

Accurate Lung Segmentation Does Not Improve Deep Learning Detection of Tuberculosis on Chest Radiographs: A Controlled U-Net and ResNet-18 Study

Adnan Haider*

Department of Computer Science, Government College University Faisalabad, Faisalabad, Pakistan Email: adnanhaiderjaferi512@gmail.com

Ikram Ullah

Data Science and Analytics, University of Hertfordshire, Hatfield, United Kingdom Email: ikramaliswabi@gmail.com

Summiya Hakeem

Department of Computer Science, Balochistan University of Engineering and Technology, Khuzdar, Pakistan Email: summiyahakeem1010@gmail.com

Kamran Ali

Department of Computer Systems Engineering and Sciences, Balochistan University of Engineering and Technology, Khuzdar, Pakistan Email: enr.kamranali.official@gmail.com

Author Details

Keywords: Nursing values, future time perspective, problem-solving proficiency, nurse interns.

Received on 25 Jan 2026

Accepted on 10 Mar 2026

Published on 19 Mar 2026

Corresponding E-mail & Author*:

Adnan Haider*

Department of Computer Science, Government College University Faisalabad, Faisalabad, Pakistan Email: adnanhaiderjaferi512@gmail.com

Abstract

Background. Lung field segmentation is routinely placed in front of convolutional neural network (CNN) classifiers for tuberculosis (TB) screening, on the assumption that removing non-lung anatomy helps the classifier concentrate on disease. This assumption has rarely been tested under matched training conditions, at clinically meaningful operating points, or in relation to the quality of the mask itself. This is the research gap the present study addresses.

Methods. We used the 138 posteroanterior chest radiographs of the Montgomery County dataset (80 normal, 58 TB) and their manual lung masks. A four-level U-Net was trained for lung segmentation and compared with a contrast-enhanced Otsu thresholding baseline. An ImageNet-pretrained ResNet-18 was then fine-tuned to detect TB from either the whole radiograph or a lung region of interest (ROI) cropped with the U-Net mask. Both

pipelines shared the same stratified five-fold cross-validation, architecture and training schedule, and all metrics were computed on out-of-fold predictions. Areas under the ROC curve (AUC) were compared with DeLong's test, operating points were judged against the World Health Organization (WHO) minimum accuracy targets for a TB triage test, and changes in predicted probability were related to mask quality, lesion type and Grad-CAM attention.

Results. The U-Net reached a mean Dice similarity coefficient (DSC) of 0.976 ± 0.015 (IoU 0.954, HD95 3.3 px) against 0.607 ± 0.243 for the baseline, with a lung-area bias

of only -0.48 percentage points. The whole-image classifier achieved an AUC of 0.946 (95% CI 0.907–0.975) and the lung-ROI classifier 0.882 (0.814–0.938), a significant difference ($\Delta\text{AUC} = -0.064$, $p = 0.017$). At 90% sensitivity, specificity was 83.8% with the whole image and 48.8% with the lung ROI, so only the whole-image model met the WHO triage target; at 5% prevalence this corresponds to about 199 versus 531 confirmatory tests per 1,000 people screened. The loss of discrimination was not explained by segmentation error: among TB images, larger score drops occurred on better masks (Spearman $\rho = 0.36$, $p = 0.006$), and most additional misses were inactive or fibrotic lesions.

Conclusions. Accurate lung segmentation did not lead to better TB detection on this dataset. Masking and cropping to the aerated lung appear to remove hilar, mediastinal, pleural and apical evidence, together with lung-size information, that the CNN uses to recognise TB. Segmentation should be validated as a design choice in each TB screening pipeline instead of being assumed to help.

Keywords: Tuberculosis; Chest Radiograph; U-Net; Lung Segmentation; Convolutional Neural Network; ResNet-18; Grad-CAM; Computer-Aided Detection; Shortcut Learning; Triage

Highlights

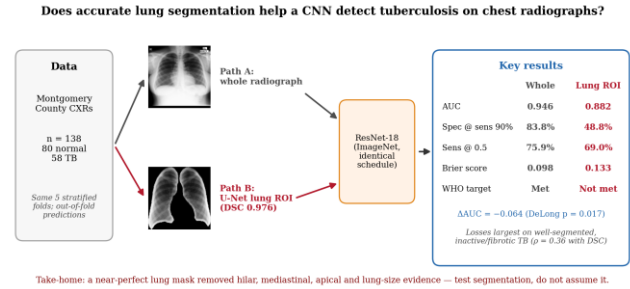
A U-Net segmented the lungs with a Dice coefficient of 0.976 and a lung-area bias of only -0.48 percentage points.

Feeding the same ResNet-18 only the segmented lungs lowered the AUC for TB from 0.946 to 0.882 (DeLong $p = 0.017$).

At 90% sensitivity, specificity fell from 83.8% to 48.8%, so only the whole-image model met the WHO triage target.

Larger score drops occurred on better masks ($\rho = 0.36$), pointing to information removed by design rather than segmentation error.

Lung segmentation should be validated as a design choice in each TB pipeline rather than assumed to help.



Graphical abstract. Study question, matched two-path design and principal findings. All values are out-of-fold estimates on the Montgomery County dataset ($n = 138$).

Introduction

Clinical context

Tuberculosis remains one of the leading infectious causes of death worldwide. In 2024 an estimated 10.7 million people fell ill with TB and 1.23 million died, while only 8.3 million were newly diagnosed and notified, leaving a gap of about 2.4 million people who were either undiagnosed or unreported [1]. The burden is concentrated: eight countries account for two-thirds of all cases, and Pakistan alone accounted for 6.3% of global incident cases and 7.2% of the global gap between estimated and notified cases [1]. Closing this gap depends on screening tools that are cheap, fast and usable where radiologists are scarce.

Chest radiography is cheap and widely available, but in many high-burden countries there are far fewer readers than radiographs. The World Health Organization has

therefore called for triage tests that can rule out TB in people who do not need confirmatory testing, with minimum targets of 90% sensitivity and 70% specificity [2]. In 2021 the WHO went further and recommended that computer-aided detection (CAD) software may be used in place of human readers to interpret digital chest radiographs for screening and triage in people aged 15 years and older [3]. Deep learning systems have since matched or exceeded radiologists on multinational test sets [4, 5], and independent evaluations of five and twelve commercial products confirmed high accuracy, but also lower performance in older people and in those with a history of TB [6, 7]. Why these subgroups are harder has not been explained.

Segmentation as a default preprocessing step

Early CAD systems for TB followed a fixed sequence of lung segmentation, hand-crafted texture and shape features, and a conventional classifier. Jaeger *et al.* [8] used graph-cut lung segmentation followed by feature extraction and reported an AUC of 0.87 on the Montgomery County set. Deep learning has since replaced most hand-crafted pipelines in medical image analysis [9]. CNNs trained end to end on chest radiographs have shown strong performance for TB [10] and for a wide range of thoracic findings [11, 12], and compact architectures have been proposed for fast TB screening in low-resource settings [13].

Segmentation did not disappear with the move to deep learning. The U-Net [14], an encoder–decoder network with skip connections designed for biomedical images, segments lung fields on chest radiographs with high overlap, and earlier atlas-based methods had already shown that the lung boundary can be recovered reliably on public datasets [15]. Several TB studies now place a U-Net in front of the classifier and feed it only the segmented lungs. Rahman *et al.* [16] reported that classification on segmented lungs outperformed classification on whole radiographs. The usual argument is that masking prevents the network from learning shortcuts, such as text markers, positioning differences or hospital-specific features, which are known to inflate the measured performance of chest radiograph models [17–19].

Problem statement

This argument has a weakness. TB is not confined to the aerated lung parenchyma. Pleural effusion, hilar and mediastinal lymphadenopathy, apical scarring partly hidden behind the clavicles, and dense consolidation that a segmentation network treats as non-lung can all lie outside a typical lung mask [20, 21]. Upper-lobe volume loss in fibrotic TB pulls the trachea and hila upward and sideways, and these signs are read from structures that a lung mask removes [21]. If masking takes away this evidence, the classifier may become worse even when the mask itself is excellent.

The problem addressed in this study is therefore the following. A widely adopted design choice, hard lung masking and cropping before classification, is justified by an argument about shortcuts but has not been weighed against the diagnostic evidence it discards. TB screening pipelines may consequently be built on a preprocessing step that lowers their accuracy at exactly the high-sensitivity operating points that WHO requires of a triage test [2, 3]. Whether segmentation helps or hurts is an empirical question whose answer may depend on the dataset, the disease spectrum and the design of the mask.

Research gap

Figure 1 maps the main lines of evidence that bear on this question, and Table 1 summarises representative studies. Four gaps stand out.

G1 – Unmatched comparisons. Studies that compare whole and segmented inputs usually differ in data splits, architectures or training schedules between conditions and rarely test the difference formally. Rahman *et al.*, for example, reported their best whole-image and best segmented-lung results from different networks on a pooled

multi-source collection [16]. Work on COVID-19 radiographs showed that dataset bias persists after segmentation [22] and that classifiers can perform well even when the lungs are blacked out [23], but these studies did not measure what masking costs for a disease whose signs extend beyond the lung.

G2 – Segmentation reported on mostly normal images. Dice scores above 0.95 are commonly reported on the Montgomery and JSRT datasets, but they are averaged over all images; few studies report abnormal radiographs separately or quantify bias in lung area, and U-Nets are known to degrade on radiographs with severe abnormalities [24–26].

G3 – Clinically irrelevant operating points. Most comparisons report accuracy at a default threshold of 0.5, whereas a triage test is used at a high-sensitivity operating point [2, 6]. With small public datasets, cross-validated estimates also carry wide error bars that are seldom reported [27, 28].

G4 – No link between performance change, mask quality and lesion type. Almost no study asks whether a change in classifier performance is caused by poor masks or by information removed by design, or which lesion types are affected. Aggregate metrics can hide clinically important failures in subgroups [29], and saliency maps alone are an unreliable guide to what a model uses [30].

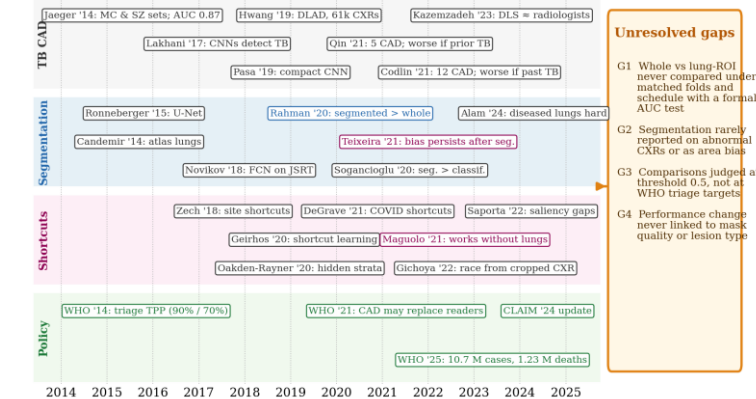


Figure 1. Evidence map, 2014–2025, of the four strands of literature that bear on segmentation before TB classification: TB CAD and public datasets [4–8, 10, 13], lung segmentation [14–16, 22, 26, 31, 32], shortcut learning and explainability [17–19, 23, 29, 30, 33] and policy and reporting [1–3, 34]. Coloured boxes mark studies that bear most directly on the present question. The panel on the right lists the gaps (G1–G4) addressed in this study.

Table 1. Representative studies on segmentation, shortcut learning and TB CAD, and the gaps that remain for the present question.

Study	Data and design	Key finding	Gap relative to the present question
Jaeger <i>et al.</i> , 2014 [8]	Montgomery and Shenzhen; graph-cut lung segmentation, hand-crafted features, classifier	AUC 0.87 on Montgomery	Segmentation assumed; no whole-image comparator
Rahman <i>et al.</i> , 2020 [16]	3,500 TB + 3,500 normal pooled from several public sources; nine CNNs on whole versus U-Net-segmented images	Best accuracy 96.47% whole (CheXNet) versus 98.6% segmented (DenseNet201)	Different best networks per condition; multi-site pooling rich in site cues; no paired test or triage operating point (G1, G3)

Study	Data and design	Key finding	Gap relative to the present question
Sogancioglu <i>et al.</i> , 2020 [32]	ChestX-ray14 cardiomegaly; segmentation-based versus image-level classification	AUC 0.977 versus 0.941 in favour of segmentation	Cardiomegaly is defined by anatomy (cardiothoracic ratio); TB signs are not confined to one structure
Teixeira <i>et al.</i> , 2021 [22]	Multi-source COVID-19, pneumonia and normal; U-Net (Dice 0.982) plus three CNNs, LIME and Grad-CAM	Strong source bias persists after segmentation	Not TB; change not related to mask quality (G4)
Maguolo and Nanni, 2021 [23]	COVID-19 radiographs with the image centre (lungs) blacked out	Similar results without the lungs: dataset cues	Shows shortcuts outside the lungs, not the evidence lost by masking
Qin <i>et al.</i> , 2021 [6]; Codlin <i>et al.</i> , 2021 [7]	Independent evaluations of 5 and 12 commercial CAD products	High accuracy; lower performance in older people and with prior TB	Black-box products; contribution of preprocessing unknown (G4)
This study	Montgomery (n = 138); identical folds, network and schedule; whole image versus fold-specific U-Net lung ROI	AUC 0.946 versus 0.882 ($p = 0.017$); losses on well-segmented fibrotic TB	Addresses G1–G4 with a paired DeLong test, WHO operating points, mask-quality and lesion-type analysis

Research questions and hypotheses

The study was designed to answer four research questions (RQs), each matched to one gap:

RQ1 (G2): How accurately does a U-Net segment the lung fields on normal and TB radiographs, and where and why does it fail?

RQ2 (G1): Under identical folds, architecture and training schedule, does a CNN trained on U-Net lung crops discriminate TB from normal radiographs better than the same CNN trained on whole radiographs?

RQ3 (G3): Does either pipeline meet the WHO minimum accuracy targets for a triage test, and what would each imply for confirmatory testing in a screening programme?

RQ4 (G4): Is any difference between the pipelines explained by segmentation error, or by evidence that masking removes by design, and which lesion types are affected?

The conventional view was framed as hypothesis H1: the lung-ROI classifier has a higher AUC than the whole-image classifier, against the null hypothesis H0 of no difference. Based on the radiology of TB [20, 21], a competing information-loss hypothesis H2 predicts that if masking lowers performance, the loss will be largest on well-segmented images of disease located near the hila, mediastinum or apices, and will not track mask error.

Contributions

The study makes five contributions:

A matched, leakage-controlled comparison of whole-image and lung-ROI inputs in which the input is the only difference between pipelines, with fold-specific masks and a formal DeLong test of correlated AUCs [35].

A segmentation evaluation stratified by normal and TB radiographs that reports overlap, boundary distance and Bland–Altman lung-area agreement together with a failure analysis [36, 37].

An evaluation at the WHO triage operating points [2], with a projection of confirmatory-test demand in the spirit of recent CAD evaluations [6]

A mechanistic analysis that links the change in predicted probability to mask quality, lesion type and Grad-CAM attention [38].

A fully reproducible protocol that runs on a single CPU core in under an hour, using only a public dataset [39]

Literature Review

Computer-aided detection of tuberculosis

Automated reading of chest radiographs for TB began with systems that described the segmented lung fields with intensity, texture and shape features and passed these features to a statistical classifier. The system of Jaeger *et al.* [8] is a representative example and, together with the release of the Montgomery County and Shenzhen datasets [39], set the reference point that many later studies compared against. Their accuracy was limited by the hand-crafted features and by the quality of the underlying lung segmentation.

Convolutional neural networks learn features directly from the pixels and removed the need for hand-crafted descriptors [9]. Lakhani and Sundaram [10] showed that CNNs pre-trained on natural images and fine-tuned on chest radiographs could classify pulmonary TB with high accuracy, and large labelled collections such as ChestX-ray8 [12] allowed models such as CheXNet [11] to detect a broad range of thoracic findings. Pasa *et al.* [13] demonstrated that much smaller networks can reach competitive TB performance while running quickly on modest hardware, and larger TB-specific collections with lesion annotations, such as TBX11K, have since been introduced [40]. At scale, systems trained on tens of thousands of radiographs have matched or outperformed radiologists [4, 5], and programmatic evaluations showed that CAD can halve confirmatory testing at above 90% sensitivity [6]. A recent systematic review confirmed the overall promise of machine and deep learning for TB detection while noting heterogeneous methods and reporting [41].

Results on the Montgomery County dataset and the comparability problem

Reported AUCs vary widely, from about 0.75 to 0.95 among the studies collated in a recent systematic review [41], because training data, pre-training and evaluation protocols differ. Table 2 lists results whose protocol can be identified. Gozes and Greenspan, who trained on Shenzhen and tested on Montgomery as an external set, showed how strongly pre-training matters: an ImageNet-initialised DenseNet-121 reached an AUC of 0.846, whereas a network pre-trained on ChestX-ray14 with age, sex and positioning metadata reached 0.928 [42]. With only 138 images, cross-validated estimates have error bars of several percentage points [27], small samples are associated with optimistic reported accuracy [28], and leakage between training and test data, for example from augmentation, preprocessing or patients with several images, is a widespread source of inflated results in machine-learning-based science [43–45]. Comparisons across studies on this dataset are therefore indicative at best, which is why the present study compares its two conditions within one protocol.

Table 2. Published AUCs on the Montgomery County dataset for which the evaluation protocol can be identified.

Study	Method	Evaluation on Montgomery	AUC
Jaeger <i>et al.</i> , 2014 [8]	Lung segmentation, hand-crafted features, classifier	Cross-validation within Montgomery	0.87
Pasa <i>et al.</i> , 2019 [13]	Compact CNN on whole radiographs	Per-dataset evaluation as reported	0.811
Gozes and Greenspan, 2019 [42]	DenseNet-121, ImageNet initialisation, fine-tuned on Shenzhen	External test, all 138 images	0.846
Gozes and Greenspan, 2019 [42]	MetaChexNet: ChestX-ray14 pre-training with age, sex and position metadata	External test, all 138 images	0.928
This study	ResNet-18, ImageNet initialisation, whole image	Five-fold out-of-fold predictions	0.946 (0.907–0.975)
This study	ResNet-18, ImageNet initialisation, U-Net lung ROI	Same folds and schedule	0.882 (0.814–0.938)

Values are as reported by the original authors; training data and protocols differ, so the figures are not directly comparable. Values in parentheses are 95% bootstrap confidence intervals.

Lung field segmentation on chest radiographs

Lung boundary detection on chest radiographs has a long history that includes rule-based thresholding, deformable and statistical shape models, atlas registration and, more recently, deep networks [24]. Candemir *et al.* [15] combined anatomical atlases with non-rigid registration and achieved high overlap on public datasets. The U-Net [14] and its variants now dominate the task because they learn both local edges and global lung shape from a few hundred annotated images; fully convolutional multi-class designs reached Jaccard overlaps of about 0.95 for the lungs on JSRT [31], and self-configuring frameworks such as nnU-Net have made strong segmentation baselines routine [46]. Most reported Dice scores on the Montgomery and JSRT datasets lie above 0.95, and lung segmentation is often regarded as a solved problem. Reported scores, however, are usually averaged over all images. When radiographs with severe abnormalities are tested separately, standard U-Nets perform noticeably worse [25], and recent work has designed architectures and augmentation specifically for diseased lungs whose outlines are distorted by opacification, effusion or volume loss [26].

Segmenting before classifying: evidence for and against

A two-stage design, in which the lungs are first segmented and the classifier then sees only the lung region, has been used in several TB studies. Rahman *et al.* [16] compared nine CNNs on whole and segmented radiographs from a pooled multi-source dataset and found that segmented inputs gave higher accuracy. In cardiomegaly, Sogancioglu *et al.* showed that a segmentation-based approach outperformed image-level classification while needing far fewer annotated images [32], but cardiomegaly is defined by an anatomical ratio that segmentation measures directly. Evidence from COVID-19 radiographs points the other way: Teixeira *et al.* found that a strong bias from the source datasets remained after U-Net segmentation [22], and Maguolo and

Nanni obtained similar results after blacking out the centre of the image, showing that classifiers were learning dataset signatures rather than lung disease [23]. Global–local architectures [47] suggest that the choice need not be binary, but no study, to our knowledge, has measured the cost of hard lung masking for TB under matched conditions.

Shortcut learning, hidden stratification and demographic encoding

A separate line of work has shown that CNNs trained on chest radiographs readily exploit cues unrelated to disease, such as the hospital of origin, portable-scanner markers, text overlays and patient positioning [17, 18]. Geirhos *et al.* [19] described this behaviour more generally as shortcut learning, and Oakden-Rayner *et al.* showed that strong average performance can hide clinically important failures in unrecognised subgroups [29]. Demographic information is also encoded in radiographs: networks can estimate age and sex [42] and even self-reported race, the latter persisting in cropped and degraded images [33]. Imbalances in such attributes can bias classifiers and produce systematic under-diagnosis in some groups [48, 49]. Lung masking is a natural defence against shortcuts outside the thorax; its cost, the loss of diagnostic evidence outside the mask, has received far less attention.

Where TB evidence lies on a frontal radiograph

The radiological appearance of pulmonary TB depends on the stage and activity of disease. Primary TB commonly shows hilar or mediastinal lymphadenopathy and pleural effusion, whereas post-primary disease favours the apical and posterior segments of the upper lobes with consolidation, cavitation, fibrosis, calcification and volume loss, often with apical pleural thickening and retraction of the hila and trachea [20, 21]. Several of these signs lie wholly or partly outside the aerated lung that a typical mask retains (Figure 2a). Cropping to a bounding box and resizing to a fixed frame also removes the absolute size of the lungs, which radiologists use to recognise chronic volume loss (Figure 2b). The weaker performance of commercial CAD in people with prior TB [6, 7] is consistent with the idea that old, fibrotic disease is harder to detect from the lung parenchyma alone. Figure 2 sets out the four mechanisms (M1–M4) that the present study examines.

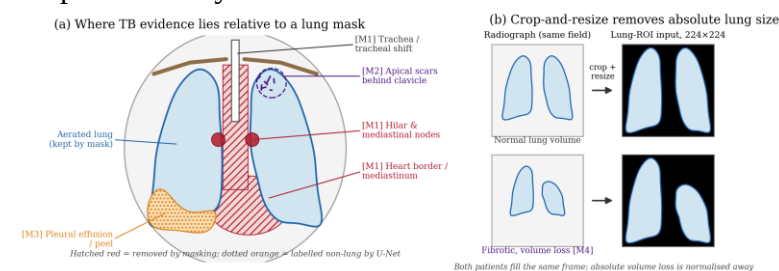


Figure 2. Hypothesised mechanisms by which hard lung masking can remove TB evidence. (a) Schematic frontal radiograph showing the aerated lung retained by a mask (blue), structures removed by masking (hatched red: trachea, hila, mediastinum and heart border; M1), apical scars at the mask border behind the clavicle (M2), and pleural fluid that a U-Net labels as non-lung (M3). (b) Cropping each mask to its bounding box and resizing to 224×224 makes a patient with volume loss fill the same frame as a patient with normal lungs, so absolute lung size is lost (M4). The schematic is illustrative and based on the radiology of TB described in [20, 21].

Visual explanation of chest radiograph classifiers

Grad-CAM [38] is widely used to show that TB classifiers attend to plausible regions [13, 16] and to assess segmentation effects [22]. A systematic benchmark, however, found that all seven saliency methods tested, including Grad-CAM, localised pathology

significantly worse than radiologists, especially for small lesions [30]. We therefore use it only qualitatively.

Synthesis

In short, masking removes site cues that inflate multi-site performance [17, 18, 23], yet TB produces signs outside the aerated lung [20, 21], segmentation is least reliable on abnormal radiographs [25, 26], and CAD already struggles with old disease [6, 7]. What is missing is a matched experiment that isolates the input, evaluates it where a triage test is used, and asks why any difference arises.

Materials and Methods

Study design and reporting

This is a retrospective diagnostic-accuracy study on a public, de-identified dataset. It is reported with reference to the Checklist for Artificial Intelligence in Medical Imaging (CLAIM) 2024 update [34]. The design isolates a single factor, the classifier input, while holding the data splits, network, initialisation, augmentation and training schedule constant (Figure 3). Every number reported in Section 4 comes from out-of-fold predictions, so it describes images that the corresponding segmentation and classification models never saw during training, a safeguard against the leakage that commonly inflates machine-learning results [43, 44].

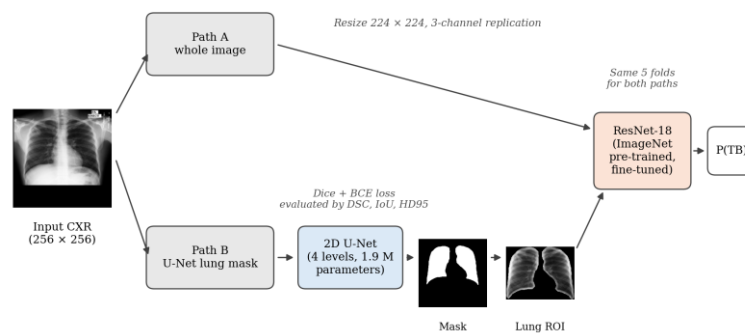


Figure 3. Study design. Path A passes the whole radiograph to the classifier. Path B first segments the lungs with a U-Net [14], masks and crops the radiograph to the lung region, and then passes the crop to an identical ResNet-18 [50]. Both paths use the same five cross-validation folds, the same architecture and the same training schedule, so the input is the only difference.

Dataset

We used the Montgomery County (MC) chest X-ray set released by the U.S. National Library of Medicine [39]. It contains 138 posteroanterior radiographs collected through the TB control program of the Department of Health and Human Services of Montgomery County, Maryland, USA. Eighty images are labelled normal and 58 show manifestations of TB. Each image is distributed as a 12-bit greyscale PNG of 4020×4892 or 4892×4020 pixels and comes with manually drawn left and right lung masks and a short clinical reading that records sex, age and the radiological finding. Table 3 summarises the cohort. Patients with TB were older than normal subjects (Mann–Whitney $p < 0.001$) and more often male (Fisher’s exact $p < 0.001$). The TB readings cover a wide clinical range, from active cavitary and miliary disease to pleural effusion and long-standing inactive scars, a spectrum that matches the primary and post-primary patterns described in the radiological literature [20, 21]. Because demographic differences between classes are a recognised source of bias in chest radiograph models [48, 49], they are reported explicitly and revisited in Section 5.4.

Table 3. Characteristics of the Montgomery County dataset.

Characteristic	Normal	TB	All
Radiographs, n	80	58	138
Male / female / not recorded	26 / 53 / 1	37 / 21 / 0	63 / 74 / 1
Age, years, mean $\pm$ SD	34.3 $\pm$ 15.4	48.1 $\pm$ 20.1	40.1 $\pm$ 18.8
Age range, years	4–70	14–89	4–89
Landscape acquisition (4892 $\times$ 4020)	29 (36.3%)	12 (20.7%)	41 (29.7%)
Readings mentioning pleural disease	—	13 (22.4%)	—
Readings describing inactive, scarred or calcified disease	—	24 (41.4%)	—

Age and sex were taken from the clinical reading supplied with each image. Reading categories were assigned by keyword search and may overlap.

The dataset is publicly available and fully de-identified; no new data were collected, and institutional ethical approval was therefore not required.

Pre-processing

Each radiograph was converted to 8-bit greyscale and resized to 256×256 pixels with bilinear interpolation. The left and right lung masks were merged into a single binary mask and resized with nearest-neighbour interpolation. Resizing to a square grid does not preserve the original aspect ratio; we accepted this because the same transformation was applied to every image and mask. No lung-specific enhancement was applied to the network inputs. Intensities were scaled to $[0, 1]$ and then standardised for each network as described below; for the classifier this used the ImageNet channel statistics that match its pre-training [51].

Lung segmentation with U-Net

The segmentation network followed the original U-Net design [14] with four encoder levels and a bottleneck (Figure 4, Table 4). Each level contains two 3×3 convolutions, each followed by batch normalisation [52] and a ReLU activation, and the number of filters doubles at each level from 16 to 256 at the bottleneck. The decoder upsamples with 2×2 transposed convolutions and concatenates the matching encoder features through skip connections, which return fine spatial detail lost during pooling. A final 1×1 convolution produces one logit per pixel. The network has 1,942,289 trainable parameters. To keep training feasible on a single CPU core, segmentation inputs were downsampled to 128×128 by area averaging; predicted probability maps were upsampled bilinearly back to 256×256 before thresholding, so all segmentation metrics are reported against the 256×256 reference masks.

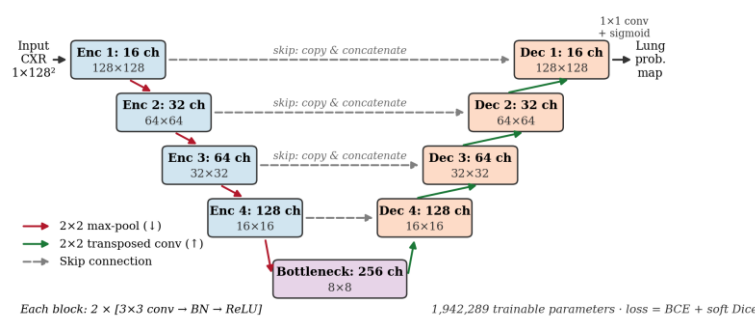


Figure 4. Architecture of the four-level U-Net used for lung segmentation, after Ronneberger *et al.* [14]. Encoder blocks (blue) halve the spatial size with 2×2 max-pooling; decoder blocks (orange) double it with 2×2 transposed convolutions and

concatenate the matching encoder features through skip connections. Each block applies two 3×3 convolutions with batch normalisation [52] and ReLU.

Table 4. U-Net architecture used for lung segmentation.

Stage	Operations	Output channels	Spatial size
Encoder 1	$2 \times (\text{Conv } 3 \times 3, \text{BN, ReLU})$	16	128×128
Encoder 2	Max-pool 2×2 , $2 \times (\text{Conv } 3 \times 3, \text{BN, ReLU})$	32	64×64
Encoder 3	Max-pool 2×2 , $2 \times (\text{Conv } 3 \times 3, \text{BN, ReLU})$	64	32×32
Encoder 4	Max-pool 2×2 , $2 \times (\text{Conv } 3 \times 3, \text{BN, ReLU})$	128	16×16
Bottleneck	Max-pool 2×2 , $2 \times (\text{Conv } 3 \times 3, \text{BN, ReLU})$	256	8×8
Decoder 4	Up-conv 2×2 , concat Enc. 4, $2 \times (\text{Conv, BN, ReLU})$	128	16×16
Decoder 3	Up-conv 2×2 , concat Enc. 3, $2 \times (\text{Conv, BN, ReLU})$	64	32×32
Decoder 2	Up-conv 2×2 , concat Enc. 2, $2 \times (\text{Conv, BN, ReLU})$	32	64×64
Decoder 1	Up-conv 2×2 , concat Enc. 1, $2 \times (\text{Conv, BN, ReLU})$	16	128×128
Output	Conv 1×1 , sigmoid	1	128×128

BN, batch normalisation; Up-conv, transposed convolution. Total trainable parameters: 1,942,289.

The loss was the sum of binary cross-entropy and soft Dice loss [53]. For a predicted probability map p and reference mask g over N pixels, the two terms are

$$L_{\text{BCE}} = -(1/N) \sum_i [g_i \log p_i + (1 - g_i) \log(1 - p_i)] \quad (1)$$

$$L_{\text{Dice}} = 1 - (2 \sum_i p_i g_i + 1) / (\sum_i p_i + \sum_i g_i + 1) \quad (2)$$

and the training objective was $L = L_{\text{BCE}} + L_{\text{Dice}}$. Binary cross-entropy gives stable pixel-wise gradients, while the Dice term directly rewards overlap and is less sensitive to the imbalance between lung and background pixels [53]. We used the Adam optimiser [54] with an initial learning rate of 10^{-3} , cosine annealing [55], a batch size of 8 and 50 epochs. Training images were augmented on the fly [56] with random rotation ($\pm 10^\circ$), isotropic scaling ($\pm 10\%$), translation ($\pm 5\%$ of image size) and brightness and contrast changes, with the same geometric transform applied to image and mask. Horizontal flipping was not used, because it would place the heart on the wrong side. Predictions were thresholded at 0.5, the two largest connected components were kept and holes were filled.

Classical segmentation baseline

To give the U-Net results a reference point, we implemented a thresholding pipeline that needs no training. Contrast-limited adaptive histogram equalisation (CLAHE) [57] was applied first. A global Otsu threshold [58] separated the body from the background, and a second Otsu threshold computed inside the body selected dark, air-filled regions. A 10-pixel border was excluded, the result was opened morphologically, components touching the image edge were removed, and the two largest remaining components were kept and hole-filled. Rule-based pipelines of this kind preceded learned methods in the lung-boundary literature [24].

Tuberculosis classification under two input conditions

We used ResNet-18 [50] initialised with ImageNet weights [51], with the final fully connected layer replaced by a single output unit. ResNet-18 was chosen because its residual blocks train stably on small datasets and because its 11.2 million parameters are modest enough for CPU fine-tuning; transfer from ImageNet is the established

approach for small TB datasets [10, 13]. The greyscale input was replicated across three channels and normalised with the ImageNet mean and standard deviation. Two input conditions were compared (Figure 5). In the *whole-image* condition, the 256×256 radiograph was resized to 224×224 . In the *lung-ROI* condition, the U-Net probability map was thresholded at 0.5, reduced to its two largest components, dilated by six pixels so that the pleural margin was not cut too tightly, and multiplied with the radiograph; the result was cropped to the bounding box of the mask and resized to 224×224 , following the two-stage design used in earlier TB work [16]. For each fold, the lung masks were produced by the U-Net trained on that fold’s training images only, so the classifier test images were never seen by the segmentation model either.

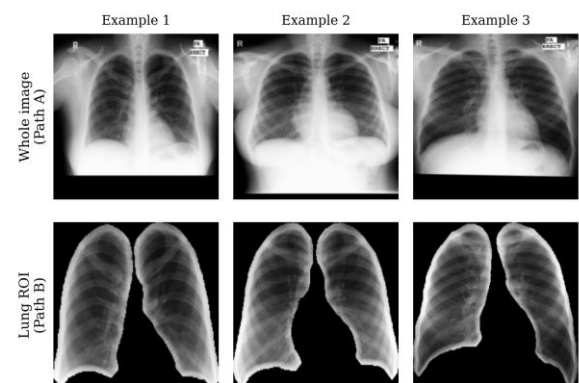


Figure 5. Classifier inputs for three radiographs. Top row: whole-image input (Path A). Bottom row: lung-ROI input (Path B), produced by masking the radiograph with the dilated U-Net mask, cropping to the mask’s bounding box and resizing to 224×224 . The lung-ROI input removes the mediastinum, hila, trachea and extrathoracic structures and rescales every pair of lungs to the same frame (mechanisms M1 and M4 in Figure 2).

Both classifiers were trained identically: binary cross-entropy loss, AdamW [59] with a learning rate of 10^{-4} and weight decay of 10^{-4} , cosine annealing [55], a batch size of 16 and 15 epochs. Augmentation [56] consisted of rotation ($\pm 7^\circ$), scaling ($\pm 8\%$), translation ($\pm 4\%$) and small brightness and contrast changes. The number of epochs was fixed in advance, and no early stopping or threshold tuning was done on the test folds. Table 5 lists the full configuration of both networks.

Table 5. Training configuration for the segmentation and classification networks.

Setting	U-Net (segmentation)	ResNet-18 (classification)
Input size	128×128 (evaluated at 256×256)	224×224 , 3 channels
Initialisation	Random	ImageNet pre-trained [51]
Trainable parameters	1.94 M	11.18 M
Loss	BCE + soft Dice [53]	BCE
Optimiser, learning rate	Adam [54], 1×10^{-3}	AdamW [59], 1×10^{-4} (wd 1×10^{-4})
Schedule, epochs, batch	Cosine [55], 50, 8	Cosine, 15, 16
Augmentation	Rotation $\pm 10^\circ$, scale $\pm 10\%$, shift $\pm 5\%$, intensity	Rotation $\pm 7^\circ$, scale $\pm 8\%$, shift $\pm 4\%$, intensity
Post-processing decision	/ Threshold 0.5, two largest components, hole filling	Threshold 0.5 on sigmoid output

BCE, binary cross-entropy; wd, weight decay.

Evaluation protocol

The 138 images were split once into five stratified folds (random seed 42), each holding 27 or 28 images with the same class ratio as the full set. The same five folds were used for the U-Net, the whole-image classifier and the lung-ROI classifier. For each fold a model was trained on the other four folds and applied to the held-out fold, and the held-out predictions from all five folds were pooled. All results therefore describe 138 out-of-fold predictions per model. Because the masks used by the lung-ROI classifier were produced by a fold-specific U-Net, neither stage was exposed to test images, avoiding a form of leakage that is easy to introduce in two-stage pipelines [43]. With 138 images, fold-to-fold variation is expected to be substantial [27], so pooled out-of-fold statistics with bootstrap intervals were used as the primary analysis and fold-wise results are reported for transparency.

Segmentation metrics

Let P be the set of pixels predicted as lung and G the set of reference lung pixels. Overlap was measured with the Dice similarity coefficient and the intersection over union [36],

$$\text{DSC} = 2|P \cap G| / (|P| + |G|), \quad \text{IoU} = |P \cap G| / |P \cup G| \quad (3)$$

together with pixel sensitivity $|P \cap G| / |G|$ and pixel precision $|P \cap G| / |P|$. Because overlap metrics are insensitive to boundary errors on large structures, boundary accuracy was measured as well, following current recommendations on metric selection [60], with the 95th-percentile Hausdorff distance [36]. If ∂P and ∂G are the boundary pixels of the two masks and $d(x, Y)$ is the Euclidean distance from pixel x to the nearest pixel of Y , then

$$\text{HD}_{95} = P_{95} \{ d(x, \partial G) : x \in \partial P \} \cup \{ d(y, \partial P) : y \in \partial G \} \quad (4)$$

where P_{95} denotes the 95th percentile of the pooled distances, measured in pixels at 256×256 resolution. Using the 95th percentile instead of the maximum makes the measure robust to a few isolated boundary pixels. Agreement between predicted and reference lung area, expressed as a percentage of the image, was examined with Pearson correlation and a Bland–Altman analysis [37]. The U-Net and the baseline were compared with the Wilcoxon signed-rank test, and U-Net DSC on normal and TB images was compared with the Mann–Whitney U test.

Classification metrics and statistical analysis

With TP, TN, FP and FN denoting true and false positives and negatives at a given threshold, we report sensitivity $\text{TP}/(\text{TP} + \text{FN})$, specificity $\text{TN}/(\text{TN} + \text{FP})$, precision $\text{TP}/(\text{TP} + \text{FP})$, accuracy and the F1-score, the harmonic mean of precision and sensitivity. The threshold-free AUC was the primary measure. Ninety-five per cent confidence intervals were obtained from 2,000 bootstrap resamples of the 138 out-of-fold predictions [61]. The two correlated AUCs were compared with DeLong’s test [35], and the paired correct/incorrect decisions at the default threshold of 0.5 were compared with an exact McNemar test [62].

Because a fixed threshold of 0.5 has no clinical meaning, we also examined three operating points: the point that maximises Youden’s index $J = \text{sensitivity} + \text{specificity} - 1$ [63]; the specificity reached when sensitivity is at least 90%, the WHO minimum for a triage test [2]; and the sensitivity reached when specificity is at least 95%. Calibration was summarised with the Brier score [64], the mean squared difference between predicted probability and true label. To translate the high-sensitivity operating points into programme terms, in the spirit of the test-demand framework used in recent CAD evaluations [6], we projected the number of confirmatory tests required per 1,000 people screened, T , and the number of tests per TB case detected, N , for a screened population with TB prevalence π :

$$T = 1000 [\pi \text{Se} + (1 - \pi)(1 - \text{Sp})] \quad (5)$$

$$N = [\pi \text{Se} + (1 - \pi)(1 - \text{Sp})] / (\pi \text{Se}) \quad (6)$$

where Se and Sp are the sensitivity and specificity at the operating point. These projections assume that the pooled out-of-fold operating characteristics transfer to the screened population and are intended for comparison between pipelines, not as estimates for a specific programme.

To ask whether classification changes were linked to mask quality, we computed for every TB image the change in predicted probability between the two models, $\Delta p = p_{\text{whole}} - p_{\text{ROI}}$, and correlated it with the U-Net DSC using Spearman's ρ . Finally, TB images were grouped by keywords in their clinical readings (pleural disease; inactive, scarred or calcified disease) to see which lesion types each model missed, with Wilson score intervals for subgroup sensitivities [65]. Subgroup analyses were planned to probe hidden stratification [29] and are treated as hypothesis-generating. Two-sided $p < 0.05$ was considered significant.

Visual explanation

Gradient-weighted class activation mapping (Grad-CAM) [38] was computed on the last residual stage of each ResNet-18, using the fold model for which the image was in the test set. If A^k is the k -th feature map of that stage and y the output logit, the channel weights and the map are

$$\alpha_k = (1/Z) \sum_{i,j} \partial y / \partial A^k_{ij}, \quad M = \text{ReLU}(\sum_k \alpha_k A^k) \quad (7)$$

where Z is the number of spatial positions. The map was upsampled to 224×224 , normalised to $[0, 1]$ and overlaid on the network input. Given the documented limitations of saliency maps for localising chest radiograph pathology [30], we used Grad-CAM qualitatively, to see whether the two models attended to plausible regions and how their attention differed.

Implementation

All code was written in Python with PyTorch [66], scikit-learn, SciPy and scikit-image, and ran on a single CPU core with 3 GB of memory. One U-Net fold took about 4.3 minutes to train and one ResNet-18 fold about 2.3 minutes, so the complete experiment can be reproduced on an ordinary laptop in under an hour. Random seeds were fixed for the data split and for network initialisation.

Results

Lung segmentation accuracy

Table 6 reports the out-of-fold segmentation results. The U-Net reached a mean DSC of 0.976 ± 0.015 and a mean HD95 of 3.3 pixels. The thresholding baseline reached a DSC of 0.607 ± 0.243 , and the U-Net was better on every one of the 138 images (Wilcoxon $p < 10^{-23}$). The baseline usually failed by merging the lungs with the dark background outside the body, or by losing the lower lung zones where the diaphragm and heart overlap the parenchyma. Performance was consistent across folds, with fold-wise mean DSC between 0.973 and 0.980 (Table 9). These values lie within the range reported for deep lung segmentation on public datasets [22, 24, 31], so the classifier comparison below was not handicapped by a weak mask.

Table 6. Out-of-fold lung segmentation performance (mean $\pm$ SD over images).

Method / subset	n	DSC	IoU	Sensitivity	HD95 (px)
Otsu + CLAHE baseline	138	0.607 ± 0.243	± 0.480 0.252	$\pm 0.709 \pm 0.226$	61.2 ± 38.7
U-Net, all images	138	0.976 ± 0.015	± 0.954 0.028	$\pm 0.967 \pm 0.025$	3.33 ± 3.57

Method / subset	<i>n</i>	DSC	IoU	Sensitivity	HD95 (px)
U-Net, normal	80	0.981 0.005	$\pm$ 0.962 0.010	$\pm$ 0.973 $\pm$ 0.011	2.22 $\pm$ 0.73
U-Net, TB	58	0.969 0.021	$\pm$ 0.941 0.038	$\pm$ 0.959 $\pm$ 0.034	4.86 $\pm$ 5.06

DSC, Dice similarity coefficient; IoU, intersection over union; HD95, 95th-percentile Hausdorff distance at 256×256 . U-Net pixel precision was 0.986 ± 0.013 overall, 0.989 ± 0.006 on normal and 0.981 ± 0.017 on TB images.

Segmentation was harder on abnormal radiographs, as anticipated from studies of severely abnormal images [25, 26]. Mean DSC fell from 0.981 on normal images to 0.969 on TB images (Mann–Whitney $p < 0.001$), and the standard deviation of HD95 rose sevenfold. All five of the lowest-scoring masks (DSC 0.880–0.922) belonged to TB patients, and their clinical readings described a large pleural effusion, a moderate effusion with cavitation, a residual pleural peel, re-accumulating pleural fluid and TB complicated by heart failure. Figure 6 shows the typical pattern. In the worst case the reference mask follows the expected lung outline down to the diaphragm, whereas the U-Net stops at the upper edge of the effusion and labels the fluid-filled region as non-lung (mechanism M3 in Figure 2). Pixel precision stayed high on TB images while sensitivity fell, which confirms that the errors were mostly missed lung and rarely false lung.

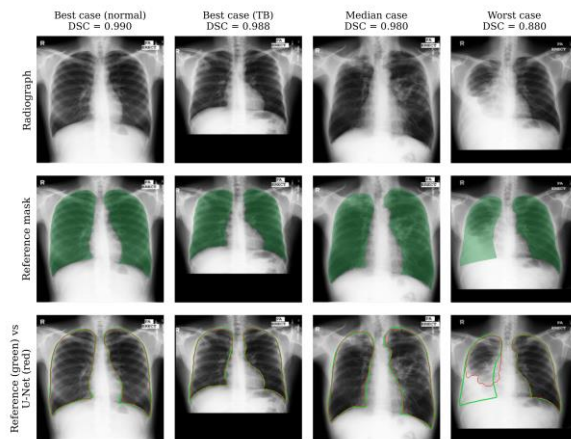


Figure 6. Representative U-Net segmentations. Columns show the best normal case, the best TB case, the case with median DSC and the worst case. Top row: radiograph; middle row: reference mask; bottom row: reference contour (green) and U-Net contour (red). In the worst case, a TB patient with a large right pleural effusion, the network excludes the opacified lower right hemithorax that the reference annotator included.

Figure 7a shows the learning curve for one representative fold. Training loss fell steeply over the first ten epochs. Held-out DSC first exceeded 0.95 at epoch 6, fluctuated during the early high-learning-rate phase of the cosine schedule [55] (with brief drops at epochs 11 and 17), stayed above 0.95 from epoch 20 onward and finished at 0.981. It did not decline late in training, so there was no sign of overfitting. Figure 7b shows the per-image DSC distributions: both U-Net subsets sit in a narrow band near 1, with the TB subset carrying a longer lower tail.

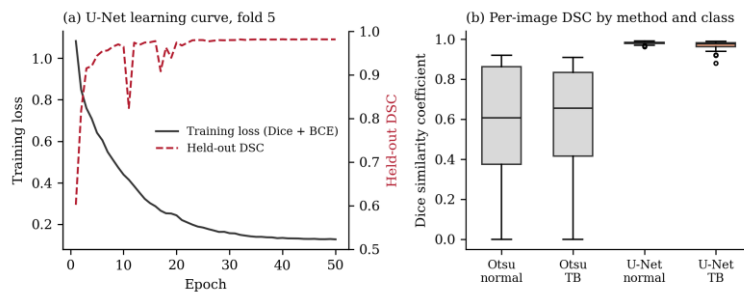


Figure 7. (a) Training loss and held-out DSC per epoch for the U-Net in fold 5. (b) Distribution of per-image DSC for the thresholding baseline and the U-Net, split by class.

Lung area agreement

Overlap scores can hide a systematic error in lung size, which matters if the mask is later used to crop the image [60]. Predicted and reference lung areas were almost perfectly correlated ($r = 0.994$; Figure 8a). On average the reference masks covered 25.5% of the image and the U-Net masks 25.1%, a bias of -0.48 percentage points with 95% limits of agreement from -1.83 to 0.87 (Figure 8b) [37]. The U-Net therefore slightly under-segmented the lungs, and slightly more so on TB images (bias -0.61) than on normal ones (-0.38). The three largest negative differences, between -3.4 and -4.1 percentage points, were all TB radiographs with pleural fluid or consolidation.

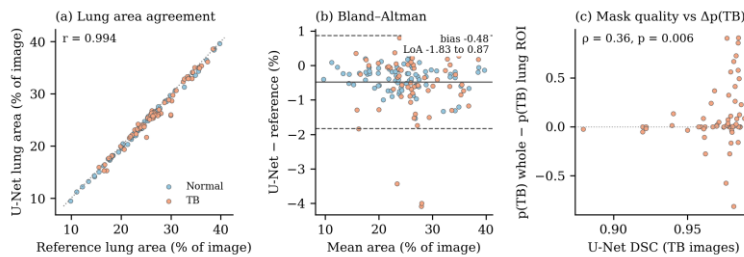


Figure 8. (a) U-Net versus reference lung area, as a percentage of the 256×256 image. (b) Bland–Altman plot with mean bias (solid line) and 95% limits of agreement (dashed lines). (c) For TB images, the change in predicted TB probability between the whole-image and lung-ROI classifiers plotted against U-Net DSC; positive values mean the lung-ROI model gave a lower probability.

Tuberculosis classification

Table 7 compares the two classifiers. The whole-image model reached an AUC of 0.946 (95% CI 0.907–0.975) and the lung-ROI model 0.882 (0.814–0.938). The difference of -0.064 was significant by DeLong’s test [35] ($p = 0.017$), so H_0 was rejected, but in the opposite direction to the conventional hypothesis H_1 . At the fixed 0.5 threshold the two models had identical accuracy, 84.1%, but they made different errors. The whole-image model detected more TB cases (sensitivity 75.9% vs 69.0%), while the lung-ROI model produced fewer false alarms (specificity 95.0% vs 90.0%). Each model was right on ten images where the other was wrong, so the McNemar test [62] found no difference at this particular threshold ($p = 1.00$). The difference between the models lies mainly in how well they rank patients across the range of thresholds, which is what the ROC curves in Figure 9a show.

Table 7. Out-of-fold TB classification performance with 95% bootstrap confidence intervals.

Metric	Whole image	U-Net lung ROI
AUC	0.946 (0.907–0.975)	0.882 (0.814–0.938)
Accuracy	0.841 (0.783–0.899)	0.841 (0.783–0.899)

Metric	Whole image	U-Net lung ROI
Sensitivity	0.759 (0.648–0.867)	0.690 (0.571–0.803)
Specificity	0.900 (0.829–0.960)	0.950 (0.893–0.988)
Precision	0.846 (0.745–0.933)	0.909 (0.810–0.979)
F1-score	0.800 (0.713–0.875)	0.784 (0.692–0.865)
Brier score	0.098	0.133

Threshold-dependent metrics use a fixed threshold of 0.5. AUC difference (lung ROI – whole image) = -0.064 , DeLong $p = 0.017$. Discordant decisions: 10 vs 10, exact McNemar $p = 1.00$. Lower Brier scores indicate better calibrated probabilities [64].

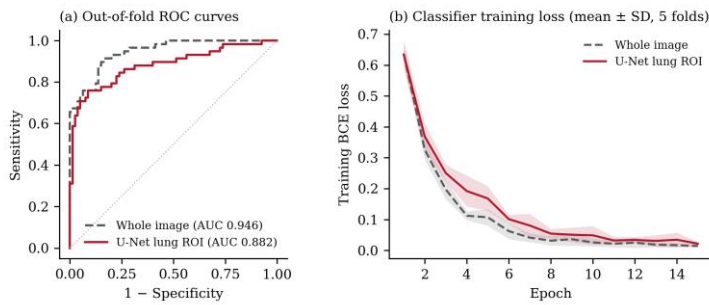


Figure 9. (a) Out-of-fold ROC curves for the whole-image and lung-ROI classifiers. (b) Training loss of the two classifiers, shown as the mean across the five folds with a shaded ± 1 SD band.

The training curves in Figure 9b show that both classifiers fitted their training data almost perfectly by the final epoch, with the lung-ROI model converging slightly more slowly. Neither model was undertrained, so the gap in AUC is unlikely to come from the optimisation schedule. Figure 10 gives the confusion matrices. The whole-image model missed 14 of 58 TB cases and the lung-ROI model 18, while false positives fell from 8 to 4. The whole-image model was also better calibrated, with a Brier score of 0.098 against 0.133 [64]. Figure 11 brings all threshold-dependent and threshold-free metrics together: only the AUC separates the models beyond overlapping intervals, while the error profile at 0.5 shows the trade of four extra missed TB cases for four fewer false alarms.

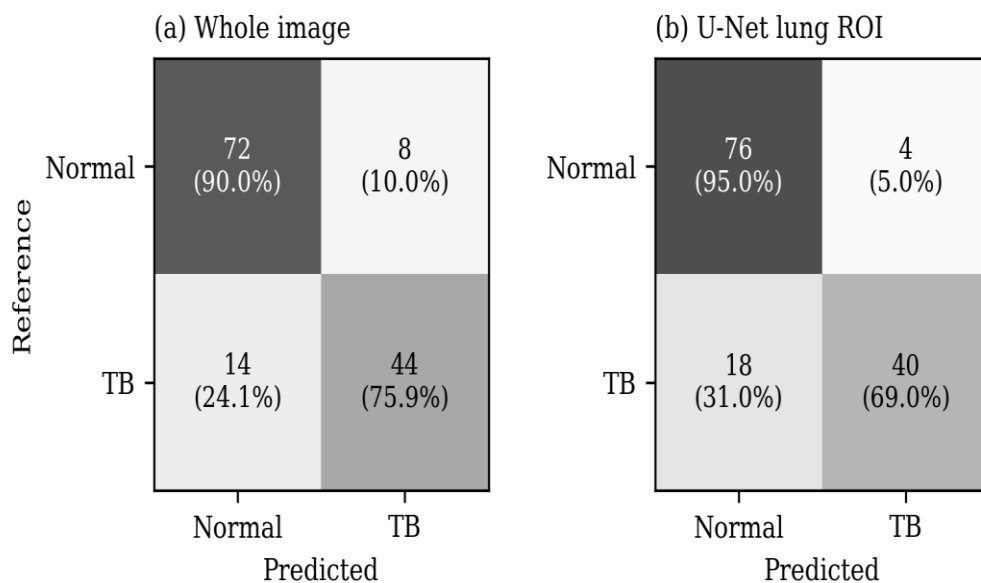


Figure 10. Confusion matrices at a threshold of 0.5 for (a) the whole-image classifier and (b) the lung-ROI classifier. Percentages are row-normalised.

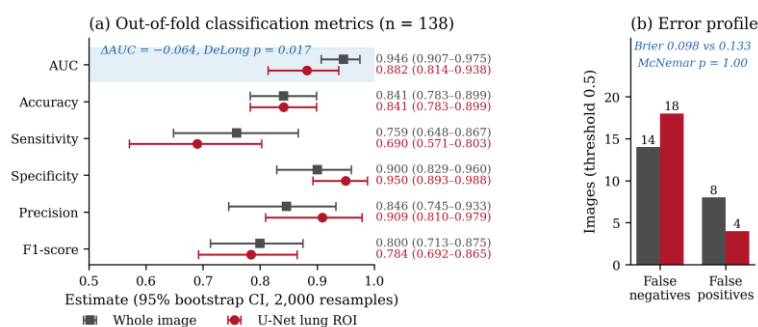


Figure 11. (a) Forest plot of out-of-fold classification metrics with 95% bootstrap confidence intervals (2,000 resamples [61]) for the whole-image (grey squares) and lung-ROI (red circles) classifiers; the shaded row marks the primary end point, compared with DeLong's test [35]. (b) Error profile at the 0.5 threshold, with Brier scores [64] and the exact McNemar test [62]. All values are taken from Table 7 and Figure 10.

Clinically relevant operating points

Figure 12a shows the out-of-fold scores. Both models placed most normal radiographs close to zero, with median probabilities of 0.010 and 0.012. The difference lay in the TB class: the whole-image model pushed most TB images towards 1, whereas the lung-ROI model left a cluster of TB images with probabilities below 0.1, where no reasonable threshold could recover them without also flagging many normal images. Figure 12b shows the consequence. As the threshold is lowered to reach high sensitivity, the specificity of the lung-ROI model collapses much earlier than that of the whole-image model.

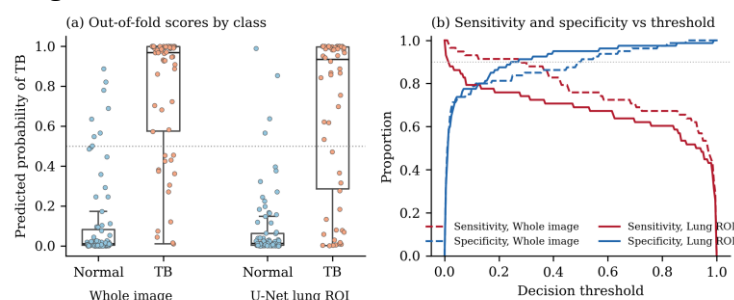


Figure 12. (a) Out-of-fold predicted probabilities of TB by true class for both classifiers, with boxes showing the median and interquartile range. (b) Sensitivity (red) and specificity (blue) as functions of the decision threshold for the whole-image (dashed) and lung-ROI (solid) classifiers; the dotted line marks 90%.

Table 8 quantifies this. When the threshold was set so that sensitivity reached at least 90%, the whole-image model kept a specificity of 83.8%, comfortably above the WHO minimum of 70% for a triage test [2], while the lung-ROI model fell to 48.8%. The bootstrap interval for the lung-ROI model was wide (24.7–82.1%), reflecting the small number of TB cases near the threshold, but its point estimate was well below the target. At the Youden-optimal threshold [63] the whole-image model reached 91.4% sensitivity with 83.8% specificity, whereas the lung-ROI model reached only 75.9% sensitivity. When specificity was held at 95%, the two models were close (72.4% vs 70.7%), which matches the observation that the lung-ROI model is competitive only when false positives are the main concern.

Table 8. Operating points derived from the out-of-fold ROC curves and projected confirmatory-test demand.

Operating point	Whole image	U-Net lung ROI
Youden-optimal threshold	0.271	0.276

Operating point	Whole image	U-Net lung ROI
Sensitivity / specificity at Youden point	0.914 / 0.838	0.759 / 0.913
Specificity at sensitivity ≥ 0.90	0.838 (0.650–0.923)	0.488 (0.247–0.821)
Sensitivity at specificity ≥ 0.95	0.724 (0.574–0.850)	0.707 (0.508–0.825)
Meets WHO triage minimum (sens ≥ 0.90 , spec ≥ 0.70)	Yes	No
Projected confirmatory tests per 1,000 screened, $\pi = 5\%$ (Eq. 5)	199	531
Projected tests per TB case detected, $\pi = 5\%$ (Eq. 6)	4.4	11.8
Projected share of confirmatory tests avoided, $\pi = 5\%$	80.1%	46.9%

Values in parentheses are 95% bootstrap confidence intervals (2,000 resamples). Thresholds were read from the pooled out-of-fold ROC curves and would need confirmation on independent data. Projections use $Se = 0.90$ and the specificity at sensitivity ≥ 0.90 for each model.

Figure 13 places these operating points in ROC space beside the WHO target region and translates them into programme terms with Eqs. (5) and (6). At 5% prevalence the whole-image pipeline would need about 199 confirmatory tests per 1,000 people screened (4.4 per case detected) and the lung-ROI pipeline about 531 (11.8 per case); the ratio holds from 1% (18.8 vs 57.3 per case) to 10% prevalence (2.6 vs 6.1). In the framework of Qin *et al.* [6], the whole-image model would avoid about 80% of confirmatory tests at 90% sensitivity, against about 47% for the lung-ROI model.

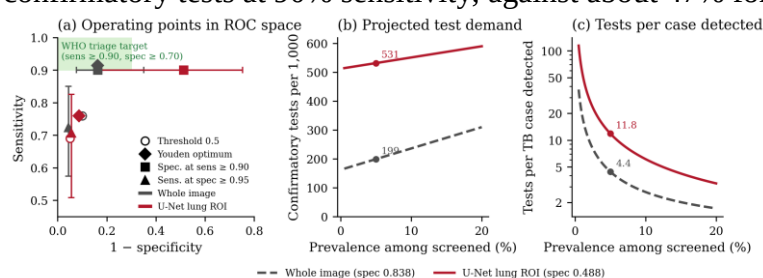


Figure 13. Clinical operating characteristics. (a) Operating points of both classifiers in ROC space: threshold 0.5 (open circles), Youden optimum [63] (diamonds), specificity at sensitivity ≥ 0.90 (squares, horizontal bars are 95% CIs) and sensitivity at specificity ≥ 0.95 (triangles, vertical bars are 95% CIs); the green region is the WHO minimum for a triage test [2]. (b) Projected confirmatory tests per 1,000 people screened (Eq. 5) and (c) tests per TB case detected (Eq. 6, log scale) as a function of prevalence, at 90% sensitivity. Projections are illustrative and assume that the pooled out-of-fold operating characteristics transfer to the screened population.

Consistency across folds

Table 9 and Figure 14a list the results of each fold. The U-Net was stable, with fold-wise DSC varying by less than 0.01. The classifiers varied more, as expected with 27 or 28 test images per fold [27]. The whole-image model had the higher AUC in four of the five folds, and the largest gap occurred in fold 4 (0.972 vs 0.795). In fold 5 the lung-ROI model was slightly better (0.909 vs 0.886). The mean fold-wise AUC was 0.942 ± 0.036 for the whole image and 0.882 ± 0.050 for the lung ROI. A Wilcoxon test on five paired fold values has little power and was not significant ($p = 0.125$), so the pooled DeLong comparison [35] remains the primary test.

Table 9. Results for each cross-validation fold.

Fold	<i>n</i> (TB)	U-Net DSC	U-Net HD95	AUC whole image	AUC lung ROI	Accuracy whole / ROI
1	28 (12)	0.974 0.021	± 3.87	0.974	0.922	0.893 / 0.929
2	28 (12)	0.978 0.012	± 2.98	0.932	0.896	0.786 / 0.786
3	28 (12)	0.973 0.017	± 3.53	0.948	0.885	0.929 / 0.857
4	27 (11)	0.975 0.015	± 3.95	0.972	0.795	0.889 / 0.852
5	27 (11)	0.980 0.006	± 2.32	0.886	0.909	0.704 / 0.778
Mean ± SD	—	0.976	3.33	0.942 0.036	± 0.882 0.050	± 0.840 / 0.840

HD95 in pixels at 256×256 . Accuracy at a threshold of 0.5.

Is the loss of discrimination caused by segmentation errors?

If poor masks were responsible, the lung-ROI model should have lost most on the images the U-Net segmented worst. The opposite was observed. Among the 58 TB images, the drop in predicted probability Δp was positively correlated with DSC (Spearman $\rho = 0.36$, $p = 0.006$; Figure 8c). The large drops, some exceeding 0.8, all occurred on images with DSC above 0.96, while the five worst-segmented images showed almost no change in probability. The subgroup analysis pointed in the same direction (Figure 14b). Among the 13 TB images whose readings mentioned pleural disease, the group with the poorest masks, both models detected 10 (76.9%). Among the 24 images described as inactive, scarred, calcified or stable disease, sensitivity fell from 66.7% with the whole image to 54.2% with the lung ROI, and among the 45 images without pleural disease it fell from 75.6% to 66.7%. The Wilson intervals [65] overlap, as expected with these subgroup sizes, so the pattern is hypothesis-generating rather than confirmatory [29]. The additional misses of the lung-ROI model were nevertheless concentrated in well-segmented images of old, fibrotic or calcified disease, which tends to lie at the apices and near the hila [20, 21], consistent with hypothesis H2.

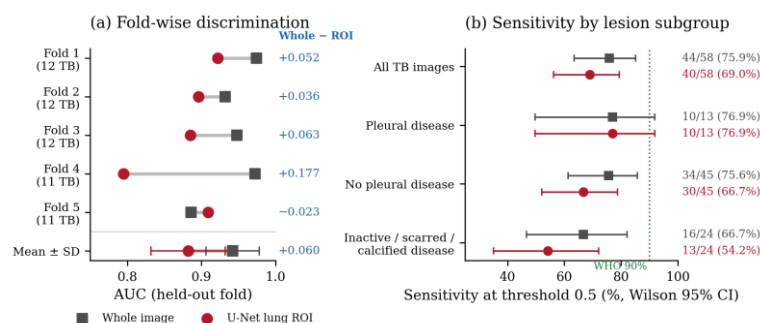


Figure 14. Robustness and subgroup analysis. (a) Fold-wise AUC for the whole-image (squares) and lung-ROI (circles) classifiers, with the per-fold difference; the bottom row shows the mean $\pm$ SD across folds. (b) Sensitivity at the 0.5 threshold by lesion subgroup with Wilson 95% score intervals [65]; counts are detected/total TB images. The dotted line marks the WHO 90% sensitivity target [2]. Subgroups were defined by keywords in the clinical readings and may overlap.

Grad-CAM analysis

Figure 15 shows Grad-CAM maps [38] for four TB radiographs. In the two cases with extensive upper-zone disease (columns a and b), both models gave a probability close to 1 and both attended to the affected upper zones. In the lung-ROI model, part of the activation also fell on the masked mediastinal gap between the lungs, which suggests that the network used the outline of the mask as a feature. Column c is a patient with inactive scarring in the left upper lobe. The whole-image model placed its activation over the left upper zone but assigned a probability just below threshold, while the lung-ROI model crossed the threshold with activation over the opposite lung. Column d, another left upper lobe scar, was missed by both models. The whole-image model at least attended to the left apex, close to the clavicle, but the lung-ROI model concentrated on a small area in the right mid-zone. These maps are coarse and should be read with the caution that benchmark studies recommend [30].

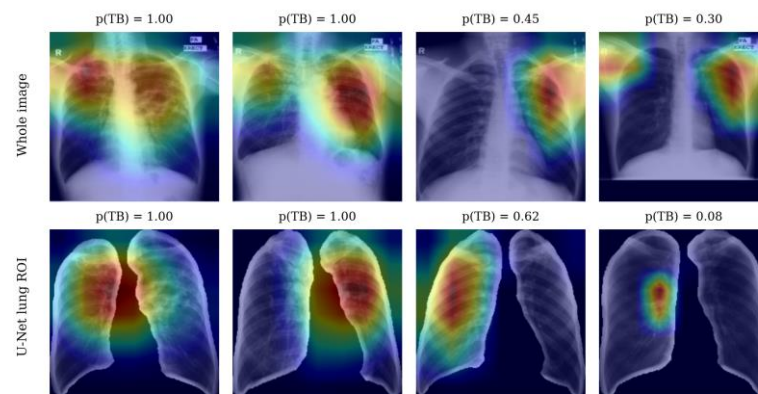


Figure 15. Grad-CAM maps [38] for four TB radiographs (columns a–d). Top row: whole-image classifier; bottom row: lung-ROI classifier. The number above each panel is the out-of-fold predicted probability of TB. Clinical readings: (a) right upper lobe fibrotic disease with calcification and extensive infiltrates; (b) inactive TB scars; (c) left upper lobe inactive scars with volume loss; (d) inactive left upper lobe scars, unchanged for two years.

Discussion

Principal findings

This study asked a narrow question: does putting a well-performing U-Net in front of a CNN improve TB detection on chest radiographs? On the Montgomery County dataset the answer was no. The U-Net segmented the lungs with a DSC of 0.976 and almost no bias in lung area, yet the classifier trained on lung crops ranked patients significantly worse than the same classifier trained on whole radiographs (AUC 0.882 vs 0.946). Accuracy at the default threshold was the same, which shows how a single accuracy figure can hide a real difference. When both models were set to the 90% sensitivity required of a triage test [2], the whole-image model kept a specificity of 83.8% and the lung-ROI model fell to 48.8%. In a screening programme this would mean sending roughly three times as many healthy people for confirmatory testing to reach the same case detection, a difference with direct cost implications in high-burden settings such as Pakistan, where the gap between incidence and notification is among the largest in the world [1]. Table 10 summarises how the findings answer the four research questions.

Table 10. Research questions, evidence and answers.

Research question	Key evidence	Answer
RQ1: How accurate is U-Net segmentation, and where does it fail?	DSC 0.976 overall; 0.981 normal vs 0.969 TB ($p < 0.001$); five worst masks all TB with pleural disease; area bias -0.48 pp	Accurate on average, but errors concentrate in abnormal radiographs, especially pleural fluid (G2)
RQ2: Does a lung-ROI input improve discrimination under matched conditions?	AUC 0.946 vs 0.882, Δ AUC -0.064 , DeLong $p = 0.017$; better in 4 of 5 folds	No. H1 is not supported; the whole image is better (G1)
RQ3: Does either pipeline meet the WHO triage target?	Specificity at $\geq 90\%$ sensitivity 83.8% vs 48.8%; projected 199 vs 531 tests per 1,000 at 5% prevalence	Only the whole-image pipeline (G3)
RQ4: Is the loss explained by mask errors?	Δp rises with DSC ($\rho = 0.36$, $p = 0.006$); pleural subgroup unchanged; extra misses in inactive or fibrotic disease	No. Evidence is removed by design, consistent with H2 (G4)

Why masking reduced performance

The obvious explanation, that poor masks cut away lesions, does not fit the data. The worst masks, in pleural effusion, showed the failure mode described for U-Nets on severely abnormal radiographs [25, 26], yet these patients were detected equally well by both classifiers, and the probability drop was larger when the mask was better. The information loss therefore comes from masking itself, by design.

Three properties of the lung-ROI input can account for this, corresponding to mechanisms M1, M2 and M4 in Figure 2. First, the mask removes the mediastinum, the hila and the trachea. Hilar lymph node enlargement and calcified hilar nodes are typical of TB, and upper-lobe volume loss in fibrotic disease shows up as tracheal deviation and hilar elevation [20, 21]. None of these can be seen in the lung-ROI input. Second, the apices are partly hidden behind the clavicles and first ribs, the mask edge runs close to them, and old apical scars therefore lie at the border of the cropped input where the convolutional features have little surrounding context. Third, cropping to the bounding box and resizing to a fixed size makes every pair of lungs the same size. A shrunken, fibrotic upper lobe looks smaller than its partner only in relative terms, and the absolute reduction in lung volume, which radiologists use as a sign of chronic disease, is normalised away. The concentration of extra misses among inactive and fibrotic cases is consistent with all three mechanisms. It also mirrors the weaker performance of commercial CAD products in people with prior TB [6, 7], which suggests that how a pipeline handles old, extrapulmonary-adjacent findings deserves more attention than it has received.

The Grad-CAM maps add a detail: in several lung-ROI cases activation spread into the masked mediastinal gap, suggesting that the network learned from the shape of the mask itself. This is another shortcut [19]; masking does not remove non-parenchymal information, it only changes which is available.

Comparison with previous studies

Our results differ from those of Rahman *et al.* [16], who found that segmented lungs improved classification. Several differences could account for this. Their study pooled several public datasets into a much larger collection, in which removing scanner- and

site-specific cues outside the lungs probably matters more than in a single-site set. Shortcut learning of this kind has been documented clearly in multi-site chest radiograph models [17, 18], and the COVID-19 literature shows that pooled sources can be distinguished even without the lungs [23] and that source bias can persist after segmentation [22]. In a single-source dataset such as Montgomery there are fewer site cues for masking to remove, and the cost of losing extrapulmonary evidence dominates. Their best whole-image and segmented results also came from different networks, and their disease spectrum may have contained more active parenchymal disease and fewer old fibrotic lesions. The cardiomegaly study of Sogancioglu *et al.* [32], in which segmentation helped, concerns a finding that is defined by the anatomical structures being segmented; TB is not. The comparison suggests that the value of segmentation depends on the data and the disease, and that it should be tested for each new dataset under matched conditions.

For context, both of our CNN pipelines exceeded the classical system of Jaeger *et al.* (AUC 0.87) [8], and the whole-image model compares favourably with other published results on this dataset (Table 2) [13, 42], although protocols differ and small-sample estimates carry wide error bars [27, 41].

Could the whole-image model be using shortcuts?

We must also consider whether the whole-image model performed well for the wrong reasons. The TB patients in this cohort were on average 14 years older and more often male, and fewer TB images had been acquired in landscape orientation (Table 3). CNNs can estimate age and sex from chest radiographs [42] and even encode attributes invisible to radiologists from cropped images [33], and demographic imbalance can bias chest radiograph classifiers [48, 49]. A whole-image CNN could therefore in principle learn some of these cues from bone density, breast shadows or image geometry. We cannot rule this out with 138 images. The Grad-CAM maps, however, placed most of the whole-image model's activation over the lung fields and upper zones and not over the shoulders, the labels or the image edges. The extra cases it detected were mostly old fibrotic lesions, which are radiologically real findings [21]. Together these observations argue against a dominant shortcut, although they do not exclude a contribution from demographic cues; and because such cues can survive cropping [33], the lung-ROI model is not immune to them either.

Implications for TB screening pipelines

Four practical points follow. First, segmentation used before TB classifiers should be evaluated separately on abnormal radiographs, with lung-area bias reported [25, 60]. Second, a high Dice score does not guarantee that the downstream classifier benefits; the effect of the mask on classification must be measured directly, in a matched design that guards against leakage [43]. Third, if masking is used to suppress shortcuts, designs that keep the hila, mediastinum and lung size are preferable, such as a wider dilation margin, a soft lung-attention map or a global–local two-stream network [47]. Fourth, models intended for triage should be compared at the high-sensitivity operating point where they will be used [2, 3], and the comparison should be translated into confirmatory-test demand [6], since accuracy at an arbitrary threshold of 0.5 concealed the main difference in this study.

Limitations

The dataset is small and single-site, so intervals are wide and generalisation to other populations, devices or disease spectra is uncertain [27]. The operating thresholds were read from the pooled out-of-fold predictions and would need confirmation on independent data, and the test-demand projections assume that these operating characteristics transfer to a screened population. The keyword-based subgroup analysis relied on short free-text readings and should be read as hypothesis-generating. A few

TB images in the dataset are follow-up studies of the same patient, and the clinical readings indicate at least two such patients whose images fell into different folds; because the published file names do not identify patients, we could not group them, and this may have slightly favoured both classifiers [43]. To fit the computational budget, the U-Net was trained at 128×128 rather than with a self-configuring framework such as nnU-Net [46], and we tested only one CNN architecture, one dilation margin and one training schedule without hyper-parameter search. Grad-CAM maps are coarse at the resolution of the last ResNet stage and show where a model looks, not what it understands [30]. Finally, the reference masks were drawn by a single group of annotators and may themselves contain variation, particularly in how they treat lung regions obscured by pleural fluid [24].

Future work

Natural next steps are to repeat the comparison on the larger Shenzhen dataset [39] and on multi-site collections such as TBX11K [40], where shortcut removal may tip the balance the other way; to vary the dilation margin systematically and compare hard masks with soft lung-attention maps and two-stream models [47]; to train segmentation networks on masks that include opacified lung, so that pleural and consolidated regions are kept [26]; and to add external validation on images from a different hospital, ideally from a high-burden setting such as Pakistan [1]. A prospective reader study, following the programmatic evaluation designs already used for commercial CAD [6, 7], would show whether either model changes radiologists' decisions in a screening workflow.

Conclusion

A U-Net segmented the lungs on the Montgomery County chest radiographs with a mean Dice coefficient of 0.976 and negligible area bias, but feeding a ResNet-18 only the segmented lungs lowered its AUC for TB detection from 0.946 to 0.882 and its specificity at 90% sensitivity from 83.8% to 48.8%, so that only the whole-image pipeline met the WHO triage target [2]. The loss did not come from segmentation errors. It was largest on well-segmented images and concentrated in inactive and fibrotic disease, which points to the removal of hilar, mediastinal and apical evidence and of absolute lung size [20, 21]. On this dataset, lung segmentation removed useful diagnostic information. As CAD is adopted for TB screening and triage [3], segmentation should be treated as a hypothesis to test in each pipeline, not as a default preprocessing step.

References

- [1] World Health Organization. *Global Tuberculosis Report 2025*. Geneva: World Health Organization; 2025.
- [2] World Health Organization. *High-Priority Target Product Profiles for New Tuberculosis Diagnostics: Report of a Consensus Meeting*. Geneva: World Health Organization; 2014.
- [3] World Health Organization. *WHO Consolidated Guidelines on Tuberculosis. Module 2: Screening – Systematic Screening for Tuberculosis Disease*. Geneva: World Health Organization; 2021.
- [4] Hwang EJ, Park S, Jin KN, Kim JI, Choi SY, Lee JH, et al. Development and validation of a deep learning-based automatic detection algorithm for active pulmonary tuberculosis on chest radiographs. *Clin Infect Dis*. 2019;69(5):739–747. doi:10.1093/cid/ciy967
- [5] Kazemzadeh S, Yu J, Jamshy S, et al. Deep learning detection of active pulmonary tuberculosis at chest radiography matched the clinical performance of radiologists. *Radiology*. 2023;306(1):124–137. doi:10.1148/radiol.212213
- [6] Qin ZZ, Ahmed S, Sarker MS, Paul K, Adel ASS, Naheyant T, et al. Tuberculosis detection from chest x-rays for triaging in a high tuberculosis-burden setting: an

- evaluation of five artificial intelligence algorithms. *Lancet Digit Health*. 2021;3(9):e543–e554.
- [7] Codlin AJ, Dao TP, Vo LNQ, Forse RJ, Van Truong V, Dang HM, et al. Independent evaluation of 12 artificial intelligence solutions for the detection of tuberculosis. *Sci Rep*. 2021;11:23895. doi:10.1038/s41598-021-03265-0
- [8] Jaeger S, Karargyris A, Candemir S, Folio L, Siegelman J, Callaghan F, et al. Automatic tuberculosis screening using chest radiographs. *IEEE Trans Med Imaging*. 2014;33(2):233–245. doi:10.1109/TMI.2013.2284099
- [9] Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, et al. A survey on deep learning in medical image analysis. *Med Image Anal*. 2017;42:60–88.
- [10] Lakhani P, Sundaram B. Deep learning at chest radiography: automated classification of pulmonary tuberculosis by using convolutional neural networks. *Radiology*. 2017;284(2):574–582.
- [11] Rajpurkar P, Irvin J, Zhu K, Yang B, Mehta H, Duan T, et al. CheXNet: radiologist-level pneumonia detection on chest X-rays with deep learning. arXiv:1711.05225; 2017.
- [12] Wang X, Peng Y, Lu L, Lu Z, Bagheri M, Summers RM. ChestX-ray8: hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In: *Proc IEEE Conf Comput Vis Pattern Recognit (CVPR)*; 2017. p. 2097–2106.
- [13] Pasa F, Golkov V, Pfeiffer F, Cremers D, Pfeiffer D. Efficient deep network architectures for fast chest X-ray tuberculosis screening and visualization. *Sci Rep*. 2019;9:6268. doi:10.1038/s41598-019-42557-4
- [14] Ronneberger O, Fischer P, Brox T. U-Net: convolutional networks for biomedical image segmentation. In: *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. LNCS 9351. Springer; 2015. p. 234–241.
- [15] Candemir S, Jaeger S, Palaniappan K, Musco JP, Singh RK, Xue Z, et al. Lung segmentation in chest radiographs using anatomical atlases with nonrigid registration. *IEEE Trans Med Imaging*. 2014;33(2):577–590.
- [16] Rahman T, Khandakar A, Kadir MA, Islam KR, Islam KF, Mazhar R, et al. Reliable tuberculosis detection using chest X-ray with deep learning, segmentation and visualization. *IEEE Access*. 2020;8:191586–191601. doi:10.1109/ACCESS.2020.3031384
- [17] Zech JR, Badgeley MA, Liu M, Costa AB, Titano JJ, Oermann EK. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. *PLoS Med*. 2018;15(11):e1002683.
- [18] DeGrave AJ, Janizek JD, Lee SI. AI for radiographic COVID-19 detection selects shortcuts over signal. *Nat Mach Intell*. 2021;3:610–619.
- [19] Geirhos R, Jacobsen JH, Michaelis C, Zemel R, Brendel W, Bethge M, et al. Shortcut learning in deep neural networks. *Nat Mach Intell*. 2020;2:665–673.
- [20] Nachiappan AC, Rahbar K, Shi X, et al. Pulmonary tuberculosis: role of radiology in diagnosis and management. *RadioGraphics*. 2017;37(1):52–72.
- [21] Burrill J, Williams CJ, Bain G, Conder G, Hine AL, Misra RR. Tuberculosis: a radiologic review. *RadioGraphics*. 2007;27(5):1255–1273.
- [22] Teixeira LO, Pereira RM, Bertolini D, Oliveira LS, Nanni L, Cavalcanti GDC, et al. Impact of lung segmentation on the diagnosis and explanation of COVID-19 in chest X-ray images. *Sensors*. 2021;21(21):7116. doi:10.3390/s21217116
- [23] Maguolo G, Nanni L. A critic evaluation of methods for COVID-19 automatic detection from X-ray images. *Inf Fusion*. 2021;76:1–7.
- [24] Candemir S, Antani S. A review on lung boundary detection in chest X-rays. *Int J Comput Assist Radiol Surg*. 2019;14(4):563–576.

- [25]Nishio M, Fujimoto K, Togashi K. Lung segmentation on chest x-ray images in patients with severe abnormal findings using deep learning. *arXiv:1908.07704*; 2019.
- [26]Alam MS, Wang D, Arzhaeva Y, Ende JA, Kao J, Silverstone L, et al. Attention-based multi-residual network for lung segmentation in diseased lungs with custom data augmentation. *Sci Rep*. 2024;14. doi:10.1038/s41598-024-79494-w
- [27]Varoquaux G. Cross-validation failure: small sample sizes lead to large error bars. *NeuroImage*. 2018;180:68–77. doi:10.1016/j.neuroimage.2017.06.061
- [28]Vabalas A, Gowen E, Poliakoff E, Casson AJ. Machine learning algorithm validation with a limited sample size. *PLoS One*. 2019;14(11):e0224365.
- [29]Oakden-Rayner L, Dunnmon J, Carneiro G, Ré C. Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. In: *Proc ACM Conf Health Inference Learn (CHIL)*; 2020. p. 151–159. doi:10.1145/3368555.3384468
- [30]Saporta A, Gui X, Agrawal A, Pareek A, Truong SQH, Nguyen CDT, et al. Benchmarking saliency methods for chest X-ray interpretation. *Nat Mach Intell*. 2022;4:867–878.
- [31]Novikov AA, Lenis D, Major D, Hladůvka J, Wimmer M, Bühler K. Fully convolutional architectures for multiclass segmentation in chest radiographs. *IEEE Trans Med Imaging*. 2018;37(8):1865–1876.
- [32]Sogancioglu E, Murphy K, Calli E, Scholten ET, Schalekamp S, van Ginneken B. Cardiomegaly detection on chest radiographs: segmentation versus classification. *IEEE Access*. 2020;8:94631–94642. doi:10.1109/ACCESS.2020.2995567
- [33]Gichoya JW, Banerjee I, Bhimireddy AR, Burns JL, et al. AI recognition of patient race in medical imaging: a modelling study. *Lancet Digit Health*. 2022;4(6):e406–e414.
- [34]Tejani AS, Klontzas ME, Gatti AA, Mongan JT, Moy L, Park SH, et al. Checklist for Artificial Intelligence in Medical Imaging (CLAIM): 2024 update. *Radiol Artif Intell*. 2024;6(4):e240300.
- [35]DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*. 1988;44(3):837–845.
- [36]Taha AA, Hanbury A. Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool. *BMC Med Imaging*. 2015;15:29.
- [37]Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet*. 1986;1(8476):307–310.
- [38]Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: visual explanations from deep networks via gradient-based localization. In: *Proc IEEE Int Conf Comput Vis (ICCV)*; 2017. p. 618–626.
- [39]Jaeger S, Candemir S, Antani S, Wang YXJ, Lu PX, Thoma G. Two public chest X-ray datasets for computer-aided screening of pulmonary diseases. *Quant Imaging Med Surg*. 2014;4(6):475–477.
- [40]Liu Y, Wu YH, Ban Y, Wang H, Cheng MM. Rethinking computer-aided tuberculosis diagnosis. In: *Proc IEEE/CVF Conf Comput Vis Pattern Recognit (CVPR)*; 2020. p. 2646–2655.
- [41]Hansun S, Argha A, Liaw ST, Celler BG, Marks GB. Machine and deep learning for tuberculosis detection on chest X-rays: systematic literature review. *J Med Internet Res*. 2023;25:e43154. doi:10.2196/43154
- [42]Gozes O, Greenspan H. Deep feature learning from a hospital-scale chest X-ray dataset with application to TB detection on a small-scale dataset. In: *Proc 41st Annu Int Conf IEEE Eng Med Biol Soc (EMBC)*; 2019. p. 4076–4079.

- [43]Kapoor S, Narayanan A. Leakage and the reproducibility crisis in machine-learning-based science. *Patterns*. 2023;4(9):100804. doi:10.1016/j.patter.2023.100804
- [44]Varoquaux G, Cheplygina V. Machine learning for medical imaging: methodological failures and recommendations for the future. *NPJ Digit Med*. 2022;5:48. doi:10.1038/s41746-022-00592-y
- [45]Roberts M, Driggs D, Thorpe M, Gilbey J, Yeung M, Ursprung S, et al. Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. *Nat Mach Intell*. 2021;3:199–217.
- [46]Isensee F, Jaeger PF, Kohl SAA, Petersen J, Maier-Hein KH. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nat Methods*. 2021;18(2):203–211.
- [47]Guan Q, Huang Y, Zhong Z, Zheng Z, Zheng L, Yang Y. Thorax disease classification with attention guided convolutional neural network. *Pattern Recognit Lett*. 2020;131:38–45.
- [48]Larrazabal AJ, Nieto N, Peterson V, Milone DH, Ferrante E. Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. *Proc Natl Acad Sci USA*. 2020;117(23):12592–12594.
- [49]Seyyed-Kalantari L, Zhang H, McDermott MBA, Chen IY, Ghassemi M. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. *Nat Med*. 2021;27(12):2176–2182.
- [50]He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: *Proc IEEE Conf Comput Vis Pattern Recognit (CVPR)*; 2016. p. 770–778.
- [51]Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L. ImageNet: a large-scale hierarchical image database. In: *Proc IEEE Conf Comput Vis Pattern Recognit (CVPR)*; 2009. p. 248–255.
- [52]Ioffe S, Szegedy C. Batch normalization: accelerating deep network training by reducing internal covariate shift. In: *Proc 32nd Int Conf Mach Learn (ICML)*; 2015. p. 448–456.
- [53]Milletari F, Navab N, Ahmadi SA. V-Net: fully convolutional neural networks for volumetric medical image segmentation. In: *Proc Fourth Int Conf 3D Vision (3DV)*; 2016. p. 565–571.
- [54]Kingma DP, Ba J. Adam: a method for stochastic optimization. In: *Int Conf Learn Represent (ICLR)*; 2015. arXiv:1412.6980.
- [55]Loshchilov I, Hutter F. SGDR: stochastic gradient descent with warm restarts. In: *Int Conf Learn Represent (ICLR)*; 2017.
- [56]Shorten C, Khoshgoftaar TM. A survey on image data augmentation for deep learning. *J Big Data*. 2019;6:60.
- [57]Zuiderveld K. Contrast limited adaptive histogram equalization. In: Heckbert PS, editor. *Graphics Gems IV*. San Diego: Academic Press; 1994. p. 474–485.
- [58]Otsu N. A threshold selection method from gray-level histograms. *IEEE Trans Syst Man Cybern*. 1979;9(1):62–66.
- [59]Loshchilov I, Hutter F. Decoupled weight decay regularization. In: *Int Conf Learn Represent (ICLR)*; 2019.
- [60]Maier-Hein L, Reinke A, Godau P, Tizabi MD, Buettner F, Christodoulou E, et al. Metrics reloaded: recommendations for image analysis validation. *Nat Methods*. 2024;21(2):195–212.
- [61]Efron B, Tibshirani RJ. *An Introduction to the Bootstrap*. New York: Chapman & Hall; 1993.
- [62]McNemar Q. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*. 1947;12(2):153–157.
- [63]Youden WJ. Index for rating diagnostic tests. *Cancer*. 1950;3(1):32–35.

- [64]Brier GW. Verification of forecasts expressed in terms of probability. *Mon Weather Rev.* 1950;78(1):1–3.
- [65]Wilson EB. Probable inference, the law of succession, and statistical inference. *J Am Stat Assoc.* 1927;22(158):209–212.
- [66]Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, et al. PyTorch: an imperative style, high-performance deep learning library. In: *Adv Neural Inf Process Syst (NeurIPS)*. 2019;32:8024–8035.