

Localization and classification of abnormalities on chest X-ray images using a Mamba-YOLOvX model

Ebrahim Khalili ^{a,b}, Daniel Sanchez-Morillo ^{a,b,*}, Blanca Priego-Torres ^{a,b}, Antonio León-Jiménez ^{b,c}

^a Department of Automation, Electronics and Computers and Network Architecture, Bioengineering, Automation and Robotics Research Group; University of Cadiz, Avda. de la Universidad de Cadiz, 10, Puerto Real (Cadiz), 11519, Spain

^b Biomedical Research and Innovation Institute of Cadiz - INIBICA, Avda. Ana de Viya, 21, Cadiz, 11009, Spain

^c Pulmonology Department, Puerta del Mar University Hospital, Avda. Ana de Viya, 21, Cadiz, 11009, Spain

ARTICLE INFO

Keywords:

Chest X-ray
Lung
YOLO
Deep learning
Mamba
Selective state space
Medical imaging
CXR

ABSTRACT

Chest X-rays (CXR) are critical diagnostic tools for detecting thoracic abnormalities. However, challenges such as overlapping anatomical structures, class imbalance, and dataset heterogeneity hinder accurate interpretation and limit model generalizability. To address these issues, a Mamba-YOLOvX model is presented in this study. It was aimed to integrate global and local lesion information to improve the detection and localization of thoracic abnormalities. The model incorporates novel architectural improvements, including combined spatial and channel attention mechanisms and selective scanning blocks, to capture fine-grained features and enhance multi-scale detection. In addition, a projection-based data augmentation strategy, leveraging rib segmentation and keypoint alignment was developed to improve the anatomical consistency and the intensity normalization across datasets. Extensive experiments were conducted on three large-scale datasets (VinDr-CXR, ChestX-ray8, and CXR-AL14), achieving state-of-the-art performance in detecting abnormalities of varying sizes. The proposed method reached an average precision at 50 % intersection over union of 0.366, 0.153, and 0.615 on the VinDr-CXR, ChestX-ray8, and CXR-AL14 datasets, respectively. Results demonstrated significant improvements in precision, recall, and mean average precision, particularly for small lesions. Cross-dataset validation confirmed the model's robustness and generalizability. This study highlights the potential of integrating advanced deep learning techniques with domain-specific augmentations to enhance clinical decision support systems for thoracic disease detection. By addressing critical challenges such as class imbalance, annotation inconsistencies, and scale variations, the enhanced Mamba-YOLOvX model is shown as a scalable, accurate, and generalizable solution for automated CXR analysis.

1. Introduction

Chest X-rays (CXR) remain the most frequently performed radiological examination worldwide, with millions of scans performed annually. In industrialized countries, an average of 238 CXR is estimated to be performed per 1000 inhabitants per year (Radiation, 2010). This underscores the widespread use of chest X-rays as an essential tool for diagnosing various thoracic diseases and highlights their status as a cornerstone of radiological imaging for decades (Çalli, Sogancioglu, van Ginneken, van Leeuwen, & Murphy, 2021; Xu & Duan, 2024).

CXR provides radiologists with vital information about the size, shape, and condition of internal organs by revealing details about the lungs, heart, spine, ribs, blood vessels, and airways. Importantly, CXR

are cost-effective and involve relatively low radiation doses, making them a primary approach to diagnosing and screening thoracic diseases (Xu & Duan, 2024). Despite this medical test's relevance, there are inherent challenges in radiological interpretation. First, interpreting chest radiographs can be difficult due to overlapping anatomic structures. Detecting small or subtle abnormalities, or accurately distinguishing between pathological patterns, can be challenging (Çalli et al., 2021). Furthermore, the inherent subjectivity of image interpretation leads to high variability between observers among radiologists when analyzing chest radiographs (Balabanova et al., 2005; Quekel, Kessels, Goei, & van Engelschoven, 2001). This discrepancy can result in inconsistent diagnoses and possible errors, with a day-to-day radiologist error rate averaging 3–5 % and a retrospective error rate among radiological studies of 30 %

* Corresponding author.

E-mail addresses: ebrahim.khalili@uca.es (E. Khalili), daniel.morillo@uca.es (D. Sanchez-Morillo), blanca.priego@uca.es (B. Priego-Torres), antonio.leon.sspa@juntadeandalucia.es (A. León-Jiménez).

<https://doi.org/10.1016/j.eswa.2025.127929>

Received 31 January 2025; Received in revised form 9 April 2025; Accepted 26 April 2025

Available online 28 April 2025

0957-4174/© 2025 The Author(s). Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

(Brady, 2017). The radiologist's experience and the cognitive process underlying the interpretation of X-rays directly impact the clinical utility and efficacy of this technique in the treatment of patients (Bruls & Kwee, 2020; Lee, Nagy, Weaver, & Newman-Toker, 2013). To come full circle, the rising demand for imaging examinations has significantly increased the workload of radiologists, potentially leading to fatigue and a higher risk of errors, particularly in emergencies (Fanni et al., 2023). Diagnostic efficiency is also influenced by technical factors related to imaging, such as image quality. Key aspects, including image resolution, noise, patient positioning, and artifacts can impact the visibility of anatomical structures and abnormalities (Pham, Le, Tran, Ngo, & Nguyen, 2021). Furthermore, variations in X-ray equipment and imaging protocols across institutions and geographic regions can create inconsistencies in radiograph appearance (Zunaed, Haque, & Hasan, 2024). Certain thoracic pathologies may exhibit similar visual characteristics on radiographs, complicating differentiation (Çalli et al., 2021). Additionally, chest radiographs often reveal multiple abnormalities simultaneously, increasing diagnostic complexity (Chen, Zhang, Lin, Chen, & Lu, 2020).

Clinical decision support systems (CDSS), which include computer-aided detection (CAD) systems, have the potential to improve the efficiency of radiological interpretation significantly. By leveraging artificial intelligence (AI) techniques, CDSS can identify subtle abnormalities that may be overlooked by fatigued radiologists (Pham et al., 2021). The research undertaken in recent years indicates that both novice and experienced radiologists achieve improved accuracy when utilizing AI as a "second reader" (Fanni et al., 2023). CDSS can integrate various data, including clinical information and prior investigations, to support diagnostic decision-making (Çalli et al., 2021). The history of CAD in radiology dates back to the 1960s, evolving through significant developments in the 1980s and 1990s, when various CAD schemes emerged for automated lesion detection (Becker, Nettleton, Meyers, Sweeney, & Nice, 1964; Heber et al., 1995; Toriwaki, Suenaga, Negoro, & Fukumura, 1973). The rise of deep learning and convolutional neural networks (CNNs) in the 2010s further revolutionized CAD systems, enhancing their accuracy and capability to detect abnormalities, supported by extensive datasets (Kallianos et al., 2019; Oakden-Rayner, 2019). Currently, there is a focus on integrating CAD systems into radiological workflows, alongside efforts to create more advanced systems capable of detecting multiple findings across different imaging modalities (Fan et al., 2024). However, challenges remain, such as the overlapping of anatomical structures, the high visual similarity between thoracic diseases, unbalanced datasets leading to overfitting, and incomplete annotations that hinder model generalization, compounded by variations in image quality that affect detection accuracy (Çalli et al., 2021; Chen et al., 2020; Fan et al., 2024; Fernández, García, & Herrera, 2011; Xu & Duan, 2024).

In recent years, there have been several approaches to address the challenge of detecting multiple abnormalities in CXR. In 2023, Euyoung K. et al. (Le et al., 2023) implemented the YOLOv5 baseline for supervised object detection models to learn from various annotators in the VinDr-CXR training set. Labels provided by multiple radiologists with varying levels of expertise were provided, and an overall AP₅₀ (average precision at intersection over union or IOU = 0.5) of 0.2 was achieved. In 2024, Weijie F. et al. (Fan et al., 2024) developed a large CXR dataset (CXR-AL14), comprising 165,988 CXR. Using this dataset, a YOLOvX framework (Ge, Liu, Wang, Li, & Sun, 2021) was implemented to identify and localize 14 common abnormalities and calculate the cardiopulmonary ratio (CTR) simultaneously. AP₅₀ values ranging from 0.572 to 0.631 were estimated for the 14 abnormalities. Also in 2024, the DualAttNet framework, a dual attention supervised module developed for detecting multiple lesions in chest radiographs, was presented (Xu & Duan, 2024). This module effectively merged global and local lesion classification data through an image-level attention block and a detailed disease attention algorithm. The performance of DualAttNet on the VinDr-CXR, ChestX-ray8, and COVID-19 datasets was evaluated, and the estimated AP₅₀ was 0.241, 0.145, and 0.064, respectively. In all these approaches,

the absence of localization information limited the diagnostic accuracy and hindered the precise identification. Without location information, deep learning models have difficulty in focusing on the relevant regions of the image that contain the anomalies. Attention mechanisms, which have been used to improve anomaly detection, heavily rely on location information to guide the model's attention (Liu et al., 2019). The lack of location information can increase the likelihood of false positives and restrict the interpretability of models (Meedeniya et al., 2022). The latest deep learning architectures, such as Mamba (Gu & Dao, 2023), are showing promise for their potential to impact notably the field of computer vision. Its computational efficiency, ability to handle long-range dependencies, and adaptability to multidimensional data make it an attractive alternative to transformers in a variety of applications.

In this study, a Mamba-YOLOvX method with combined channel and spatial attention mechanisms was implemented and evaluated. Inspired by FER-YOLO-Mamba (Ma, Lei, Celik, & Li, 2024), the conventional attention mechanism was adjusted to address more effectively the challenges of detecting multiple abnormalities in low-resolution X-ray images. A combined channel and spatial attention mechanism was integrated. Unlike the FER-YOLO-Mamba model, which utilized channel attention through adaptive average pooling to highlight informative feature channels, this approach incorporated both average and max pooling, capturing a richer set of global contextual features. Moreover, a spatial attention mechanism was embedded. It focused on key spatial regions by leveraging a combination of pooled feature maps and convolutional filtering, emphasizing subtle pixel variations that help distinguish pathological areas from surrounding lung tissue. In addition, dilated convolutions were integrated to expand the receptive field, allowing the model to capture multiscale contextual information without sacrificing resolution. These modifications improved the model's ability to capture both local and global relationships, resulting in enhanced discrimination between abnormalities and lung tissue. In summary, the primary contributions of this paper are the following:

- Development of a Mamba-YOLOvX model to locate and classify lesion data in chest X-rays. We leveraged the long-range dependency modeling of Mamba, a recent state-space model, integrating it into the backbone of YOLOvX.
- Selective scanning mechanisms implemented in the YOLOvX backbone were used to focus on critical regions of the input data, and were shown effective in extracting features associated with both large and small abnormalities.
- A novel rib-guided projection-based data augmentation strategy was introduced. This method used rib segmentation and keypoint alignment to generate anatomically consistent augmentations, preserving structural integrity and improving the model's generalization.
- An ablation analysis was conducted to evaluate the contribution of the different modules (attention mechanism and data augmentation) to detection performance, demonstrating that both components significantly improve the results.
- Exhaustive validation of the proposed model using three large-scale datasets (VinDr-CXR (Nguyen et al., 2022), CXR-AL14 (Fan et al., 2024), and ChestX-ray8 (Wang et al., 2017)), achieving state-of-the-art performance in detecting abnormalities of varying sizes. Cross-dataset validation was also performed to confirm the model's robustness and generalizability.

2. Materials and methods

2.1. Key challenges

The development of a robust model for anomaly detection in medical imaging is based heavily on the diversity and quality of the datasets used for training and validation. Considering multiple datasets during algorithm development is crucial for enhancing model generalization, ensuring robustness, and reducing biases. Single-dataset approaches often fail

to capture the wide variability inherent in real-world clinical settings, where factors such as patient demography, imaging equipment, acquisition protocols, and disease presentation differ. Using publicly available datasets, to assess models can better address this heterogeneity, ensuring a better performance and a broader applicability in practical scenarios.

However, considering multiple datasets raises challenges that confound the training of effective algorithms. Key issues include significant sample distribution imbalances, variability in the scale and spatial distribution of abnormalities, and inconsistencies in labelling practices across datasets (Oksuz, Cam, Kalkan, & Akbas, 2020). In the following lines, these challenges are discussed, and strategies in the context of this study to mitigate them are proposed.

2.1.1. Class imbalance

Class imbalance is one of the most significant challenges in medical imaging, particularly when integrating datasets from diverse sources. Rare anomalies may be well represented in one dataset, but present scarcely in others, leading to biases during training. To address this issue, Mamba-YOLOvX, a modified version of FER-YOLO-Mamba, was integrated to enhance the detection of small and irregular objects in complex backgrounds. By incorporating enhancements such as multi-scale detection, attention mechanisms, and an improved feature extraction pipeline, this model focused on subtle features associated with underrepresented classes. These modifications allowed Mamba-YOLOvX to achieve high detection performance, even for classes with limited representation across diverse datasets.

2.1.2. Scale variation

The size of the pulmonary regions under study, which contain the lesions or anomalies of interest, can vary significantly not only within a single dataset but also across datasets due to differences in imaging protocols, resolution, and equipment. To address this, an image projection approach was applied, utilizing rib detection as a reference structure and leveraging the anatomical consistency of ribs to define and align regions of interest. Additionally, a state-of-the-art matching method, LightGlue (Lindenberger, Sarlin, & Pollefeys, 2023), was used to accurately calculate the transformations required for the projection. This combined strategy ensures precise normalization of the region size, enabling the model to maintain consistent focus and improve detection performance across datasets.

2.1.3. Annotation variability

Combining datasets from different sources often introduces inconsistencies in annotations. In some datasets, multiple radiologists independently annotate the same images, leading to overlapping or conflicting definitions of bounding boxes caused by differences in interpretation. To address these inconsistencies, Weighted Box Fusion (WBF) (Solovyev, Wang, & Gabruseva, 2021) was used. WBF combines multiple overlapping predictions into a single, more accurate bounding box by weighting them based on their confidence scores, providing a unified spatial representation.

2.1.4. Intensity distribution imbalance

Objective imbalance stems from variations in imaging equipment, acquisition protocols, and the inherent heterogeneity of patient populations across datasets. These differences can significantly affect image quality and intensity distributions, complicating the detection of anomalies. To address this, rib segmentation was used to normalize the intensity distribution within images. The ribs serve as consistent anatomical markers that enclose the lung and thoracic spaces in CXR. These areas typically appear as the densest parts of the image, with the highest pixel intensities. By detecting the ribs, they were used as a reference to apply an intensity offset that adjusts the overall brightness and contrast of the image. This normalization of the image histogram distribution reduced aliasing effects, such as excessive brightness in non-bone areas, and ensured that features in the lungs and thoracic region were consistently represented.

Table 1

Summary of the characteristics of datasets containing anomalies.

Dataset	Image size	Images	Classes
CXR-AL14 (training set)	Variable	98,124	14
VinDr-CXR (training set)	Variable	4394	14
VinDr-CXR (test set)	Variable	948	14
ChestX-ray8	1024x1024	880	8

2.2. Dataset description

In this study, four publicly available CXR datasets were used to train and evaluate the proposed anomaly detection framework. These datasets, namely VinDr-CXR, ChestX-ray8, CXR-AL14, and VinDr-RibCXR, differ in their size, diversity, labelling practices, and the types of anomalies included. The VinDr-RibCXR dataset, specifically, was employed for rib segmentation, serving as a foundational step in the proposed augmentation pipeline to normalize intensity distributions and improve anatomical consistency across images.

Table 1 delivers additional details about their characteristics, including image size, total number of bounding boxes, number of images, and number of classes.

Fig. 1 shows a Venn diagram comparing the abnormalities included in the three datasets, and Table 2 details the distribution of the different classes (labels) along the four datasets used in this study.

2.2.1. VinDr-CXR

The VinDr-CXR dataset (Nguyen et al., 2022) is a valuable resource for CXR imaging research, comprising 18,000 scans, divided into 15,000 for training and 3000 for testing. The dataset includes detailed information on the localization of abnormal regions and the classification of common thoracic conditions. Annotations were conducted by a team of 17 radiologists, each with at least eight years of professional experience. For the training set, each scan was independently reviewed by three radiologists, while the testing set underwent a more rigorous process, with labels determined by consensus among five radiologists. The training set includes 4394 images with a total of 22 pathological findings. The testing dataset includes 948 abnormal CXR. Since some of the 22 categories contained a very low number of instances, eight classes were grouped under the label of "other lesions". Thus, the final dataset contained labels for 14 different radiological findings. This approach was followed in Lin, Huang, Wang, Feng, and Huang (2023) and Xu and Duan (2024). This was the database used for the Kaggle 2021 VinBigData Chest X-ray Abnormalities Detection Competition.

This dataset exhibits a significant class imbalance, as detailed in the Table 2. Pleural thickening, aortic enlargement, fibrosis and car-

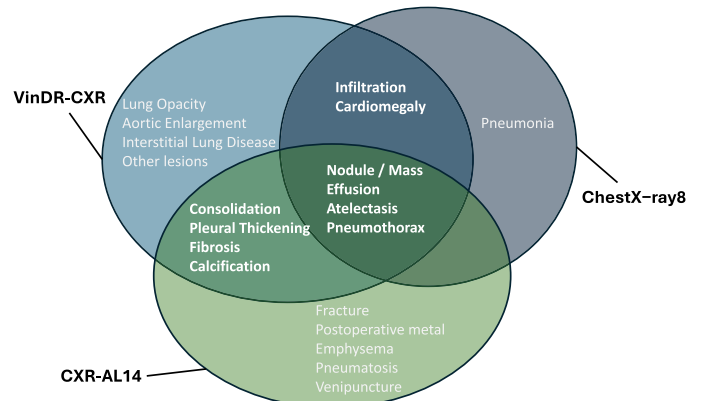


Fig. 1. Venn diagram illustrating the overlap of abnormalities included across the three chest X-ray datasets: VinDr-CXR, CXR-AL14, and ChestX-ray8.

Table 2
Distribution of the pathological findings in the CXR datasets used in this study.

Label	VinDR Training Set	VinDR Test Set	CXR-AL14	ChestX-ray8
Aortic enlargement	14.6 %	14.4 %		
Cardiomegaly	10.9 %	10.6 %		14.8 %
Atelectasis	1.0 %	1.2 %	0.1 %	18.3 %
Calcification	3.3 %	3.0 %	13.4 %	
Consolidation	1.9 %	2.1 %	8.2 %	
Pleural effusion	6.5 %	6.6 %	17.5 %	15.5 %
Infiltration	4.1 %	3.6 %		12.5 %
Pulmonary fibrosis	13.2 %	12.2 %	6.6 %	
Interstitial lung disease	3.1 %	2.6 %		
Nodule/Mass	8.2 %	9.5 %	19.7 %	16.7 %
Pleural thickening	16.2 %	16.6 %	3.0 %	
Lung opacity	8.5 %	8.4 %		
Pneumothorax	0.5 %	0.7 %	2.7 %	10.0 %
Other lesion	7.9 %	8.3 %		
Emphysema			9.8 %	
Postoperative metal			7.4 %	
Pneumatosis			2.4 %	
Venipuncture			4.1 %	
Fracture			5.1 %	
Pneumonia				12.2 %

diomegaly are overrepresented compared to other findings such as pneumothorax, consolidation, atelectasis, calcification, and infiltration.

2.2.2. CXR-AL14

The CXR-AL14 dataset (Fan et al., 2024) is a comprehensive resource for CXR imaging research, comprising a total of 165,988 scans, including both normal images and those with thoracic abnormalities. Among these, 149,425 scans are included in the training set, with 92,620 scans containing abnormalities, which are the sole focus of this analysis. The dataset was created using a human-in-the-loop approach, combining automated methods with expert intervention to reduce the annotation burden on radiologists. This process enabled the generation of a large number of ground truth bounding boxes for 14 types of abnormalities.

Again, this dataset reveals unbalanced, as detailed in the Table 2. Nodules and masses and pleural effusion classes are overrepresented while atelectasis, pneumatosis, pneumothorax, pleural thickening, venipuncture and fracture, among others, appears to be underrepresented.

2.2.3. ChestX-ray8

The ChestX-ray8 includes 108,948 images, of which 24,636 are abnormal CXRs. As part of the ChestX-ray8 database, a small number of pathological images are provided with hand-labeled bounding boxes of 14 types of thoracic abnormalities. Annotations were performed by board-certified radiologists.

Compared to VinDr-CXR and CXR-AL14, ChestX-ray8 exhibits a less severe class imbalance, with findings such as atelectasis and pleural effusion having reasonable representation. Nodules and masses labels were group for the fair comparison with the rest of datasets.

2.2.4. VinDr-RibCXR

VinDr-RibCXR (Nguyen, Le, Pham, & Nguyen, 2021) consists of 245 CXRs, each segmented and annotated for rib identification tasks. An expert radiologist manually annotated all 20 ribs in each image, categorizing them as left (L1 to L10) or right (R1 to R10) ribs. This dataset was used in our study to capture and utilize the structural features of each CXR. Fig. 2 (a) shows an example of the original image, while Fig. 2 (b) illustrates the 20 ribs, each labelled in a different colour.

2.3. Image augmentation

Data augmentation is a widely used technique to enhance the diversity of training datasets. Traditional approaches include basic pixel-level

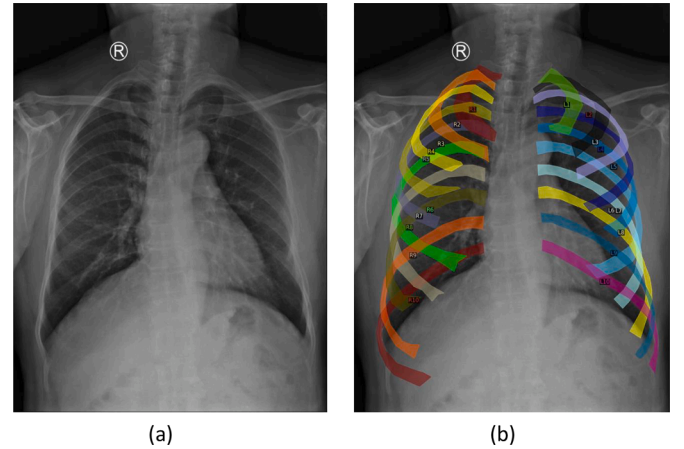


Fig. 2. Example of a chest X-ray image available in the VinDr-RibCXR dataset: a) original image, b) original image with the ribs labelled in colours.

transformations such as rotation, HSV saturation, perspective, and translation (Krizhevsky, Sutskever, & Hinton, 2012). More advanced methods like AutoAugment (Cubuk, Zoph, Mane, Vasudevan, & Le, 2018) leverage reinforcement learning to identify optimal combinations of augmentations. Although these methods improve data diversity, they often fail to preserve the realistic anatomical details and structural integrity required for high-stakes medical imaging tasks. In addition, they struggle to capture critical features such as realistic edges, depth, and anatomical correspondence.

To address these limitations, a projection-based data augmentation strategy is proposed. It uses anatomical features, specifically the ribs, as reference structures. By segmenting and aligning rib structures, the method generates anatomically consistent projections and augmentations that maintain clinical relevance and improve model generalization. This approach integrates rib segmentation, keypoint matching, and advanced contrast enhancement into a unified workflow.

The proposed augmentation workflow consists of four key steps, each designed to ensure that the augmented images preserve anatomical integrity and reflect the realistic variations in medical imaging data.

2.3.1. Workflow overview

a) *Rib Segmentation.* The process begins with rib segmentation using YOLOv8m-seg, a segmentation-enhanced version of the YOLO architecture (Jocher, Chaurasia, Qiu, 2023). This model produces precise rib

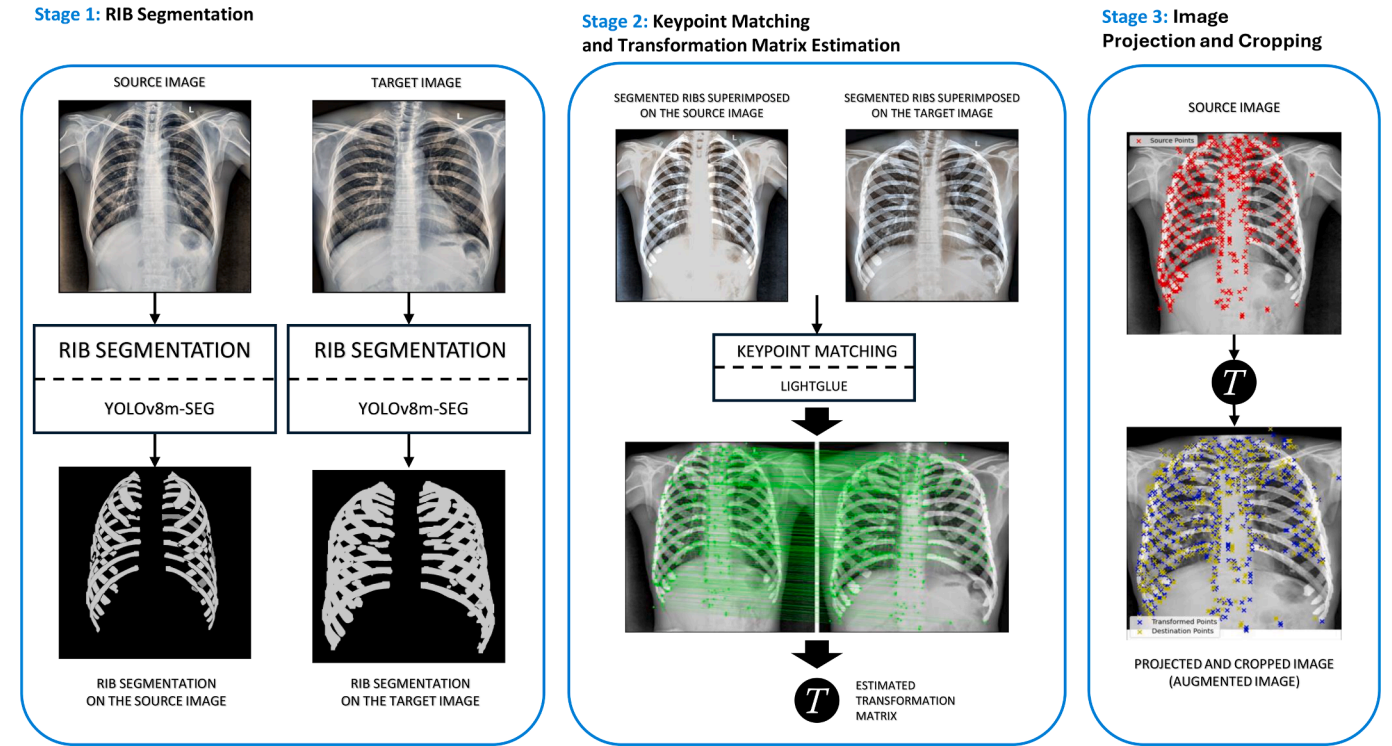


Fig. 3. Workflow of the image augmentation method: Stage 1 shows rib segmentation using YOLOv8, highlighting the ribs as key anatomical markers in X-ray images. Stage 2 applies the LightGlue matching method to identify target points. Stage 3 transforms each point from the source image to align with the target image.

masks by combining a CNN backbone for feature extraction with a segmentation head for detailed delineation. The fine-tuned YOLOv8m-seg was trained on the VinDr-RibCXR dataset, annotated specifically for rib structures, ensuring robust performance across diverse imaging conditions. Images were resized to 1024×1024 , and a 5% offset was added around the rib area to capture the total thoracic region for transformation (stage 2 in Fig. 3).

b) Keypoint Matching Using LightGlue. The LightGlue algorithm efficiently identified key points within rib structures and establishes precise correspondences between source and target images using a transformer-based architecture. This process ensured accurate transformation estimation and alignment of thoracic regions while preserving anatomical consistency (stage 2 in Fig. 3).

c) Projection and Cropping. Using the transformation matrix derived from LightGlue, source images were projected onto target images. The lung regions were delineated and cropped using rib edges as boundaries, ensuring precise focus on regions of interest and removing irrelevant areas (stage 3 in Fig. 3).

d) Histogram Normalization and Contrast Enhancement. To account for intensity variations across datasets, histogram normalization was applied, followed by three variations of contrast-limited adaptive histogram equalization (CLAHE) (Zuiderveld, 1994) with clip limits of $\{1, 1.5, 2\}$. This step enhanced contrast, preserved anatomical consistency, and lessened aliasing effects, ensuring a consistent image quality across the augmented dataset.

2.3.2. Projection-based data augmentation strategy

For each image in the dataset, the augmentation process generated N new augmented images. To create these, a different image from the dataset was selected as the target for the projection each time. This ensures that each original image was projected onto the anatomical framework of a distinct target image, resulting in unique variations that re-

flect diverse spatial alignments and perspectives. The detailed steps of the augmentation process are outlined in Algorithm 1.

2.4. Mamba-YOLOvX framework for multi-abnormality detection and localization

The Mamba-YOLOvX framework, as illustrated in Fig. 4, consisted of four main modules, namely: a) Preprocessing; b) Cross-Stage Partial connections Darknet (CSP Darknet); c) Selective Scanning 2D Block (SS2D) and Combined Spatial and Attention Dynamics (CSAD); d) You Only Look Once (YOLOvX) Head.

CLAHE with three clip limits of 1, 1.5, 2 was applied to the original CXR as a preprocessing technique to generate a new image with three enhanced contrast channels which fed the CSP Darknet stage. CSP Darknet (Bochkovskiy, Wang, & Liao, 2020) serves as the primary feature extraction network. It downscales the input image through convolutional layers and pooling operations to generate feature maps at multiple resolutions. These layers use kernels of varying sizes to capture complementary information: larger kernels focus on broader structures and gradual transitions across wider regions, while smaller kernels emphasize finer details such as edges and contours. The resulting feature maps are organized into three scales, capturing hierarchical features essential for multi-scale detection. These multi-scale feature maps are refined by the SS2D block, which integrates State Space Models (SSM) to capture both global and local spatial dependencies. By leveraging the mathematical foundations of SSM, the SS2D block models relationships between adjacent patches and improves multi-dimensional feature representation across scales. This upsampling mechanism ensures comprehensive coverage of the input image regions, enhancing robustness in abnormality detection. Finally, the architecture integrates three dedicated fully connected heads for bounding box regression, mask segmentation, and label classification tasks. These heads operate on the refined feature maps, enabling precise localization and classification of abnormalities. The following sections provide a more straightforward description of each component.

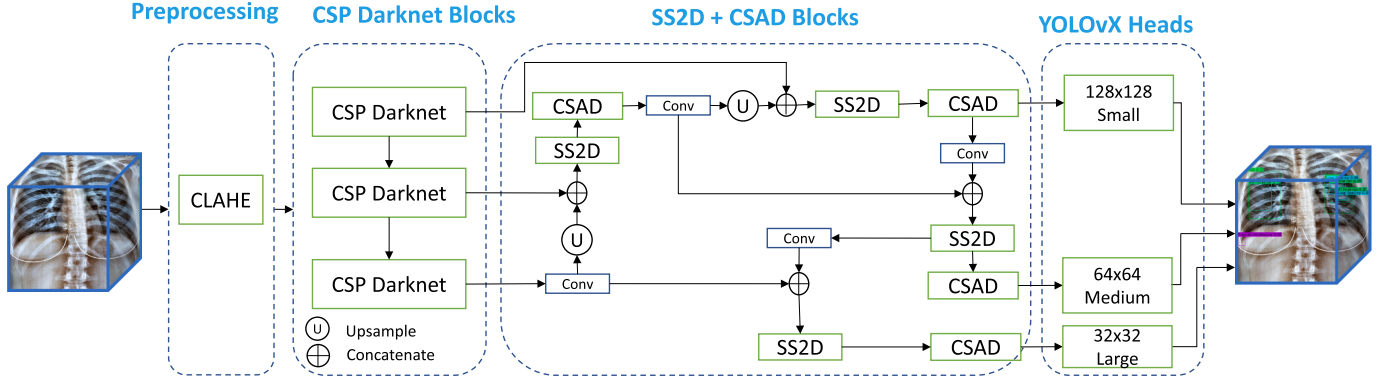


Fig. 4. Overall architecture of the proposed Mamba-YOLOvX framework.

Algorithm 1 Projection-Based Data Augmentation Strategy.

Require: Image dataset D , number of augmentations per image $N = 10$, CLAHE parameters, resizing dimension 1024×1024

- 1: **Input:** Original images $I \in D$, bounding box annotations B , rib-segmentation model $RibSeg$
- 2: **Output:** Augmented images $\tilde{I}$ with updated annotations $\tilde{B}$
- 3: **for** each image I_i in D **do**
- 4: Perform rib segmentation on I_i ; $RibMask_i = RibSeg(I_i)$
- 5: Resize I_i and $RibMask_i$ to 1024×1024
- 6: **for** $k = 1$ to N **do**
- 7: Randomly select a target image I_j from D , where $j \neq i$
- 8: Perform rib segmentation on I_j ; $RibMask_j = RibSeg(I_j)$
- 9: Resize I_j and $RibMask_j$ to 1024×1024
- 10: Add a 5% offset to rib masks:
- 11: $RibMask_i^{offset} = RibMask_i + 0.05 \times RibMask_i$
- 12: $RibMask_j^{offset} = RibMask_j + 0.05 \times RibMask_j$
- 13: Use LightGlue to match key points and compute the transformation matrix:
- 14: $T = LightGlue(RibMask_i^{offset}, RibMask_j^{offset})$
- 15: Apply transformation T to project I_i onto I_j :
- 16: $\tilde{I}_{ij} = T(I_i)$
- 17: Apply CLAHE to enhance contrast on $\tilde{I}_{ij}$:
- 18: $\tilde{I}_{ij} = CLAHE(\tilde{I}_{ij}, \{1, 1.5, 2\})$
- 19: Update bounding box annotations B_i to align with transformation T :
- 20: $\tilde{B}_{ij} = T(B_i)$
- 21: Exclude bounding boxes outside the valid rib area in $RibMask_j^{offset}$
- 22: Store augmented image and updated annotations:
- 23: $\tilde{I} \leftarrow \tilde{I}_{ij}, \tilde{B} \leftarrow \tilde{B}_{ij}$
- 24: **end for**
- 25: **end for**

A: SSM

SSM are grounded in the control theory and have been widely applied to represent linear time-invariant (LTI) systems. These models excel at capturing the dynamic relationships between input sequences and output variables through a compact representation of latent states. The equations governing an SSM are divided into the state equation, which models the temporal evolution of the latent states $\mathbf{h}(t)$, and the output equation, which relates these states to the system's observable output $\mathbf{y}(t)$.

With p defined as the number of inputs, q the number of outputs and n the number of state variables, the state equation describes the temporal evolution of the latent state vector, $\mathbf{h}(t) \in \mathbb{R}^n$, where the rate of change of the state depends on both the current state and the system input. Specifically, the latent state evolves according to the state transition

matrix, $\mathbf{A}$, ($\dim[\mathbf{A}(\cdot)] = n \times n$), which captures the intrinsic dynamics of the system, and the input projection matrix, $\mathbf{B}$, ($\dim[\mathbf{B}(\cdot)] = n \times p$), which describes how the input sequence, $\mathbf{x}(t) \in \mathbb{R}^p$, influences the state. This relationship is captured as follows:

$$\dot{\mathbf{h}}(t) = \mathbf{A}\mathbf{h}(t) + \mathbf{B}\mathbf{x}(t) \quad (1)$$

The output equation defines how the latent states are translated into observable outputs. The output vector, $\mathbf{y}(t) \in \mathbb{R}^q$, is expressed as a linear combination of the current state, modulated by the output projection matrix, $\mathbf{C}$, ($\dim[\mathbf{C}(\cdot)] = q \times n$), and the input sequence, scaled by the direct feedthrough matrix, $\mathbf{D}$, with $\dim[\mathbf{D}(\cdot)] = q \times p$.

$$\mathbf{y}(t) = \mathbf{C}\mathbf{h}(t) + \mathbf{D}\mathbf{x}(t) \quad (2)$$

In this formulation, the matrix $\mathbf{A}$ governs the internal dynamics of the system, while $\mathbf{B}$ regulates the influence of the input on the latent states. The matrices $\mathbf{C}$ and $\mathbf{D}$ determine how the states and inputs are translated into observable outputs, respectively.

Mamba adapts this continuous-time system for discrete-time sequence data by utilizing fixed discretization rules, commonly the Zero-Order Hold (ZOH). This method converts $\mathbf{A}$ and $\mathbf{B}$ into their discrete equivalents $\bar{\mathbf{A}}$ and $\bar{\mathbf{B}}$, respectively. This makes the system compatible with deep learning architectures that rely on discrete sequences. The discrete-time SSM can then be expressed in the following form:

$$\mathbf{h}_k = \bar{\mathbf{A}}\mathbf{h}_{k-1} + \bar{\mathbf{B}}\mathbf{x}_k \quad (3)$$

$$\mathbf{y}_k = \bar{\mathbf{C}}\mathbf{h}_k + \bar{\mathbf{D}}\mathbf{x}_k \quad (4)$$

where $\bar{\mathbf{A}} = e^{\Delta\mathbf{A}}$, $\bar{\mathbf{B}} = (e^{\Delta\mathbf{A}} - \mathbf{I})\mathbf{A}^{-1}\mathbf{B}$, $\bar{\mathbf{C}} = \mathbf{C}$, with ($\dim[\mathbf{B}(\cdot)]$, $\dim[\mathbf{C}(\cdot)] = d \times n$, and $\Delta \in \mathbb{R}^d$, with d associated with the number of features in the image patches.

These discrete SSM serve as the mathematical backbone of SS2D block, enabling it to process and enhance feature sequences derived from image patches. By adopting control theory principles, the SS2D block leverages this mathematical framework to systematically refine multi-dimensional features within the architecture.

B: SS2D

The SS2D mechanism, introduced by Liu et al. (2024), builds upon the foundations of SSM to address limitations in traditional feature extraction methods. SSM provide a structured approach to modelling sequential data, with each patch in the input image treated as either a temporal or spatial sequence. By discretizing the continuous SSM equations and utilizing modern GPU-based computation, the state $\mathbf{h}$ is efficiently represented within the memory hierarchy. This design enables the SS2D block to extract spatially-aware features by processing sequences derived from image patches. The block works across multiple orientations, ensuring a comprehensive analysis of the input image.

In the proposed model, each image is divided into 16 equally sized patches, which serve as the input for further processing. These patches

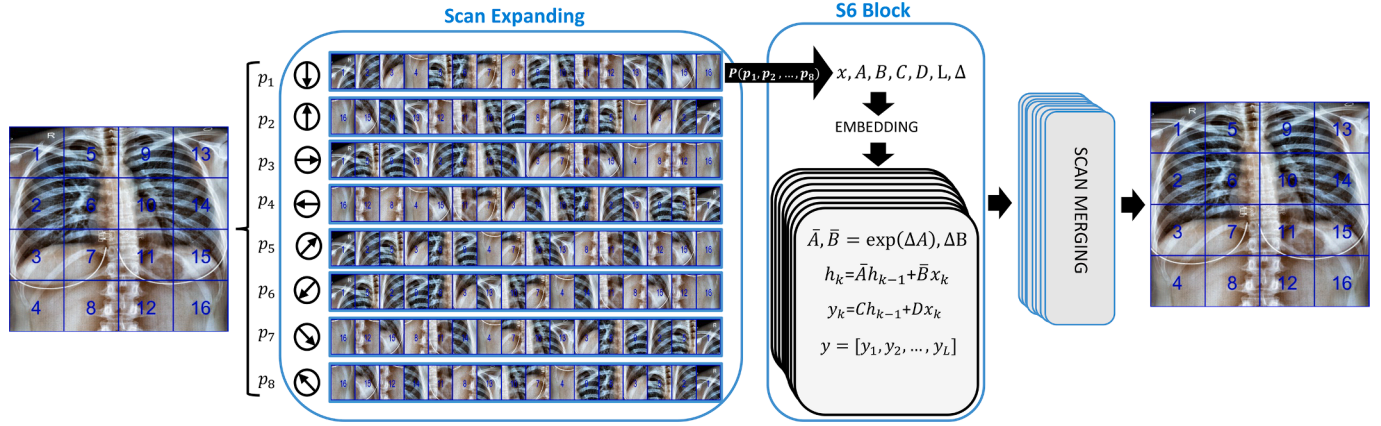


Fig. 5. Architecture of the SS2D module.

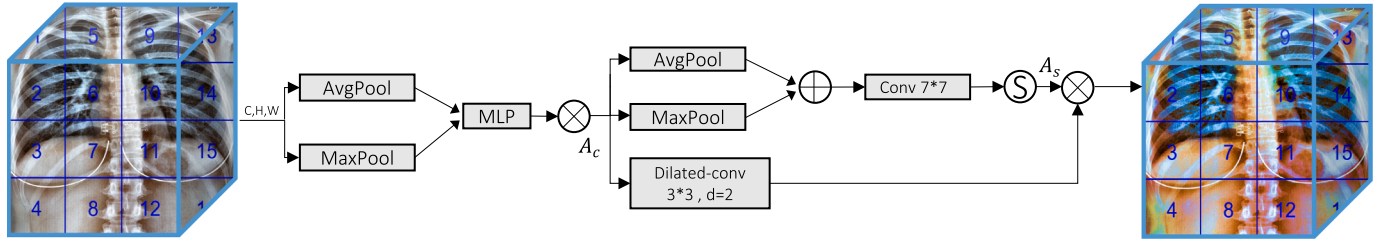


Fig. 6. Architecture of the CSAD module.

are systematically expanded to generate eight distinct sequences by applying selective scanning in various orientations: horizontal, vertical, diagonal, and reverse diagonal. This directional expansion ensures that the spatial relationships within the image are captured from multiple perspectives. The SS2D framework is illustrated in Fig. 5.

Each of the eight feature sequences is treated as a one-dimensional vector, representing the feature information extracted along its respective direction. These sequences are then independently processed through dedicated S6 Blocks (Gu & Dao, 2023), leveraging the efficiency and scalability of the SSM-based architecture. These blocks independently handle each sequence, reshaping and merging the outputs to form a unified feature representation. The resulting sequences are subsequently passed to the CSAD module, where they are refined to generate the final output map. This pipeline ensures that the SS2D block captures intricate spatial relationships within the input image while maintaining computational efficiency and scalability.

C: CSAD

The CSAD module uses selective scanning to focus on critical regions of the input data, thereby enhancing the detection of abnormalities. This is achieved through channel and spatial attention blocks that capture fine-grained variations across features and spatial dimensions. Fig. 6 shows the detailed structure of this module.

The input to CSAD is the output of SS2D, represented as $SS2D(x)$, with size $C \times H \times W$, where C , H , and W denote the number of channels, height, and width of the feature map, respectively.

Channel Attention: The input feature map $X = SS2D(x)$ undergoes max pooling and average pooling along spatial dimensions, producing two feature maps of size $C \times 1 \times 1$. These are passed through a multi-layer perceptron (MLP), combined via element-wise addition, and processed through a sigmoid activation to generate channel-specific attention weights W_c , with size $C \times 1 \times 1$. These weights are then forwarded as inputs to both the spatial attention mechanism and the dilated convolution layer.

Spatial Attention: For spatial attention, the pooled features are merged into two 2D maps, concatenated, and processed by a standard convolution layer to produce a 2D spatial attention map. A sigmoid activation is then applied to generate spatial attention weights.

Final Output: The outputs of the dilated convolution and spatial attention are combined via element-wise multiplication to produce the final output.

The operations can be summarized as follows, where $f_{3 \times 3}^{\text{dil}=2}$ is a dilated convolution with a kernel size of 3 and dilation factor of 2, and $f_{7 \times 7}$ represents a convolution with a kernel size of 7:

$$X_{\text{pool}} = \text{AvgPool}(X) + \text{MaxPool}(X) \quad (5)$$

$$\text{MLP}(X_{\text{pool}}) = \sigma(W_1(W_0(X_{\text{pool}}))) \quad (6)$$

$$A_c = W_c \cdot \text{MLP}(X_{\text{pool}}) \quad (7)$$

$$A_s = \sigma(f_{7 \times 7}([\text{AvgPool}(A_c); \text{MaxPool}(A_c)])) \quad (8)$$

$$\text{output} = A_s \cdot f_{3 \times 3}^{\text{dil}=2}(A_c) \quad (9)$$

D: Heads

The heads in Mamba-YOLOvX leverage a decoupled approach inspired by YOLOvX, where classification and bounding box regression tasks are separated into independent heads. This design ensures specialized optimization for each task, enhancing both classification accuracy and localization precision. Furthermore, the heads operate at multiple scales, similar to YOLOvX, enabling the effective detection of abnormalities across various sizes. Large, medium and small head outputs had size a 32×32 , 64×64 , and 128×128 respectively. Each head processes refined feature maps from the CSPDarknet backbone and the SS2D and CSAD modules, producing outputs for class prediction, bounding box regression, and confidence estimation. Notably, this architecture supports multi-scale detection without relying on an NMS-free approach (Li & Xiao, 2024). This design inherits YOLOvX's efficiency while incorporating Mamba-specific innovations, making it suitable for detecting subtle anomalies.

2.5. Validation schemes

Recent studies on anomaly detection in specialized datasets have employed random splits between training and testing sets at varying ratios to evaluate model performance. Following these practices, this criterion was adopted in this study to validate the presented approach across different datasets. Furthermore, to ensure the generalizability of the proposed method, performed cross-dataset validation was also carried out.

For intra-dataset validation, the standard practice of randomly splitting each dataset into training and testing sets was followed, ensuring no overlap of patient data. Tests with various split ratios were conducted, such as 90-10 % and 70-30 %, with the 90-10 % split serving as the primary configuration.

To further assess the generalizability of the proposed model, cross-dataset validation was conducted by training the model on one dataset and testing it on another. During this process, the focus was placed on pathologies common to both datasets to ensure a fair evaluation. As shown in Fig. 1, The datasets with the greatest class overlap are VinDr-CXR and ChestX-ray8. Accordingly, the model was trained on the VinDr-CXR training dataset and validated on the ChestX-ray8 dataset. All labels in the ChestX-ray8, with the exception of pneumonia, overlap with the VinDr-CXR dataset. To settle this single dissimilarity, the pneumonia label was not considered for the comparison of results. Furthermore, the mass and nodule classes that appear as separate findings in the ChestX-ray8 were merged to enable the comparison of results with the VinDr-CXR dataset. Nodules and masses refer to lung lesions, but they differ mainly in size. Nodules are typically smaller (less than 3 cm), while masses are larger (more than 3 cm). In some cases, distinguishing between nodules and masses may be difficult on a radiograph, especially when image resolution is insufficient or if the lesion characteristics are ambiguous. Grouping these classes enable a fair comparison between datasets and also backed the reduction of class imbalance.

2.6. Ablation Study

An ablation study was conducted on the VinDr-CXR dataset to analyze the contribution of different modules within the proposed approach to detection performance. The model was trained using the VinDr-CXR training set, while the testing dataset was reserved for evaluation. To assess the impact of the attention mechanism, two versions of the model were trained and compared: one with the attention block and one without. This comparison allowed for the evaluation of the attention mechanism's influence on overall performance.

Additionally, the impact of data augmentation was examined by training a separate model using augmented images. This step was aimed at determining how the inclusion of augmented data influenced the model's detection capabilities. The results from these experiments were compared to assess the effectiveness of each module in improving detection performance.

2.7. Performance metrics

Multiple metrics were estimated to evaluate the performance of the presented approach, including precision, average precision (AP), average recall (AR), mean average precision (mAP), the area under the receiver operating curve (AUC), and the Intersection over Union (IoU) accuracy (Everingham, Gool, Williams, Winn, & Zisserman, 2010).

To calculate the IoU accuracy, a localization is classified as "correct" only if $\text{IoU} > T(\text{IoU})$, where $T(\text{IoU})$ represents the threshold assigned for localization accuracy. The ground truth was evaluated against various thresholds, specifically $T = \{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7\}$, following the guidelines in Rezatofighi et al. (2019).

In the context of object detection tasks, predicted bounding boxes are classified into three categories based on their alignment with the ground truth: true positive (TP), false positive (FP), and false negative

(FN). The precision was estimated using an IoU threshold of 0.5.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (10)$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (11)$$

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (12)$$

AP_{50} and AP_{75} or average precision at IoU threshold of 0.5 and 0.75 respectively, were estimated to assess model performance. These metrics measure the precision of a model's predictions at a specific level of overlap (IoU) between the predicted bounding boxes and the ground truth bounding boxes.

The mean Average Precision (mAP) was used to evaluate the performance of a model in detecting and segmenting objects across multiple classes. It was calculated by averaging AP values across the different classes.

Finally, the model's performance was assessed using average recall and precision metrics to account for the substantial differences in lesion sizes across various diseases: AP_S and AR_S for small lesions (0–32 pixels), AP_M and AR_M for medium lesions (32–96 pixels), AP_L and AR_L for large lesions (over 96 pixels).

In assessing computational resources, we quantified the complexity of the architecture using metrics such as the number of trainable parameters and the floating point operations (FLOPS).

2.8. Experimental Environment

All models were implemented using the PyTorch framework, with training performed on an Ubuntu 20.04 system, equipped with an A100 GPU (40GB) and CUDA version 11.7.0. The YOLO model was trained for 200 epochs, employing an input image resolution of 640 and a batch size of 32. The confidence threshold and IoU threshold were both set at 0.5. The Mamba model was initialized from the pre-trained YOLOvX architecture. A learning rate of 0.001 was used and systematically decreased through a scheduled learning rate decay approach. The model was optimized using the Stochastic Gradient Descent (SGD) algorithm and trained over 100 epochs.

3. Results and discussion

In this section, the results of the proposed framework for abnormality detection are presented and discussed. First, the results obtained from the rib segmentation model used as the basis of the data augmentation technique are qualitatively illustrated. Subsequently, the results derived from the ablation studies performed are described thoroughly. Next, the results obtained in the evaluation of the final pipeline derived from the ablation studies are presented. Metrics estimated for all the validation schemes are presented. Finally, the calculated results are compared with those described in the state of the art in the context of this research.

3.1. Ribs segmentation

The proposed augmentation technique, based on anatomical alignment and intensity normalization, was applied to improve robustness. An example of the segmented ribs is presented in Fig. 7 to illustrate the effectiveness of the rib segmentation process. It highlights the accuracy of the YOLOv8m-seg model in identifying and delineating the rib structures, which serve as the foundation for subsequent steps in the augmentation pipeline. The segmented ribs are accurately marked, demonstrating the model's ability to handle complex anatomical features and ensure precise alignment for further image transformations.

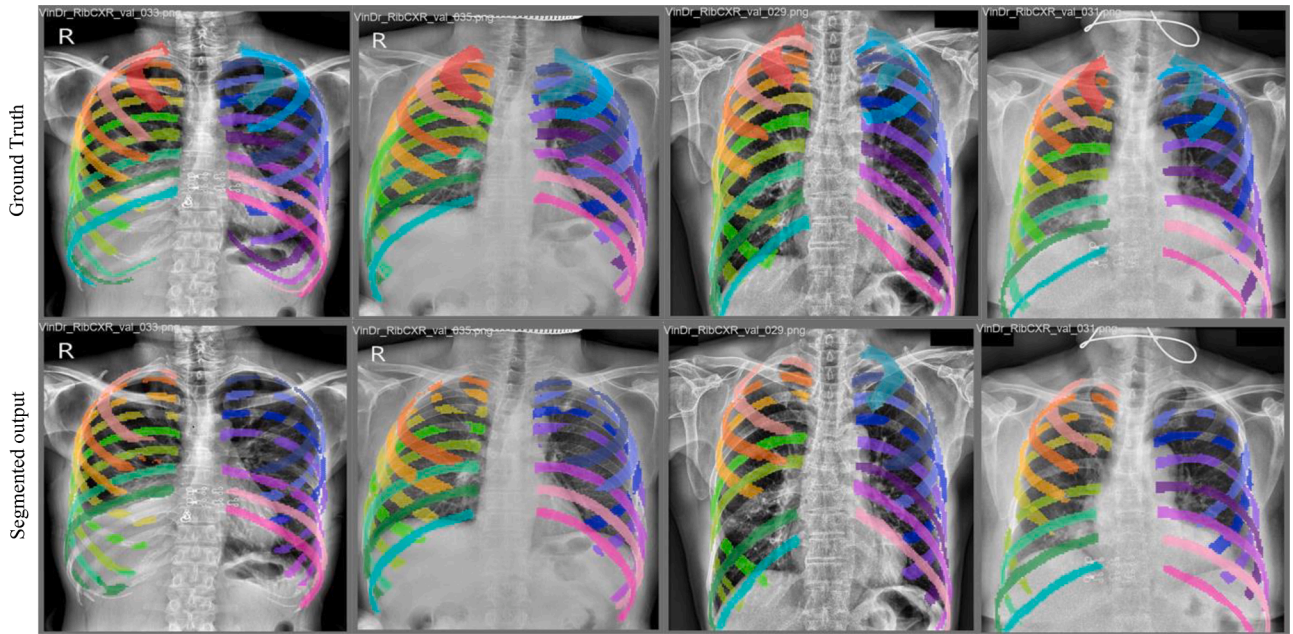


Fig. 7. Ground truth annotations of segmented ribs in chest X-ray images, and predicted rib segmentation output from the YOLOv8m-seg model.

Table 3

Ablation study. Results in the VinDr-CXR test dataset. M1: Mamba-YOLOvX; M2: Mamba-YOLOvX + CSAD; M3: Mamba-YOLOvX + CSAD + Image Augmentation. Best results are in bold.

Model	Precision	AP ₅₀
YolovX	0.267	0.160
Retinanet-R-50-FPN-3x	0.209	0.147
Faster-Rcnn-R-50-FPN-3x	0.235	0.142
M1	0.321	0.163
M2	0.303	0.171
M3	0.301	0.174

3.2. Ablation study

The ablation study was conducted on the VinDr-CXR test set to evaluate the influence of the attention mechanism and data augmentation on detection performance by comparing models trained with and without these modules. YOLOvX model was considered as the baseline. Model M1 was built by integrating Mamba and YOLOvX. M2 model was built using M1 and adding CSAD. Finally, M3 consisted of M2 evaluated applying the image augmentation technique proposed in this study. Table 3 summarizes the achieved results.

AP₅₀ measures how accurately the model localizes lesions with sufficient overlap. As shown, the AP₅₀ metric for M1, M2 and M3 models improved significantly this metric for the baseline model. This suggests that Mamba-based models outperformed YOLOvX in the precision of their detections, and underscores the capability of Mamba in managing the state space modelling component. Moreover, large-scale models such as RetinaNet and Faster R-CNN as popularized by the Detectron2 library available in Wu, Kirillov, Massa, Lo, and Girshick (2019), despite their scale, demonstrated lower precision and AP₅₀ scores compared to YOLOvX-based methods, as evidenced in Table 3.

Differences in anatomy and lung-related conditions can create multiple sources of imbalance in CXR datasets. The Mamba-YOLOvX model addressed this challenge by effectively recognizing objects of varying sizes. Input patches are scanned along three distinct paths, each processed independently through separate blocks. While the individual sequences in each branch alone do not yield significant performance gains,

the integration of the CSAD module enables a more thorough extraction of relevant features. This enhancement led to an increase of 0.79% in AP₅₀, resulting from modifications made to M1 to create M2. The AP₅₀ value for M3 suggests it excels in accurately localizing and segmenting lesions.

As shown in Fig. 8, the CSAD module (M2) enhanced the ability to capture fine-grained features. This attention block enhances feature representation by combining channel and spatial attention, leveraging both global and local information. Unlike self-attention, which has quadratic complexity and only captures global relationships, it remains computationally efficient using pooling operations and a lightweight MLP for channel attention, and dilated convolutions to capture fine-grained and long-range dependencies. The dilated convolution expands the receptive field without significant computational cost, benefiting the object detection task. By adaptively weighting feature maps, it reduces noise, improves focus on discriminative regions, and enhances generalization and robustness. This hybrid approach achieved a balanced trade-off between performance and efficiency, making it ideal for real-time and resource-constrained applications.

Fig. 9 illustrates how objects of varying sizes were effectively recognized across different heads. This supports the conclusion that the heads generate independent outputs in the heatmap, highlighting the effectiveness of the proposed technique. Notably, the small head output excelled at detecting most small lesions. Mamba-based models are more conservative than YOLOvX models in generating predictions. The progressive improvements in precision introduced in M1, M2, and M3 derive from the inclusion of mechanisms such as attention mechanisms and architectural changes (Mamba integration) what reduced false positive predictions, as detailed in Table 3.

That is, the models M1, M2 and M3 were more selective in their predictions. The fact that M1 has the highest precision suggests that the improvements introduced in this stage effectively reduced false positives without yet introducing the trade-offs in M2 and M3. M2 and M3 slightly reduced precision as these models aimed to generalize better and capture a broader range of lesions.

Fig. 10 illustrates the AP₅₀ values for various abnormalities across different models. High identification rates were achieved for classes like cardiomegaly, likely due to the distinctive features of these abnormalities, that facilitate more accurate detection. Notably, the most significant improvements over the baseline were observed in detect-

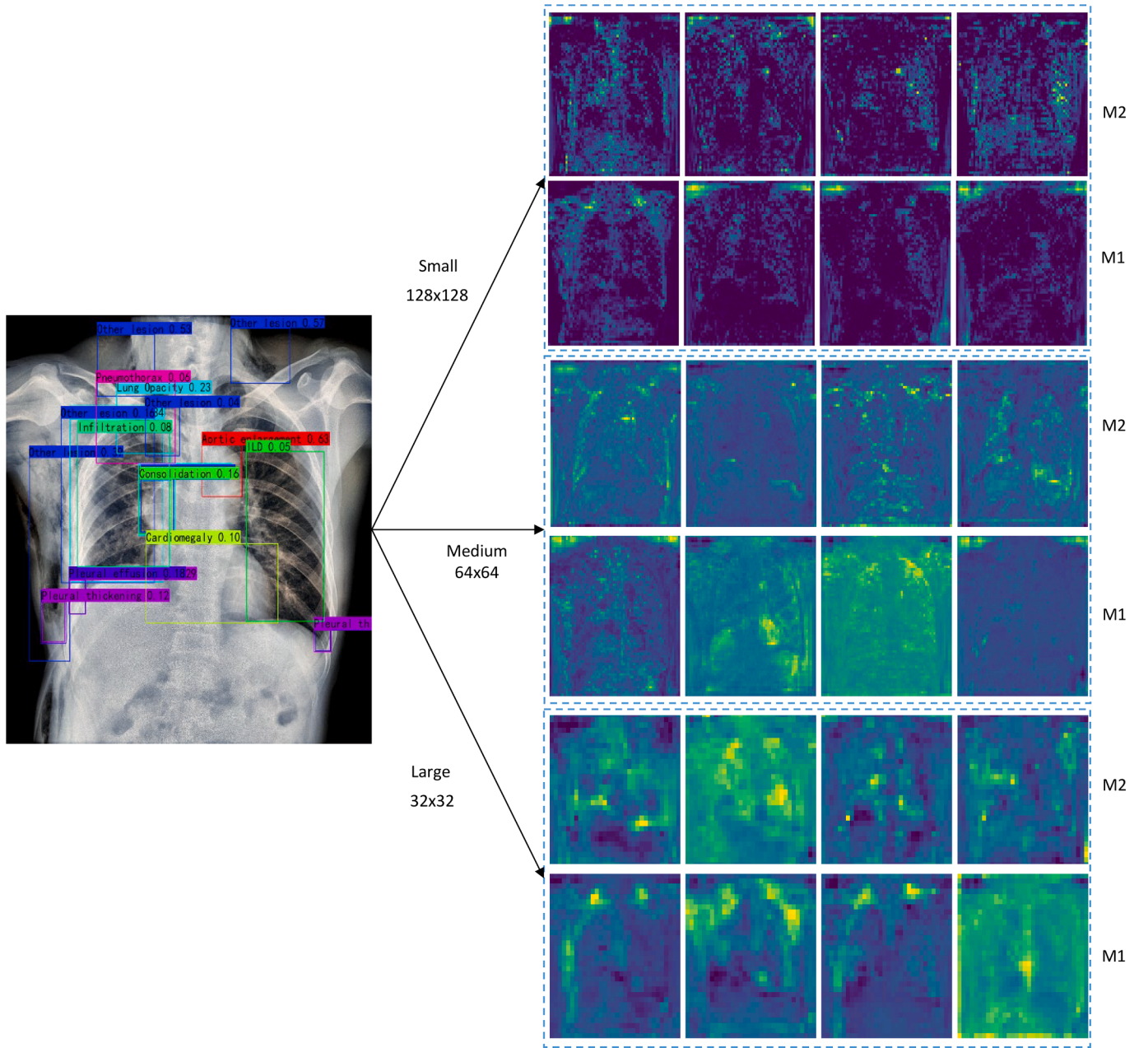


Fig. 8. Comparison of feature maps between M1 and M2 models for different size of heads, with images uniformly resized to ensure precise side-by-side analysis. M1: Mamba-YOLOvX; M2: Mamba-YOLOvX + CSAD.

ing small lesions, such as aortic enlargement, consolidation, infiltration, nodule/mass and pleural thickening.

For the task of segmentation of the lesion in chest radiography, the spatial quality of detections (AP_{50}) and the precision are crucial. Therefore, as a result of the ablation study, the Mamba-YOLOvX enhanced with CSAD was selected within the scope of this study to be evaluated using the proposed image augmentation technique. Its best performance in AP_{50} metric reflected its ability to localize accurately and segment lesions in CXR. Additionally, it presented a good balance across the general precision metric, prioritizing detection accuracy without significantly compromising the ability to identify lesions in minority classes.

To evaluate the performance of our model on a large-scale dataset, we utilized AL14 dataset consisting of 119,600 bounding box annotations, split evenly into training and testing sets (50-50). The per-class performance metrics are illustrated in Fig. 11, showing the mAP, Recall, Precision, and F1 Score across 14 diagnostic categories. The average estimated value of AP_{50} was 61.47%. Classes such as 'Postoperative metal'

and 'Venipuncture' showed exceptional performance across all metrics, while more challenging categories like 'Atelectasis' and 'Nodule' yielded significantly lower scores, primarily due to variations in position and size.

3.3. Performance

The proposed framework was rigorously evaluated using a comprehensive array of datasets and multiple performance metrics to ensure robustness and generalizability. This subsection details the accuracy assessments conducted at varying IoU thresholds, stratified into two distinct evaluation paradigms: intra-dataset and cross-dataset validation.

3.3.1. Intra-Dataset Validation

The images from each dataset were randomly divided into a 90% training set and a 10% testing set, ensuring that no patients were present in both sets. Tables 4, 5, and 6 display the results calculated for the

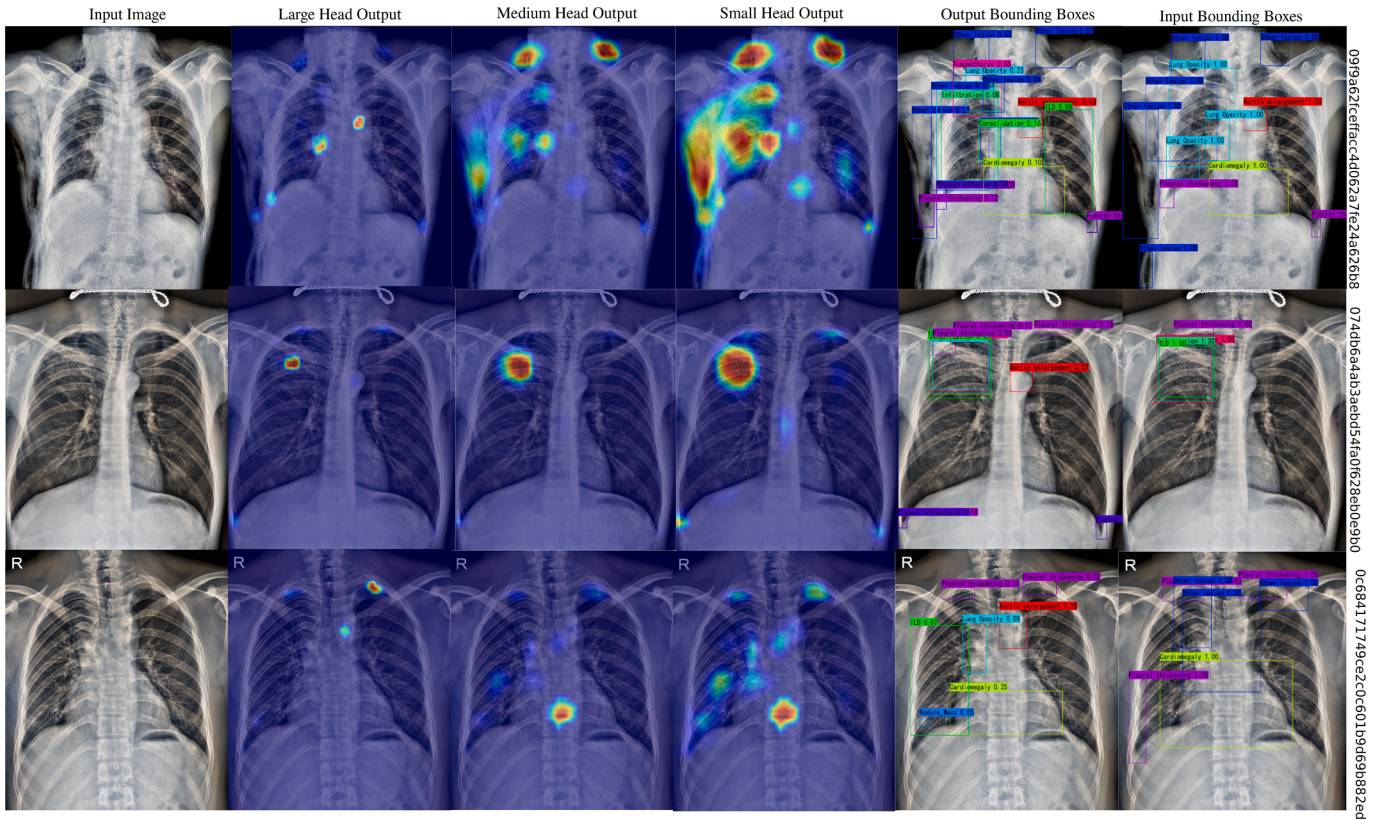


Fig. 9. Heatmap comparison of different heads across M2 model.

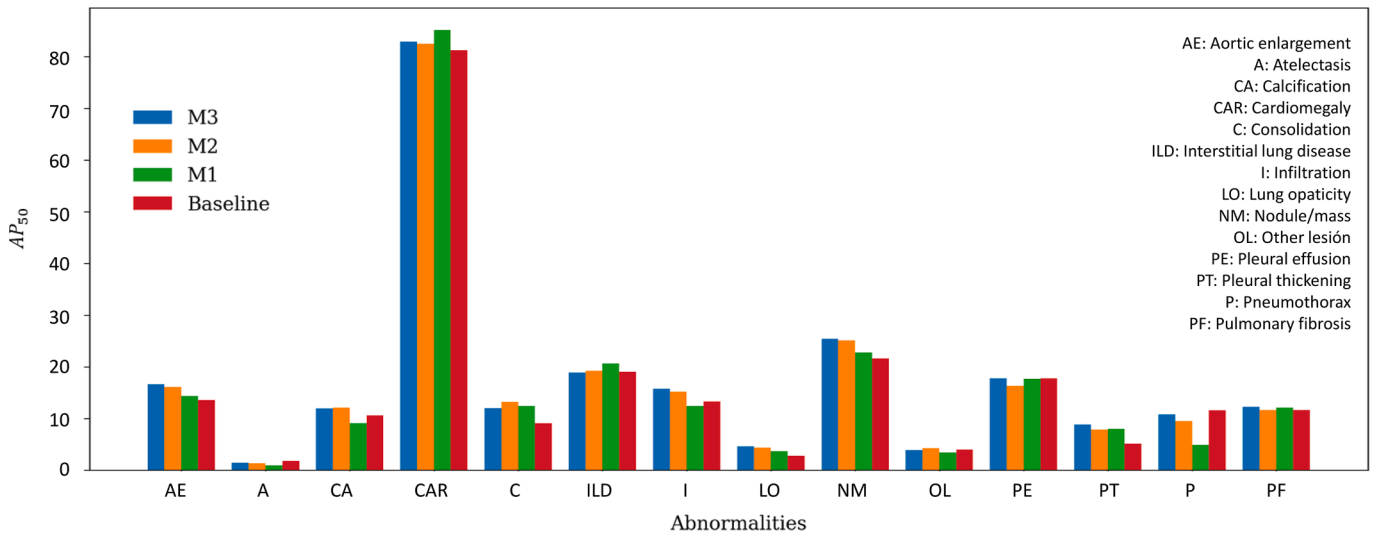


Fig. 10. Ablation study. Comparison of AP_{50} metric across the four proposed models in the VinDr-CXR test dataset. M1: Mamba-YOLOvX; M2: Mamba-YOLOvX + CSAD; M3: Mamba-YOLOvX + CSAD + Image Augmentation.

VinDr CXR (training set), CXR AL14, and ChestX-ray8 datasets at varying IoU thresholds. As shown in the quantitative results presented in Tables 4 and 6, the findings highlight the effectiveness of the proposed framework in accurately localizing abnormalities.

Concerning the VinDr-CXR (training) dataset, the model performed robustly, achieving high accuracy at IoU of 0.5 for labels like aortic enlargement (0.977) and cardiomegaly (0.991), as detailed in Table 4. Performance was lower for underrepresented or visually complex conditions such as atelectasis and pleural thickening, with an accuracy at 0.5 IoU of 0.417 and 0.585, respectively.

The results for the CRX-AL4 dataset, as summarized in Table 5. Misdetction was predominantly driven by class imbalance within the dataset. Findings like pleural thickening and atelectasis were poorly detected. Only a 0.1 % and 3.0 % of instances labelled as atelectasis and pleural thickening are present in the dataset. Pathological findings with distinct imaging characteristics, such as emphysema, achieved perfect detection at lower IoU thresholds. However, the model encountered difficulties differentiating between nodules and masses, which exhibit correlated image patterns and remains a challenge for segmentation models.

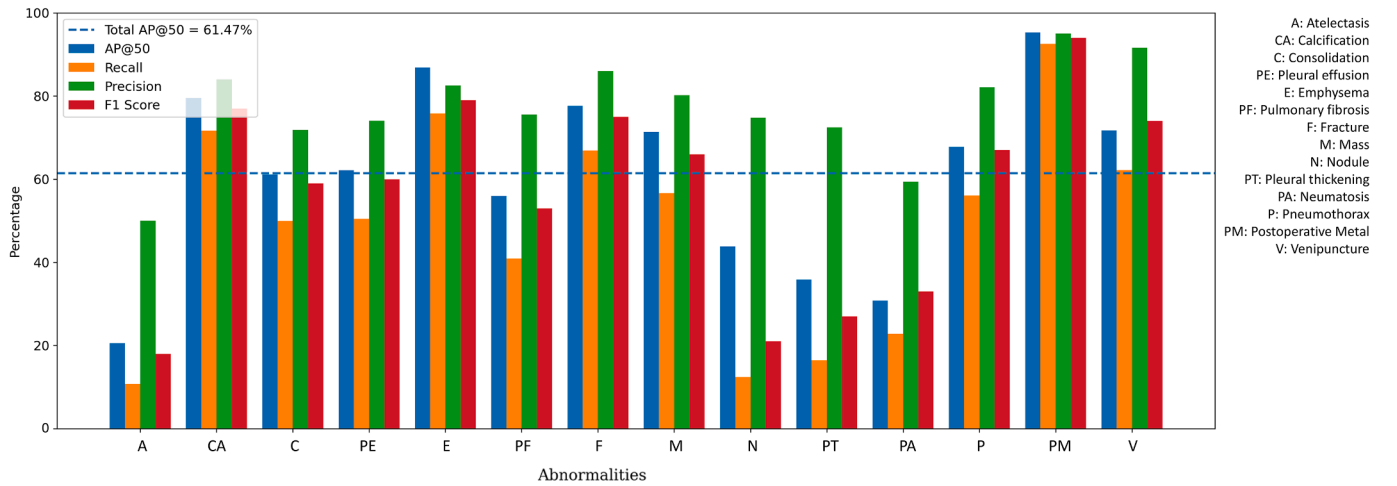


Fig. 11. Performance metrics (AP₅₀, Recall, Precision, and F1 Score) for each class in the large dataset AL14, with a 50-50 % training-validation split.

Table 4

IoU accuracy at a fixed threshold along the different labels in the VinDr-CXR train dataset (90/10 % split).

Label	0.1 IoU	0.2 IoU	0.3 IoU	0.4 IoU	0.5 IoU	0.6 IoU	0.7 IoU
Consolidation	0.937	0.823	0.792	0.750	0.651	0.526	0.369
Infiltration	0.898	0.875	0.802	0.687	0.557	0.369	0.200
Nodule Mass	0.819	0.798	0.774	0.745	0.703	0.591	0.456
Lung Opacity	0.886	0.870	0.830	0.707	0.571	0.391	0.230
Interstitial Lung Disease	0.966	0.911	0.893	0.804	0.693	0.449	0.304
Aortic Enlargement	0.998	0.998	0.995	0.987	0.977	0.955	0.877
Cardiomegaly	0.996	0.996	0.996	0.996	0.991	0.951	0.902
Pleural Thickening	0.961	0.946	0.859	0.720	0.585	0.369	0.186
Pulmonary Fibrosis	0.901	0.834	0.766	0.655	0.485	0.364	0.212
Other Lesion	0.751	0.619	0.526	0.404	0.320	0.203	0.097
Pleural Effusion	0.982	0.958	0.861	0.783	0.699	0.512	0.289
Calcification	0.821	0.715	0.632	0.549	0.410	0.286	0.148
Atelectasis	0.555	0.528	0.528	0.472	0.417	0.250	0.111
Pneumothorax	0.666	0.667	0.555	0.4074	0.370	0.333	0.333
Average	0.867	0.824	0.772	0.690	0.602	0.468	0.337

Table 5

IoU accuracy at a fixed threshold along the different labels in the CXR-AL14 dataset (90/10 % split).

Label	0.1 IoU	0.2 IoU	0.3 IoU	0.4 IoU	0.5 IoU	0.6 IoU	0.7 IoU
Consolidation	0.620	0.397	0.227	0.115	0.045	0.016	0.007
Nodule	0.106	0.062	0.036	0.026	0.013	0.004	0.001
Mass	0.500	0.379	0.258	0.193	0.129	0.089	0.024
Pleural Thickening	0.372	0.270	0.183	0.108	0.052	0.018	0.006
Pulmonary Fibrosis	0.269	0.205	0.159	0.106	0.062	0.026	0.004
Effusion	0.765	0.649	0.539	0.405	0.254	0.117	0.038
Calcification	0.586	0.538	0.424	0.248	0.096	0.035	0.008
Atelectasis	0.750	0.583	0.417	0.167	0.167	0.167	0.000
Pneumothorax	0.738	0.593	0.427	0.252	0.155	0.076	0.038
Emphysema	0.997	0.996	0.995	0.998	0.942	0.801	0.601
Fracture	0.344	0.327	0.295	0.203	0.099	0.031	0.009
Postoperative Metal	0.838	0.766	0.639	0.472	0.270	0.107	0.025
Pneumotosis	0.472	0.302	0.151	0.067	0.011	0.007	0.003
Venipuncture	0.445	0.350	0.198	0.096	0.026	0.009	0.002
Average	0.557	0.458	0.353	0.247	0.166	0.107	0.055

Finally, in the ChestX-ray8 dataset (Table 6), a high accuracy was observed for findings like cardiomegaly (0.900 at 0.5 IoU) and infiltration (0.625 at 0.5 IoU). Certain findings such as nodule and mass were consistently challenging (e.g., nodules segmentation achieved an IoT of 0.200 at 0.5 IoU).

In general, the proposed model demonstrated strong localization performance in controlled settings, such as the VinDr-CXR dataset. However, performance diminished in datasets with significant class imbalances, such as CXR-AL4 and ChestX-ray8. Conditions with low representation, such as pleural thickening and atelectasis, contributed to higher rates of misdetection across datasets. Conditions with overlapping ra-

diographic appearances, such as nodules and masses, remain a key challenge. The disparity in performance across datasets underscores the importance of evaluating models under diverse settings to ensure generalizability, particularly for clinical applications.

3.3.2. Cross-dataset validation

The model trained on the VinDr-CXR train dataset was evaluated on the ChestX-ray8 dataset. Results are detailed in Table 7. The model achieves high IoU at lower thresholds (0.1–0.3) but significantly drops as the threshold increases. Cardiomegaly detection achieved the highest IoU at lower thresholds, with a perfect score (1.000) at 0.1 IoU and still

Table 6

IoU accuracy at a fixed threshold along the different labels in the ChestX-ray8 dataset (90/10 % split).

Label	0.1 IoU	0.2 IoU	0.3 IoU	0.4 IoU	0.5 IoU	0.6 IoU	0.7 IoU
Infiltration	0.937	0.875	0.875	0.750	0.625	0.562	0.375
Nodule	0.400	0.400	0.400	0.400	0.200	0.200	0.200
Mass	0.444	0.444	0.444	0.444	0.444	0.333	0.222
Cardiomegaly	1.000	1.000	1.000	1.000	0.900	0.700	0.700
Effusion	0.900	0.850	0.750	0.650	0.500	0.400	0.200
Atelectasis	0.619	0.571	0.476	0.381	0.238	0.238	0.047
Pneumonia	0.857	0.857	0.857	0.714	0.571	0.286	0.143
Pneumothorax	0.461	0.461	0.385	0.308	0.231	0.154	0.077
Average	0.702	0.682	0.648	0.581	0.464	0.359	0.246

Table 7

Comparison of IoU accuracy at a fixed threshold on the ChestX-ray8 as validation dataset. The model was trained with the VinDr - CXR train dataset.

Label	0.1 IoU	0.2 IoU	0.3 IoU	0.4 IoU	0.5 IoU	0.6 IoU	0.7 IoU
Infiltration	0.837	0.732	0.545	0.374	0.171	0.090	0.057
Nodule-Mass	0.591	0.537	0.476	0.445	0.305	0.177	0.091
Cardiomegaly	1.000	0.993	0.979	0.630	0.329	0.082	0.000
Effusion	0.699	0.490	0.340	0.163	0.065	0.013	0.006
Atelectasis	0.083	0.039	0.011	0.006	0.006	0.000	0.000
Pneumothorax	0.429	0.316	0.245	0.143	0.092	0.061	0.031
Average	0.607	0.518	0.433	0.294	0.161	0.071	0.031

Table 8

Results of state-of-the-art methods using the ChestX-ray8 dataset for validation.

Study	Model	0.1 IoU	0.2 IoU	0.3 IoU	0.4 IoU	0.5 IoU	0.6 IoU	0.7 IoU
Zhu et al. (2022)	PCAN	0.778	0.574	0.364	0.207	0.103	–	–
Han et al. (2022)	RGT	0.591	0.424	0.283	0.173	0.090	0.041	0.015
Kang, Kim, and Ryu (2024)	ADNet	< 0.6	< 0.5	< 0.3	< 0.2	< 0.1	< 0.05	–
Wang, Huang, Xu, and Huang (2024)	Weakly Supervised	0.476	–	0.248	–	–	–	–
Ours (2025)	Mamba-YOLOvX	0.607	0.518	0.433	0.294	0.161	0.071	0.031

strong at 0.3 (0.979). However, its performance drops sharply beyond 0.4. Atelectasis performed poorly across all thresholds, with an IoU of only 0.083 at 0.1 and reaching 0 at higher thresholds. Effusion and Pneumothorax both showed weak segmentation ability, with IoUs dropping quickly beyond 0.2 IoU. The average IoU decreased from 0.607 at 0.1 to 0.031 at 0.7, highlighting that the model struggled to maintain accurate segmentation at higher precision levels.

The model was trained on the VinDr-CXR dataset and validated on ChestX-ray8, which come from different sources. This means it underwent cross-dataset validation between datasets from different clinical settings. Differences in image acquisition, study quality, clinical protocols, and disease distribution may impact the model's accuracy. Some conditions (e.g., cardiomegaly and pneumonia) retain high IoU scores at lower thresholds, while others (e.g., atelectasis and pneumothorax) show a sharp decline, suggesting challenges in knowledge transfer across datasets. In addition, labels may be inconsistent across datasets due to different expert opinions, labelling protocols, or clinical definitions. Indeed, cross-dataset generalization remains a significant challenge in medical AI, especially for segmentation models in radiology.

3.4. Comparison to state-of-the-art models

In the task of localization of abnormal radiological findings, various studies have utilized distinct validation metrics. To facilitate a fair comparison, these commonly used metrics were adopted. However, direct comparison of results remains challenging due to the diversity in training and validation strategies employed across studies.

Table 8 summarizes the methodologies and results from recent works that presented models trained on different datasets and validated on the complete ChestX-ray8 dataset. In this case, the Mamba-YOLOvX model was trained in the VinDr-CXR training dataset.

The results demonstrated that the Mamba-YOLOvX model delivered noteworthy efficiency results. While it showed a slight shortfall for IoU

> 0.2 compared to the findings of Zhu et al. (2022), the approach presented in this study consistently achieved robust performance across a broad range of IoU thresholds, what reveals the model generalization ability. When compared to the results obtained by Zhu et al. (2022), Han et al. (2022), Wang et al. (2024) and Kang et al. (2024), which rely on weakly-supervised techniques that train on unlabeled pathology locations, the Mamba-YOLOvX with CSD model performed notably well.

Table 9 details the results of other state-of-the-art models that were assessed using an intra-dataset evaluation. As shown, the proposed model performed exceptionally well when compared to very recent approaches both in AR and AP metrics. Estimating the AR_S and AP_S metrics for the ChestX-ray8 dataset was not feasible due to the absence of target sizes smaller than 32 pixels according to COCO standards. The proposed model improved the results presented by Xu and Duan (2024), that incorporated attention blocks within RetinaNet to focus the network on relevant areas, reducing distractions from irrelevant detections.

Recently, Fan et al. (2024) employed the extensive CXR-AL14 dataset, achieving a remarkable average precision ranging from 0.572 % to 0.631 %. This performance exceeds that of the presented Mamba-YOLOvX framework. However, the Mamba-YOLOvX model could only be trained and validated on a smaller subset of the whole dataset, since the validation set was not publicly available. Despite this limitation, the proposed model attained state-of-the-art results that compete very well with existing benchmarks. In comparison to the YOLO-CXR model presented in Hao et al. (2024), which reports a mAP of 0.167 and AP_{50} of 0.338, our proposed network achieved a mAP of 0.175 and an AP_{50} of 0.366 (Table 10).

In the context of abnormality detection in radiography, where the primary challenge lies in offline processing, memory efficiency is at least as critical as computing time. The proposed model, with 10.8 million parameters, reduced memory occupation while maintaining competitive computational performance (32.2 GFLOPS) when compared to models described in Kang et al. (2024) and Zhu et al. (2022). In addition, the

Table 9

Results of state-of-the-art methods evaluated using intra-dataset validation.

Study	Training/Validation	Model	mAP	AP ₅₀	AP ₇₅	AP _S	AP _M	AP _L	AR _S	AR _M	AR _L	AUC(%)
Le et al. (2023)	VinDr-CXR-70/30 %	YOLOv5-S		0.158								
Lin et al. (2023)	VinDr-CXR-70/30 %	SAR-CNN		0.157	0.132							
Ours (2025)	VinDr-CXR-70/30 %	Mamba-YOLOvX	0.163	0.355	0.134	0.016	0.071	0.186	0.034	0.134	0.284	0.869
Xu and Duan (2024)	VinDr-CXR-90/10 %	DualAttNet	0.116	0.241	0.114	0.005	0.098	0.121	0.027	0.177	0.257	
Hao et al. (2024)	VinDr-CXR-90/10 %	YOLO-CXR	0.167	0.338								
Ours (2025)	VinDr-CXR-90/10 %	Mamba-YOLOvX	0.175	0.366	0.158	0.024	0.103	0.203	0.051	0.174	0.302	0.884
Xu and Duan (2024)	ChestX-ray8-90/10 %	DualAttNet	0.071	0.145	0.076		0.026	0.076		0.056	0.161	
Ours (2025)	ChestX-ray8-90/10 %	Mamba-YOLOvX	0.086	0.153	0.094		0.037	0.086		0.042	0.169	0.850

Table 10

The number of parameters of the Mamba-YOLOvX model in a million (M) and FLOPS in billion (G).

Study	Model	Parameters (M)	FLOPS (G)
Zhu et al. (2022)	PCAN	11.2	27.5
Kang et al. (2024)	ADNet	39	9.7
Xu and Duan (2024)	DualAttNet	-	53.7
Wu et al. (2019)	Faster - Rcn - R - 50 - FPN - 3x	41.3	140.2
Wu et al. (2019)	Retinanet - R - 50 - FPN - 3x	37.9	98.5
Ours	Mamba-YOLOvX	10.8	32.2

presented approach achieved a favorable trade-off by lowering both parameter count and computational demand in comparison to the model presented in Xu and Duan (2024), which required 53.7 GFLOPS.

Despite the promising results, this study present some limitations that must be considered. The variety of validation metrics employed in different studies makes direct comparisons of results difficult. Although the study attempted to standardize the metrics, variations in training and validation strategies may still influence the conclusions. In addition, the Mamba-YOLOvX model with CSD, like many deep learning models, relies heavily on the quality and accuracy of the training data annotations. Errors or inconsistencies in the annotations can affect the model performance and its ability to generalize to new data.

To address these limitations and strengthen our approach, our future work will focus on integrating hybrid methods, particularly promptable models that leverage expert decision-making. This approach can enable domain experts to guide the model's predictions, enhancing interpretability and robustness for real-world applications. Additionally, we recognize the need to extend our model's applicability to abnormalities with limited or no publicly available datasets, such as disease specific. Since such conditions often lack large-scale labeled datasets, our future research will explore semi-supervised and self-supervised learning approaches to train the model with minimal annotated data. We also plan to develop synthetic data augmentation techniques to simulate real-world variations, ensuring better generalization to rare medical conditions. In addition, we plan to explore comprehensive clinical validation in collaboration with medical experts to validate the model in a real clinical environment. By pursuing these directions, we aim to make our model not only more robust but also more practical for deployment in real-world medical and industrial scenarios.

4. Conclusion

This study introduced an improved Mamba-YOLOvX-based model, which integrates attention blocks with dilated CNN, for anomaly detection in chest radiographs. The main objective was to address the challenges associated with multiple datasets, including class imbalance, scale variation, annotation variability, and intensity distribution imbalance.

A projection-based data augmentation strategy using anatomical features has been proposed. This method involves rib segmentation, key point matching, and contrast enhancement to generate anatomically consistent projections and magnifications. Several models were evaluated using publicly available chest X-ray datasets, which varied in

size, diversity, labelling practices, and types of abnormalities included. The results show that the application of the Mamba-YOLOvX model enhanced the performance metrics, showing improvements in robustness.

The ablation study demonstrated that the key components of the proposed model, including the attention mechanism and augmentation techniques, significantly improved the detection performance compared to the YOLOvX-based reference model. As a consequence, the proposed framework achieved promising results in intra-and cross-dataset validation, outperforming state-of-the-art models in most cases. The next steps include conducting prospective clinical studies to evaluate its efficacy in clinical practice and its impact on medical decision-making.

Data availability

Data will be made available on request.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

CRediT authorship contribution statement

Ebrahim Khalili: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Writing - original draft, Writing - review & editing; **Daniel Sanchez-Morillo:** Conceptualization, Methodology, Investigation, Validation, Visualization, Writing - original draft, Writing - review & editing, Project administration, Funding acquisition, Resources, Supervision; **Blanca Priego-Torres:** Formal analysis, Investigation, Validation, Visualization, Writing - review & editing; **Antonio León-Jiménez:** Validation, Writing - review & editing.

Acknowledgements

This work was funded in the "Convocatoria 2021 de Ayudas a Proyectos de Excelencia, en régimen de concurrencia competitiva, destinadas a entidades calificadas como Agentes del Sistema Andaluz del Conocimiento, en el ámbito del Plan Andaluz de Investigación, Desarrollo e Innovación (PAIDI 2020). Consejería de Universidad, Investigación e Innovación de la Junta de Andalucía," under grant number ProyExcel_00942; and by grant PID2021-126810OB-I00 funded by MICIU/AEI/10.13039/501100011033 and by ERDF/EU.

References

- Balabanova, Y., Coker, R., Fedorin, I., Zakharova, S., Plavinskij, S., Krukov, N., Atun, R., & Drobniowski, F. (2005). Variability in interpretation of chest radiographs among Russian clinicians and implications for screening programmes: Observational study. *BMJ (Clinical research ed.)*, 331(7513), 379–382.
- Becker, H. C., Nettleton, W. J., Meyers, P. H., Sweeney, J. W., & Nice, C. M. (1964). Digital computer determination of a medical diagnostic index directly from chest X-ray images. *IEEE Transactions on Biomedical Engineering*, (3), 67–72.
- Bochkovskiy, A., Wang, C.-Y., & Liao, H.-Y. M. (2020). YOLOv4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934*.

- Brady, A. P. (2017). Error and discrepancy in radiology: Inevitable or avoidable? *Insights into imaging*, 8, 171–182.
- Bruls, R., & Kwee, R. M. (2020). Workload for radiologists during on-call hours: Dramatic increase in the past 15 years. *Insights into Imaging*, 11, 1–7.
- Çalli, W., Sogancioglu, E., van Ginneken, B., van Leeuwen, K. G., & Murphy, K. (2021). Deep learning for chest X-ray analysis: A survey. *Medical Image Analysis*, 72, 102125.
- Chen, B., Zhang, Z., Lin, J., Chen, Y., & Lu, G. (2020). Two-stream collaborative network for multi-label chest X-ray image classification with lung segmentation. *Pattern Recognition Letters*, 135, 221–227.
- Cubuk, E. D., Zoph, B., Mane, D., Vasudevan, V., & Le, Q. V. (2018). AutoAugment: Learning augmentation policies from data. *arXiv preprint arXiv:1805.09501*.
- Everingham, M., Gool, L. V., Williams, C. K. I., Winn, J., & Zisserman, A. (2010). The PASCAL visual object classes (VOC) challenge. *Int J Comput Vis*, 88, 303–338.
- Fan, Y., Yang, Y., Qi, J., Zhang, Q., Liao, C., Wen, L., Wang, S., Wang, G., Xia, Y., Wu, Q. et al. (2024). A deep-learning-based framework for identifying and localizing multiple abnormalities and assessing cardiomegaly in chest X-ray. *Nature Communications*, 15(1), 1347.
- Fanni, S. C., Marcucci, A., Volpi, F., Valentino, S., Neri, E., & Romei, C. (2023). Artificial intelligence-based software with CE mark for chest X-ray interpretation: Opportunities and challenges. *Diagnostics*, 13(12), 2020.
- Fernández, A., García, S., & Herrera, F. (2011). Addressing the classification with imbalanced data: open problems and new challenges on class distribution. In *Hybrid artificial intelligent systems: 6th international conference, HAIS 2011, Wroclaw, Poland, May 23–25, 2011, proceedings, Part i* 6 (pp. 1–10). Springer.
- Ge, Z., Liu, S., Wang, F., Li, Z., & Sun, J. (2021). YOLOX: Exceeding YOLO series in 2021. *arXiv preprint arXiv:2107.08430*.
- Gu, A., & Dao, T. (2023). Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*.
- Han, Y., Holste, G., Ding, Y., Tewfik, A., Peng, Y., & Wang, Z. (2022). Radiomics-guided global-local transformer for weakly supervised pathology localization in chest X-rays. *IEEE Transactions on Medical Imaging*, 42(3), 750–761.
- Hao, S., Li, X., Peng, W., Fan, Z., Ji, Z., & Ganchev, I. (2024). YOLO-CXR: A novel detection network for locating multiple small lesions in chest X-ray images. *IEEE Access*, 12, 156003–156019.
- Heber, M. et al. (1995). Recent progress in development of computer-aided diagnostic (CAD) schemes in radiology. *Medical Imaging Technology*, 13(6), 822.
- Jocher, G., Chaurasia, A., & Qiu, J. (2023). YOLO by Ultralytics. Accessed: 2025-04-08 <https://github.com/ultralytics/ultralytics>.
- Kallianos, K., Mongan, J., Antani, S., Henry, T., Taylor, A., Abuya, J., & Kohli, M. (2019). How far have we come? Artificial intelligence for chest radiograph interpretation. *Clinical Radiology*, 74(5), 338–345.
- Kang, H., Kim, N., & Ryu, J. (2024). Attentional decoder networks for chest X-ray image recognition on high-resolution features. *Computer Methods and Programs in Biomedicine*, 251, 108198.
- Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). ImageNet classification with deep convolutional neural networks. *Commun. ACM*, 60 (6), 84–90.
- Le, K. H., Tran, T. V., Pham, H. H., Nguyen, H. T., Le, T. T., & Nguyen, H. Q. (2023). Learning from multiple expert annotators for enhancing anomaly detection in medical image analysis. *IEEE Access*, 11, 14105–14114.
- Lee, C. S., Nagy, P. G., Weaver, S. J., & Newman-Toker, D. E. (2013). Cognitive and system factors contributing to diagnostic errors in radiology. *American Journal of Roentgenology*, 201(3), 611–617.
- Li, Y., & Xiao, J. (2024). DS MYOLO: A reliable object detector based on SSMS for driving scenarios. *arXiv preprint arXiv:2409.01093*.
- Lin, C., Huang, Y., Wang, W., Feng, S., & Huang, M. (2023). Lesion detection of chest X-ray based on scalable attention residual CNN. *Mathematical Biosciences and Engineering: MBE*, 20(20), 1730–1749.
- Lindenberger, P., Sarlin, P.-E., & Pollefeys, M. (2023). LightGlue: Local feature matching at light speed. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 17627–17638).
- Liu, J., Zhao, G., Fei, Y., Zhang, M., Wang, Y., & Yu, Y. (2019). Align, attend and locate: Chest X-ray diagnosis via contrast induced attention network with limited supervision. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 10632–10641).
- Liu, Y., Tian, Y., Zhao, Y., Yu, H., Xie, L., Wang, Y., Ye, Q., & Liu, Y. (2024). VMamba: Visual state space model 2024. *arXiv preprint arXiv:2401.10166*.
- Ma, H., Lei, S., Celik, T., & Li, H.-C. (2024). FER-YOLO-Mamba: Facial expression detection and classification based on selective state space. *arXiv preprint arXiv:2405.01828*.
- Meedeniya, D., Kumarasinghe, H., Kolonne, S., Fernando, C., De la Torre Díez, I., & Marques, G. (2022). Chest X-ray analysis empowered with deep learning: A systematic review. *Applied Soft Computing*, 126, 109319.
- Nguyen, H. C., Le, T. T., Pham, H. H., & Nguyen, H. Q. (2021). VinDr-RibCXR: A benchmark dataset for automatic segmentation and labeling of individual ribs on chest X-rays. *arXiv preprint arXiv:2107.01327*.
- Nguyen, H. Q., Lam, K., Le, L. T., Pham, H. H., Tran, D. Q., Nguyen, D. B., Le, D. D., Pham, C. M., Tong, H. T. T., Dinh, D. H. et al. (2022). VinDr-CXR: An open dataset of chest X-rays with radiologist's annotations. *Sci Data*, 9(1), 429.
- Oakden-Rayner, L. (2019). The rebirth of CAD: How is modern AI different from the CAD we know?
- Oksuz, K., Cam, B. C., Kalkan, S., & Akbas, E. (2020). Imbalance problems in object detection: A review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(10), 3388–3415.
- Pham, H. H., Le, T. T., Tran, D. Q., Ngo, D. T., & Nguyen, H. Q. (2021). Interpreting chest X-rays via CNNs that exploit hierarchical disease dependencies and uncertainty labels. *Neurocomputing*, 437, 186–194.
- Quekel, L. G., Kessels, A. G., Goel, R., & van Engelshoven, J. M. (2001). Detection of lung cancer on the chest radiograph: A study on observer performance. *European Journal of Radiology*, 39(2), 111–116.
- Radiation, U. N. S. C. o. t. E. o. A. (2010). Sources and effects of ionizing radiation, united nations scientific committee on the effects of atomic radiation (UNSCEAR) 2008 report, volume i. United Nations.
- Rezatoghhi, H., Tsoi, N., Gwak, J., Sadeghian, A., Reid, I., & Savarese, S. (2019). Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 658–666).
- Solovyyev, R., Wang, W., & Gabruseva, T. (2021). Weighted boxes fusion: Ensembling boxes from different object detection models. *Image and Vision Computing*, 107, 104117.
- Toriwaki, J.-I., Suenaga, Y., Negoro, T., & Fukumura, T. (1973). Pattern recognition of chest X-ray images. *Computer Graphics and Image Processing*, 2(3–4), 252–271.
- Wang, T., Huang, K., Xu, M., & Huang, J. (2024). Weakly supervised chest X-ray abnormality localization with non-linear modulation and foreground control. *Scientific Reports*, 14(1), 29181.
- 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. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2097–2106).
- Wu, Y., Kirillov, A., Massa, F., Lo, W.-Y., & Girshick, R. (2019). Detectron2. Accessed: 2025-04-08 <https://github.com/facebookresearch/detectron2>.
- Xu, Q., & Duan, W. (2024). DualAttNet: Synergistic fusion of image-level and fine-grained disease attention for multi-label lesion detection in chest X-rays. *Computers in Biology and Medicine*, 168, 107742.
- Zhu, X., Pang, S., Zhang, X., Huang, J., Zhao, L., Tang, K., & Feng, Q. (2022). PCAN: Pixel-wise classification and attention network for thoracic disease classification and weakly supervised localization. *Computerized Medical Imaging and Graphics*, 102, 102137.
- Zuiderveld, K. (1994). Contrast limited adaptive histogram equalization. In P. S. Heckbert (Ed.), *Graphics gems IV* (pp. 474–485). Academic Press Professional, Inc.
- Zunaed, M., Haque, M. A., & Hasan, T. (2024). Learning to generalize towards unseen domains via a content-aware style invariant model for disease detection from chest X-rays. *IEEE Journal of Biomedical and Health Informatics*, 28(6), 3626–3636.