Exploratory Analysis of Deep Learning Model for Non Invasive Classification of Pulmonary Hypertension Based On Chest X-Ray Images

Estela Ribeiro[1], Gabriella G. Carvalho[1,2] Diego A. C. Cardenas[1], Rogério de Souza[1,2], Marco A. Gutierrez[1,2]

[1]Heart Institute (InCor) – Clinics Hospital
University of Sao Paulo Medical School (HCFMUSP)
Sao Paulo – SP – Brazil

[2]University of Sao Paulo Medical School (FMUSP)
Sao Paulo – SP – Brazil

estela.ribeiro@hc.fm.usp.br, gabriella.gcarvalho@hc.fm.usp.br

diego.cardona@hc.fm.usp.br, rogerio.souza@hc.fm.usp.br

marco.gutierrez@hc.fm.usp.br

Abstract. *Pulmonary Hypertension (PH) is a progressive condition in which early detection is essential for improving outcomes. Chest X-rays (CXR) may contain patterns associated with PH. This study evaluated automated PH detection from CXRs. A retrospective private dataset of 1,354 exams (1,138 PH) was analyzed. Multiple CNNs and Transformer-based architectures were trained and tested. CNNs achieved AUROC values from 0.71 to 0.76, outperforming Transformers (AUROC $\leq$ 0.60). Performance was mainly driven by PH cases with mPAP $>$ 25 mmHg, while sensitivity decreased in normal and mildly elevated ranges. These results support the feasibility of non-invasive PH screening, although larger and more balanced datasets are needed for clinical validation.*

1. Introduction

Pulmonary Hypertension (PH) is a progressive pulmonary vascular disorder characterized by right ventricular dysfunction, which may evolve to right heart failure and, in advanced cases, lead to death [Ferreira et al. 2014, Luna-López et al. 2022]. Despite advances in diagnostic strategies, early detection of PH remains a challenge worldwide. Timely diagnosis is essential for initiating appropriate treatment and improving both prognosis and quality of life. In contrast, delayed diagnosis is associated with more advanced and often irreversible vascular and cardiac remodeling, significantly increasing morbidity and mortality [McLaughlin et al. 2002, Khou et al. 2020].

The gold standard for the diagnosis of PH is right heart catheterization (RHC) [Luna-López et al. 2022, Becerra-Muñoz et al. 2024]. However, RHC is invasive, costly, and technically demanding procedure, making it unsuitable as an initial screening tool. Transthoracic Echocardiography (TTE) is widely recommended for the initial evaluation of suspected PH cases, but its accuracy may be limited, particularly in early-stage disease or in patients with suboptimal acoustic windows [Luna-López et al. 2022]. With recent advances in Artificial Intelligence (AI), especially in the field of Deep

Learning (DL), promising tools have emerged to address the challenges associated with PH diagnosis. These tools leverage widely available clinical data, such as electrocardiography (ECG) and Chest X-Ray (CXR), to support early detection of PH [Fadilah et al. 2024, Imai et al. 2024].

The CXR remains one of the most accessible and low-cost imaging examinations. According to the 2023 Clinical Protocol and Therapeutic Guidelines (PCDT) for PH [Magalhães Junior and Safatle 2023], radiographic signs suggestive of PH include pulmonary trunk enlargement, right ventricular border–spinal column distance greater than 44 mm, pulmonary hilum-to-thoracic width ratio above 0.44, right descending artery diameter greater than 16 mm and left greater than 18 mm, as well as retrosternal space filling by the cardiac silhouette [Magalhães Junior and Safatle 2023]. Right atrium, right ventricular enlargement and pericardial effusion may be seen in more advanced cases and other signs of the underlying cause of PH, such as left heart disease or lung disease, may be found [Humbert et al. 2022]. However, in early stage disease, structural and compensatory changes are often subtle and may escape visual detection, even by experienced clinicians. The diagnostic challenge may contribute to delayed recognition and treatment initiation [Peña et al. 2012]. Despite its potential, only a limited number of studies have investigated DL-based approaches for PH detection using CXR images.

[Kusunose et al. 2020] developed a DL algorithm to predict elevated mean pulmonary artery pressure (mPAP > 20 mmHg) from CXR images in 900 patients with suspected PH, achieving an area under the under the receiver operating characteristic curve (AUROC) of 0.71 and a negative predictive value (NPV) of 95.0%. Similarly, [Zou et al. 2020] applied an InceptionV3 DL model to classify PH using CXR paired with TTE measurements. In their dataset, 405 cases were defined as PH (pulmonary arterial systolic pressure ≥ 40 mmHg), while the remaining cases were normal. The model achieved an AUROC of 0.97.

[Imai et al. 2024] proposed a ResNet-based DL model to predict pulmonary arterial hypertension (PAH), a subtype of PH, from CXR images, including 145 patients with PAH and 260 controls, reporting an AUROC of 0.99. More recently, [Li et al. 2024] trained a ResNet-based model on 831 patients with ventricular septal defect (VSD), including 161 with PAH-VSD and 670 controls. Their model achieved an AUROC of 0.82 and outperformed human observers (0.82 vs. 0.65). Class activation mapping (CAM) analysis demonstrated that the model focused on the pulmonary artery region, supporting the interpretability its predictions. Table 1 summarizes the main studies employing CXR for PH detection.

Overall, these studies demonstrate the potential of CXR as a low-cost and widely available imaging modality for PH detection when combined with DL approaches. However, important limitations remain. Most prior investigations are based on relatively small, single-center datasets, and adopted heterogeneous definitions of PH, frequently relying on echocardiographic rather than invasive hemodynamic criteria. These factors limit reproducibility, external generalization, and clinical applicability of the proposed models.

In this context, our study aims to develop and evaluate DL models for the classification of PH using CXR images, with hemodynamic confirmation as the reference standard. We performed a systematic comparison with multiple convolutional neural network

Table 1. Summary of studies using chest X-ray (CXR) and deep learning for pulmonary hypertension (PH) detection.

Study	Performance (AUROC)	Dataset	Class Definition	Methodology
Kusunose et al. 2020	0.71 (NPV=95%)	439 high mPAP / 461 normal mPAP	High mPAP $>$ 20 mmHg vs. $\leq$ 20 mmHg	CNN with residual blocks; pre-trained on a pneumonia dataset and fine-tuned on study data
Zou et al. 2020	0.97 (internal), 0.96 (external)	405 PH / 357 controls	PH defined as PASP $\geq$ 40 mmHg (TTE)	Transfer learning using ResNet50, Xception, and InceptionV3 pre-trained on ImageNet
Imai et al. 2024	0.99	259 PAH / 260 controls	PAWP $>$ 20 mmHg vs. controls without PAH suspicion	ResNet50 pre-trained on ImageNet
Li et al. 2024	0.82	161 PAH-VSD / 670 controls	PAH diagnosed by Doppler echocardiography in VSD patients	ResNet50 with class activation mapping highlighting pulmonary artery regions

(CNN) architectures to identify the most effective approach under real-world constraints. The goal is to provide a robust non-invasive screening tool, based on a universally accessible imaging modality, supporting early identification of PH and assisting clinical decision-making pathways. The main contributions of our study are as follows:

1. A curated dataset of CXRs exams paired with RHC measurements as the gold standard.
2. A standardized preprocessing and lung segmentation pipeline to enhance signal extraction.
3. A comprehensive comparative evaluation of multiple DL architectures, providing insights into model performance in moderately sized and imbalanced clinical datasets.

2. Methods

This section describes the dataset, preprocessing pipeline, and the DL architectures employed for binary classification of CXR images into PH or Non-PH. The overall workflow is illustrated in Figure 1. First, we present the dataset and data curation procedures. Subsequently, we detail the preprocessing pipeline, including image resizing, lung segmentation, and contrast enhancement. Finally, we describe the model training strategy, validation protocol, and evaluation metrics used to assess performance.

All experiments were conducted on a FOXCONN High-Performance Computing (HPC) M100-NHI system equipped with eight NVIDIA Tesla V100 GPUs (32 GB memory each). The methodology was implemented in Python (version 3.11.13) using PyTorch framework (version 2.5.1).

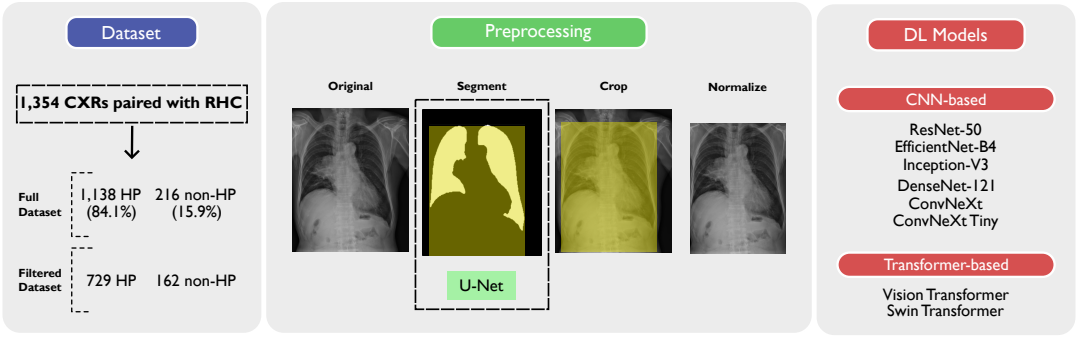

Figure 1. Overview of the proposed methodology.

2.1. Dataset

We used a private retrospective dataset, hereafter referred to as InCor-PH-CXR, collected at the Heart Institute (InCor), a tertiary cardiopneumology hospital in Sao Paulo, Brazil, between 2015 and 2025. From the institutional Picture Archiving and Communication System (PACS), we identified 20,725 RHC exams performed in approximately 13,325 unique individuals. To avoid intra-patient correlation and repeated-measure bias, only the first RHC exam per individual was retained for analysis. Subsequently, we searched for CXR exams performed within $\pm$ 6 month window relative to the RHC. This temporal pairing was defined to maximize the likelihood of hemodynamic and radiographic correspondence while preserving sample size.

A major limitation encountered during cohort construction was the lack of structured reporting for many RHC examinations. In a substantial proportion of cases, hemodynamic parameters were documented exclusively in free-text reports, limiting automated extraction of key hemodynamic parameters. As a result, although a larger number of CXR exams were initially paired, only 3,224 cases had structured and extractable hemodynamic data available for analysis.

We retrieved the corresponding Digital Imaging and Communications in Medicine (DICOM) files for all eligible CXRs examinations. A filtering process was applied to ensure dataset consistency: only adult patients ($\geq$ 18 years) were included; only posteroanterior (PA) or anteroposterior (AP) projections were retained; and examinations with severe artifacts or insufficient image quality were excluded. Moreover, 35.9% of the paired CXRs were acquired at bedside (portable AP projections), meaning that their positioning and quality may not follow standard acquisition protocols. After this refinement, the dataset consisted of 1,354 paired CXR–RHC exams.

For labeling, mean pulmonary arterial pressure (mPAP) values obtained from the RHC was used as the reference standard. Cases with mPAP $\leq$ 20 mmHg were classified as Non-PH, whereas while cases with mPAP $>$ 20 mmHg were labeled as PH. Figure 2 presents the distribution of mPAP values, and Table 2 summarizes the final dataset composition.

Additionally, we constructed a filtered subset by excluding bedside CXRs with suboptimal positioning, as well as examinations containing artifacts or medical devices that could interfere with image interpretation. This refinement resulted in a database of 891 CXRs examinations. Figure 3 illustrates representative examples from this filtered

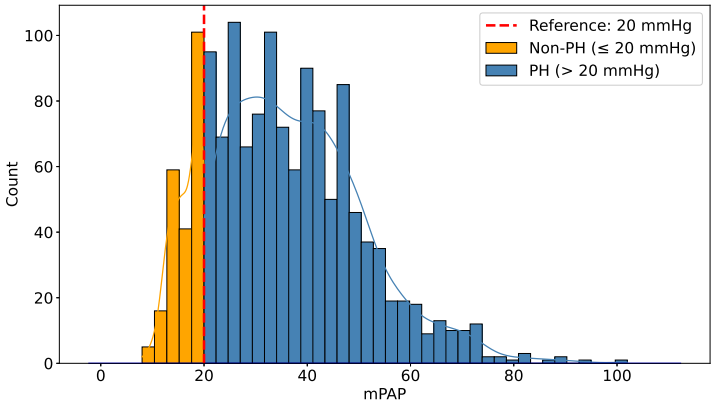

Figure 2. Histogram of the mPAP distribution for the full dataset.

Table 2. Demographic and clinical characteristics of the full dataset.

Characteristic	**Global** 1,354	**PH** 1,138 (84.1%)	**Non-PH** 216 (15.9%)
Age (years)	51.76 ± 13.78	52.65 ± 13.33	47.07 ± 15.14
Male	648	547	101
Female	706	591	115
Technical characteristics			
Bedside CXR	487	435	52
Pacemaker	176	158	18
Electrodes (artifacts)	335	307	28

cohort. The filtered subset comprised 729 PH cases, and 162 Non-PH cases. This subset was used to compare model performance against that obtained using the full dataset, enabling the assessment of whether the models were inadvertently leveraging acquisition-related artifacts or device-related features rather than pathophysiological imaging patterns associated with PH. This comparison provides insight into model robustness and potential shortcut learning behavior.

This research protocol was submitted through the national research platform Plataforma Brasil and approved by the Research Ethics Committee of the Clinics Hospital, University of São Paulo Medical School (CAAE: 94820726.7.0000.0068). As this study involved a secondary analysis of retrospectively collected and fully anonymized data stored in the InCor-PH-DB, the requirement for informed consent was waived by the Research Ethics Committee. All procedures were conducted in accordance with the approved ethical guidelines and institutional regulations.

2.2. Preprocessing

The original CXR images had a resolution of 3,480 × 4,248 pixels. All images were first resized to 768 × 768 pixels to standardize input dimensions and reduce computational burden. Images were converted to single-channel grayscale intensity format prior to model input. A U-Net–based segmentation model [Ribeiro et al. 2024] previously trained on publicly available CXR datasets was employed to generate binary lung masks. These

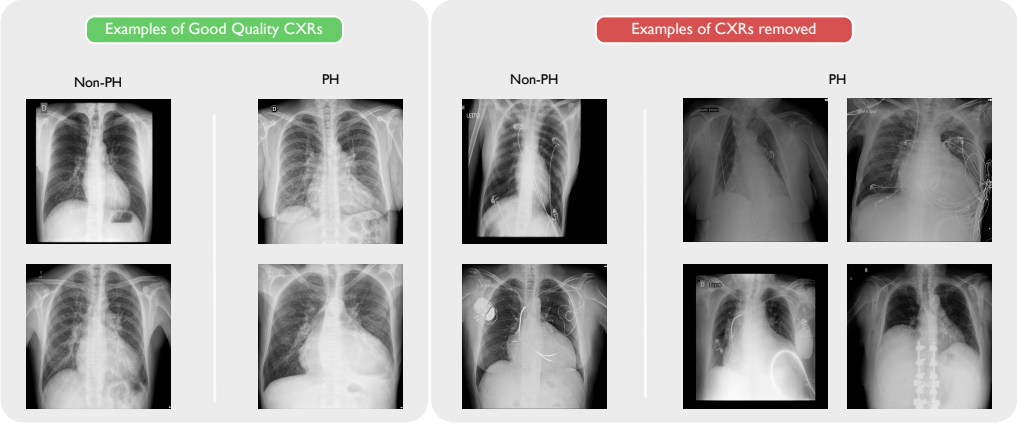

Figure 3. Examples of CXRs with good quality an CXRs with artifacts and devices that were removed after filtering.

masks were used to crop the resized images, removing non-relevant upper and lateral regions while preserving the lung fields and lower cardiomediastinal area. This step aimed to reduce background noise and focus feature extraction on anatomically relevant regions. The cropped images were subsequently normalized, and contrast stretching was applied using the 2nd and 98th percentiles of the pixel intensity distribution to enhance dynamic range and improve inter-exam visual consistency.

To increase data variability and mitigate overfitting, offline data augmentation was performed using the *Albumentations* Python library exclusively on the training set, with no augmentation applied to the validation or test sets. For each original training image, five augmented images were generated. The applied transformations included small rotations ($\pm$ 5), affine translations (up to 10%), random scaling ($\pm$ 10%), and slight brightness and contrast adjustments ($\pm$ 5%). Horizontal and vertical flips were deliberately disabled to prevent anatomically implausible representations and to preserve the physiological orientation of thoracic structures.

2.3. Deep Learning Models

We selected a set of convolutional and transformer-based architectures implemented using the *PyTorch* and *timm* libraries. All models were initialized with ImageNet pretrained weights to leverage transfer learning. The CNN-based architectures included: (i) ResNet-50 [He et al. 2015], (ii) EfficientNet-B4 [Tan and Le 2019], (iii) Inception-V3 [Szegedy et al. 2014], (iv) DenseNet-121 [Huang et al. 2017], (v) ConvNeXt [Liu et al. 2022], and (vi) ConvNeXt Tiny. The transformer-based architectures included: Swin Transformer [Liu et al. 2021], and Vision Transformer (ViT) [Dosovitskiy et al. 2021]. All models were trained for binary classification (PH vs Non-PH), with a batch size of 16 and learning rate of $1e^{-3}$. We retained the fully connected layers of the original models, only modifying the last layer with a single output using a sigmoid activation.

Each model underwent training for up to 50 epochs. The loss function was binary cross-entropy with logits, incorporating class balancing through a positive weight factor. Optimization was carried out with the AdamW optimizer, coupled with a learning rate scheduler that reduced the learning rate by a factor of 0.5 after 3 epochs without improve-

ment. Early stopping with a patience of 9 epochs was employed to prevent overfitting.

All models were evaluated using a group-wise 5-fold cross-validation strategy at the patient level. In each fold, one partition was held out as the test set, while the remaining data were used for training and validation. Within each fold, a group-aware split was applied to further separate training and validation sets, ensuring that samples from the same patient were never shared across splits. Model performance was assessed using Accuracy (Acc), Sensitivity (Se), Specificity (Spe), F1-score (F1, and the AUROC.

3. Results

Table 3 summarizes the classification performance of all evaluated models on the full dataset, whereas Table 4 presents the corresponding results on the filtered subset, in which low-quality artifacts containing CXR examinations were excluded. Results are reported as mean ± standard deviation across the five cross-validation folds. Models in the global analysis were ranked according to AUROC, as it provides a threshold-independent measure of overall discriminative performance.

Table 3. Performance of DL models for PH detection using CXR images, evaluated with 5-fold cross-validation on the full dataset (n = 1,354).

Model	Acc	Sensitivity	Specificity	F1-score	AUROC
Inception	0.6943 ± 0.0561	0.6986 ± 0.0736	0.6712 ± 0.0923	0.7918 ± 0.0475	0.7620 ± 0.0432
EfficientNet	0.7430 ± 0.0423	0.7752 ± 0.0730	0.5743 ± 0.1268	0.8338 ± 0.0357	0.7458 ± 0.0344
ResNet-50	0.6959 ± 0.1050	0.7199 ± 0.1628	0.5693 ± 0.2162	0.7914 ± 0.0886	0.7316 ± 0.0127
DenseNet	0.6781 ± 0.1408	0.6856 ± 0.1965	0.6387 ± 0.1825	0.7665 ± 0.1490	0.7197 ± 0.0652
ConvNeXt	0.1595 ± 0.0013	0.0000 ± 0.0000	1.0000 ± 0.0000	0.0000 ± 0.0000	0.6009 ± 0.0920
ViT	0.4522 ± 0.1202	0.3992 ± 0.1636	0.7315 ± 0.1192	0.5354 ± 0.1489	0.5972 ± 0.0517
Swin	0.2958 ± 0.3046	0.2000 ± 0.4472	0.8000 ± 0.4472	0.1827 ± 0.4085	0.4085 ± 0.5240
ConvNeXt Tiny	0.1595 ± 0.0013	0.0000 ± 0.0000	1.0000 ± 0.0000	0.0000 ± 0.0000	0.4991 ± 0.0163

Table 4. Performance of DL models for PH detection using CXR images, evaluated with 5-fold cross-validation on the filtered subset (n = 891).

Model	Acc	Sensitivity	Specificity	F1-score	AUROC
Inception	0.7262 ± 0.0559	0.7503 ± 0.0928	0.6159 ± 0.1379	0.8150 ± 0.0521	0.7423 ± 0.0542
EfficientNet	0.7340 ± 0.0408	0.7792 ± 0.0537	0.5307 ± 0.0738	0.8267 ± 0.0306	0.7349 ± 0.0460
ResNet-50	0.6844 ± 0.0811	0.7077 ± 0.1442	0.5820 ± 0.2510	0.7797 ± 0.0767	0.7219 ± 0.0490
DenseNet	0.6628 ± 0.1375	0.6909 ± 0.2027	0.5392 ± 0.1803	0.7555 ± 0.1412	0.6600 ± 0.0732
ConvNeXt	0.3509 ± 0.2472	0.2452 ± 0.3612	0.8288 ± 0.2661	0.2802 ± 0.3939	0.5616 ± 0.0818
ViT	0.4131 ± 0.1519	0.3458 ± 0.2509	0.7176 ± 0.3165	0.4441 ± 0.2741	0.5393 ± 0.0642
Swin	0.1818 ± 0.0024	0.0000 ± 0.0000	1.0000 ± 0.0000	0.0000 ± 0.0000	0.5230 ± 0.0654
ConvNeXt Tiny	0.1841 ± 0.0050	0.0028 ± 0.0062	1.0000 ± 0.0000	0.0054 ± 0.0122	0.5600 ± 0.0503

For the full dataset (Table 3), Inception achieved the highest discriminative performance, with an AUROC of 0.7620 ± 0.0432, followed closely by EfficientNet (0.7458 ± 0.0344). EfficientNet obtained the highest accuracy (0.7430 ± 0.0423) and F1 (0.8338 ± 0.0357), reflecting a favorable balance between Se (0.7752 ± 0.0730) and Spe (0.5743

$\pm$ 0.1268). Inception also demonstrated competitive performance, with balanced Se (0.6986 ± 0.0736) and Spe (0.6712 ± 0.0923). ResNet-50 and DenseNet showed slightly lower AUROC values (0.7316 ± 0.0127 and 0.7197 ± 0.0652, respectively). In contrast, transformer-based models (ViT, Swin) and lightweight ConvNeXt variants showed less stable behavior, with substantial variability across folds and performance approaching random classification in some evaluation metrics.

After applying dataset filtering (Table 4), performance differences among CNN-based architectures became more consistent. Inception achieved the highest AUROC (0.7423 ± 0.0542), while EfficientNet maintained the highest F1 (0.8267 ± 0.0306) and comparable Acc (0.7340 ± 0.0408). ResNet-50 showed similar Se (0.7077 ± 0.1442) compared to the full dataset, although specificity remained variable (0.5820 ± 0.2510). DenseNet presented a slight reduction in AUROC (0.6600 ± 0.0732) relative to the full dataset. Transformer-based models continued to exhibit limited robustness, characterized by extreme sensitivity–specificity trade-offs and high inter-folds variance, reinforcing the hypothesis that such architectures may require larger and more diverse datasets to generalize effectively in medical imaging tasks.

To further investigate model behavior across disease severity, the full dataset was stratified into three clinically relevant mPAP ranges: (i) mPAP $\leq$ 20 (Non-PH), (ii) 20 $<$ mPAP $\leq$ 25 (borderline range); and (iii) mPAP $>$ 25 (established PH). Table 5 reports the Se and Spe of the four models with the highest AUROC values, stratified by mPAP category. In this analysis, Se was used to assess detection capability across increasing hemodynamic burden. Metrics such as overall Acc and AUROC were not considered appropriate within individual range due to class imbalance and reduced sample size. Moreover, since each range contains only a single true class, sensitivity and specificity mathematically reduce to accuracy for the corresponding positive (PH) or negative (Non-PH), respectively.

The stratified analysis by mPAP ranges (Table 5) demonstrated that all four top-performing CNN models achieved higher sensitivity in patients with established PH (mPAP $>$ 25 mmHg), whereas consistently lower performance and greater variability were observed in the mildly elevated (20–25 mmHg) and in normal cases (mPAP $\leq$ 20 mmHg).

Table 5. Stratified Se and Spe (mean $\pm$ std) across mPAP ranges for the model trained on the full dataset.

Model	mPAP $\leq$ 20 Specificity	mPAP 20–25 Sensitivity	mPAP $>$ 25 Sensitivity
Inception	0.6712 ± 0.0923	0.5112 ± 0.1010	0.7298 ± 0.0775
EfficientNet	0.5743 ± 0.1268	0.6635 ± 0.1512	0.7952 ± 0.0683
ResNet-50	0.5693 ± 0.2132	0.6263 ± 0.2391	0.7374 ± 0.1507
DenseNet	0.6387 ± 0.1825	0.6267 ± 0.1919	0.6950 ± 0.1998

4. Discussion

Automated analysis of CXR using AI-based algorithms has the potential to serve as an accessible and scalable screening tool, facilitating earlier identification of patients who

may require further cardiopulmonary evaluation. This approach has the potential to reduce reliance on invasive diagnostic procedures, such as RHC, and help optimize referral pathways. Earlier detection and timely intervention may slow disease progression, improve quality of life and enhance survival, positioning CXR based AI tools as promising adjuncts in PH screening strategies.

Regarding model performance (Tables 3 and 4), CNN-based architectures (ResNet-50, Inception, EfficientNet, and DenseNet) consistently outperformed Transformer-based models (Swin and ViT) and ConvNeXt variants. On the full dataset, the best-performing CNN models achieved AUROC values between 0.72 and 0.76, with Inception reaching the highest discriminative performance, whereas Transformer models remained below 0.60. The ConvNeXt and Swin models failed to converge properly, often predicting only the majority class (specificity = 1.0, sensitivity = 0.0). These findings highlight the limited suitability of Transformer-based approaches for this data regime and imaging configuration.

Although overall AUROC values were satisfactory, stratified analysis (Table 5) revealed that for most architectures, classification errors were predominantly driven by false-positive predictions in patients with normal and borderline mPAP values, indicating reduced specificity in these subgroups. These results suggest that models preferentially learned features associated with more advanced pulmonary vascular remodeling, whereas subtle or early-stage disease patterns remain challenging.

The moderate size of our dataset (1,354 CXR exams) may partly explain the relative underperformance of Transformer-based architecture. Transformers typically require substantially larger datasets to effectively model global spatial dependencies and long-range relationships. In contrast, CNNs rely on localized feature extraction and inductive biases that are better suited to smaller medical imaging datasets, enabling more stable learning under limited data conditions.

Additionally, the pronounced class imbalance in our dataset, approximately 84% PH and 16% non-PH cases, further constrains generalization. Without extensive data augmentation, domain-specific pretraining, or tailored loss functions, large-capacity models are prone to overfitting and biased predictions toward the majority class.

Filtering the dataset (Table 4) did not result in systematic performance improvements. While CNN-based models maintained relatively stable AUROC values (approximately 0.72–0.74 for the top-performing architectures), improvements were modest and not uniformly observed across metrics. These findings highlight the trade-off between data quality and dataset size and emphasize the importance of preserving sufficient data diversity to ensure robust model training. Consequently, future work should prioritize dataset expansion, rigorous image quality control, and domain-adaptive pretraining strategies to better exploit modern architectures while maintaining generalization across diverse clinical settings.

When compared with previous studies (Table 1), our results are broadly consistent with the literature. [Kusunose et al. 2020] obtained an AUROC of 0.71, closely aligning with our best-performing models (AUROC $\approx$ 0.72–0.76). Studies reporting substantially higher AUROC values often relied on smaller or more curated datasets, potentially amplifying disease–control differences. In contrast, our dataset includes heterogeneous

real-world CXRs, such as bedside acquisitions and lower-quality images, better reflecting routine clinical practice.

Our study has some limitations. First, the dataset size, though larger than many prior works, remains relatively small for DL applications, particularly for Transformer-based models. Second, the data were collected from a single center, which limits generalizability to other populations and acquisition settings. Third, label definition was based on RHC, but no stratification by PH subtype or severity was performed. Fourth, due to bedside acquisitions, artifacts, and heterogeneous positioning, image variability may have introduced confounding factors. Finally, no external validation was performed.

Overall, our study represents an exploratory investigation into the feasibility of using CXR for PH detection through DL, evaluating whether CXRs contain sufficient information to distinguish PH from non-PH cases when paired with invasive reference measurements such as RHC. Our results suggest that while CXR may encode meaningful pulmonary and cardiac morphological cues, its diagnostic performance remains limited when used in isolation. However, this modality could serve as a complementary, low-cost screening tool or as an input feature within multimodal diagnostic frameworks combining imaging, electrocardiograms, echocardiograms, and clinical data.

5. Conclusion

In this study, we investigated the feasibility of using DL architectures to detect PH based on CXR images. Among the evaluated approaches, CNNs models demonstrated the most stable and consistent performance. In contrast, Transformer-based models exhibited limited convergence and lower discriminative capability, likely reflecting the constraints imposed by dataset size and class imbalance. Our findings suggest that CXR images encode subtle morphological patterns associated with established PH, particularly in patients with elevated mPAP. However, overall diagnostic performance remained moderate. Reduced Se and Spe were observed inpatients with normal or mildly elevated mPAP values, highlighting the limited ability of CXR-based models to reliably detect early or borderline disease. Taking together, these results indicate that, although CXR-derived DL models may capture meaningful cardiopulmonary structural cues, their performance is insufficient for standalone diagnostic. Future work should prioritize the expansion of datasets through multicenter and prospective studies, the integration of multimodal clinical and imaging data, and the application of explainable AI methods to elucidate the anatomical regions and features driving model predictions. Such advances may enhance model robustness, improve interpretability, and support the development of scalable and accessible tools for PH screening and risk stratification in clinical practice.

Acknowledgements

This study was supported in part by the Zerbini Foundation, and Bayer, as part of the research project “Modelo Multimodal de Aprendizado Profundo para Diagnóstico não invasivo de Hipertensão Pulmonar baseado na análise de Eletrocardiograma e Radiografia de Tórax”.

References

Becerra-Muñoz, V. M., Gómez Sáenz, J. T., and Escribano Subías, P. (2024). The importance of data in pulmonary arterial hypertension: From international registries to machine learning. *Medicina Clínica (English Edition)*, 162(12):591–598.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, pages 1–21.

Fadilah, A., Putri, V. Y. S., Puling, I. M. D. R., and Willyanto, S. E. (2024). Assessing the precision of machine learning for diagnosing pulmonary arterial hypertension: a systematic review and meta-analysis of diagnostic accuracy studies. *Frontiers in Cardiovascular Medicine*, 11:1–17.

Ferreira, M. I. P., Oliveira, H. G., Krug, B. C., Gonçalves, C. B. T., Amaral, K. M., Mosca, M., Picon, P. D., M., R. R., and Schneiders, R. E. (2014). Hipertensão arterial pulmonar - protocolo clínico e diretrizes terapêuticas.

He, K., Zhang, X., Ren, S., and Sun, J. (2015). Deep residual learning for image recognition. *arXiv*, pages 1–12.

Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q. (2017). Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708.

Humbert, M., Kovacs, G., Hoeper, M. M., Badagliacca, R., Berger, R. M. F., Brida, M., Carlsen, J., Coats, A. J. S., Escribano-Subias, P., Ferrari, P., Ferreira, D. S., Ghofrani, H. A., Giannakoulas, G., Kiely, D. G., Mayer, E., Meszaros, G., Nagavci, B., Olsson, K. M., Pepke-Zaba, J., Quint, J. K., Rådegran, G., Simonneau, G., Sitbon, O., Tonia, T., Toshner, M., Vachiery, J. L., Vonk Noordegraaf, A., Delcroix, M., Rosenkranz, S., and Group, E. S. D. (2022). 2022 esc/ers guidelines for the diagnosis and treatment of pulmonary hypertension: Developed by the task force for the diagnosis and treatment of pulmonary hypertension of the european society of cardiology (esc) and the european respiratory society (ers). endorsed by the international society for heart and lung transplantation (ishlt) and the european reference network on rare respiratory diseases (ern-lung). *European Heart Journal*, 43(38):3618–3731.

Imai, S., Sakao, S., Nagata, J., Naito, A., Sekine, A., Sugiura, T., Shigeta, A., Nishiyama, A., Yokota, H., Shimizu, N., Sugawara, T., Nomi, T., Honda, S., Ogaki, K., Tanabe, N., Baba, T., and Suzuki, T. (2024). Artificial intelligence-based model for predicting pulmonary arterial hypertension on chest x-ray images. *BMC Pulmonary Medicine*, 24:101.

Khou, V., Anderson, J. J., Strange, G., Corrigan, C., Collins, N., Celermajer, D. S., Dwyer, N., Feenstra, J., Horrigan, M., Keating, D., Kotlyar, E., Lavender, M., McWilliams, T. J., Steele, P., Weintraub, R., Whitford, H., Whyte, K., Williams, T. J., Wrobel, J. P., Keogh, A., and Lau, E. M. (2020). Diagnostic delay in pulmonary arterial hypertension: Insights from the australian and new zealand pulmonary hypertension registry. *Respirology*, 25:863–871.

Kusunose, K., Hirata, Y., Tsuji, T., Kotoku, J., and Sata, M. (2020). Deep learning to predict elevated pulmonary artery pressure in patients with suspected pulmonary hypertension using standard chest x ray. *Scientific Reports*, 10:19311.

Li, Z., Luo, G., Ji, Z., Wang, S., and Pan, S. (2024). Explanatory deep learning to predict elevated pulmonary artery pressure in children with ventricular septal defects using standard chest x-rays: a novel approach. *Frontiers in Cardiovascular Medicine*, Volume 11 - 2024.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. *arXiv*, pages 1–14.

Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., and Xie, S. (2022). A convnet for the 2020s. *arXiv*, pages 1–15.

Luna-López, R., Ruiz Martín, A., and Escribano Subías, P. (2022). Pulmonary arterial hypertension. *Medicina Clínica (English Edition)*, 158(12):622–629.

Magalhães Junior, H. M. and Safatle, L. P. (2023). Hipertensão pulmonar - protocolo clínico e diretrizes terapêuticas.

McLaughlin, V. V., Shillington, A., and Rich, S. (2002). Survival in primary pulmonary hypertension. *Circulation*, 106(12):1477–1482.

Peña, E., Dennie, C., Veinot, J., and Muñiz, S. H. (2012). Pulmonary hypertension: How the radiologist can help. *RadioGraphics*, 32(1):9–32.

Ribeiro, E., Cardenas, D. A. C., Dias, F. M., Krieger, J. E., and Gutierrez, M. A. (2024). Explainable artificial intelligence in deep learning–based detection of aortic elongation on chest x-ray images. *European Heart Journal - Digital Health*, 5(5):524–534.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. (2014). Going deeper with convolutions. *arXiv*, pages 1–12.

Tan, M. and Le, Q. (2019). Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR.

Zou, X.-L., Ren, Y., Feng, D.-Y., He, X.-Q., Guo, Y.-F., Yang, H.-L., Li, X., Fang, J., Li, Q., Ye, J.-J., Han, L.-Q., and Zhang, T.-T. (2020). A promising approach for screening pulmonary hypertension based on frontal chest radiographs using deep learning: A retrospective study. *PLOS ONE*, 15(7):1–13.