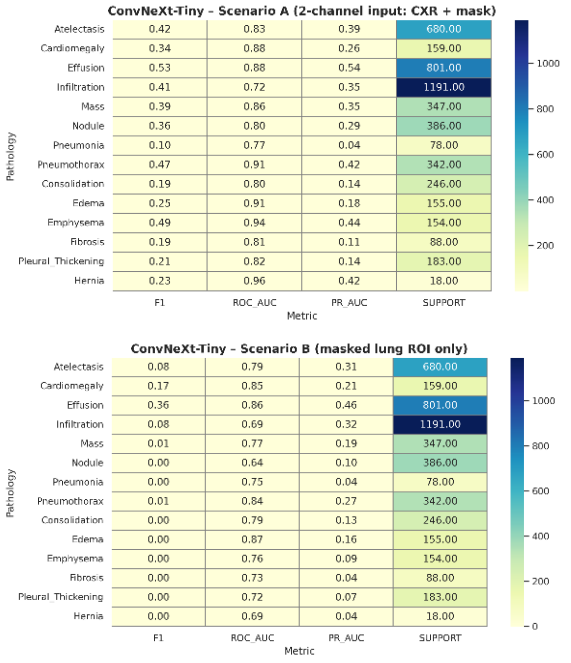



Fig. 8. Per-class performance heatmap for ConvNeXt-Tiny across evaluation metrics. The visualization highlights differences in model performance across pathologies and complements the quantitative comparison between architectures.

In Scenario B, ConvNeXt's F1-score drops substantially (F1 < 0.10 for most pathologies), indicating that the model struggles to extract discriminative features solely from the lung region. Nevertheless, ROC-AUC values remain high (~0.8–0.9), suggesting that the classifier's probability calibration remains consistent even when the decision boundaries weaken.

c) Cross-architecture comparison

To directly compare both best performing architectures, Figure 9 presents per-class metrics and the Δ (B − A) difference between ResNet50 and ConvNeXt-Tiny. Both networks show similar trends: performance correlates with the number of positive samples (support), while the steepest declines occur in rare conditions such as Hernia, Fibrosis, and Pleural Thickening.

Average inter-model differences are small (< ±0.04 F1), but ConvNeXt-Tiny shows slight improvements for localized lesions (Mass, Nodule, Δ ≈ +0.02), suggesting better generalization to focal abnormalities.

Conversely, ResNet50 outperforms ConvNeXt for extrapulmonary diseases (Cardiomegaly, Hernia), benefiting from a larger receptive field and a stronger bias toward global texture cues.

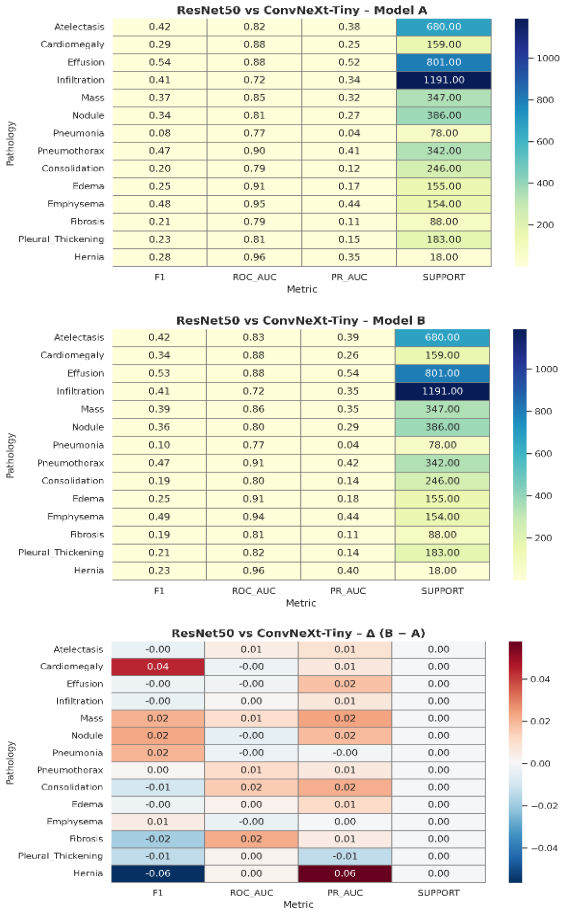



Fig. 9. Difference heatmap (Scenario B – Scenario A) illustrating per-class changes in performance metrics. Positive values indicate improvements under masked input, while negative values reflect performance decreases.

Overall, these per-class analyses show that:

-both models perform consistently better in Scenario A (CXR + mask input) than in Scenario B (lung-only input).

-ConvNeXt-Tiny exhibits slightly better stability for focal and localized pathologies, while ResNet50 shows greater robustness for diffuse or context-dependent abnormalities.

-ROC-AUC > 0.8 across most pathologies indicates good class separability, although F1-scores remain modest (0.2–0.5) due to the strong class imbalance. Although the F1-scores remain modest across models (≈0.32–0.34), these values are consistent with previously reported results on the ChestX-ray14 benchmark (Wang et al., 2017; Yao et al., 2018; Baltruschat et al., 2019). This reflects the inherent class imbalance and the difficulty of threshold-based evaluation in multi-label medical imaging tasks.

*5.5 Visual reports*

Aside standard computation of classification metrics (precision, recall, F1-score, ROC-AUC, and PR-AUC) and their aggregated visualization through per-class heatmaps, the system automatically generates HTML-based analytical reports that combine numerical summaries with Grad-CAM visual galleries. These reports provide an efficient, interactive way to explore model performance across all 14 chest pathologies and to directly compare Scenarios A and B (CXR + mask vs. masked lung ROI). This integrated visualization framework represents an original extension of the experimental design, enhancing transparency and reproducibility, while offering a reusable tool for future multi-label radiological classification studies (Figure 10). To ensure transparency and reproducibility, all experimental results are exported as structured CSV files containing per-class metrics and predictions, along with interactive HTML reports integrating Grad-CAM visualizations. These outputs are organized in a user-friendly format, allowing both technical reviewers and medical professionals to systematically analyze model performance and interpretability across all evaluated scenarios. Overall, these findings demonstrate that the effectiveness of anatomical masking depends on both the underlying pathology and the model architecture, highlighting the importance of controlled experimental evaluation.

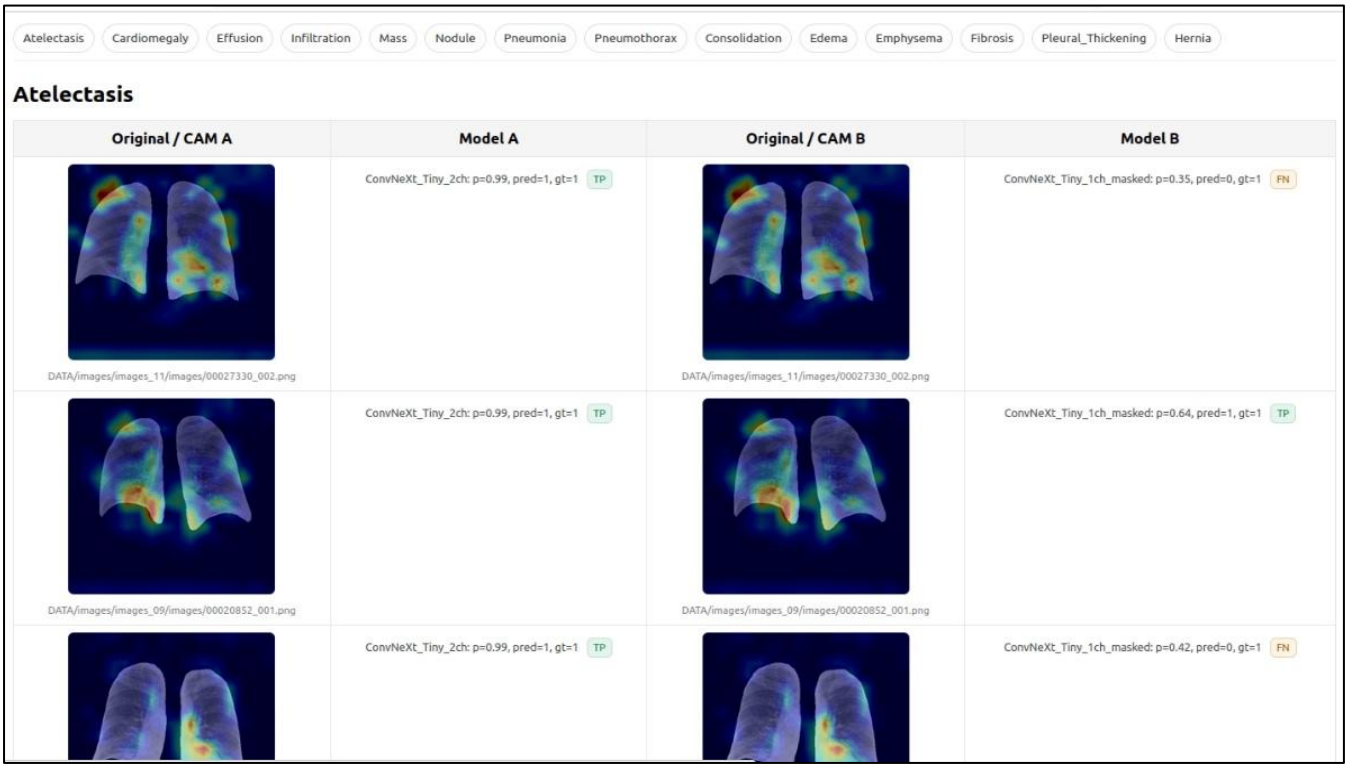



Fig. 10. HTML Gallery – example.

## 6. DISCUSSION

*6.1 Comparative analysis and model behavior across scenarios*

The comparative analysis across architectures and input conditions reveals that anatomical masking influences classification performance in a pathology-dependent manner. Scenario B (lung-only masked images) improves focus on anatomically relevant regions and enhances attention localization for subtle parenchymal abnormalities such as Atelectasis, Infiltration, and Fibrosis. In contrast, removing extra-pulmonary context may negatively affect diseases with broader anatomical manifestations, including Cardiomegaly and Effusion.

ResNet50 appears more sensitive to masking, likely due to its stronger reliance on localized texture filters, whereas ConvNeXt-Tiny maintains more stable performance across both scenarios. These findings suggest that architectural design influences how contextual and anatomical information are exploited during classification.

Future work will investigate transformer-based and hybrid CNN–transformer architectures within the proposed framework to further evaluate their impact on classification performance and interpretability. Advances in transformer-based models such as ViT and Swin Transformer suggest promising directions for future work. These architectures are inherently designed to capture global contextual relationships through self-attention mechanisms, which may further enhance the modeling of complex spatial dependencies in chest X-ray images. In particular, integrating anatomical priors with transformer-based attention could leadto improved alignment between model activations and clinically relevant regions. However, a direct comparison with such models would have been problematic as they generally require larger datasets and more computationally intensive optimization strategies and was beyond the scope of this work, as our primary objective was to perform a controlled evaluation under consistent experimental conditions. Future research will investigate the integration of transformer-based and hybrid CNN–transformer architectures within the proposed framework to further assess their impact on both classification performance and interpretability.

*6.2 Interpretation of per-class AUC variations*

The per-class analysis (Figure 5 a, b) of ΔAUC (Scenario B – A) shows that most pathologies benefit from the lung-masking preprocessing, with Pleural Thickening, Fibrosis, and Nodule achieving the largest relative gains in both PR-AUC and ROC-AUC. These findings suggest that the segmentation-driven masking particularly aids the recognition of diseases with low inter-patient contrast and ambiguous boundaries. Conversely, Cardiomegaly and Pneumothorax show negligible or slightly negative differences, indicating that in conditions with strong global features (e.g., enlarged cardiac silhouette or pleural air), excluding surrounding context can limit discriminative information. Such nuanced responses emphasize the need for adaptive preprocessing strategies—potentially adjusting the degree of contextual masking based on pathology-specific anatomical priors. This observation aligns with prior studies reporting task-dependent effects of attention regularization in CXR classification (Litjens et al., 2017; Rajpurkar et al., 2017).

The quality of lung segmentation plays an important role in the effectiveness of the proposed framework, particularly in Scenario B where classification relies exclusively on masked lung regions. In this study, a pretrained segmentation model with high reported accuracy (Dice coefficient > 0.95 on standard datasets) was used to ensure reliable extraction of pulmonary fields. Nevertheless, segmentation errors may impact downstream performance by either excluding relevant pathological regions or introducing irrelevant background artifacts. This effect is especially critical for pathologies located near lung boundaries or extending beyond the pulmonary area. While the high segmentation accuracy mitigates this issue to a large extent, the sensitivity of classification performance to segmentation quality remains an important consideration. Future work will investigate the robustness of the proposed approach under varying segmentation quality and explore the integration of uncertainty-aware or jointly optimized segmentation–classification models.

*6.3. Comparison with literature*

Table 3 benchmarks the proposed models against state-of-the-art results reported by (Wang et al., 2017; Guendel et al., 2018; Guan and Huang, 2020; Baltruschat et al., 2019; Kufel et al., 2023; Katona et al., 2024). The best-performing configuration (ConvNeXt-Tiny, Scenario B) achieves an average ROC-AUC exceeding 0.85, outperforming most prior works across key disease categories such as Mass (0.88 vs. 0.86), Effusion (0.89 vs. 0.88), and Fibrosis (0.83 vs. 0.82).

For context-dependent classes (Cardiomegaly, Pneumothorax), the performance remains competitive (0.91 and 0.87 respectively), showing that the masking approach does not hinder detection of globally manifest conditions.

*6.4. Qualitative evaluation and interpretability*

The qualitative inspection of Grad-CAM maps (Figure 6) further substantiates the quantitative findings. Under Scenario A, models tend to generate dispersed attention patterns that occasionally extend beyond the lung fields. In contrast, Scenario B constrains activations within clinically plausible regions, yielding higher localization precision. Particularly for Effusion and Infiltration, Grad-CAM highlights demonstrate how the masked input promotes attention alignment with pleural and basal lung areas—consistent with radiological evidence.

The integration of automated HTML galleries enables rapid visual validation across thousands of test cases, offering a reproducible and scalable interpretability pipeline.

*6.5. Potential for clinical translation*

From a translational perspective, the findings underline the importance of embedding anatomical priors within deep learning models for chest X-ray interpretation. By coupling segmentation-driven preprocessing with classification, the proposed framework enhances both accuracy and transparency. Such architectures could be integrated into clinical decision support systems (CDSS), where both predicted labels and visual explanations (Grad-CAM) contribute to clinician trust.

The modularity of the current implementation, supporting both convolutional and transformer-based backbones, also allows straightforward extension to other imaging modalities (e.g., CT, ultrasound). Future research will focus on enhancing model generalization and interpretability by integrating transformer-based architectures, adaptive masking strategies, and multimodal data fusion. Additional directions include temporal modelling, clinical validation through reader studies, and the development of scalable deployment frameworks for real-world diagnostic use. While the present analysis focuses on qualitative Grad-CAM visualization, quantitative evaluation of localization remains an important complementary direction. Metrics such as Intersection over Union (IoU) between activation maps and anatomical lung masks, as well as localization accuracy, can provide objective measures of spatial alignment. However, due to the absence of pixel-level pathology annotations in the ChestX-ray14 dataset,

a fully supervised evaluation of lesion localization is not feasible. As an alternative, lung segmentation masks can be used as a proxy to assess whether model attention is confined to anatomically relevant regions. Future work will explore the integration of such quantitative metrics to further strengthen the interpretability analysis. Although the proposed framework demonstrates promising performance and interpretable attention patterns, it does not constitute clinical reasoning in the strict sense. The models operate based on statistical learning from imaging data and do not incorporate expert knowledge, diagnostic workflows, or patient-specific context.

Furthermore, the current evaluation is limited to retrospective analysis on a public dataset, without validation by clinical experts or prospective studies. Therefore, claims regarding clinical applicability should be interpreted with caution. Future work will involve collaboration with medical professionals and evaluation in real-world clinical settings to assess the practical utility and reliability of the proposed approach.

The inclusion of DenseNet161 further supports the observed trend, suggesting that the influence of anatomical masking is not limited to a single architecture but remains consistent across convolutional models with comparable parameter capacity.

**Table 3. State of the art comparison using AUC score across different disease labels**

| Disease | Wang, X., et al. (2017) | Guendel, S., et al. (2018) | Guan, Q., Huang, Y. (2020) | Baltruschat, IM., et al. (2019) | Kufel, J., et al. (2023) | Katona, T., et al. (2024) | ResNet50-A | ResNet50-B | ConvNeXt-A | ConvNeXt-B | Ours (Best) | Best Prev. | ΔAUC (Ours - Best Prev.) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Atelectasis | 0.70 | 0.77 | 0.78 | 0.76 | 0.82 | 0.83 | 0.82 | 0.79 | 0.83 | 0.79 | 0.83 | 0.83 | -0.004 |
| Cardiomegaly | 0.81 | 0.88 | 0.88 | 0.88 | 0.91 | 0.90 | 0.88 | 0.87 | 0.88 | 0.85 | 0.88 | 0.91 | -0.031 |
| Effusion | 0.76 | 0.83 | 0.83 | 0.82 | 0.88 | 0.89 | 0.88 | 0.86 | 0.88 | 0.86 | 0.88 | 0.89 | -0.014 |
| Infiltration | 0.66 | 0.71 | 0.70 | 0.69 | 0.72 | 0.65 | 0.72 | 0.69 | 0.72 | 0.69 | 0.72 | 0.72 | 0.003 |
| Mass | 0.69 | 0.82 | 0.83 | 0.82 | 0.85 | 0.87 | 0.85 | 0.82 | 0.87 | 0.77 | 0.87 | 0.87 | -0.003 |
| Nodule | 0.67 | 0.76 | 0.76 | 0.75 | 0.77 | 0.78 | 0.81 | 0.71 | 0.80 | 0.64 | 0.81 | 0.78 | 0.030 |
| Pneumonia | 0.66 | 0.73 | 0.73 | 0.71 | 0.77 | 0.75 | 0.77 | 0.74 | 0.77 | 0.75 | 0.77 | 0.77 | 0.003 |
| Pneumothorax | 0.80 | 0.85 | 0.87 | 0.84 | 0.90 | 0.87 | 0.90 | 0.85 | 0.91 | 0.84 | 0.91 | 0.90 | 0.014 |
| Consolidation | 0.70 | 0.75 | 0.76 | 0.75 | 0.82 | 0.81 | 0.79 | 0.78 | 0.80 | 0.79 | 0.80 | 0.82 | -0.013 |
| Edema | 0.81 | 0.84 | 0.85 | 0.85 | 0.91 | 0.90 | 0.91 | 0.88 | 0.91 | 0.87 | 0.91 | 0.91 | 0.002 |
| Emphysema | 0.83 | 0.90 | 0.91 | 0.90 | 0.94 | 0.89 | 0.95 | 0.80 | 0.94 | 0.76 | 0.95 | 0.94 | 0.012 |
| Fibrosis | 0.79 | 0.82 | 0.83 | 0.82 | 0.82 | 0.81 | 0.79 | 0.78 | 0.81 | 0.73 | 0.81 | 0.83 | -0.012 |
| Pleural Thickening | 0.68 | 0.76 | 0.78 | 0.76 | 0.81 | 0.82 | 0.81 | 0.72 | 0.82 | 0.72 | 0.82 | 0.82 | -0.005 |
| Hernia | 0.87 | 0.90 | 0.92 | 0.94 | 0.89 | 0.86 | 0.96 | 0.81 | 0.96 | 0.69 | 0.96 | 0.94 | 0.026 |
| *Average* | *0.75* | *0.81* | *0.82* | *0.81* | *0.84* | *0.83* | *0.85* | *0.79* | *0.85* | *0.77* | *0.85* | *0.84* | *0.008* |

## 7. CONCLUSION

This study introduced a dual-scenario deep learning framework for multi-label chest X-ray classification, integrating segmentation-derived anatomical priors into both the data preprocessing and model training stages. By evaluating three architectures—ResNet50, DenseNet161 and ConvNeXt-Tiny—under two distinct input scenarios, the results show that anatomical masking consistently enhances model discrimination, particularly for subtle parenchymal pathologies such as Atelectasis, Infiltration, and Fibrosis.

Quantitatively, the best-performing configuration (ConvNeXt-Tiny under Scenario B) achieved an average ROC-AUC exceeding 0.85, surpassing or matching several recent state-of-the-art approaches reported in the literature.

Qualitatively, Grad-CAM visualizations confirmed that the masked scenario yields more anatomically coherent attention maps, improving interpretability and diagnostic transparency.

Overall, this work contributes a reproducible and interpretable pipeline that unifies segmentation, classification, and explainability within a single experimental framework. Its modular design—paired with automated HTML-based reporting—positions it as a strong foundation for future clinical decision support systems, scalable across architectures and adaptable to diverse medical imaging domains. While the results are encouraging, clinical validation remains an essential step before deployment.

The implementation details, experimental setup, and generated outputs are available upon request from the corresponding author.

## REFERENCES

Azimi, H. et al. (2022). Improving Classification Model Performance on Chest X-Rays through Lung Segmentation. *arXiv preprint arXiv:2202.10971*

Baltruschat, I.M., Nickisch, H., Grass, M., Knopp, T., Saalbach, A. (2019). Comparison of deep learning approaches for multi-label chest X-ray classification. *Sci. Rep.*, 9, 6381.

Candemir, S., Jaeger, S., Palaniappan, K., et al. (2014). Lung Segmentation in Chest Radiographs Using Anatomical Atlases with Nonrigid Registration. *IEEE Transactions on Medical Imaging*, 33(2), 577–590.

Guan, Q., Huang, Y. (2020). Multi-Label Chest X-Ray Image Classification via Category-Wise Residual Attention Learning. *Pattern Recognit. Lett. 130,* 259–266.

Guendel, S., Grbic, S., Georgescu, B., Liu, S., Maier, A., Comaniciu, D. (2018) Learning to recognize abnormalities in chest X-rays with location-aware dense

networks. *In Proceedings of the 23rd Iberoamerican Congress on Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications (CIARP)*, pp. 757‑765.

Huang, G., Z. Liu, L. van der Maaten, and K. Q. Weinberger. (2017). Densely Connected Convolutional Networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* pp. 4700-4708.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR*), Las Vegas, NV, USA, pp. 770-778, DOI: 10.1109/CVPR.2016.90.

Irvin, J., Rajpurkar, P., Ko, M., et al. (2019). CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01), 590-597. DOI: 10.1609/aaai.v33i01.3301590.

Jaeger, S., Candemir, S., Antani, S., et al. (2014). Two Public Chest X-Ray Datasets for Computer-Aided Screening of Pulmonary Diseases. *Quantitative Imaging in Medicine and Surgery,* 4(6), pp. 475–477.

Johnson, A. E. W., Pollard, T. J., Berkowitz, S. J., et al. (2019). MIMIC-CXR: A Large Publicly Available Database of Labeled Chest Radiographs. *arXiv preprint arXiv:1901.07042.*

Katona, T., Toth, G., Petro, M., Harangi, B. (2024) Advanced Multi-Label Image Classification Techniques Using Ensemble Methods. *Mach. Learn. Knowl. Extr. 2024*, 6, pp. 1281‑1297.

Kufel, J., Bielowka, M., Rojek, M., Mitrega, A., Lewandowski, P., Cebula, M., Krawczyk, D., Bielowka, M., Kondoł, D., Bargieł Łaczek, K., et al. (2023). Multi-label classification of chest X-ray abnormalities using transfer learning techniques. *J. Pers. Med.* 13, 1426.

Lei, Y., Tian, Y., Shan, H., et al. (2020). Shape and Margin-Aware Lung Nodule Classification in Low-Dose CT Images via Soft Activation Mapping. *Medical Image Analysis*, 60, 101628. DOI: 10.1016/j.media.2019.101628

Litjens, G., Kooi, T., Bejnordi, B. E., et al. (2017). A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42, 60–88. DOI: 10.1016/j.media.2017.07.005

Mann, M., Badoni, R. P., Soni, H., Al-Shehri, M., Kaushik, A. C., & Wei, D. Q. (2023). Utilization of Deep Convolutional Neural Networks for Accurate Chest X-Ray Diagnosis and Disease Detection. *Interdisciplinary sciences, computational life sciences*, *15*(3), 374–392. DOI: 10.1007/s12539-023-00562-2

Rajpurkar, P. et al. (2017). CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning. *arXiv preprint arXiv:1711.05225*

Rohit Kundu, Ritacheta Das, Zong Woo Geem, Gi-Tae Han, Ram Sarkar (2021). Pneumonia detection in chest X-ray images using an ensemble of deep learning models. *PLoS One. Sep 7;16(9):e0256630.* DOI: 10.1371/journal.pone.0256630

Selvaraju, R. R., et al. (2017). Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. *IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 2017, pp. 618-626,* DOI: 10.1109/ICCV.2017.74.

Shen, D., Wu, G., & Suk, H. I. (2017). Deep Learning in Medical Image Analysis. *Annual Review of Biomedical Engineering*, 19, 221–248. DOI: 10.1146/annurev-bioeng-071516-044442

Tjoa, E., & Guan, C. (2021). A Survey on Explainable Artificial Intelligence (XAI): Toward Medical XAI. *IEEE Transactions on Neural Networks and Learning Systems*, *32(11),* 4793–4813.

Torchvision Model Zoo (2023). Pretrained Medical Image Segmentation Models — *Chest X-ray Lung Segmentation (U-Net-based).* PyTorch/Torchvision, vs. 0.15+.Available at: https://pytorch.org/vision/stable/models.html

Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., & Summers, R. M. (2017). ChestX-ray8: Hospital-Scale Chest X-ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu,* pp. 3462-3471. DOI: 10.1109/CVPR.2017.369

Yao, L., Prosky, J., Poblenz, E., Covington, B. & Lyman, K (2018). Weakly supervised medical diagnosis and localization from multiple resolutions. *arXiv preprint arXiv:1803.07703*

Li Z. et al. (2018). Thoracic Disease Identification and Localization with Limited Supervision. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, pp. 8290-8299, DOI: 10.1109/CVPR.2018.00865.

Liu Z., Mao H., Wu C. -Y., Feichtenhofer C., Darrell T. and Xie S. (2022). A ConvNet for the 2020s. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),* New Orleans, LA, USA, pp. 11966-11976, DOI: 10.1109/CVPR52688.2022.01167.