Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

NEP-TB-1.1K: A finding-level benchmark and comparative evaluation of modern deep learning detectors for tuberculosis-related radiographic findings on chest X-ray

Anurag Timilsina
Kathmandu University
Nigam Niraula
nn32021320@student.ku.edu.np
Kathmandu University
Paribartan Timalsina
Kathmandu University
Shubekshya Dhakal
Nepal Medical College Teaching Hospital
Sagar Wagle
Stanford Medicine

Article

Keywords:

Posted Date: November 24th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11404289/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

NEP-TB-1.1K: A finding-level benchmark and comparative evaluation of modern deep learning detectors for tuberculosis-related radiographic findings on chest X-ray

Anurag Timilsina[1,+], **Nigam Niraula**[1,+,*], **Paribartan Timalsina**[1,+], **Shubekshya Dhakal**[2], and **Sagar Wagle**[3, 4]

[1]Department of Computer Science and Engineering, Kathmandu University, Dhulikhel, Nepal
[2]Nepal Medical College and Teaching Hospital, Jorpati, Nepal
[3]Radiology Associates of Wausau, Wausau, WI, USA
[4]Department of Radiology, Stanford University School of Medicine, Stanford, CA, USA
[*]nn32021320@student.ku.edu.np, nigam21nir@gmail.com
[+]These authors contributed equally to this work.

ABSTRACT

Chest radiography (CXR) is considered a cost-effective method for the detection of pulmonary tuberculosis (TB). However, current computer-aided diagnosis (CAD) systems focus primarily on image-level classification, providing limited information on the specific radiological abnormalities underlying their predictions. Here we introduce an expert-curated benchmark, NEP-TB-1.1K, that consists of 1,034 frontal CXRs with 1,735 bounding boxes for two cardinal radiological manifestations, airspace opacity (AO) and pleural effusion (PE). These findings represent radiological abnormalities that may occur in pulmonary TB and overlap with other pulmonary conditions, making their reliable localisation relevant to interpretable chest radiography analysis. We systematically evaluate four representative modern object detection models on this benchmark. These include the attention-centric YOLOv12, the transformer-based Real-Time DEtection TRansformer (RT-DETR), the two-stage Faster R-CNN, and the lightweight SSD-MobileNet. Because diffuse infectious exudates lack sharp radiometric margins, model localization is assessed using both standard (intersection-over-union [IoU] = 0.5) and clinically tolerant (IoU = 0.3) spatial thresholds. RT-DETR ranked first with a mean Average Precision (mAP) of 0.578 (AO: 0.606, PE: 0.550) at IoU = 0.5, and was followed by YOLOv12 (0.517) and Faster R-CNN (0.514). Relaxed to IoU = 0.3, YOLOv12 achieved the best overall performance with mAP of 0.662 (AO: 0.688, PE: 0.636), showing a significant ability to capture true pathological regions under the boundary ambiguity. Our results show that finding-level localization bridges the gap between generic CAD detections and actionable finding-level triage support, providing a solid basis for granular, interpretable automated thoracic screening.

Introduction

Tuberculosis (TB), an airborne infection caused by *Mycobacterium tuberculosis*, remains a major global health challenge, with approximately 10.7 million new cases and 1.23 million deaths in 2024[1]. The burden falls disproportionately on low-resource countries[2], where specialist radiological expertise and diagnostic infrastructure are limited. Timely identification of active pulmonary TB is essential for treatment and limiting transmission, making accessible, scalable screening particularly important.

The most widely used tests for TB infection are the Tuberculin Skin Test (TST) and Interferon-Gamma Release Assays (IGRAs). TST is prone to administration errors and false positives after Bacillus Calmette–Guérin (BCG) vaccination, while IGRAs are more specific but require laboratory infrastructure and cost considerably more[3]. More fundamentally, both measure immune response to exposure and can not confirm active disease[4], so active TB detection relies on imaging screening followed by microbiological confirmation, such as molecular testing or culture[5].

Chest radiography (CXR) is one of the most accessible, rapid, and affordable screening tools for pulmonary disease, and it can help identify patients sooner and limit transmission[6–9]. Interpretation remains difficult, however, because of inter-observer variability and subtle, overlapping radiographic signs. Even experienced radiologists reach only about 68.7% classification accuracy on complex chest radiographs[7,8], largely because the radiographic signs of pulmonary disease are subtle and often overlap.

For this reason, combining Computer-Aided Diagnosis (CAD) systems with modern computer vision has become a promising way to improve radiographic screening workflows. The World Health Organization (WHO) conditionally recommends

computer-aided detection (CAD) software as an alternative to human readers for interpreting digital chest radiographs in TB screening and triage among individuals aged 15 years and older[4], and in 2025 it identified six CAD products that met its performance standards[5]. This recommendation, however, has important limitations. The certainty of evidence supporting it is low, and it concerns systems that output an image-level TB likelihood score without indicating which radiographic abnormality produced it. Such an undifferentiated score is difficult for clinicians to verify against the image, and screen-positive individuals still require confirmatory testing.

Accurately localizing pathological findings can support early clinical intervention, allow patients to be isolated when necessary, and reduce the delays of conventional diagnostic procedures. Clinically, radiologists examine specific regions of the chest radiograph for characteristic findings such as pulmonary opacities, pleural effusions, and cavitary lesions[8]. Recent progress in deep learning and computer vision now makes it possible to detect, localize, annotate, and segment these abnormalities automatically[7]. Such methods could make diagnoses more consistent and deliver clinically actionable information from chest radiographs at scale and low cost.

Existing CAD screening outputs are often undifferentiated, whereas frontline clinicians need structured diagnostic information. To bridge this gap, this study establishes a finding-level localization benchmark for pulmonary tuberculosis on frontal chest radiographs[7,8]. We introduce NEP-TB-1.1K, a clinically curated benchmark that provides expert-verified bounding-box annotations for 1,034 standardized frontal chest radiographs drawn from the TBX11K cohort[7]. The benchmark covers two cardinal radiological manifestations: airspace opacity and pleural effusion. Using it, we systematically evaluate four representative modern detection paradigms: YOLOv12[10], RT-DETR[11], Faster R-CNN[12], SSD-MobileNet[13,14]. Because pulmonary exudates have diffuse radiometric margins rather than sharp borders, we evaluate each model at both a conventional spatial threshold (IoU $= 0.5$) and a clinically tolerant one (IoU $= 0.3$)[15], so that the results better reflect diagnostic localization utility. The models localize opacity and effusion in TB-related images; they detect non-specific radiographic findings and do not distinguish TB from other causes. Pairing finding-level detection with a TB versus non-TB classifier could also support screening-oriented triage.

The remainder of this manuscript is organized as follows. The Related Work section reviews existing image-level and lesion-level deep learning approaches to chest radiograph screening and highlights the limitations of current benchmarks. The Results section compares localization accuracy, precision–recall behavior, and training stability across the four architectures. The Discussion section considers the architectural factors behind model performance, the clinical implications of finding-specific localization, and the limitations of the study, along with directions for future research. Finally, the Methods section describes the clinical curation workflow and split protocol for NEP-TB-1.1K, presents an exploratory analysis of finding distributions, formalizes the network architectures, and specifies the training parameters and the dual-threshold evaluation framework.

Related Work

Most recent work on TB screening from chest radiographs focuses on classification and coarse localization, and it often relies on weakly annotated or image-level labeled datasets rather than fine-grained, lesion-level supervision. Acharya et al.[16], for example, fine-tune ImageNet-pretrained Normalization-Free Networks (NFNets) and use Score-CAM to highlight the regions that drive each prediction. Their method performs strongly in both binary and multiclass settings, reaching 96.91% accuracy, 99.38% AUC, 91.81% sensitivity, and 98.42% specificity on the multiclass tasks[16].

Nafisah and Muhammad[17] similarly proposed a deep learning framework for automated TB screening that segments the lung region before classification. Isolating the pulmonary region of interest from the full radiograph reduces background interference and makes the features passed to the classifier more relevant. They evaluated several convolutional neural network (CNN) architectures across benchmark datasets, and the best configuration reached 99.1% accuracy, showing that models trained on segmented lung regions can outperform those trained on unprocessed radiographs[17].

Unlike these image-level classification approaches, Liu et al.[7] introduced the Tuberculosis X-ray (TBX11K) dataset to support both classification and object detection. TBX11K contains 11,200 high-resolution chest radiographs (3000×3000 pixels) with a clinically realistic distribution: 5,000 healthy cases, 5,000 non-TB abnormal cases, 924 active TB cases, and 212 latent TB cases. The 924 active TB and 212 latent TB images carry bounding boxes marking TB-related abnormalities, which were drawn by professional radiologists and then verified by a senior expert panel. The authors evaluated several detection and classification architectures on this dataset: Single Shot MultiBox Detector (SSD) with a VGGNet-16 backbone, RetinaNet with a ResNet-50 backbone and Feature Pyramid Network (FPN), and Faster R-CNN with a ResNet-50 backbone and FPN.

For classification, transfer learning with pretrained backbones consistently helped the deeper architectures. With pretrained initialization, RetinaNet reached 87.4% accuracy and 91.8% AUC, while Faster R-CNN reached 89.7% accuracy and 93.6% AUC. Without pretraining, their accuracies dropped to 79.0% and 81.3%, respectively. SSD showed the opposite trend: training from scratch gave better results (88.2% accuracy, 93.8% AUC) than the pretrained configuration (84.7% accuracy, 93.0% AUC).

The study also assessed object detection for localizing TB manifestations. On the full dataset split (ALL), Faster R-CNN achieved the best active TB localization, with an average precision AP^{bb}_{50} of 53.3% and an AP^{bb} of 21.9%[7]. On the TB-only

subset with models trained from scratch, RetinaNet reached an AP^{bb}_{50} of 37.8% and an AP^{bb} of 12.7%[7].

Building on these findings, an earlier work examined a single-stage detection framework using YOLOv7 on TBX11K[18]. To handle the imbalance between active and non-active cases, targeted minority-class augmentation and focal loss-based optimization was applied so that training becomes stable. Because it avoids the computational overhead of multi-stage detectors, the YOLOv7-based model improved inference efficiency while keeping detection accuracy competitive. It achieved a mean Average Precision (mAP@0.5) of 0.587 for both active and latent TB cases, demonstrating effective real-time lesion localization[18].

More recently, transformer-based architectures have been explored as a way around the limitations of generic object detectors for computer-aided tuberculosis diagnosis (CTD), by exploiting the bilateral symmetry of chest radiographs. Global self-attention across all spatial locations can be overwhelmed by the many uninformative regions in CXR images. Symmetric Search Attention (SymAttention) was therefore proposed to restrict each spatial position to a learned neighborhood around its bilaterally symmetric counterpart. Symmetric Positional Encoding (SPE) complements it by recalibrating positional information through a learnable spatial transformation, which compensates for imperfect patient positioning. Together, these innovations allow TB lesion localization and CXR image classification to be performed jointly through a two-stage training strategy, and they yield notable improvements over CNN-based baselines such as Faster R-CNN and SSD on the TBX11K benchmark[19].

Despite this substantial progress in CNN-based and transformer-based detection for chest radiograph analysis, existing TB detection benchmarks and the models evaluated on them share a key limitation. The widely used TBX11K dataset[7] provides bounding boxes for TB-related abnormalities but does not distinguish between specific radiological findings. Clinically, however, radiologists interpret chest radiographs by identifying distinct manifestations such as airspace opacity, pleural effusion, and cavitary lesions. Each has different diagnostic implications and can shape subsequent clinical decisions[8]. Without finding-level annotations, current detection models are trained and evaluated on a simplified picture of disease appearance that does not fully match clinical interpretation standards[7,20,21].

In addition, attention-based and transformer-based detectors have performed strongly on general object detection tasks[19], but how different detector families compare on fine-grained, finding-specific TB localization has not been systematically studied in medical imaging. In particular, the effect of modern detection paradigms on structured, clinically meaningful lesion categorization remains insufficiently explored.

To address these gaps, this work introduces NEP-TB-1.1K, a finding-level annotated dataset derived from TBX11K[7] which contains 1,034 TB-positive chest radiographs with radiologist-verified bounding boxes for two clinically relevant findings: airspace opacity and pleural effusion. All annotations were produced under expert clinical supervision to ensure diagnostic consistency.

To establish a comprehensive benchmark on this dataset, we evaluate four representative object detection paradigms. This design allows a systematic comparison across the major modern detection families, rather than assuming that performance transfers from existing TBX11K-based evaluations.

Following established practice in medical object detection evaluation[18,19], we report performance at both a standard (IoU = 0.5) and a relaxed (IoU = 0.3) threshold. Recent findings by Radke et al.[15] further support this choice, suggesting that stricter IoU thresholds may not fully capture clinically relevant localization performance in medical imaging. Lower thresholds, by contrast, can give a more representative assessment of diagnostic utility when lesion boundaries are ambiguous or diffuse.

Overall, the proposed dataset and evaluation framework are intended to support more clinically aligned, fine-grained TB detection research, helping object detection models move beyond coarse lesion localization toward finding-specific, clinically interpretable predictions.

Results

We benchmarked four object-detection architectures across distinct design paradigms: YOLOv12 [10], RT-DETR [11], Faster R-CNN[12], and SSD-MobileNet[13,14]. The localization and classification accuracy were simultaneously measured using Average Precision (AP) and mean Average Precision (mAP) for the detection of airspace opacity and pleural effusion.

Because pulmonary abnormalities manifest with diffuse, irregular margins rather than sharp radiometric borders, performance is reported at both the conventional object-detection threshold (IoU $= 0.5$) and a relaxed threshold (IoU $= 0.3$). As established in medical localization benchmarks[15], this lower threshold credits approximate lesion localization without disproportionately penalizing boundary ambiguity. The quantitative test results are summarized in Table 1, with corresponding precision–recall trajectories detailed in Figure 1. In the following subsections, we examine the comparative detection accuracy across architectures, characterize precision–recall dynamics across operational thresholds, and analyze the training convergence of the leading models.

Detection accuracy across models

At an IoU threshold of 0.5, RT-DETR showed the highest performance with an mAP of 0.578 and was the strongest detector, leading on both classes (Airspace Opacity AP = 0.606, Pleural Effusion AP = 0.550). This result was closely followed by YOLOv12 (mAP = 0.517) and Faster R-CNN (mAP = 0.514). SSD-MobileNet showed the weakest performance, yielding an mAP of 0.410. On the other hand, when the IoU threshold was relaxed to 0.3, YOLOv12 achieved the maximum mAP of 0.662 (Airspace Opacity AP = 0.688, Pleural Effusion AP = 0.636), outperforming RT-DETR (0.633) and Faster R-CNN (0.628). As observed at IoU 0.5, SSD-MobileNet remained the lowest-performing model, with an mAP of 0.610.

When the IoU threshold was relaxed, every model demonstrated improved performance (+0.145 mAP for YOLOv12, +0.119 for RT-DETR, +0.114 for Faster R-CNN, and +0.200 for SSD-MobileNet). This behavior indicates that the models successfully identify the correct pathology regions but frequently predict bounding boxes that loosely encompass reference annotations. Across findings, Airspace Opacity was easier to detect for all four models at both thresholds. One exception was RT-DETR, where the AP for Pleural Effusion was identical (0.550) across both thresholds, indicating that the transformer detector either localized target effusions tightly or missed them entirely, rather than generating borderline detections recovered only by the looser threshold.

	IoU = 0.3			IoU = 0.5		
Model	AO	PE	mAP	AO	PE	mAP
YOLOv12	0.688	0.636	**0.662**	0.553	0.481	0.517
RT-DETR	0.716	0.550	0.633	0.606	0.550	**0.578**
Faster R-CNN	0.647	0.609	0.628	0.560	0.468	0.514
SSD-MobileNet	0.640	0.500	0.610	0.440	0.270	0.410

Table 1. Per-finding average precision (AP) and mean average precision (mAP) on the held-out test partition at IoU acceptance thresholds of 0.3 and 0.5. AO, airspace opacity; PE, pleural effusion. The best mAP at each threshold is shown in bold.

Precision–recall behaviour

The differences in accuracy are reflected in the precision–recall (PR) curves shown in Figure 1. YOLOv12 and RT-DETR sustain high precision across a broad recall range, maintaining precision above 0.80 up to a recall of approximately 0.40–0.50, after which precision declines gradually. Relaxing the threshold to 0.3 extends this high-precision plateau further across higher recall values. Faster R-CNN exhibits precision close to 1.0 at low recall (< 0.30), followed by a sharp decline and a long, low-confidence tail at higher recall. In contrast, SSD-MobileNet showed an early drop in performance on the pleural effusion at IoU 0.5, collapsing below 0.10 precision past 0.40 recall. However, at the relaxed threshold, the performance improved and maintained an overall mAP of 0.6.

Training dynamics of the leading detectors

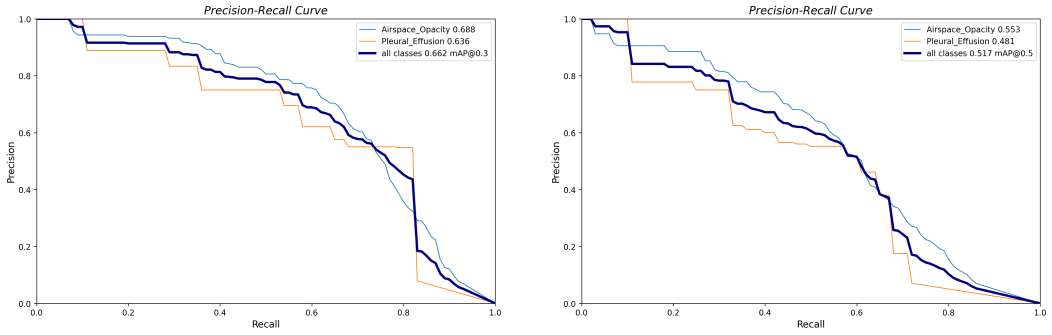
The training and validation loss curves of the leading detectors, YOLOv12 and RT-DETR, are presented in Figure 2. The three native loss curves for both training and validation are shown for each detector.

For YOLOv12 (Figure 2), training box loss, classification loss, and distribution focal loss (DFL) fall steeply across the initial 10 epochs before declining smoothly. On the validation curves, an initial transient spike drops rapidly within the early epochs. After approximately 20 epochs, the classification loss and DFL plateau, whereas the validation box loss reaches its minimum near epochs 20–25 before exhibiting slight upward drift, indicating mild overfitting in the localization branch. Early stopping halted training at optimal validation mAP, and the best-performing checkpoint was retained for test evaluation.

For RT-DETR (Figure 2), training generalized IoU (GIoU), classification, and L_1 losses decreased steadily, aside from a transient between epochs 15–20 coinciding with backbone unfreezing. On the validation curves, GIoU and L_1 losses achieved early minima before stabilizing into a noisy plateau. Validation classification loss rose initially during the first 10 epochs but declined and settled at a lower level for the remainder of optimization. Early stopping ensured retention of the checkpoint with optimal validation accuracy.

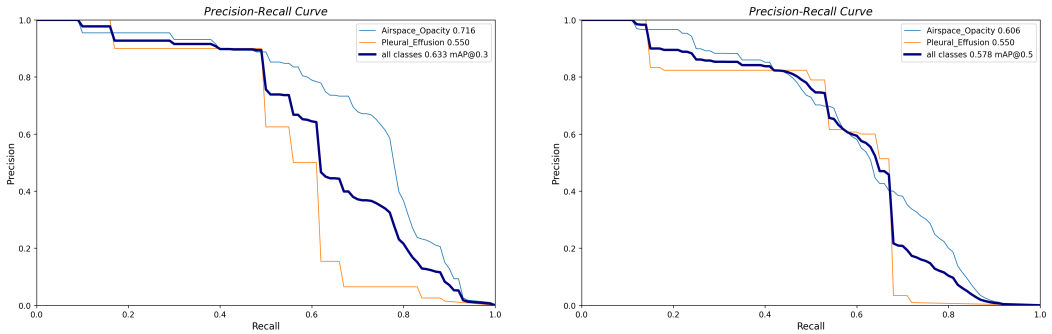
Discussion

In this study, the object detection for computer-aided tuberculosis screening is investigated, moving beyond the traditional coarse-grained disease classification towards the precise localization of individual radiological manifestations [7,8]. The NEP-TB-1.1K dataset is used for experiments, and four object-detection approaches are benchmarked to examine the impacts of architectural design, receptive field characteristics, and loss functions on overall detection performance in the context of diffuse



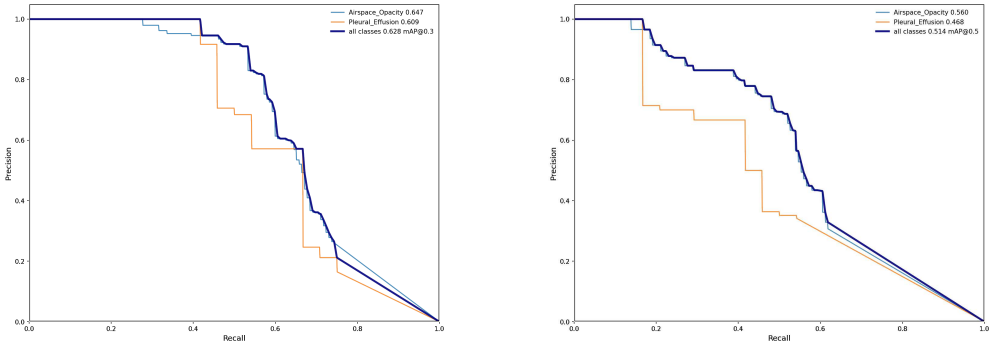
(a) YOLOv12, IoU = 0.3

(b) YOLOv12, IoU = 0.5



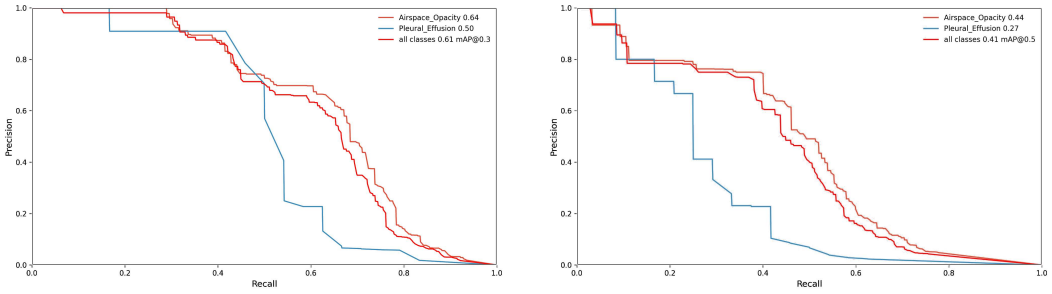
(c) RT-DETR, IoU = 0.3

(d) RT-DETR, IoU = 0.5



(e) Faster R-CNN, IoU = 0.3

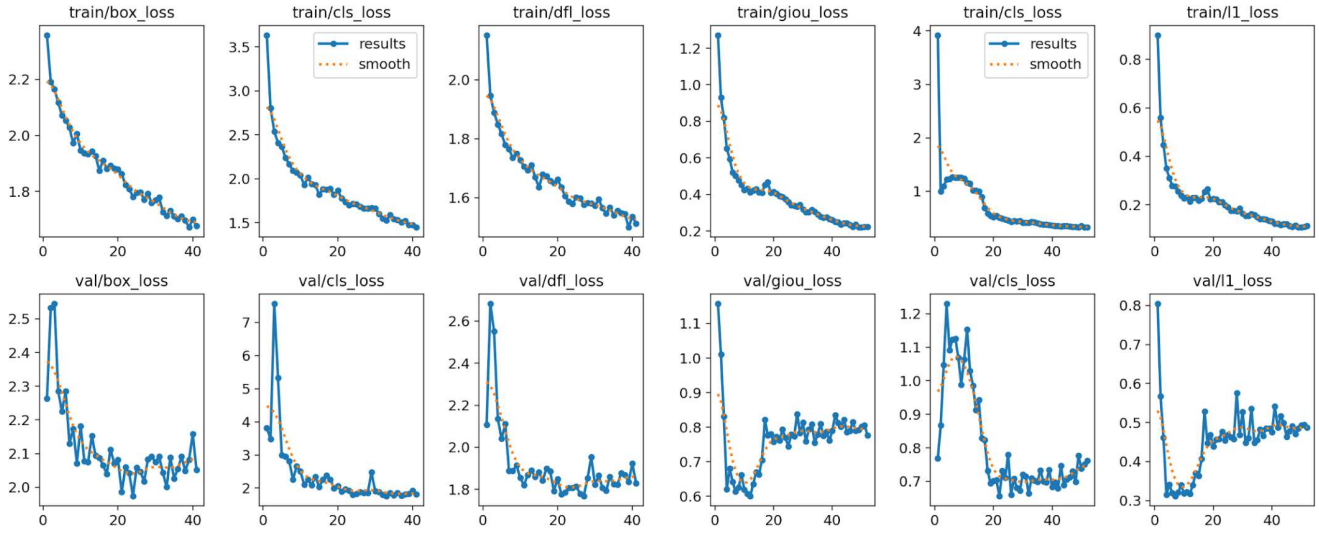
(f) Faster R-CNN, IoU = 0.5



(g) SSD-MobileNet, IoU = 0.3

(h) SSD-MobileNet, IoU = 0.5

Figure 1. Precision–recall curves of the four detectors on the test set at IoU = 0.3 (left column) and IoU = 0.5 (right column): (a,b) YOLOv12, (c,d) RT-DETR, (e,f) Faster R-CNN and (g,h) SSD-MobileNet. Each panel shows the per-class curves for airspace opacity and pleural effusion together with the mean over both classes; the per-class AP and the mAP are given in the legend.



(a) YOLOv12 (b) RT-DETR

Figure 2. Training and validation loss curves of the two leading detectors. (a) YOLOv12: top row, training box (CIoU), classification (binary cross-entropy) and distribution-focal (DFL) losses; bottom row, the corresponding validation losses. (b) RT-DETR over the two-stage schedule: top row, training GIoU, classification (IoU-aware focal) and L_1 losses; bottom row, the corresponding validation losses; the transient rise near epoch 20 coincides with unfreezing of the backbone. In all panels, the solid line shows the per-epoch values and the dotted line a smoothed trend.

and highly imbalanced thoracic abnormalities. The following subsections analyze the architectural factors that contributed to the difference in detection performance, clinical implications of distinguishing airspace opacity from pleural effusion, and limitations of the current experiment along with the potential future directions.

Architectural factors in detection performance

RT-DETR and YOLOv12 emerged as the leading detectors overall, achieving top mAP at IoU = 0.5 (0.578) and IoU = 0.3 (0.662), respectively, with Faster R-CNN closely trailing and SSD-MobileNet consistently weakest (Table 1).

The superior performance of the attention-centric models aligns with the diffuse, ill-defined boundaries characteristic of airspace opacity and pleural effusion, where expanded spatial context aids delineation over purely local receptive fields[19]. In RT-DETR, set-based bipartite matching avoids heuristic Non-Maximum Suppression (NMS), preventing the premature suppression of overlapping or ill-bordered lesions that blend into surrounding parenchyma[22]. This property is reflected in RT-DETR's precision–recall behavior (Figure 1), where pleural effusion AP remains constant (0.550) across both IoU thresholds (0.3 and 0.5), demonstrating that bipartite matching yields discrete, well-localized predictions rather than the borderline, low-overlap proposals typical of dense anchor grids. Similarly, YOLOv12's area attention achieves comparable receptive-field advantages while maintaining high sensitivity at relaxed spatial thresholds with minimal parameter overhead.

Conversely, Faster R-CNN's competitive performance without self-attention indicates that sequential region proposal and RoI-level feature refinement remain robust mechanisms for capturing irregular anatomical margins[7]. Finally, SSD-MobileNet's lagging performance stems primarily from limited backbone capacity and the challenge of adapting general-domain representations to low-prevalence, subtle thoracic targets under a dense single-stage regime[23], rather than the absence of attention mechanisms alone.

Clinical implications of finding-level localization

Transitioning computer-aided analysis of TB-related chest radiographs from undifferentiated, image-level classifications or generic lesion bounding boxes toward finding-specific localization may help narrow a longstanding mismatch between artificial intelligence benchmarks and clinical decision-making. Conventional CAD platforms that aggregate all abnormal manifestations under a uniform "TB-positive" or "abnormal" label treat tuberculosis as a single radiological entity[7,16]. In clinical practice, however, radiologists and frontline physicians do not manage thoracic abnormalities monolithically[8]. Identifying whether a pulmonary manifestation represents an airspace opacity or a pleural effusion fundamentally alters the downstream diagnostic, infection-control, and therapeutic trajectory.

Airspace opacity indicates active parenchymal disease—such as exudative alveolar infiltration, caseous pneumonia, or lobar

consolidation[8]. In high-burden settings, localizing an opacity prompts immediate airborne infection isolation precautions and prioritizes rapid sputum-based microbiological workups, including acid-fast bacilli (AFB) smear microscopy and automated nucleic acid amplification tests (e.g., GeneXpert MTB/RIF)[24,25]. Conversely, pleural effusion reflects fluid accumulation within the pleural space, representing tuberculous pleurisy or reactive parapneumonic effusions[8,26]. Because tuberculous pleural fluid is characteristically paucibacillary, sputum examinations alone frequently yield false negatives in isolated pleural involvement[26,27]. Detecting an effusion directs the clinician toward diagnostic thoracentesis to evaluate pleural fluid biochemistry, differential cell counts, and elevated adenosine deaminase (ADA) activity, or to arrange closed pleural biopsies when microbiology remains inconclusive[27,28]. By disaggregating findings, finding-level detection transforms CAD outputs from a non-specific triage alert into an actionable clinical summary that actively informs procedural pathways.

This diagnostic relevance is reinforced by the spatial distributions identified during our exploratory data analysis (Figure 4). The model predictions and ground-truth annotations reflect classic pathophysiological principles of thoracic tuberculosis[8,29]. Airspace opacities are concentrated bilaterally across the mid-to-upper lung zones, exhibiting a prominent density hotspot in the right upper zone that mirrors the high oxygen tension and impaired lymphatic clearance typical of post-primary reactivation TB[8,29]. In contrast, pleural effusions localized almost exclusively to the dependent left and right costophrenic angles, reflecting the gravitational pooling of pleural exudate on standard erect posteroanterior radiographs (Figure 4). The detectors' ability to preserve this spatial structuring confirms that their learned representations reflect genuine anatomical pathology rather than spurious radiographic artifacts.

Furthermore, evaluating performance across both standard (IoU $= 0.5$) and relaxed (IoU $= 0.3$) thresholds reflects the perceptual realities of chest radiography. Unlike rigid everyday objects evaluated in generic computer vision benchmarks, pulmonary opacities lack crisp radiometric boundaries; inflammatory exudates grade continuously from dense central consolidation into surrounding aerated parenchyma and interstitial stroma[8]. This anatomical continuity generates well-documented inter-observer variability among experienced radiologists when delineating ground-truth bounding box margins[7,8]. Enforcing a rigid IoU ≥ 0.5 criterion disproportionately penalizes models for boundary ambiguity even when the underlying pathological focus has been correctly identified[15].

The notable elevation in performance across all architectures when evaluating at IoU $= 0.3$—demonstrated by YOLOv12 improving from 0.517 to 0.662 mAP, Faster R-CNN reaching 0.628, and SSD-MobileNet recovering by $+0.200$ mAP (Table 1)—proves that these detectors reliably suggest the true pathological focus. In screening and second-reader triage environments, where CAD is utilized to mitigate diagnostic oversight or triage urgent cases in high-throughput settings, directing a healthcare provider's attention to the correct anatomical zone and pathology class provides substantial diagnostic utility without requiring exact coordinate-level boundary alignment.

Limitations and future work

NEP-TB-1.1K benchmark introduces finding-level annotations that help improve overall diagnosis. Notably, several constraints should be considered during interpretation of these results:

- **Evaluation Footprint and Statistical Power:** The NEP-TB-1.1K benchmark contains 1,735 expert-verified lesion instances within 1,034 chest radiographs. For the held-out evaluation, 126 images are separated, having 211 annotated targets (Table 2). Using stratified sampling helped to get a proportional distribution, but the presence of only a few instances of individual classes, like 26 instances of pleural effusion, can increase sensitivity and variability of the resulting estimates. Therefore, expanding the number of samples can increase the reliability of performance estimation.

- **Prevalence Imbalance and Optimization:**In the benchmark, the current underlying clinical prevalence distribution is preserved with a 7.5:1 ratio, resulting in substantial imbalance between the two targets. To alleviate this imbalance, the targeted loss reweighting was incorporated during training, but the limited exposure to pleural effusion patterns during optimization remains a challenge. As a result, the recall of effusion targets can be reduced across multiple detection architectures.

- **Cohort Diversity and Generalization:** During this study, the training and evaluation data were used from the single parent repository. However, in the real clinical screening environment, there is variability in imaging equipment, acquisition protocols, patient populations, and institutional calibration. Incorporating data from various resources with broader demographic, technical, and operational variability can be important to improve the cross-domain robustness in practical clinical settings.

- **Expanding Finding Lexicons:**To overcome the severe sample sparsity during optimization, this benchmark focused on two targets, namely airspace opacity and pleural effusion. These two targets are the most prevalent findings in this cohort. Expanding the target to additional radiological presentations, particularly cavitary tuberculosis, can broaden the clinical scope. The annotation of cavitary lesions was also performed in our experiment, but when it was modeled jointly with

two existing findings, there was a severe class imbalance. Obtaining more samples of cavitary tuberculosis can facilitate more balanced multi-finding detection and reduce the risk of unstable optimization.

This study transforms the focus from coarse, disease-level classification towards the clinically interpretable localization of individual findings. We established the baseline performance across four different object detection models on this restructured formulation. Future research should focus on increasing the dataset size, especially pleural effusion cases, to mitigate the class disparities during optimization. In addition, creating a broader unified benchmark with data from multi-center repositories will help to address clinical, demographic, and hardware heterogeneity. Further investigation into robust architectures, task-specific inductive biases, and optimized training objectives for highly imbalanced and diffuse thoracic pathologies may also contribute to advancing automated tuberculosis detection and clinical intervention.

Methods

To advance automated tuberculosis screening beyond monolithic, binary disease tagging toward clinically actionable radiological interpretation, this study evaluates finding-level object detection across frontal chest radiographs[7,8]. Using the curated NEP-TB-1.1K benchmark, we formulate a multi-class localization framework targeting two principal pathological manifestations: airspace opacity and pleural effusion. In the following subsections, we describe the clinical re-annotation workflow and split protocol, detail the morphological and spatial characteristics revealed through exploratory data analysis, formalize the four benchmarked detector architectures across one-stage, two-stage, and transformer paradigms, and delineate the training configurations, anatomy-preserving data augmentation policies, and dual-threshold evaluation metrics.

As four detectors are different in architecture, capacity and their optimisation behaviour we can't impose a same training paradigm across all models. Therefore, each model was tuned independently in terms of varying augmentation policy, training length, and classification-loss reweighting so that the selected configuration gives the highest validation mAP for each model.

Task formulation and dataset curation

We establish NEP-TB-1.1K as a finding-level benchmark formulated for multi-class object detection on frontal chest radiographs, evaluating localized pathological manifestations alongside verified true-negative background cases. Rather than adopting the coarse, disease-level bounding boxes that collapse distinct entities into a monolithic label, our dataset establishes a finding-specific re-annotation protocol targeting airspace opacity and pleural effusion across 1,034 radiographs curated from the publicly accessible Tuberculosis X-ray (TBX11K) repository[7] (licensed under CC BY 4.0).

The resulting dataset, designated NEP-TB-1.1K, was curated through a multi-stage clinical review workflow. Initial axis-aligned bounding boxes were generated and curated under expert clinical review by two co-authors—an experienced, American Board of Radiology (ABR) certified radiologist and a licensed medical doctor from Nepal. To ensure diagnostic fidelity, visual consistency, and minimize inter-observer subjectivity, all annotated radiographs subsequently underwent independent verification by a panel of clinical practitioners and radiologists from the Kathmandu University School of Medical Sciences (KUSMS). Rare cavitary lesions were excluded to prevent extreme sample sparsity during training and evaluation.

Following curation, the benchmark comprised 1,034 radiographs (maintained at a uniform resolution of 512×512 pixels) containing 1,735 finding-level bounding boxes. To ensure unbiased evaluation and prevent distribution shifts, the dataset was randomly partitioned into training, validation, and test splits following an approximate 70:20:10 ratio (Table 2), while strictly preserving the relative proportions of both finding classes and background cases across all three subsets.

Exploratory data analysis

An exploratory analysis of the curated bounding boxes revealed three key characteristics that directly shape the detection problem across these splits (Figure 3-4).

First, the annotations exhibit severe *class imbalance* (Table 2). Airspace opacity accounts for 1,532 of the 1,735 boxes, whereas pleural effusion accounts for only 203, yielding an imbalance ratio of roughly 7.5:1. This disparity remains consistent across the training (1,059 vs. 141), validation (288 vs. 36), and test (185 vs. 26) splits, representing the primary factor making pleural effusion harder to detect across all evaluated architectures.

Second, the data are *sparsely annotated*. A total of 113 images ($\sim$11%) are true-negative controls containing no findings, and radiographs containing abnormalities feature a steep distribution (Figure 3): over 400 images exhibit exactly one box, roughly 310 contain two, and the count declines sharply toward the maximum of seven (median 2, mean 1.88). Radiographs showing co-occurrence of both findings are uncommon (110 images in total).

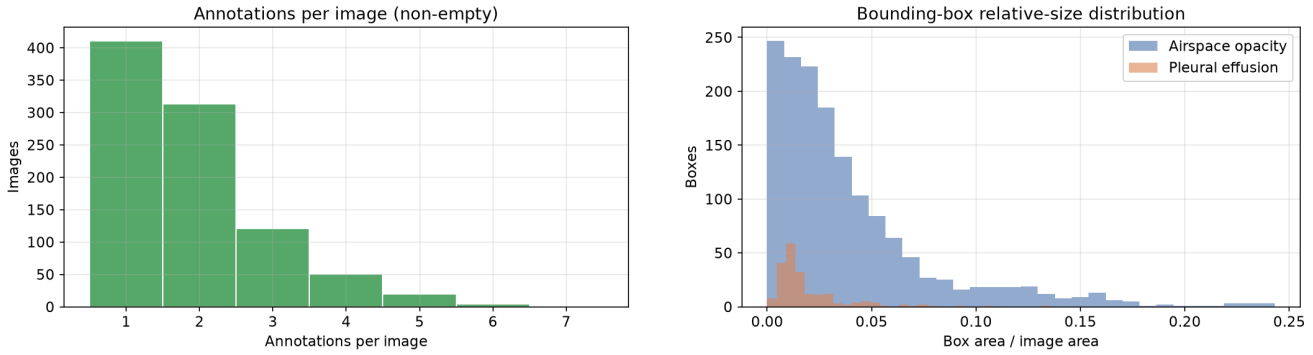
Third, the targets are systematically *small and right-skewed* (Figure 3). The median bounding box occupies only 2.7% of the total image area for airspace opacity and 1.3% for pleural effusion, with the vast majority falling into the COCO small-to-medium brackets. Airspace opacity boxes peak below 0.05 (5%) relative area with an extended tail reaching $\sim$0.25, whereas pleural effusion boxes concentrate almost exclusively below 0.02 (2%). In terms of aspect ratio, both manifestations

cluster along the unit diagonal (median width-to-height ratios of 1.0 and 0.9, respectively; Figure 4), reflecting roughly isotropic dimensions. However, their spatial scale varies substantially: pleural effusion boxes remain strictly confined within 0.20 of the relative image width and height, whereas airspace opacities display wider anatomical extent, spanning up to 0.45 in width and 0.70 in height.

Finally, the spatial distribution of box centers reveals a strong, anatomically coherent structure that corroborates diagnostic fidelity (Figure 4). Airspace opacity concentrates bilaterally across the mid-to-upper lung zones, displaying a dominant hotspot in the right upper zone (exceeding 60 instances). Conversely, pleural effusion is restricted almost exclusively to the dependent costophrenic angles—where fluid naturally layers in upright projections—with a sharper cluster at the right base (up to 9 instances) than the left. This strong spatial prior provides an anatomical rationale for why coarse regional localization (IoU = 0.3) is more readily attained than strict bounding boundary alignment (IoU = 0.5) on diffuse pulmonary opacities.

	Radiographs (Images)			Finding Annotations (Boxes)		
Split	Background	Positive	Total	Airspace opacity	Pleural effusion	Total boxes
Train	67	640	707	1,059	141	1,200
Validation	30	171	201	288	36	324
Test	16	110	126	185	26	211
Total	113	921	1,034	1,532	203	1,735

Table 2. Dataset composition of the NEP-TB-1.1K benchmark across training, validation, and test splits. Background radiographs denote true-negative control images containing no annotated findings. Positive radiographs contain one or more localized finding instances (airspace opacity or pleural effusion). Finding counts reflect individual bounding-box annotations, which sum to the total box count per split.



(a) (b)

Figure 3. Annotation statistics of the dataset after removing cavitation. (a) Distribution of the number of annotations per (non-empty) image; (b) distribution of bounding-box size as a fraction of image area, by finding.

Detection architectures

We compared four detector families that span the main design paradigms in modern object detection: a one-stage anchor-free convolutional detector (YOLOv12), a transformer-based end-to-end detector (RT-DETR), a two-stage region-proposal detector (Faster R-CNN), and a lightweight one-stage anchor-based detector (SSD-MobileNet).

YOLO

YOLOv12[10,30] is a one-stage, anchor-free detector built around an attention-centric backbone. Rather than relying purely on convolutional aggregation, as earlier YOLO generations did, it incorporates an area-attention mechanism and residual efficient layer-aggregation (R-ELAN) blocks, which widen the effective receptive field while keeping inference cheap. Multi-scale features from the backbone are fused by a path-aggregation neck and handed to a decoupled detection head, which predicts a class score and a distribution-based box offset at each spatial location across three output strides. Because the head is anchor-free, box coordinates are regressed directly as a discretised distribution rather than as offsets from predefined anchors. Fine-tuned through Ultralytics' standard trainer, YOLOv12 minimises a weighted sum of a CIoU bounding-box regression loss[31], a binary

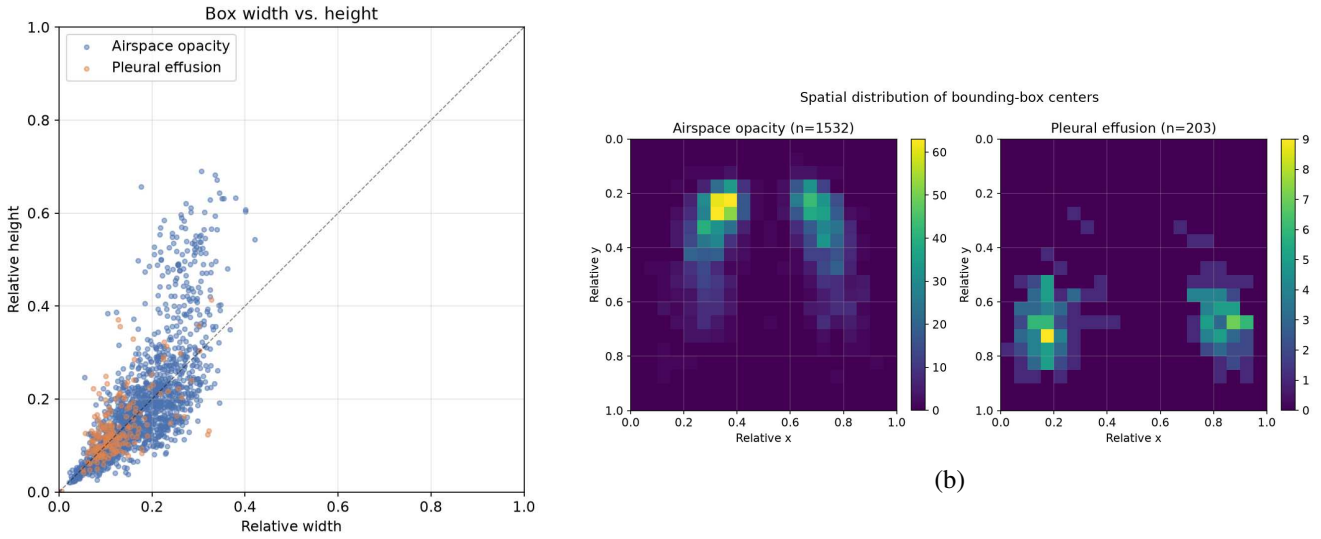


Figure 4. (a) Relative bounding-box width versus height (the dashed line marks square boxes); (b) spatial heat maps of box centres, showing airspace opacity in the mid/upper lung zones and pleural effusion at the lower costophrenic angles.

cross-entropy classification loss, and a distribution focal loss (DFL)[32] that supervises the discretised box-coordinate distribution. We raised the classification weight above its default of 1.0, combining the three terms as $\lambda_{\text{box}} = 7.5$, $\lambda_{\text{cls}} = 2.5$ and $\lambda_{\text{dfl}} = 1.5$, so that misclassifying airspace opacity or pleural effusion is penalised more heavily relative to localisation error.

RT-DETR

RT-DETR[11] is a real-time variant of the DEtection TRansformer (DETR)[33]. A convolutional backbone produces multi-scale feature maps, which an efficient hybrid encoder processes by decoupling intra-scale self-attention from cross-scale feature fusion, cutting down the usual quadratic cost of attention. An IoU-aware query-selection step seeds the object queries from the encoder features, and a transformer decoder with deformable attention then refines a fixed-size set of predictions. The detector is end-to-end and Non-Maximum Suppression(NMS)-free: predictions come directly out of bipartite matching, with no hand-tuned post-processing step. RT-DETR follows the standard DETR-style set-prediction objective: predictions are matched one-to-one to ground-truth boxes via the Hungarian algorithm[33], and each matched pair then contributes an IoU-aware focal classification loss[34,35], a generalised IoU (GIoU) box loss[36], and an L_1 box-regression loss. As with YOLOv12, the classification term was weighted at 2.5 rather than the default of 1.0, for the same reason. RT-DETR's attention-driven, query-based design lets it reason globally over the image and avoid the anchor design choices and NMS post-processing.

Faster R-CNN

Faster R-CNN[12] is the canonical two-stage detector. We utilize a MobileNetV3-Large backbone with a feature-pyramid network (FPN)[37] to build a multi-scale feature hierarchy, rather than the residual-network backbone[38] more commonly paired with this architecture. A region-proposal network (RPN) scans these features to generate class-agnostic object proposals from a dense set of anchors, which are then pooled to a fixed size via RoI-Align and passed to a second-stage head, which performs per-region classification and box regression. This two-stage design generally trades some inference speed for stronger localisation, in contrast to the single-pass formulations of YOLOv12 and SSD-MobileNet. The RPN keeps its standard loss structure, which is a binary cross-entropy objectness loss paired with a smooth-L_1 anchor-regression loss, both at their torchvision defaults, but in the second-stage head we replaced the default softmax cross-entropy classification loss with a focal loss ($\gamma = 2.0$), using per-class α weights of 1.0 (background), 1.0 (airspace opacity), and 3.0 (pleural effusion) to upweight the rarer finding.

SSD-MobileNet

Our SSD-MobileNet detector follows the single-shot, anchor-based philosophy of the original SSD[13], but with a custom architecture built around a MobileNetV3-Large backbone[14]. Rather than predicting directly off the backbone's native multi-scale outputs, we added a top-down FPN with six detection feature maps plus extra convolutional layers, giving finer control over the scales and aspect ratios used at each level. As in the original SSD, detection happens in a single forward pass, with class confidences and box offsets predicted jointly for every default box; of the four detectors, this is the most lightweight,

owing to MobileNetV3's depthwise-separable convolutions. Because this is a fully custom build with no off-the-shelf loss implementation to inherit, classification uses a focal loss ($\gamma = 2.0$, per-class $\alpha = [0.25, 1.0, 3.0]$ for background, airspace opacity and pleural effusion) in place of the original SSD's MultiBox confidence loss, paired with a smooth-L_1 localisation loss over matched anchors, combined as $\lambda_{\text{box}} = 2.0$ and $\lambda_{\text{cls}} = 1.0$. Since the focal term already addresses class imbalance through its weighting, hard-negative mining used in the original MultiBox loss to control the ratio of negative to positive anchors was not needed here.

Training configuration and hyperparameters

All four detectors were trained with AdamW[39] and a cosine-annealed learning-rate schedule[40] preceded by a short linear warm-up, but the implementation, scale and exact settings differed by architecture, as summarised in Table 3. YOLOv12 and RT-DETR were fine-tuned through Ultralytics' standard trainer, with COCO-pretrained initialisation, 512×512 input resolution, automatic mixed-precision training, a fixed seed of 42 and deterministic data ordering throughout. YOLOv12 was trained for up to 200 epochs at batch size 32 with a base learning rate of 1×10^{-3}, weight decay 5×10^{-4}, a 3-epoch warm-up and an early-stopping patience of 30 epochs. RT-DETR was trained for up to 100 epochs at the smaller batch size of 8, with a lower base learning rate of 1×10^{-4} and a heavier weight decay of 1×10^{-2}, a longer 5-epoch warm-up and an early-stopping patience of 15 epochs. Faster R-CNN and SSD-MobileNet were trained with custom loops rather than the Ultralytics trainer. Faster R-CNN, initialised from the full COCO-pretrained torchvision checkpoint, was trained for up to 150 epochs at batch size 16, with a single learning rate of 1×10^{-4} applied across all parameters, weight decay 1×10^{-4}, a 3-epoch warm-up, gradient clipping at a maximum norm of 1.0, and early stopping after 20 epochs without improvement in validation mAP (minimum delta 1×10^{-4}). SSD-MobileNet, whose custom architecture has no COCO-pretrained equivalent, was initialised from an ImageNet-pretrained backbone only and trained for up to 150 epochs at batch size 16, with separate learning rates for the head and backbone (5×10^{-4} and 5×10^{-5}, respectively), weight decay 1×10^{-4}, a 5-epoch warm-up, gradient clipping at a maximum norm of 10.0, and an early-stopping patience of 25 epochs.

Across all four models, the best checkpoint on validation mAP was retained rather than the final epoch's weights; all results reported in this paper therefore reflect each model's best-checkpoint performance.

Hyperparameter	YOLOv12	RT-DETR	Faster R-CNN	SSD-MobileNet
Backbone initialisation	COCO-pretrained	COCO-pretrained	COCO-pretrained	ImageNet-pretrained
Optimiser	AdamW	AdamW	AdamW	AdamW
Learning rate	1×10^{-3}	1×10^{-4}	1×10^{-4}	5×10^{-4} (head) 5×10^{-5} (backbone)
LR schedule	cosine	cosine	cosine	cosine
Warm-up epochs	3	5	3	5
Weight decay	5×10^{-4}	1×10^{-2}	1×10^{-4}	1×10^{-4}
Batch size	32	8	16	16
Epochs	200	100	150	150
Input resolution	512×512	512×512	512×512	512×512
Backbone freezing	none	none	—	—
Gradient clipping (max norm)	—	—	1.0	10.0
Mixed precision	yes	yes	—	—
Early-stopping patience	30	15	20	25
Random seed	42	42	—	—

Table 3. Training hyperparameters for all four detectors. YOLOv12 and RT-DETR were trained with Ultralytics' built-in trainer; Faster R-CNN and SSD-MobileNet used custom training loops with comparable optimisation settings adapted to each architecture.

Data augmentation

Since radiographic findings are sensitive to position, orientation and intensity, our overall design intent was a conservative, anatomy-preserving augmentation policy, though the exact settings varied by model. Across all four detectors, shear, perspective warping, and vertical flipping were disabled throughout, as they would produce anatomically improbable radiographs. Horizontal flipping was applied with probability 0.5 wherever it was used, since the lungs are approximately bilaterally symmetric.

Because the detectors have different capacities and a tendency to overfit on small datasets, dataset augmentation was performed in a way that each model gets their optimal performance. Therefore, we selected different augmentation policies for each selector. At first, the detectors were trained with only a horizontal flip, and starting from this baseline, we varied rotation,

translation, resizing, and brightness one at a time, and the policy with the highest mAP was retained. The final policies are listed in Table 4.

Augmentation	YOLOv12	RT-DETR	Faster R-CNN	SSD-MobileNet
Rotation (degrees)	±5	±10	±15	±10
Translation (% of image)	±8	±8	±10	±10
Scale (resize factor)	0.85–1.15	0.5–1.5	0.9–1.5	0.9–1.1
Horizontal flip	0.5	0.5	0.5	0.5
Brightness/contrast jitter	1.0	1.0	0.5	0.3
Copy-paste	0.3	0.3	0	0

Table 4. Training-time augmentation for each detector.

Evaluation protocol

Hyperparameter selection and early-stopping criteria were determined exclusively on the validation set ($N = 201$ images), whereas all reported quantitative metrics (AP, mAP, and precision–recall curves) were evaluated strictly on the unseen, held-out test split ($N = 126$ images containing 211 annotated findings). Each prediction was matched to the ground-truth annotations on the basis of the intersection-over-union (IoU), defined for a predicted box B_p and a ground-truth box B_{gt} as the ratio of the area of their overlap to the area of their union,

$$\mathrm{IoU}(B_p, B_{gt}) = \frac{|B_p \cap B_{gt}|}{|B_p \cup B_{gt}|}. \tag{1}$$

A prediction of the correct class was counted as a true positive (TP) when its IoU with an unmatched ground-truth box met or exceeded the acceptance threshold τ; otherwise it was a false positive (FP), and any ground-truth box left unmatched was a false negative (FN). From these counts, precision and recall were computed as

$$\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}, \qquad \mathrm{Recall} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}, \tag{2}$$

That is, precision is the fraction of predictions that are correct, and recall is the fraction of ground-truth findings that are retrieved. For each finding, we computed the average precision (AP) as the area under the precision–recall curve $p(r)$,

$$\mathrm{AP} = \int_0^1 p(r)\,\mathrm{d}r, \tag{3}$$

and the mean average precision (mAP) as the mean of the per-finding APs over the C classes,

$$\mathrm{mAP} = \frac{1}{C}\sum_{c=1}^{C} \mathrm{AP}_c, \tag{4}$$

with $C = 2$ here (airspace opacity and pleural effusion), following the standard detection-evaluation methodology[41]. To probe sensitivity to localisation tightness, AP and mAP were reported at two IoU acceptance thresholds, $\tau = 0.3$ and $\tau = 0.5$: the conventional 0.5 threshold and a looser 0.3 threshold that reflects the clinically acceptable tolerance for the diffuse, ill-bordered findings studied here. Precision-recall curves were generated at both thresholds for every model to characterise the full precision/recall trade-off rather than a single operating point.

Acknowledgements

The authors appreciate the support from the volunteers of Kathmandu University School of Medical Sciences and Kathmandu University Dhulikhel Hospital for verifying the annotated radiographs. We also acknowledge the Supercomputer Center at Kathmandu University (SCKU) for providing computing resources used in our experiments.

Author contributions statement

S.W. and A.T. formulated the idea. S.W. and S.D. annotated the dataset and verified the test results. A.T., P.T., and N.N. performed exploratory data analysis, executed the experiments, and wrote the manuscript. All authors reviewed the manuscript.

Data Availability

The datasets used and/or analyzed during the current study are available from the corresponding author upon reasonable request.

Research Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

References

1. World Health Organization. *Global Tuberculosis Report 2025* (World Health Organization, Geneva, 2025).
2. Wu, J. & Dalal, K. Tuberculosis in asia and the pacific: the role of socioeconomic status and health system development. *Int. J. Prev. Med.* **3**, 8–16 (2012).
3. Zellweger, J. P., Sotgiu, G., Corradi, M. & Durando, P. The diagnosis of latent tuberculosis infection (ltbi): Currently available tests, future developments, and perspectives to eliminate tuberculosis (tb). *La Medicina del Lavoro* **111**, 170–183 (2020).
4. World Health Organization. *WHO Consolidated Guidelines on Tuberculosis. Module 2: Screening – Systematic Screening for Tuberculosis Disease* (World Health Organization, Geneva, 2021).
5. World Health Organization. *Use of Computer-Aided Detection Software for Tuberculosis Screening: WHO Policy Statement* (World Health Organization, Geneva, 2025).
6. Van Cleeff, M. R. A., Kivihya-Ndugga, L. E., Meme, H., Odhiambo, J. A. & Klatser, P. R. The role and performance of chest x-ray for the diagnosis of tuberculosis: A cost-effectiveness analysis in nairobi, kenya. *BMC Infect. Dis.* **5**, 111, DOI: 10.1186/1471-2334-5-111 (2005).
7. Liu, Y., Wu, Y. H., Ban, Y., Wang, H. & Cheng, M. M. Rethinking computer-aided tuberculosis diagnosis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2646–2655 (2020).
8. Bhalla, A. S., Goyal, A., Guleria, R. & Gupta, A. K. Chest tuberculosis: Radiological review and imaging recommendations. *Indian J. Radiol. Imaging* **25**, 213–225, DOI: 10.4103/0971-3026.161431 (2015).
9. Chauhan, A., Chauhan, D. & Rout, C. Role of gist and phog features in computer-aided diagnosis of tuberculosis without segmentation. *PLoS ONE* **9**, e112980, DOI: 10.1371/journal.pone.0112980 (2014).
10. Tian, Y., Ye, Q. & Doermann, D. YOLOv12: Attention-centric real-time object detectors. *arXiv preprint arXiv:2502.12524* (2025).
11. Zhao, Y. *et al.* DETRs beat YOLOs on real-time object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 16965–16974 (2024).
12. Ren, S., He, K., Girshick, R. & Sun, J. Faster R-CNN: Towards real-time object detection with region proposal networks. *Adv. Neural Inf. Process. Syst. (NeurIPS)* **28** (2015).
13. Liu, W. *et al.* SSD: Single shot MultiBox detector. In *European Conference on Computer Vision (ECCV)*, 21–37 (Springer, 2016).
14. Howard, A. G. *et al.* MobileNets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861* (2017).
15. Radke, K. L. *et al.* Adaptive IoU thresholding for improving small object detection: A proof-of-concept study of hand erosions classification of patients with rheumatic arthritis on X-ray images. *Diagnostics* **13**, 104, DOI: 10.3390/diagnostics13010104 (2023).
16. Acharya, V. *et al.* Ai-assisted tuberculosis detection and classification from chest x-rays using a deep learning normalization-free network model. *Comput. Intell. Neurosci.* **2022**, 2399428, DOI: 10.1155/2022/2399428 (2022).
17. Nafisah, S. I. & Muhammad, G. Tuberculosis detection in chest radiograph using convolutional neural network architecture and explainable artificial intelligence. *Neural Comput. Appl.* **19**, 1–21, DOI: 10.1007/s00521-022-07905-2 (2022).
18. Bista, R. *et al.* Advancing tuberculosis detection in chest x-rays: A yolov7-based approach. *Information* **14**, 655, DOI: 10.3390/info14120655 (2023).
19. Liu, Y. *et al.* Revisiting computer-aided tuberculosis diagnosis. *IEEE Transactions on Pattern Analysis Mach. Intell.* **46**, 2316–2332, DOI: 10.1109/TPAMI.2023.3330825 (2024).

20. Irvin, J. *et al.* Chexpert: a large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'19/IAAI'19/EAAI'19, DOI: 10.1609/aaai.v33i01.3301590 (AAAI Press, 2019).

21. Johnson, A. E. W. *et al.* Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Sci. Data* **6**, 317, DOI: 10.1038/s41597-019-0322-0 (2019).

22. Bodla, N., Singh, B., Chellappa, R. & Davis, L. S. Soft-NMS–improving object detection with one line of code. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 5561–5569 (2017).

23. Raghu, M., Zhang, C., Kleinberg, J. & Bengio, S. Transfusion: Understanding transfer learning for medical imaging. In *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 32, 3347–3357 (2019).

24. World Health Organization. WHO consolidated guidelines on tuberculosis: module 2: screening: systematic screening for tuberculosis disease. *World Heal. Organ.* (2021). Licence: CC BY-NC-SA 3.0 IGO.

25. Dorman, S. E. *et al.* Xpert MTB/RIF Ultra for detection of Mycobacterium tuberculosis and rifampicin resistance: a prospective multicentre diagnostic accuracy study. *The Lancet Infect. Dis.* **18**, 76–84 (2018).

26. Vorster, M. J., Allwood, B. W., Diacon, A. H. & Bateman, E. D. Tuberculous pleural effusions: advances and controversies. *Afr. J. Thorac. Critical Care Medicine* **21**, 34–38 (2015).

27. Light, R. W. Update on tuberculous pleural effusion. *Respirology* **15**, 451–458 (2010).

28. Shaw, J. A., Diacon, A. H. & Koegelenberg, C. F. N. Tuberculous pleural effusion. *Respiration* **97**, 489–500 (2019).

29. Jeong, Y. J. & Lee, K. S. Pulmonary tuberculosis: up-to-date imaging and management. *Am. J. Roentgenol.* **191**, 834–844 (2008).

30. Jocher, G., Chaurasia, A. & Qiu, J. Ultralytics YOLO. https://github.com/ultralytics/ultralytics (2023).

31. Zheng, Z. *et al.* Distance-IoU loss: Faster and better learning for bounding box regression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, 12993–13000 (2020).

32. Li, X. *et al.* Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection. *Adv. Neural Inf. Process. Syst. (NeurIPS)* **33**, 21002–21012 (2020).

33. Carion, N. *et al.* End-to-end object detection with transformers. In *European Conference on Computer Vision (ECCV)*, 213–229 (Springer, 2020).

34. Lin, T.-Y., Goyal, P., Girshick, R., He, K. & Dollár, P. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2980–2988 (2017).

35. Zhang, H., Wang, Y., Dayoub, F. & Sünderhauf, N. VarifocalNet: An IoU-aware dense object detector. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 8514–8523 (2021).

36. Rezatofighi, H. *et al.* 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 (CVPR)*, 658–666 (2019).

37. Lin, T.-Y. *et al.* Feature pyramid networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2117–2125 (2017).

38. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778 (2016).

39. Loshchilov, I. & Hutter, F. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)* (2019).

40. Loshchilov, I. & Hutter, F. SGDR: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations (ICLR)* (2017).

41. Everingham, M., Van Gool, L., Williams, C. K. I., Winn, J. & Zisserman, A. The PASCAL visual object classes (VOC) challenge. *Int. J. Comput. Vis.* **88**, 303–338 (2010).