IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 3 December 2024, accepted 17 January 2025, date of publication 22 January 2025, date of current version 28 February 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3532631

RESEARCH ARTICLE

Tuberculosis Lesion Segmentation Improvement in X-Ray Images Using Contextual Background Label

SAHASAT KHUMANG[1], SUPAPORN KANSOMKEAT[1], WIWATANA TANOMKIAT[2], AND SATHIT INTAJAG[1]

[1]Division of Computational Science, Faculty of Science, Prince of Songkla University, Songkhla 90110, Thailand
[2]Department of Radiology, Faculty of Medicine, Prince of Songkla University, Songkhla 90110, Thailand

Corresponding author: Sathit Intajag (sathit.i@psu.ac.th)

This work was supported by the Faculty of Science Research Fund, Prince of Songkla University, under Contract 1-2565-02-009.

ABSTRACT Pulmonary tuberculosis (PTB) is a serious, potentially fatal, infectious disease that primarily affects the lungs, and poses a significant threat to public health. To detect PTB at an early stage by screening chest X-Ray (CXR) images for tuberculosis (TB) lesions, we propose a semantic segmentation scheme that uses a deep learning algorithm. However, this scheme requires high-quality training data. To improve the TB-lesion segmentation model performance, a contextual background label process was designed for decomposing the heterogeneous CXR image background. From the designed process, five background subclasses consisting of: lung, mediastinum, body, doc, and background provided to modify the ground truth data for use in training the segmentation models, which was designed in four models to assess the performance improvement of the proposed scheme. The experimental results confirmed the applicability of the designed schemes. The TB-lesion segmentation models demonstrated improvements in terms of reduced false positives and better visualizations of the shape and location of TB lesions than the visualized approximation from the classification methods. The proposed model demonstrated highest scores of 88.68% on Dice, 83.55% on Intersection over Union, and 98.64% on precision for TB-lesion detection. The proposed models were externally validated to demonstrate their generalizability. They returned sensitivity, specificity and accuracy scores of 89.00%, 95.00% and 90.00%.

INDEX TERMS TB-lesion segmentation, contextual background label, CXR image, decomposing heterogeneous background.

I. INTRODUCTION

In 2022, tuberculosis (TB) was ranked as the world's second leading cause of mortality after coronavirus disease 2019. TB remains the foremost cause of death attributable to a single infectious agent. Despite being a preventable disease, over 10 million individuals contract TB annually. All Member States of the United Nations (UN) and the World Health Organization (WHO) have set themselves the goal of ending the global TB epidemic by 2030. To achieve the goal, urgent action is imperative [1].

The associate editor coordinating the review of this manuscript and approving it for publication was Essam A. Rashed.

TB primarily affects the lungs (pulmonary tuberculosis: PTB) but can also involve other organs. Caused by *Mycobacterium tuberculosis*, it is transmitted through the air when individuals with active TB expel the bacterium, typically by coughing. Consequently, early screening and diagnosis are critical for improving survival rates and patient outcomes. Chest X-Ray (CXR) remains a cornerstone of TB screening and diagnosis due to its cost-effectiveness and widespread availability [2]. However, interpreting CXR images is challenging and necessitates the expertise of experienced radiologists to read the radiologic signs [3], [4], [5]. To mitigate these challenges, computer-aided diagnosis (CAD) systems have been developed to assist radiologists

in analyzing CXR images from large-scale PTB screenings [6], [7], [8], [9]. In recent years, deep learning (DL) techniques have gained prominence in the field of medical image segmentation [10], [11], [12], [13]. Employing DL-based image segmentation to delineate critical structures in medical images aids clinicians in making accurate diagnoses and formulating effective treatment plans.

Image segmentation is a pixel level classification that maps a corresponding category to each pixel in an image. It plays critical roles in medical image analysis, extracting brain tumor boundaries [14], [15], skin cancer boundaries [16], [17], identifying retinal vessels [18], [19], [20], and in the segmentation of tuberculosis-consistent lesions [21]. Automatic detection of TB lesions based on DL has demonstrated especially good performances in detecting and identifying PTB in CXR.

Most research into TB diagnostics has involved TB classification [22], [23], and lung segmentation [24]. While Teixeira et al. demonstrated the impact of lung segmentation in COVID-19 identification [25], Rahman et al. improved the performance of TB detection with DenseNet201 by applying the U-net model to lung segmentation [26]. Some research works have been dedicated to TB-lesion segmentation, such as in [21], [27], and [28]. Griffin et al. used region-based CNN to detect and segment four TB-lesions [27]. Rajaraman et al. proposed a model for semantic segmentation of TB images that was trained by using coarse bounding box annotations. However, training with the bounding box annotations could impact overall semantic segmentation performance [21]. Moreover, Rajaraman et al. trained models on fine-grained annotations of TB-consistent lesions and constructed their ensembles for semantically segmenting TB-consistent lesions [28]. However, they employed several assembled networks, which might influence system performances more than a single network.

In this paper, TB-lesion segmentation was designed to improve the CAD system. Motivated by work in [25] and [26] on lung segmentation to outline regions of interest before classification tasks, the proposed method achieves enhanced segmentation performances using contextual background subclasses. Furthermore, Sogancioglu et al. [29] concluded that segmentation-based models required 100 times fewer annotated chest radiographs to achieve a substantially better performance, while also producing more interpretable results. Nonetheless, the developed segmentation model used a lot of resources to modify the training dataset.

The proposed method is based on two processes: the generation of contextual background subclasses and the segmentation of images of TB lesions (Figure 1). To create contextual background subclasses, 50 images from the Shenzhen dataset [30], [31] were manually re-annotated, placing their background ground-truths into five subclasses that included lung, mediastinum, body, doc, and background. The re-annotated images were employed to design the background subclass generator model with the DeepLabV3+ [32]. The generator model established the background subclasses for the remaining CXR images in the dataset. Finally, the TB lesions from the dataset were modified with the background subclasses for use in the TB-lesion segmentation process.

From the developed background subclasses of the Shenzhen dataset, 330 TB images and 73 non-TB images were modified. The modified data were employed to design the TB-lesion segmentation by using a regular semantic network implemented with DeepLabV3+ architecture. The network architecture was investigated with different sizes of encoding backbones: ResNet18, ResNet50, [33] and InceptionResNetV2 [34]. Internal evaluation demonstrated that our proposed models consistently outperformed state-of-the-art TB lesion segmentation techniques. To assess external generalizability, the models were validated on the Montgomery County and TBX11K datasets at the image level. The results were impressive, with sensitivity, specificity, and accuracy exceeding 71% in all cases.

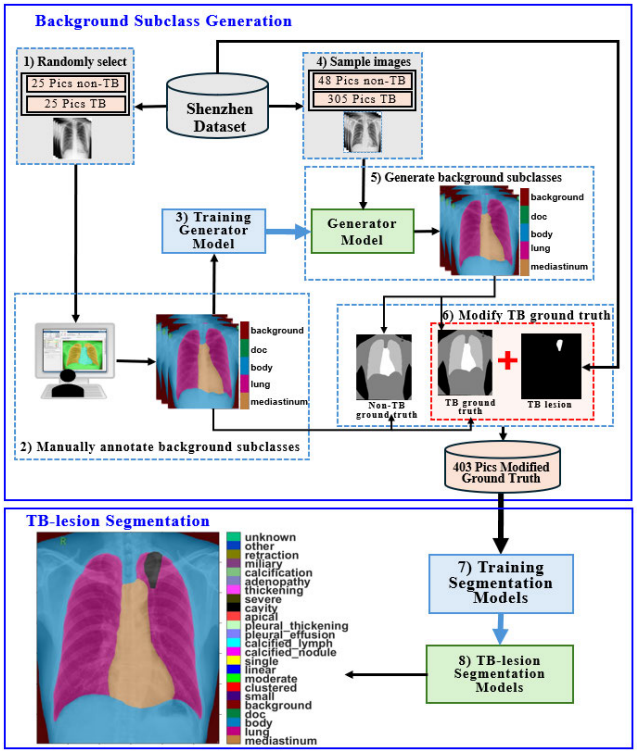


FIGURE 1. Contextual background labelling pipeline for TB-lesion segmentation.

The contributions of this study can be summarized as follows:

- A CXR background generator model was constructed to modify the Shenzhen dataset by re-annotating the background ground truth based on structural anatomy into five subclasses, consisting of lung, mediastinum, body, doc, and background.
- TB-lesion segmentation models were constructed for CXR images by utilizing the modified background

subclasses. The proposed models can improve segmentation performance for both the TB lesions and the background subclasses.

II. RELATED WORKS

The challenge in TB screening using CXR images is the difficulty of disease detection. Experienced radiologists achieve an accuracy of only 68.7% when compared to the ground truth produced by the gold standard [35]. DL algorithms may identify TB lesions better than the human eye since the gray levels of CXR images include many details. Recently, they showed their significant potential in medical image analysis [12], [36], [37]; especially, when large training datasets are available [38], [39], [40].

In screening TB, accuracy and visual interpretability are crucial factors that must be provided by DL algorithms. Neither classification [26], [39], [40], [41], nor detection [35], [42], [43] methods provide the shape and location of TB lesions directly. They approximate the discriminative regions of an input image from a specific class prediction. The location approximation may utilize gradient-weighted class activation mapping (Grad-CAM) [44], score class activation (score-CAM) [45], or saliency maps [46]. The one method that locates TB lesions directly is segmentation.

Rajaraman and Folio [21] used the TBX11K CXR dataset [35] with weak TB bounding-box annotations to segment TB-consistent lesions using a U-Net architecture. However, their model, when evaluated on other cross-institutional datasets such as Shenzhen and Montgomery [31], exhibited a higher rate of false positives and false negatives. To address these weaknesses, they developed a method of TB-consistent lesion segmentation that used fine-grained annotations to create ensembles of U-Net model variants. These models semantically segmented TB-consistent lesions in both original and bone-suppressed frontal chest images [28]. The best ensemble model, which was a stacked ensemble of the top three retrained models [23], achieved superior segmentation performances, returning Intersection over Union (IoU) and Dice scores of 0.4028 and 0.5743 [28], respectively.

To improve the robustness and accuracy of medical image analysis, ensemble learning has been used to integrate information from multiple machine learning models [47]. Narayanan et al. [48] used an ensemble method to combine lung segmentation results from two segmentation models, U-Net and DeepLabV3+, which provided good results when evaluated on the Japanese Radiological Scientific Technology (JRST) [49], [50], and Shenzhen datasets. Ou et al. [64] also employed an ensemble method to combine five segmentation models for the detection of two TB lesion types: infiltration/bronchiectasis and opacity/consolidation. Their dataset, annotated by two CXR experts, comprised 222 radiographs, which they partitioned into training, validation, and testing sets containing 110, 14, and 98 images, respectively. The ensemble model achieved a maximum mean IoU of 0.70, a mean precision of 0.88, a mean recall of 0.75, a mean F1-score of 0.81, and an accuracy of 1.0.

The development of segmentation methods for TB lesion screening necessitates high-quality data to train DL models, as shown by the work of Yang et al. [51], who assessed the inter-annotator agreement between cervical colposcopy images and the Shenzhen TB CXR dataset [30]. Nowadays, the available clean data is quite limited because producing consistent annotations throughout the labeling process requires substantial time and effort. The proposed method revealed some labeling errors in the Shenzhen dataset [30], which was annotated at pixel-level for 19 TB lesions, shown here in the discussion of the results.

One additional challenge for lesion segmentation in CXR images is the difficulty of identifying TB lesions due to the heterogeneous background of lungs and mediastinum. The CXR image background itself presents a challenge for segmentation models, as it includes information from diverse anatomical structures, such as the lungs, mediastinum, body wall, and even items on clothing.

Training a segmentation model on datasets with a highly heterogeneous background class can make it difficult to distinguish between regions of interest and similar features in the background, leading to false positives, underfitting, and degraded segmentation performance. Many researchers have addressed this challenge by segmenting the lungs first [25], [26], [52], [53], [54]. For example, Rahman et al. [26] and Teixeira et al. [25] developed training data that included segmented lung regions to reduce background heterogeneity before disease detection. However, observations from the Shenzhen dataset show that some lesions can appear in the mediastinum.

Li et al. [55] proposed an algorithm called ''context label learning'' (CoLab) to address this issue by dividing the background class into several subclasses. This algorithm has been successfully applied to segment 3D data, such as computed tomography and magnetic resonance images. Inspired by this approach, the proposed method aims to utilize contextual background annotation to improve the performance of TB-lesion segmentation for TB diagnosis.

III. MATERIALS AND METHODS

The TB-lesion segmentation scheme shown in Figure 1 was designed to improve the performance of lesion-region segmentation by reducing false segmentations. The proposed scheme solved the falsehood, especially false positives, by designing contextual background subclasses, and using the semiautomatic process shown in Figure 1 to modify the Shenzhen dataset by generating image background subclasses. The results provided modified ground truth data for training the TB-lesion segmentation model.

A. DATASET

The proposed model was designed, trained and tested on the Shenzhen, Japanese Society of Radiological Technology (JSRT) [49], [50], Montgomery County (MC) [31], and

TB11X [35] datasets. The Shenzhen dataset was employed in the design of the proposed method, while the MC and TB11X datasets were used as external data to evaluate the segmentation models.

The Shenzhen dataset [31] was collected and annotated at the image level for TB classification. The dataset was developed for pixel annotation [30] by a junior radiologist and validated by a senior radiologist affiliated with the Chinese University of Hong Kong. This comprehensive annotation effort transformed the Shenzhen dataset into a valuable resource for semantic segmentation tasks. This dataset contains 662 chest X-ray images of 336 TB cases and 326 normal (non-TB) cases. Although the 336 TB images all carried TB labels, the radiological signs indicative of TB were detected in only 330 images. The TB images were labeled with 19 tuberculosis lesions, as detailed in Table 1. Table 1, column 2, reports the frequency of each TB lesion's appearance across the image dataset. Column 3 presents the pixel proportion of each lesion relative to the total pixel area occupied by all TB lesions. The distribution of these proportions is visualized in Figure 2(a). Figure 2(b) illustrates the proportion of TB lesions relative to the total background pixels. Figure 2(c) presents deeper, presenting the proportion of TB lesions within each of the five background subclasses, as delineated by the overlay in Figure 1, block 6.

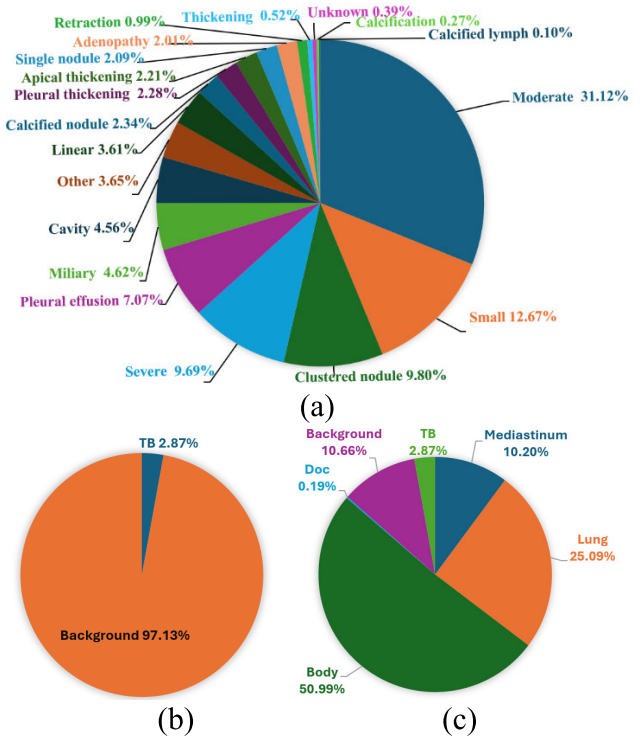


FIGURE 2. Statistics of 19 TB lesions in Shenzhen dataset. (a) Labeled pixel proportion per lesion. (b) All TB lesion proportion in a background. (c) All TB lesion proportion in five subclasses.

The publicly available JSRT dataset [50] comprises 247 radiographs, consisting of 93 without a lung nodule and 154 with a lung nodule. The image size is 2,048 × 2,048 pixels and their ground truth, deposited in the SCR (Segmentation in Chest Radiographs) dataset [56], contains multiple classes from lungs, clavicles and heart.

The MC dataset contains 138 frontal CXRs from the Montgomery County Tuberculosis Screening Program, Maryland, USA. The dataset contains images of 80 normal cases and 58 cases with manifestations of TB [31]. The image size is 4, 892 × 4.020 pixels.

The TBX11K dataset [35] was collected from several leading hospitals in China. The dataset consists of 11,200 radiographs from five categories of condition: 5,000 healthy, 5,000 sick but non-TB, 924 active TB, 212 latent TB, and 10 uncertain TB. The ground truth data annotated TB without defining lesions with bounding boxes. The image size is 512 × 512 pixels. The images in the dataset were split 6,600/1,800/2,800 into train/validate/test datasets, but the test set did not include the annotation data. Thus, the validation set was used to evaluate the proposed models.

TABLE 1. 19-lesion TB annotation of 330 images summarized with number of images per lesion and pixel proportion lesions of 19 TB lesions.

Lesions	Number of images	Lesion proportion (% pixels)
Moderate infiltrate (non-linear)	103	31.12
Small infiltrate (non-linear)	120	12.67
Clustered nodule (2 mm–5 mm apart)	94	9.80
Severe infiltrate (consolidation)	26	9.69
Pleural effusion	48	7.07
Miliary TB	4	4.62
Cavity	35	4.56
Other	13	3.65
Linear density	91	3.61
Calcified nodule	56	2.34
Pleural thickening (non-apical)	44	2.28
Apical thickening	50	2.21
Single nodule (non-calcified)	86	2.09
Adenopathy	18	2.01
Retraction	9	0.99
Thickening of the interlobar fissure	14	0.52
Unknown	10	0.39
Calcification (other than nodule and lymph node)	16	0.27
Calcified lymph node	2	0.10

B. CONTEXTUAL BACKGROUND SUBCLASS ANNOTATION

The contextual background subclasses of TB lesions were designed as shown in Figure 1. The design process started with a random selection of 25 radiographs from the subset of non-TB images and 25 radiographs from the subset of TB images. The background subclasses, categorized as lung, mediastinum, body, doc, and background, were manually labeled by using the Image Labeler App of MATLAB version R2022b [57]. A radiologist from Songklanagarind Hospital, Hat Yai, Thailand, with 30 years of experience, was consulted to help with the manual labeling. During the labeling process, the background subclass was labeled first to ensure all pixels were assigned a category. Next, a polygon was used to label all pixels belonging to the body subclass, creating the "body"

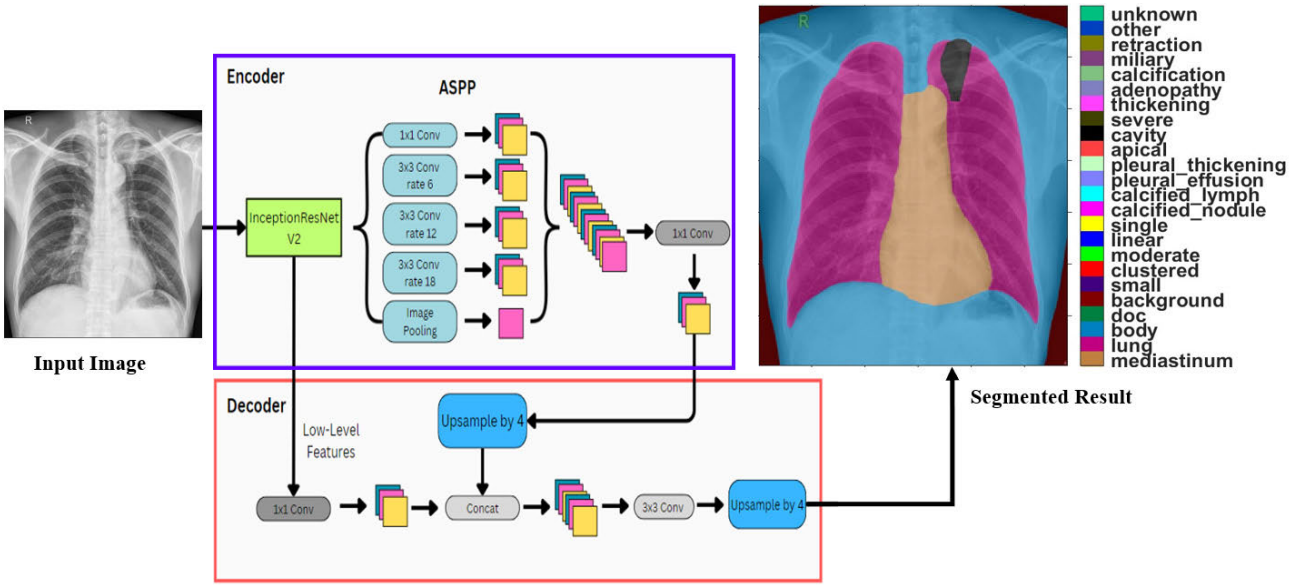


FIGURE 3. DeepLabV3+ with InceptionResNetV2 encoder.

class. Following this, both the left and right lungs, then the mediastinum subclasses were labeled, and finally, the ''doc'' subclass was annotated.

The selected CXR images with their labeled regions were used to train DeepLabV3+ with the RestNet18 backbone to design the background subclass generator model (Figure 1, block 3). Then, the trained generator model was used to create the background subclasses for the remaining CXRs of the Shenzhen dataset. Some difficult images, for which the model failed to generate complete background subclasses, were manually corrected with the Image Labeler App. The complete set of background subclasses was employed to modify the Shenzhen ground truth by overlaying TB lesions on their background subclasses, as illustrated in Figure 1, block 6.

C. SEGMENTATION NETWORKS

The semantic segmentation network used to evaluate the modified background subclasses was DeepLabV3+ [32]. DeepLab can encode with several backbones, which can be chosen on demand. In the proposed scheme, DeepLabV3+ was encoded with three backbones: RestNet18, RestNet50, and InceptionResNetV2. Figure 3 illustrates how the DeepLabV3+ was encoded with InceptionResNetV2 (the modified version of Chen et al. [32]). The DeepLabV3+ model was employed for all the proposed processes but encoded with different backbones depending on the task. The contextual background generator network (Figure 1, block 5) used DeepLabV3+ encoded by RestNet18.

In the training processes, the input images had to be resized to 320 × 320 pixels to fit the computer. Since the input data of the DeepLabV3+ architecture requires three channels, the CXR images were rendered in grayscale. Therefore, an augmentation method was used to modify the images so that they had three channels. The first channel was the original grayscale image. The second channel was the first channel enhanced using Contrast Limited Adaptive Histogram Equalization [58]. The third channel was the complement of the first channel. After modifying the images to three channels, geometric augmentations consisting of translation, reflection and rotation, were applied to simulate the variations of the captured images.

D. EVALUATION METRICS

The segmentation results of each model were evaluated from six indexes: Dice coefficients, IoU, precision, sensitivity, specificity, and accuracy, formulated as follows:

$$Dice = \frac{2TP}{2TP + FP + FN} \tag{1}$$

$$IoU = \frac{TP}{TP + FP + FN} \tag{2}$$

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

$$Sensitivity = \frac{TP}{TP + FN} \tag{4}$$

$$Specificity = \frac{TN}{TN + FP} \tag{5}$$

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN} \tag{6}$$

In the expressions above, TP (true positive) means the number of overlapping pixels between a TB classification and a ground truth TB lesion region; TN (true negative) means non-TB pixels correctly segmented as non-TB; FP (false positive) denotes the number of pixels incorrectly segmented as TB pixels; and FN (false negative) means TB pixels incorrectly segmented as non-TB pixels. In classical

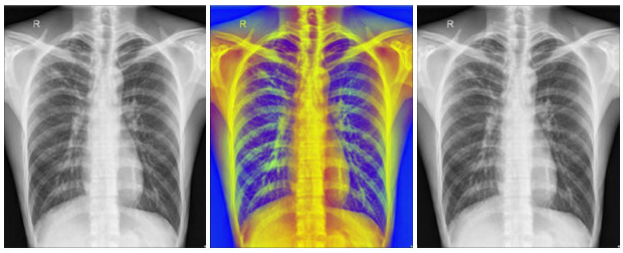


FIGURE 4. Setting three channels of the CXR image for feeding into the input layer of DeepLabV3+ model. (a) Original grayscale image. (b) Three channels of the proposed method. (c) Three channels with the same grayscale image.

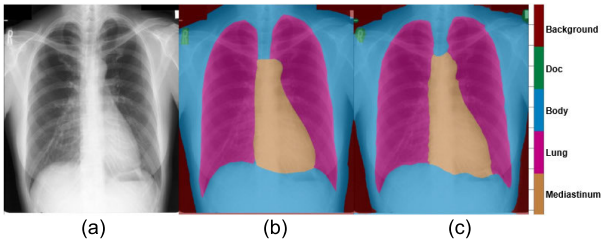


FIGURE 5. Example of data annotation: (a) Original image from JSRT dataset, (b) Ground truth edited from SCR dataset, (c) Five background subclasses from the generator model.

segmentation, the Dice coefficient and Jaccard index are used to evaluate model performances [59]. Both metrics indicate the rate of similarity between the segmented results and ground truths. If the models no longer produce FN and FP, the indexes approach unity. In the proposed scheme, contextual background labels were designed to decrease FP, resulting in higher values for these indexes. Therefore, when the ground truth is clean [60], both indexes will exhibit satisfactory responsivity. Sensitivity, specificity, precision, and accuracy metrics are insensitive to class imbalance, particularly when the background class dominates in the baseline model (as detailed in [61]). The impact of these limitations will be explored in the experimental section.

IV. EXPERIMENTATION

All model assessments employed DeepLabV3+ encoded by the pre-trained InceptionResNetV2, as shown in Figure 3. To investigate the reduction of false TB-lesion segmentations, two tasks were assigned for assessment: (1) the ''baseline'', or TB segmentation between TB lesions and background annotated in the original Shenzhen dataset [30], and (2) the ''5Bg-class'', or the contextual background segmented to five subclasses and used to modify 403 CXR images of the Shenzhen dataset.

A. EXPERIMENTAL SETUP

In the evaluation process, the 5Bg-class was divided into 6-class and 24-class models, which were compared with the baseline class that was divided into 2-class and 20-class models.

The baseline models were detailed the following subtasks:

1) 2-class model: This model was trained to differentiate between background and a single class encompassing all 19 annotated TB lesions.
2) 20-class model: This model employed a more granular approach, distinguishing between 19 individual TB classes and a background class.

The 5Bg-class models utilized two segmentation models:

1) 6-class model: This model differentiated between five background subclasses and a single TB class for comparison with the 2-class baseline model.
2) 24-class model: This model employed a more detailed approach, distinguishing between the five background subclasses and all 19 individual TB lesions (24 classes in total) and the results were compared with the baseline 20-class model.

All the four models were designed with DeepLabV3+ encoded with the pre-trained InceptionResNetV2 and were used on the 403 modified images (330 TB and 73 non-TB). From the TB set, 19 images were meticulously selected to encompass all 19 annotated lesions, ensuring a comprehensive evaluation of each TB class, isolated for the testing process. Thus, all baseline, and 5Bg-class models were trained on 353 images and validated on 31 images. The learnable model parameters were optimized using the Adam optimizer with an initial learning rate of 2×10^{-3}. The focal loss function [62] was employed to manage the imbalance data by setting the following parameters: alpha $=$ 0.5 and gamma $= 2$. The loss function was minimized with a batch size of 28 images. Training proceeded for 500 epochs, and model weights were saved whenever the validation loss decreased. The model with the minimum validation loss was then used to predict on the test set. The models were implemented on a workstation equipped with two NVIDIA Tesla M40 GPUs, utilizing the deep learning and parallel toolboxes of MATLAB version R2022b [57].

The performance of all the four models was evaluated on the internal and external data. The internal data consisted of 19 isolated TB images and 253 images from the remaining non-TB set. The evaluation results are shown in Section IV-C. Furthermore, the proposed 6-class and 24-class models were investigated for generalizability, evaluated on external data from the MC and TBX11K datasets, as shown in Section IV-D.

To further support the findings, ablation studies were conducted to complement the initial research of input image configuration and background subclass creation for the proposed models.

B. ABLATION STUDY OF THE DESIGN PROCESS

1) THREE CHANNELS OF INPUT LAYER

To confirm the effectiveness and advantage of setting images for the input layer of DeepLabV3+ models, two experimental configurations were provided for investigation. 1) The first

TABLE 2. Quantitative metrics between the duplicated images and the proposed method of the segmentation models trained with semi-automatic labels, comparing the utilization of duplicated values on three color channels and the enhanced values implemented in this study.

Methods	Dice	IoU	Precision	Accuracy
Three-channels duplication	0.4478	0.2885	0.3334	0.9526
Our three channels	**0.6080**	**0.4367**	**0.5926**	**0.9773**

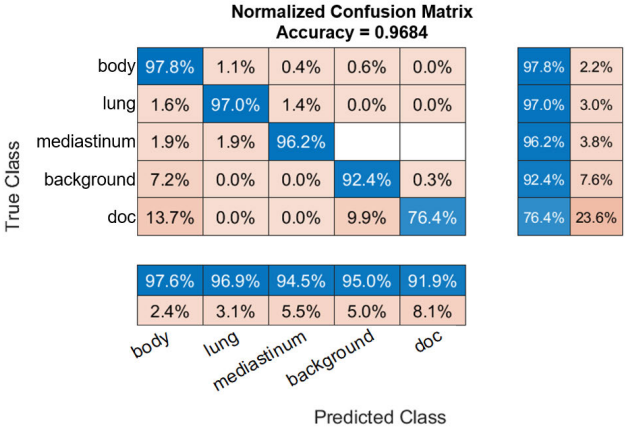


FIGURE 6. Confusion matrix of the generator model to segment five background subclasses.

channel was set to the original grayscale image. The second channel was assigned to a version of the grayscale image enhanced using CLAHE, and the third channel was set to the complement of the grayscale image (as shown in Figure 4(b)). 2) The three channels of the same image were duplicated in the original grayscale [7], [63] as shown in Figure 5(c). In both experiments, the DeepLabV3+ model was trained with the same parameters as the 6-class model. Thus, the two trained models were evaluated on the 19 images of the test set. The evaluation results of TB segmentation presented in Table 2 demonstrate that the proposed method outperformed three-channel duplication across all metrics.

2) BACKGROUND SUBCLASSES

To confirm the effectiveness of the background subclass generator model, the model performance was evaluated on the external dataset, JSRT [50], and the ground truth data in the SCR dataset [56]. The test set consisted of 30 images selected from JSRT and their corresponding labels from SCR. However, the SCR data had to be edited by adding the background subclasses: body, doc, mediastinum, and keeping the lung and heart subclasses. However, the heart subclass was included in the mediastinum subclass. An example of the edited SCR ground truth is presented in Figure 5(b).

Figure 5(c) exemplifies the test result of the input image in Figure 5(a). The IoU and Dice similarity coefficients of the 30 tested image pairs are presented in Table 3. Figure 6 presents the confusion matrix summarizing the segmentation performance. The model achieved promising results for the lung and mediastinum subclasses, with IoU scores of 93.61% and 90.56%, respectively. However, the IoU score for the Doc subclass was significantly lower at 65.86%; caused by the small rectangular artifact in the upper right which produced an FP. This discrepancy can likely be attributed to the inherent absence of the doc subclass in typical JSRT dataset images. Furthermore, the generator background model demonstrated an accuracy of 96.84% on a set of unseen, foreign-tested data, as illustrated in Figure 6. The confusion matrix in Figure 6 presents the percentage breakdown of correctly and incorrectly segmented pixels for each background subclass.

TABLE 3. Quantitative metrics evaluated background subclass of the generator model.

Class names	Dice	IoU
mediastinum	0.9503	0.9056
lung	0.9669	0.9361
body	0.9749	0.9511
doc	0.7809	0.6586
background	0.9103	0.8432

TABLE 4. Quantitative indexes of TB segmentation method.

Methods	Dice	IoU	Precision	Accuracy
Baseline (2-class model)	0.6778	0.5954	0.9779	0.9627
Baseline (20-class mode)	0.2673	0.1543	0.9763	0.9435
Rajaraman et al.	0.5743	0.4028	-	-
Ou et al.	-	0.7000	0.8800	**1.000**
5Bg-class (6-class model)	**0.8868**	**0.8355**	**0.9864**	0.9523
5Bg-class (24-class model)	0.2526	0.2292	0.9743	0.9582

C. INTERNAL TEST SET

To investigate the effectiveness of contextual background labeling in improving segmentation performance, the proposed scheme was assessed by using the baseline models as a reference to compare with the 5Bg-class models. Table 4 presents the assessment of the baseline and 5Bg-class models and contrasts their performances with the methods of Rajaraman et al. [28] and Ou et al. [64]. Rajaraman et al. segmented CXR images into two classes (Background and TB), the same as the 2-class baseline model. However, their method obtained Dice and IoU indexes higher than our 20-class, and 24-class models. The method of Ou et al. used ensemble learning to integrate five models of semantic segmentation to segment two types of TB lesion: infiltration/bronchiectasis and opacity/consolidation. The proposed models were assessed on 19 images of the test set. The 6-class model returned the highest Dice, IoU, and precision scores but the method of Ou et al. achieved a 100% accuracy score.

The best IoU and Dice scores, respectively 0.6944 and 0.8196, were derived from the first image in the test set (CHNCXR_0332_1 in the Shenzhen dataset). The CHNCXR_0332_1 image is presented in Figure 7 on the top row. Figure 7 illustrates the false segmentation improvement realized with the 6-class model compared with the 2-class model. The second and third rows show the results from images CHNCXR_0456_1, and CHNCXR_0557_1,

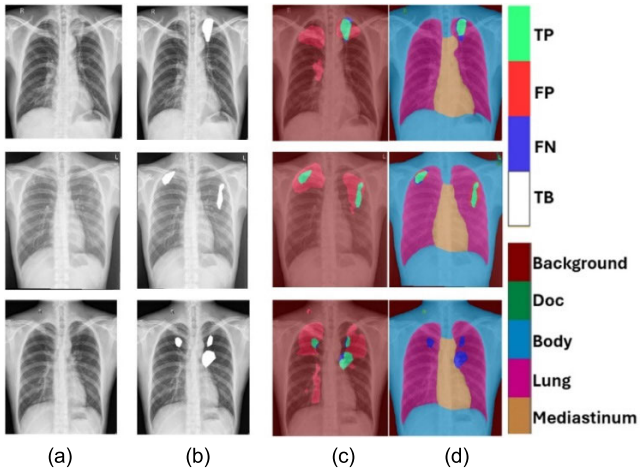


FIGURE 7. Illustration of TB segmented results of 6-class with respect to 2-class model for considering TB segmentation improvement by decreasing the FP value (Red color region). (a) Original images from top to bottom are CHNCXR_0332_1, CHNCXR_0456_1, and CHNCXR_0557_1, respectively. (b) Ground truth of (a), (c) 2-class baseline model, (d) 6-class model.

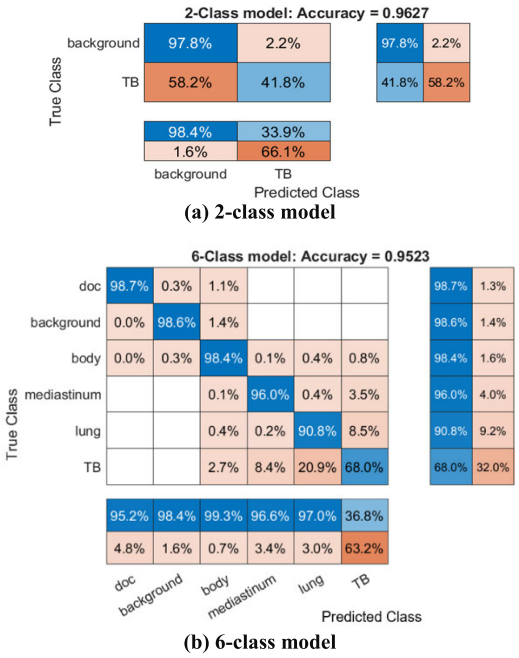


FIGURE 8. Confusion matrixes of the baseline 2-class model and the 6-class of 5Bg-class model.

respectively. The third column presents the results from the 2-class model, which segmented a lot of FP regions (colored red). The results of the 6-class model in the last column of the figure show no red FP regions but classed the lesions in CHNCXR_0557_1 as FN, colored blue.

Figure 8 presents the confusion matrixes for the baseline and 5Bg-class models. As shown in Figure 8(a), the confusion matrix of the 2-class model shows that 41.8% of the TB segmentation was correctly labeled (TP), and 58.2% falsely labeled (FP). This result corresponds to Figure 7(c), which shows large red areas of FP in all three images. The confusion matrix of the 6-class model in Figure 8(b) shows that 68.0% of the TB segmentation was classed correctly (TP) and 32.0% incorrectly (FP), with 2.7% labeled as body, 8.4% as mediastinum and 20.9% as lung classes. According to Figure 7(d), the 6-class model improved the segmentation performance by decreasing FP.

TABLE 5. Results of 4 segmentation models to assess 253 radiographs of Non-TB set.

Models	Corrected segmentation	Corrected percentage
Baseline (2-class model)	30	11.86
Baseline (20-class mode)	59	23.32
5Bg-class (6-class model)	161	63.64
5Bg-class (24-class model)	216	85.38

Figure 9 presents the 20-class and 24-class confusion matrixes of 19 TB lesions. When the 19 classes of TB lesions were overlaid on a background class, the baseline 20-class model (Figure 9(a)) could segment four lesions: small (Small infiltrate), moderate (Moderate infiltrate), apical (Apical thickening), and clustered (Clustered nodule) with TP values of 61.4%, 44.0%, 31.6%, and 2.2%, respectively. The model returned low precision scores, as shown in the first of the two rows at the bottom of the matrix. Figure 9(b) presents the 24-class confusion matrix of 19 TB lesions overlaid on the five background subclasses. The model could segment only 5 TB lesions: single (Single nodule), small (Small infiltrate), moderate (Moderate infiltrate), apical (Apical thickening) and cavity, with TP values of 87.6%, 52.6%, 47.2%, 13.2%, and 12.8%, respectively. For each segmented class, the model returned higher precision scores than the 20-class model. From the segmentation results, the five background subclasses could reduce FP values more than the baseline 20-class model. The visualization results of the 20-class and 24-class models are presented in Figure 10, using the CHNCXR_0497_1 image as the original input. Figure 10(b) and (c) illustrate the ground truth and the corresponding segmentation results from the 20-class model, while Figure 10(d) and (e) show the ground truth and the corresponding segmentation results from the 24-class model. Figure 10(f) was formulated from Figure 10(b) and (c) to analyze the areas of TP, FP and FN for measuring the falsehoods, IoU = 0.0340. Figure 10(d) and (e) were edited by grouping the five subclasses to the background class, the same as with the baseline model, and were then used to assess the regions of TP, FP, and FN as shown in Figure 10(g), giving an IoU value of 0.2473.

The evaluation results of the four TB segmentation models are presented in Figure 11 and Table 5. In Figure 11, the input CHNCXR_0002_0 image was classed as non-TB in the dataset. However, the baseline 2-class model detected some TB lesion, and the baseline 20-class model identified a small infiltrate lesion, both labeled pink in the figure. For the proposed 5Bg-class models, both the 6-class and 24-class models

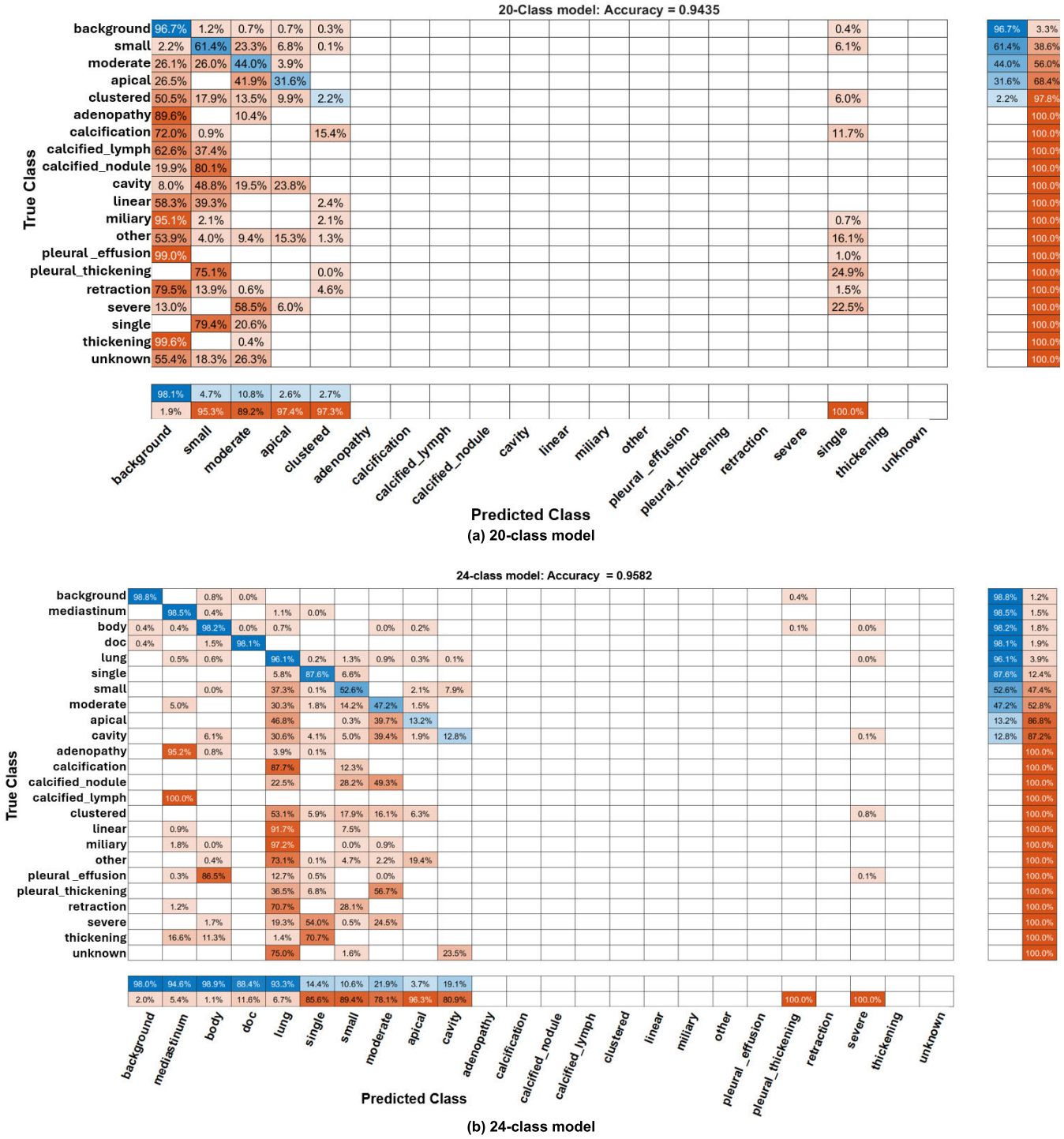


FIGURE 9. Confusion matrixes of the baseline 20-class model and the 24-class of 5Bg-class model.

correctly segmented the image, labeling only the background subclasses. The 24-class model produced the highest number (216) of correctly segmented radiographs from the total of 253 (Table 5). The lowest number of correct segmentations was 30 radiographs, or only 11.86%. This part of the experiment provided good evidence that heterogeneous background decomposition reduced false segmentations. If the same objects in images are more homogeneous, the segmentation method can provide more correct predictions.

D. EXTERNAL TEST SETS

The MC and TBX11K datasets were utilized for the external validation of the 6-class and 24-class of 5Bg-class models. The MC dataset comprised 138 radiographs, which were all

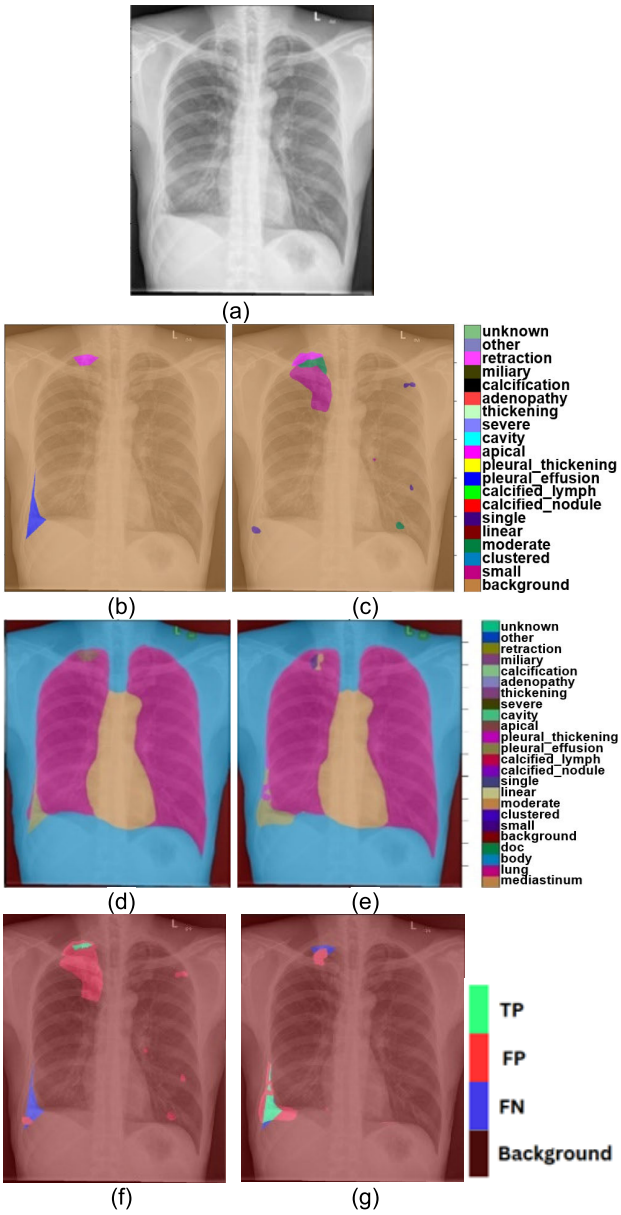


FIGURE 10. 19-TB-lesion segmentation, (a) original image (CHNCXR_0497_1), (b) ground truth of 20-class model, (c) segmented result of 20-class model, (d) ground truth of 24-class model, (e) segmented result of 24-class model, (f) TP, FP and FN of 20-class model provided IoU = 0.0340, (g) TP, FP and FN of 24-class model provided IoU = 0.2473.

used for testing. The TBX11K dataset images were selected from the validation subset and contained 800 normal images and 200 TB images. Table 6 summarizes the tested results. The MC set was annotated at the image level. In the TB segmentation of the MC set, the number of events denotes the number of images. In the TBX11K set, TB lesions were annotated with bounding boxes, so the counting event for TB detection was the number of boxes in the 200 TB images, which was 309 boxes. The healthy events of the TBX11K set were shown as Normal and the count indicates the number of images in the set. The performance of each model

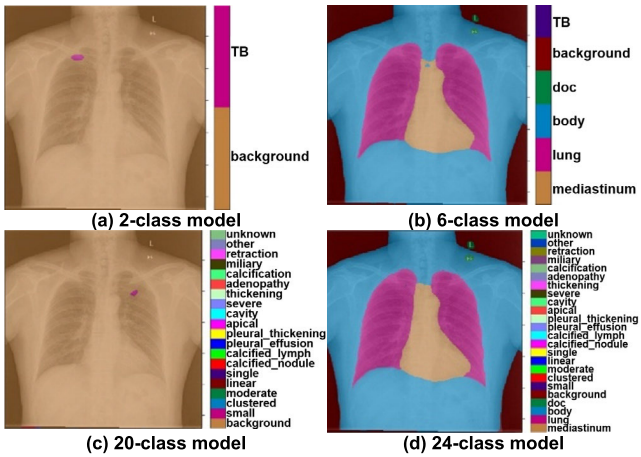


FIGURE 11. Non-TB segmentation results, 2-class model detected TB, 6-class model detected only the background subclasses, 20-class model detected small infiltrate lesion, and 24-class model detected only the background subclasses.

was measured from TP, FP, FN, and TN scores. On the MC set, the 6-Class model achieved 89.00% for sensitivity, specificity, and accuracy. On the TBX11K set, the 24-class model achieved 71.00%, 94.00%, and 86.00% for sensitivity, specificity, and accuracy, respectively. Figure 12 shows some examples of the segmented results from the external datasets.

Figure 12(a)-(f) presents the segmentation results of the MC set. The image MCUCXR_0001_0 (a) was annotated non-TB, and both the 6-class and the 24-class detected no sign of any TB lesions. The image MCUCXR_0104_1 (d) was annotated TB, and the 6-class model segmented a TB region on the left lung (e), while the 24-class model provided more detail of TB lesions (f), identifying moderate (Moderate infiltrate), linear (Linear density), and apical (Apical thickening) in the same region as the 6-class model. The segmentation results of the TBX11K set are presented in Figure 12(g)-(l). Image h0004 (g) was annotated healthy. The 6-class model detected no TB signs (h) but the 24-class model detected a moderate infiltrate lesion (i), which was a FP. The image tb0014 (j) was annotated with two bounding boxes of TB. The 6-class model segmented the lesions inside and outside both boxes (k), returning two TPs. The 24-class model segmented four lesions on the right lung (l): a moderate infiltrate, a single nodule, a calcified nodule and pleural thickening, and three lesions on the left lung (l): a clustered nodule, a calcified nodule and a single nodule which were also intersected by the bounding boxes. The result was also two TPs.

V. DISCUSSION

The precise segmentation of TB lesions in CXR images is crucial for identifying pathologies. The task presents several challenges, which include complex textured lesions in grayscale images, heterogeneous backgrounds, confusion between different types of lesions, and class imbalances. All these factors significantly impact the performance of the segmentation model.

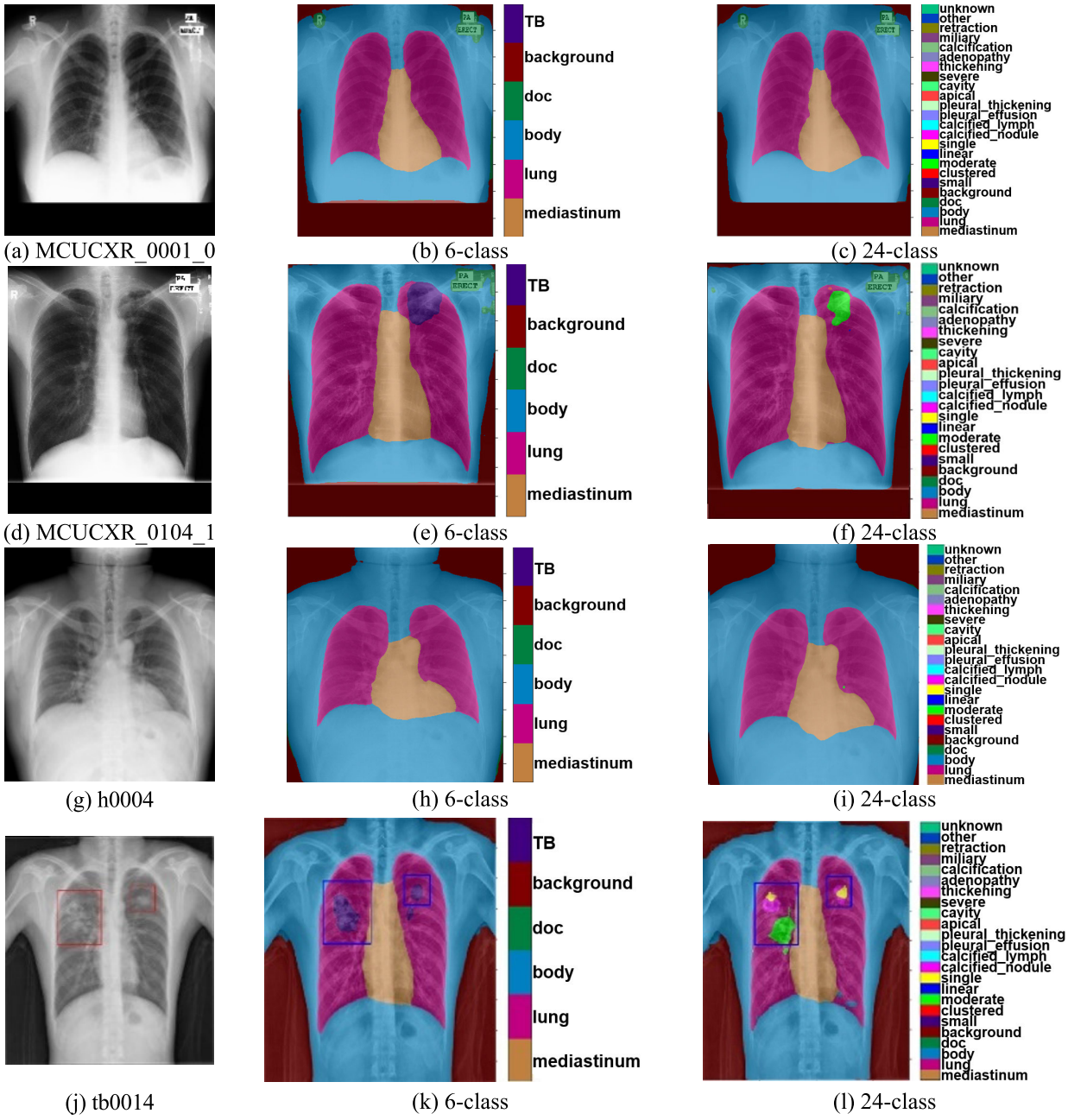


FIGURE 12. External test sets from MC and TBX11K, the first column showed original images of (a) and (g) no sign of TB; (d) and (j) annotated TB. The second column showed the detected results of 6-class model. The last column showed the detected of 24-class model.

TABLE 6. External dataset evaluations by MC and TBX11K.

Datasets	#Sample (TB/Normal)	Models	TP	FP	FN	TN	Sensitivity	Specificity	Accuracy
MC	138 (58/80)	6-class	49	9	6	74	**0.89**	0.89	0.89
		24-class	55	3	19	61	0.74	**0.95**	0.84
TBX11K	1109 (309/800)	6-class	261	48	61	739	0.81	0.94	**0.90**
		24-class	262	47	108	692	0.71	0.94	0.86

To address the problem of background heterogeneity, contextual background annotations were developed to specifically reduce background variability, minimizing the high correlation observed between background and TB lesions. In the proposed scheme, the background generator model could generate five background subclasses of CXR images to fulfill the ground truth data of the Shenzhen dataset [30], setting a new standard for future studies.

From the Shenzhen dataset, 403 modified images were used to create four TB segmentation models. The segmentation results provided the contextual background labels used to decompose the heterogeneous CXR background and improve the TB-lesion segmentation models.

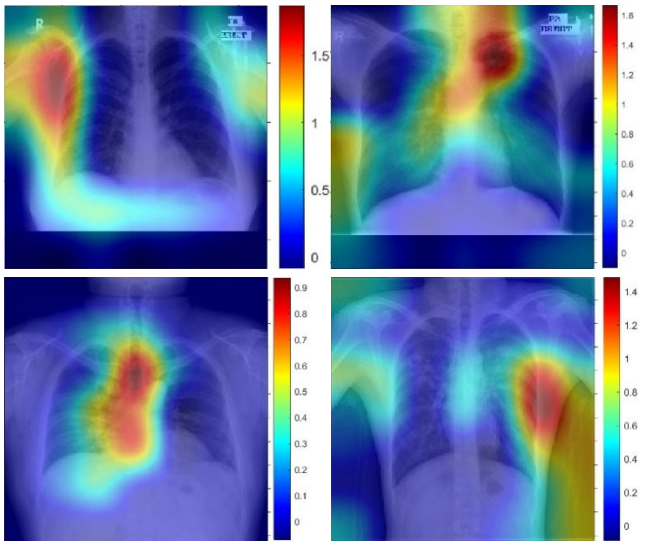
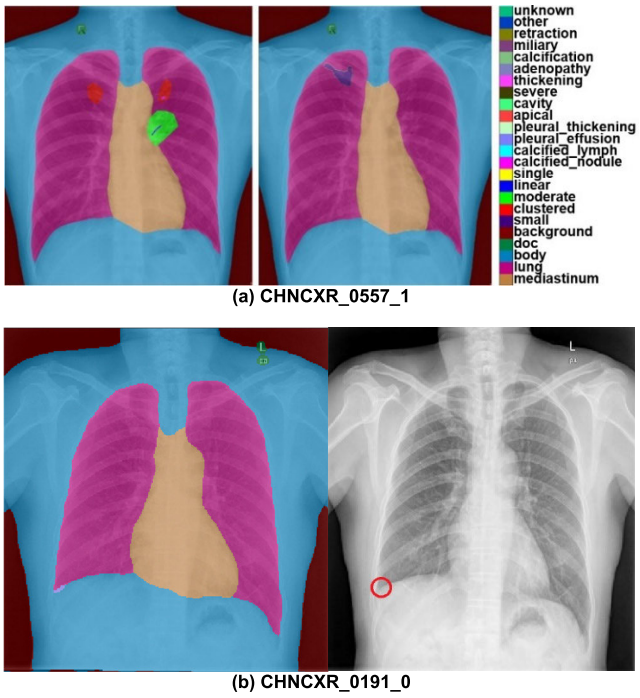


FIGURE 13. Identifying TB lesions by heat map using the original images from Figure 12. The heat map derived from a classification method, which reimplement the method of Rahman et al. 26.



(a) CHNCXR_0557_1

(b) CHNCXR_0191_0

FIGURE 14. (a) Labeling errors of CHNCXR_0557_1 on the left, and corrected annotation showed on the right, (b) the 24-class model detected pleural effusion in CHNCXR_0191_0, and the red cycle on the right confirmed by radiologist.

In the case of TB-lesion segmentation performed on the TB and non-TB internal test sets, when considering the main diagonal of 24-class confusion matrixes, the proposed scheme could only segment five classes of TB lesions because of class imbalance. However, the model successfully reduced both false positives and false negatives. The empirical evidence consisted of the precision and sensitivity indices, entered in the two rows at the bottom and the two columns beside the confusion matrix, respectively. Another significant index was the IoU, which showed good results in the 6-class model, as seen in Table 4, and Figure 7 and 10, respectively. Furthermore, the proposed 24-class model produced a corrected segmentation percentage of 85.38% on 253 non-TB radiographs.

On external datasets, the 6-class model achieved the highest accuracy score of 90.00% when tested on the TBX11K dataset and provided good sensitivity and specificity scores at 81.00% and 94.00%, respectively. On the MC dataset, the 24-class model achieved a high sensitivity score of 89.00% and the highest specificity score of 95.00%. The experimental results obtained on the external datasets showed the good generalizability of the proposed models, which returned a positive predictive value (PPV = TP/(TP + FP)) higher than 84.47%, and a negative predictive value (NPV = TN/(TN + FN)) more than 86.50%; except in the case of the 24-class model tested on the MC set, which returned an NPV = 76.25%.

For visualization, the proposed model provided exactly both the shape and location of TB lesions. Moreover, the 24-class model, as seen in Figure 12(f) revealed three lesions: a moderate infiltrate, linear density, and apical thickening, and in Figure 12(l), the model revealed five lesions: a clustered nodule, a moderate infiltrate, a single nodule, a calcified nodule, and pleural thickening. Although the 6-class model could not provide lesion details, it revealed the shape and location of TB lesions rather than the heat map obtained from classification or detection methods, as seen in the images in Figure 13, which were derived from the work of Rahman et al. who used DenseNet201 [26]. Several studies have attempted to directly detect the shape and location of TB lesions with different techniques [28], [64]. However, these techniques, like the proposed scheme, have been hindered by a common limitation: the scarcity of high-quality training data. The proposed models demonstrate a tendency to produce numerous false segmentations when applied to normal radiographs within the TBX11K dataset. This limitation may be attributed to the relatively small number of normal radiographs (73 images) included in the training set. Additionally, a radiologist with 30 years of experience reviewed a subset of radiographs from the internal test set. The ground truth annotations were found to have some labeling errors. For instance, CHNCXR_0362_1 was annotated with a calcification lesion that was not found, and CHNCXR_0557_1 was annotated with four different lesions, but only one was found (Figure 14(a)), which was different from the annotated lesions. CHNCXR_0191_0 was annotated to non-TB but the 6-class model segmented the CXR to the TB class and the 24-class model classified a pleural effusion lesion, which the radiologist confirmed, as presented in Figure 14(b).

While annotating background subclasses is undoubtedly time-consuming, the annotation of TB lesions demands significantly more effort and time compared to background

annotation. Moreover, lesion-level annotations are considerably more labor-intensive than image-level labeling, which can often be automatically generated from radiology reports. The manual annotation of 50 images with background subclasses took an average of 3 minutes per image, whereas the background generator model needed only approximately 1 minute per image to correct labeling errors. Finally, the trained segmentation models have the potential to serve as a valuable tool for guiding the clinical interpretation of abnormalities in CXR images. Future research will focus on integrating TB segmentation techniques into clinical workflows to evaluate their impact. This endeavor will follow the correction of labeling errors in the training set.

VI. CONCLUSION

This paper proposed the contextual background subclass annotation of CXR images to improve TB-lesion segmentation. The backgrounds of CXR images from the Shenzhen dataset were annotated at the pixel level with 19 TB lesions. The backgrounds were decomposed in a contextual subclass operation called the background subclass generator model, which generated five background subclasses: lung, mediastinum, body, doc, and background. The TB lesion annotations of the dataset were overlaid on the background subclasses to improve the ground truth data. The improved dataset, consisting of 403 images, was designed to train DeepLabV3+ encoded with InceptionResNetV2 for the construction of TB-lesion segmentation models. To confirm the contribution of the new contextual background subclasses to any improvements in image segmentation, four TB segmentation models were designed for evaluation. The models were evaluated on both internal and external datasets. The internal set was annotated with pixel levels, and the models achieved 83.55%, 98.64%, and 95.23% for IoU, precision, and accuracy, respectively. The external set was annotated with image levels, and the models achieved 89.00%, 95.00%, and 90.00% for sensitivity, specificity, and accuracy, respectively. Furthermore, the proposed models provided direct TB lesion visualization to support clinical diagnosis. In future works, the labeling errors will be corrected, a clinical impact assessment will be conducted, and a background generator model will be developed to generate background subclasses aligned with CXR anatomy. This project includes adding a heart class inside the mediastinum class, which can be utilized to detect an enlarged heart.

REFERENCES

[1] World Health Organization. (2023). *Global Tuberculosis Report 2023*. [Online]. Available: https://www.who.int/publications/i/item/9789240083851

[2] World Health Organization. (Apr. 22, 2022). *WHO Operational Handbook on Tuberculosis: Module 2: Screening: Systematic Screening for Tuberculosis Disease*. [Online]. Available: https://iris.who.int/bitstream/handle/10665/340256/9789240022614-eng.pdf

[3] D. P. Carmody, C. F. Nodine, and H. L. Kundel, "An analysis of perceptual and cognitive factors in radiographic interpretation," *Perception*, vol. 9, no. 3, pp. 339–344, Jun. 1980.

[4] R. Fitzgerald, "Error in radiology," *Clin. Radiol.*, vol. 56, no. 12, pp. 938–946, Dec. 2001.

[5] L. Delrue, R. C. Gosselin, B. Ilsen, A. V. Landeghem, J. D. Mey, and W. Duyck, "Difficulties in the interpretation of chest radiography," in *Comparative Interpretation CT Standard Radiography Chest*. Berlin, Germany: Springer, 2010, pp. 27–49.

[6] Y.-D. Zhang, Z. Dong, S.-H. Wang, and C. Cattani, "IEEE access special section editorial: Deep learning for computer-aided medical diagnosis," *IEEE Access*, vol. 8, pp. 96804–96810, 2020.

[7] L. Visuña, D. Yang, J. Garcia-Blas, and J. Carretero, "Computer-aided diagnostic for classifying chest X-ray images using deep ensemble learning," *BMC Med. Imag.*, vol. 22, no. 1, p. 178, Oct. 2022.

[8] C. Geric, Z. Z. Qin, C. M. Denkinger, S. V. Kik, B. Marais, A. Anjos, P.-M. David, F. Ahmad Khan, and A. Trajman, "The rise of artificial intelligence reading of chest X-rays for enhanced TB diagnosis and elimination," *Int. J. Tuberculosis Lung Disease*, vol. 27, no. 5, pp. 367–372, May 2023.

[9] A. Nath, Z. Hashim, S. Shukla, P. A. Poduvattil, Z. Neyaz, R. Mishra, M. Singh, N. Misra, and A. Shukla, "A multicentre study to evaluate the diagnostic performance of a novel CAD software, DecXpert, for radiological diagnosis of tuberculosis in the northern Indian population," *Sci. Rep.*, vol. 14, no. 1, p. 20711, Sep. 2024.

[10] J. Peng and Y. Wang, "Medical image segmentation with limited supervision: A review of deep network models," *IEEE Access*, vol. 9, pp. 36827–36851, 2021.

[11] R. Wang, T. Lei, R. Cui, B. Zhang, H. Meng, and A. K. Nandi, "Medical image segmentation using deep learning: A survey," *IET Image Process.*, vol. 16, no. 5, pp. 1243–1267, Apr. 2022.

[12] P.-H. Conze, G. Andrade-Miranda, V. K. Singh, V. Jaouen, and D. Visvikis, "Current and emerging trends in medical image segmentation with deep learning," *IEEE Trans. Radiat. Plasma Med. Sci.*, vol. 7, no. 6, pp. 545–569, Apr. 2023.

[13] M. F. Sohan and A. Basalamah, "A systematic review on federated learning in medical image analysis," *IEEE Access*, vol. 11, pp. 28628–28644, 2023.

[14] M. Havaei, A. Davy, D. Warde-Farley, A. Biard, A. Courville, Y. Bengio, C. Pal, P. Jodoin, and H. Larochelle, "Brain tumor segmentation with deep neural networks," *Med. Image Anal.*, vol. 35, pp. 18–31, May 2016.

[15] A. Ali, M. Sharif, C. Muhammad Shahzad Faisal, A. Rizwan, G. Atteia, and M. Alabdulhafith, "Brain tumor segmentation using generative adversarial networks," *IEEE Access*, vol. 12, pp. 183525–183541, 2024.

[16] M. Zahangir Alom, T. Aspiras, T. M. Taha, and V. K. Asari, "Skin cancer segmentation and classification with NABLA-N and inception recurrent residual convolutional networks," 2019, *arXiv:1904.11126*.

[17] M. D. Alahmadi, "Multiscale attention U-Net for skin lesion segmentation," *IEEE Access*, vol. 10, pp. 59145–59154, 2022.

[18] J. Son, S. J. Park, and K.-H. Jung, "Retinal vessel segmentation in fundoscopic images with generative adversarial networks," 2017, *arXiv:1706.09318*.

[19] A. Khanal and R. Estrada, "Dynamic deep networks for retinal vessel segmentation," 2019, *arXiv:1903.07803*.

[20] K.-B. Park, S. H. Choi, and J. Y. Lee, "M-GAN: Retinal blood vessel segmentation by balancing losses through stacked deep fully convolutional networks," *IEEE Access*, vol. 8, pp. 146308–146322, 2020.

[21] S. Rajaraman, L. R. Folio, J. Dimperio, P. O. Alderson, and S. K. Antani, "Improved semantic segmentation of tuberculosis—Consistent findings in chest X-rays using augmented training of modality-specific U-Net models with weak localizations," *Diagnostics*, vol. 11, no. 4, p. 616, Mar. 2021.

[22] E. Kotei and R. Thirunavukarasu, "A comprehensive review on advancement in deep learning techniques for automatic detection of tuberculosis from chest X-ray images," *Arch. Comput. Methods Eng.*, vol. 31, no. 1, pp. 455–474, Jan. 2024.

[23] S. Rajaraman and S. K. Antani, "Modality-specific deep learning model ensembles toward improving TB detection in chest radiographs," *IEEE Access*, vol. 8, pp. 27318–27326, 2020.

[24] T. Bansal, S. Gupta, and N. Jindal, "Segmentation techniques for detection of tuberculosis using deep learning: A review," in *Proc. 3rd Int. Conf. Smart Gener. Comput., Commun. Netw. (SMART GENCON)*, Bengaluru, India, Dec. 2023, pp. 1–6.

[25] L. O. Teixeira, R. M. Pereira, D. Bertolini, L. S. Oliveira, L. Nanni, G. D. C. Cavalcanti, and Y. M. G. Costa, "Impact of lung segmentation on the diagnosis and explanation of COVID-19 in chest X-ray images," *Sensors*, vol. 21, no. 21, p. 7116, Oct. 2021.

[26] T. Rahman, A. Khandakar, M. A. Kadir, K. R. Islam, K. F. Islam, R. Mazhar, T. Hamid, M. T. Islam, S. Kashem, Z. B. Mahbub, M. A. Ayari, and M. E. H. Chowdhury, "Reliable tuberculosis detection using chest X-ray with deep learning, segmentation and visualization," *IEEE Access*, vol. 8, pp. 191586–191601, 2020.

[27] T. Griffin, Y. Cao, B. Liu, and M. J. Brunette, "Object detection and segmentation in chest X-rays for tuberculosis screening," in *Proc. 2nd Int. Conf. Transdisciplinary AI (TransAI)*, Sep. 2020, pp. 34–42.

[28] S. Rajaraman, F. Yang, G. Zamzmi, Z. Xue, and S. K. Antani, "A systematic evaluation of ensemble learning methods for fine-grained semantic segmentation of tuberculosis-consistent lesions in chest radiographs," *Bioengineering*, vol. 9, no. 9, p. 413, Aug. 2022.

[29] E. Sogancioglu, K. Murphy, E. Calli, E. T. Scholten, S. Schalekamp, and B. Van Ginneken, "Cardiomegaly detection on chest radiographs: Segmentation versus classification," *IEEE Access*, vol. 8, pp. 94631–94642, 2020.

[30] F. Yang, P. X. Lu, M. Deng, Y. X. J. Wáng, S. Rajaraman, Z. Xue, L. R. Folio, S. K. Antani, and S. Jaeger, "Annotations of lung abnormalities in the Shenzhen chest X-ray dataset for computer-aided screening of pulmonary diseases," *Data*, vol. 7, no. 7, p. 95, Jul. 2022.

[31] S. Jaeger, S. Candemir, S. Antani, Y. Wáng, P.-X. Lu, and G. R. Thoma, "Two public chest X-ray datasets for computer-aided screening of pulmonary diseases," *Quantum Imag. Med. surgery*, vol. 4, no. 6, p. 475, Dec. 2014.

[32] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. Euro. conf. Comput. Vis.*, Munic, Germany, Jan. 2018, pp. 833–851.

[33] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[34] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-ResNet and the impact of residual connections on learning," in *Proc. AAAI conf. Artif. Intell.*, Feb. 2017, vol. 31, no. 1.

[35] Y. Liu, Y.-H. Wu, Y. Ban, H. Wang, and M.-M. Cheng, "Rethinking computer-aided tuberculosis diagnosis," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 2643–2652.

[36] F. Altaf, S. M. S. Islam, N. Akhtar, and N. K. Janjua, "Going deep in medical image analysis: Concepts, methods, challenges, and future directions," *IEEE Access*, vol. 7, pp. 99540–99572, 2019.

[37] M. E. Rayed, S. M. S. Islam, S. I. Niha, J. R. Jim, M. M. Kabir, and M. F. Mridha, "Deep learning for medical image segmentation: State-of-the-art advancements and challenges," *Informat. Med. Unlocked*, vol. 47, Apr. 2024, Art. no. 101504.

[38] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, "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)*, Jul. 2017, pp. 3462–3471.

[39] P. Rajpurkar, J. Irvin, K. Zhu, B. Yang, H. Mehta, T. Duan, D. Ding, A. Bagul, C. Langlotz, K. Shpanskaya, M. P. Lungren, and A. Y. Ng, "CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning," 2017, *arXiv:1711.05225*.

[40] A. Bustos, A. Pertusa, J.-M. Salinas, and M. De La Iglesia-Vayá, "PadChest: A large chest X-ray image dataset with multi-label annotated reports," *Med. Image Anal.*, vol. 66, Dec. 2020, Art. no. 101797.

[41] J. Devasia, H. Goswami, S. Lakshminarayanan, M. Rajaram, and S. Adithan, "Deep learning classification of active tuberculosis lung zones wise manifestations using chest X-rays: A multi label approach," *Sci. Rep.*, vol. 13, no. 1, p. 887, Jan. 2023.

[42] Y. Liu, Y.-H. Wu, S.-C. Zhang, L. Liu, M. Wu, and M.-M. Cheng, "Revisiting computer-aided tuberculosis diagnosis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 4, pp. 2316–2332, Apr. 2024.

[43] H. H. Pham, H. Q. Nguyen, H. T. Nguyen, L. T. Le, and L. Khanh, "An accurate and explainable deep learning system improves interobserver agreement in the interpretation of chest radiograph," *IEEE Access*, vol. 10, pp. 104512–104531, 2022.

[44] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," 2016, *arXiv:1610.02391*.

[45] H. Wang, Z. Wang, M. Du, F. Yang, Z. Zhang, S. Ding, P. Mardziel, and X. Hu, "Score-CAM: Score-weighted visual explanations for convolutional neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2020, pp. 111–119.

[46] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," 2013, *arXiv:1312.6034*.

[47] D. Müller, I. Soto-Rey, and F. Kramer, "An analysis on ensemble learning optimized medical image classification with deep convolutional neural networks," *IEEE Access*, vol. 10, pp. 66467–66480, 2022.

[48] B. N. Narayanan, M. S. De Silva, R. C. Hardie, and R. Ali, "Ensemble method of lung segmentation in chest radiographs," in *Proc. IEEE Nat. Aerosp. Electron. Conf. (NAECON)*, Aug. 2021, pp. 382–385.

[49] J. Shiraishi, S. Katsuragawa, J. Ikezoe, T. Matsumoto, T. Kobayashi, K.-I. Komatsu, M. Matsui, H. Fujita, Y. Kodera, and K. Doi, "Development of a digital image database for chest radiographs with and without a lung nodule: Receiver operating characteristic analysis of Radiologists' detection of pulmonary nodules," *Amer. J. Roentgenol.*, vol. 174, no. 1, pp. 71–74, Jan. 2000.

[50] JSRT. *Japanese Society of Radiological Technology*. [Online]. Available: http://db.jsrt.or.jp/eng.php

[51] F. Yang, G. Zamzmi, S. Angara, S. Rajaraman, A. Aquilina, Z. Xue, S. Jaeger, E. Papagiannakis, and S. K. Antani, "Assessing inter-annotator agreement for medical image segmentation," *IEEE Access*, vol. 11, pp. 21300–21312, 2023.

[52] A. Iqbal, M. Usman, and Z. Ahmed, "Tuberculosis chest X-ray detection using CNN-based hybrid segmentation and classification approach," *Biomed. Signal Process. Control*, vol. 84, Jul. 2023, Art. no. 104667.

[53] I. Ullah, F. Ali, B. Shah, S. El-Sappagh, T. Abuhmed, and S. H. Park, "A deep learning based dual encoder–decoder framework for anatomical structure segmentation in chest X-ray images," *Sci. Rep.*, vol. 13, no. 1, p. 791, Jan. 2023.

[54] S. I. Nafisah and G. Muhammad, "Tuberculosis detection in chest radiograph using convolutional neural network architecture and explainable artificial intelligence," *Neural Comput. Appl.*, vol. 36, no. 1, pp. 111–131, Jan. 2024.

[55] Z. Li, K. Kamnitsas, C. Ouyang, C. Chen, and B. Glocker, "Context label learning: Improving background class representations in semantic segmentation," *IEEE Trans. Med. Imag.*, vol. 42, no. 6, pp. 1885–1896, Feb. 2023.

[56] B. van Ginneken, M. B. Stegmann, and M. Loog, "Segmentation of anatomical structures in chest radiographs using supervised methods: A comparative study on a public database," *Med. Image Anal.*, vol. 10, no. 1, pp. 19–40, Feb. 2006.

[57] *MATLAB*, MathWorks, Natick, MA, USA. [Online]. Available: https://www.mathworks.com/products/MATLAB.html

[58] K. Zuiderveld, "Contrast limited adaptive histograph equalization," in *Graphic Gems IV*. San Diego, CA, USA: Academic, 1994, pp. 474–485.

[59] J. Bertels, T. Eelbode, M. Berman, D. Vandermeulen, F. Maes, R. Bisschops, and M. B. Blaschko, "Optimizing the dice score and Jaccard index for medical image segmentation: Theory and practice," in *Proc. Int. Conf. Med. Image Comput. Comput. Assist. Intervent.*, Jan. 2019, pp. 92–100.

[60] J. Shi, K. Zhang, C. Guo, Y. Yang, Y. Xu, and J. Wu, "A survey of label-noise deep learning for medical image analysis," *Med. Image Anal.*, vol. 95, Jul. 2024, Art. no. 103166.

[61] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognit. Lett.*, vol. 27, no. 8, pp. 861–874, Jun. 2006.

[62] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 2, pp. 318–327, Feb. 2020.

[63] A. Trivedi, C. Robinson, M. Blazes, A. Ortiz, J. Desbiens, S. Gupta, R. Dodhia, P. K. Bhatraju, W. C. Liles, J. Kalpathy-Cramer, A. Y. Lee, and J. M. L. Ferres, "Deep learning models for COVID-19 chest X-ray classification: Preventing shortcut learning using feature disentanglement," *PLoS ONE*, vol. 17, no. 10, Oct. 2022, Art. no. e0274098.

[64] C.-Y. Ou, I.-Y. Chen, H.-T. Chang, C.-Y. Wei, D.-Y. Li, Y.-K. Chen, and C.-Y. Chang, "Deep learning-based classification and semantic segmentation of lung tuberculosis lesions in chest X-ray images," *Diagnostics*, vol. 14, no. 9, p. 952, Apr. 2024.

SAHASAT KHUMANG received the B.S. degree in computer science from the Prince of Songkla University, Songkhla, Thailand, in 2021, where he is currently pursuing the M.S. degree. His current research interests include medical image processing and machine learning.

SUPAPORN KANSOMKEAT received the B.S. degree in mathematics and the M.S. degree in computer science from the Prince of Songkla University, Songkhla, Thailand, in 1991 and 1995, respectively, and the Ph.D. degree in computer engineering from Chulalongkorn University, in 2007. Since 1996, she has been an Instructor with the Division of Computational Science, Faculty of Science, Prince of Songkla University. Her research interests include software testing, image processing, and machine learning.

WIWATANA TANOMKIAT is currently an Associate Professor with the Department of Radiology, Faculty of Medicine, Prince of Songkla University, Songkhla, Thailand. He is also the President of the Radiological Society of Thailand. His current research interests include tuberculosis, occupational lung disease, interstitial lung disease, and artificial intelligence. He is the Editor-in-Chief of *The ASEAN Journal of Radiology*.

SATHIT INTAJAG received the M.S. and Ph.D. degrees in electrical engineering from the King Mongkut's Institute of Technology Ladkrabang (KMITL), Thailand, in 1998 and 2005, respectively. He was an Assistant Professor and an Associate Professor, in 2003 and 2006, respectively. Since 2010, he has been an Instructor with the Division of Computational Science, Faculty of Science, Prince of Songkla University, Songkhla, Thailand. His research interests include signal processing, statistical analysis, and computer vision. He is a Reviewer of IEEE ACCESS.