

Improved TB classification using bone-suppressed chest radiographs

Sivaramakrishnan Rajaraman ^{1,*}, Ghada Zamzmi ¹, Les Folio ², Philip Alderson ³ and Sameer Antani ¹

¹United States National Library of Medicine, United States

²National Institutes of Health Clinical Center, United States

³Saint Louis University, United States

Abstract

Chest X-rays (CXRs) are the most commonly performed diagnostic examination to detect cardiopulmonary abnormalities. However, the presence of bony structures such as ribs and clavicles can obscure subtle abnormalities resulting in diagnostic errors. This study aims to build a deep learning (DL)-based bone suppression model that identifies and removes these occluding bony structures in frontal CXRs to assist in reducing errors in radiological interpretation, including DL workflows, related to detecting manifestations consistent with Tuberculosis (TB). Several bone suppression models with various deep architectures are trained and their performances are evaluated in a cross-institutional test setting. The best-performing model (ResNet-BS) is used to suppress bones in the Shenzhen and Montgomery TB CXR collections. A VGG-16 model is pretrained on a large collection of publicly available CXRs. The CXR-pretrained model is then fine-tuned individually on the non-bone-suppressed and bone-suppressed CXRs of Shenzhen and Montgomery TB CXR collections to classify them as showing normal lungs or TB manifestations. The performances of these models are compared using several performance metrics, analyzed for statistical significance, and their predictions are qualitatively interpreted through class-selective relevance maps (CRM). It is observed that the models trained on bone-suppressed CXRs significantly outperformed the models trained individually on the non-bone-suppressed CXRs ($p < 0.05$) in the Shenzhen and Montgomery TB collections. Models trained on bone-suppressed CXRs improved detection of TB-consistent findings and resulted in compact clustering of the data points in the feature space signifying that bone suppression improved the model sensitivity toward TB classification.

Introduction

Several thousand people die every year from lung-related diseases and their complications¹. Chest X-rays (CXRs) are the most frequently performed diagnostic examination that helps detect various cardiopulmonary abnormalities. However, it can be difficult for radiologists and computer-aided diagnostic (CADx) systems to detect and localize subtle findings related to tuberculosis (TB) when they occur in apical regions in which lung parenchyma is obscured by overlying ribs and the clavicles². Due to the two-dimensional nature of image projection, the posterior and anterior bony structures on a

typical CXR overlap with the lung tissues, resulting in a cross-hatching pattern. Further, the resulting strong edges from ribs and clavicles may occlude abnormalities in the lung regions thereby complicating diagnosis. Therefore, removing the superimposing bony structures could assist in reducing interpretation errors and enhance the value of screening digital chest radiography in underserved and remotely located populations³.

Bone suppression involves subtracting the bones from the CXRs to create a soft-tissue image. It would be of potential use to radiologists and CADx systems in screening for subtle lung abnormalities by increasing the quality of soft-tissue visibility. A common practice for suppressing bony structures involves the use of dual-energy subtraction (DES) chest radiography. The DES-based radiographic acquisition is performed to improve diagnosis by producing two different images, thereby separating the bony structures from the soft-tissues. However, compared to conventional CXRs, DES has several limitations: (a) DES radiography exposes the subject to slightly higher radiation doses compared to conventional CXR acquisition protocols and is not recommended for patients younger than 16 years of age⁴; (b) DES is not used for portable chest radiography, which limits its use in low and middle resource regions (LMRR); and, (c) DES is performed only on the posterior-anterior view.

Literature review reveals several image processing techniques for automated detection and removal of bony structures in CXRs^{5,6}. In one study⁷, the authors used a multi-resolution artificial neural network to generate bony structures, subtracted these from the original CXRs to suppress the clavicles and ribs, and generated soft-tissue images. Another study⁸ used independent component analysis to separate the ribs and soft-tissues in CXRs to increase the visibility of lung nodules. Following this study, subsequent research adopted bone suppression to improve the detection of lung nodules and other pulmonary abnormalities⁹⁻¹¹ including pneumonia detection¹².

Inspired by their superior performance in natural and medical image recognition tasks, convolutional neural networks (ConvNets) have supplanted traditional techniques to perform bone suppression in CXRs. In one study¹³, the authors used a cascade of ConvNets to predict bony structures at multiple resolutions and fused them to produce the final estimate of the bone image. The fused images are subtracted from their respective CXRs to produce soft-tissue images. In another study¹⁴, the authors used a custom ConvNet model to classify the original, lung-segmented, and bone-suppressed versions of the Japanese Society of Radiological Technology (JSRT) CXR dataset¹⁵. It was observed that the model trained on the bone-suppressed dataset offered superior performance toward nodule detection, compared to those trained on the original and lung-segmented datasets.

Bone suppression would help detect TB-consistent findings that often manifest in the apical lung regions so that their visibility is not obstructed by the occlusion of ribs and clavicles^{2,16}. The effect of bone suppression on improving TB detection is discussed in the literature. For instance, the authors¹⁷ fused information from local and global texture descriptors and a clavicle detection system toward detecting TB manifestations in CXRs. The performance with the fused detection system was observed to be superior (area under the

curve (AUC) = 0.86) compared to using only the textural features (AUC = 0.67). In another study¹⁸, the authors compared the performance of two CADx systems toward detecting TB-consistent findings in CXRs. One of the systems was trained on bone-suppressed images generated by commercially available software and the other was trained using original CXRs. It was observed that the CADx system trained on bone-suppressed images delivered superior performance in classifying CXRs as showing TB-consistent findings or normal lungs compared to the other CADx system trained on the original CXRs. CXRs were digitally reconstructed from CT images in another study¹⁹. The authors suppressed bones in these reconstructed CXRs by leveraging a bone decomposition model that was trained on unpaired CT images. A ConvNet-based model was proposed²⁰ to extract bones from CXRs and subtract them from the original input CXRs to generate bone-subtracted images. In another study²¹, the authors performed multi-level wavelet-based decomposition to predict bone images and subtract them from the original CXRs to produce bone-suppressed images. Other than these studies, literature that discusses the effect of bone suppression on TB detection is limited. Also, these methods involve multiple steps including predicting bony structures and then subtracting them from the original image to create bone-suppressed images. However, the literature is limited considering the availability of a bone suppression approach that would directly produce a bone-suppressed image from the input CXR. At the time of writing this manuscript, there is no literature available that evaluates the use of ConvNet-based bone suppression models toward improving automated detection of TB-consistent findings in CXRs.

In this study, we propose a systematic methodology toward training customized ConvNet-based bone suppression models and evaluating their performance toward classifying and detecting TB-consistent findings in CXRs: First, we retrain an ImageNet-trained VGG-16²² model on a large-scale collection of publicly available CXRs from varied sources, where images were acquired for different clinical goals, to help it learn CXR modality-specific features. This model is hereafter referred to as the *CXR-VGG-16* model. We use the VGG-16 model as it has demonstrated superior classification and localization performances in CXR classification tasks²³. Next, we assess the performance of the CXR-VGG-16 model toward classifying CXRs in the Shenzhen and Montgomery TB CXR collections²⁴ as showing normal lungs or pulmonary TB manifestations. These are referred to as the *baseline* models. Then, we train several customized ConvNet-based bone suppression models with varying architecture on the JSRT CXR dataset¹⁵ and its bone-suppressed counterpart²⁵. We conduct cross-institutional testing using the National Institutes of Health (NIH) clinical center (CC) dual-energy subtraction (DES) CXR test set²⁶. The best performing model is then used to suppress the bones in the Shenzhen and Montgomery TB CXR collections. The CXR-VGG-16 model is individually fine-tuned on the bone-suppressed images of the Shenzhen and Montgomery TB CXR collections toward classifying them as showing normal lungs or pulmonary TB manifestations. They are referred to as *bone-suppressed* models. Finally, the performance of the baseline and bone-suppressed models is quantitatively compared through several performance metrics and analyzed for statistically significant differences. Also, the predictions of the baseline and bone-suppressed models are qualitatively

interpreted using class-selective relevance maps (CRM) visualization²⁷. Supplementary Figure 1 shows the stages involved in the proposed approach.

The contributions of this retrospective study are highlighted as follows:

- This is the first study to propose and compare the performance of several customized ConvNet-based bone suppression models with a diversified architecture, including a sequential ConvNet model, an autoencoder (AE) model, a residual learning (RL) model, and a residual network (ResNet) model toward suppressing bones in CXRs.
- This study performs rigorous empirical evaluations, statistical significance analysis, and qualitative evaluation of the bone suppression and classification models.
- The models proposed in this study are not limited to the task of CXR bone suppression but can potentially be extended to other image denoising problems.

The rest of the study is organized as follows: Section II discusses the datasets and methods used, Section III interprets the results, and Section IV discusses and concludes this study.

Materials and Methods

Datasets and their characteristics

The following CXR collections are used in this study:

JSRT CXR: The JSRT¹⁵ released a set of 247 CXR images with and without lung nodules. The collection includes 154 images with a nodule, of which 100 nodules are malignant, 54 are benign, and 93 images are without nodules.

NIH-CC-DES CXR: A set of 27 DES CXRs is acquired as a part of routine clinical care using the GE Discovery XR656 digital radiography system²⁶. The DES images were taken at 120 and 133 Kilovoltage-peak (kVp) to respectively capture the soft-tissue images and bony structures. This dataset is used as the cross-institutional test set to evaluate the performance of the bone suppression models proposed in this study.

Shenzhen TB CXR: This de-identified dataset contains 326 CXRs with normal lungs and 336 abnormal CXRs showing various TB manifestations²⁴. The CXRs are collected from Shenzhen No.3 hospital in Shenzhen, China. It is exempted from institutional review board (IRB) review (OHSRP#5357) by the National Institutes of Health (NIH) Office of Human Research Protection Programs (OHSRP) and made publicly available by the National Library of Medicine (NLM). An equal number of normal and abnormal CXRs (n = 326) is used in this study.

Montgomery TB CXR: The CXR images and their associated radiology reports in this collection are acquired through the TB control program of the Department of Health and Human Services of Montgomery County, Maryland, USA²⁴. The collection includes 58 CXRs showing TB-consistent findings and 80 CXRs with normal lungs. The CXRs are de-identified

to ensure patient privacy and are made publicly available. An equal number of normal and abnormal CXRs ($n = 58$) is used in this study.

Radiological Society of North America (RSNA) CXR: A subset of the NIH CXR dataset²⁸ is curated by the RSNA²⁹ and made publicly available. The collection includes 17,833 frontal CXRs showing various lung abnormalities and 8851 CXRs showing normal lungs.

Pediatric pneumonia CXR: A collection of 4273 CXRs, acquired from children 1 to 5 years of age, showing bacterial and viral pneumonia manifestations, and 1493 normal CXRs is publicly available³⁰. Supplementary Table 1 provides the demographic details of the datasets used in this study.

Bone suppression models

The researchers from the Budapest University of Technology and Economics used their in-house clavicle and rib-shadow removal algorithms to suppress the bones in the 247 JSRT CXRs and made the bone-suppressed soft-tissue images publicly available²⁵. Affine transformations including rotations (-10 to 10 degrees), horizontal and vertical shifting (-5 to 5 pixels), horizontal mirroring, zooming, median, maximum, and minimum, and unsharp masking are used to generate 4500 image pairs from this initial set of CXRs and their bone-suppressed counterparts. The augmented images are resized to 256×256 spatial resolution. The image contrast is enhanced by saturating the bottom and top 1% of all image pixel values. The grayscale pixel values are then normalized.

Several ConvNet-based bone suppression models with varying architecture are trained on this augmented dataset. We evaluated their performance with the cross-institutional NIH-CC-DES test set. During training, we allocated 10% of the training data for validation using a fixed seed. Four different model architectures are proposed toward the task of bone suppression in CXRs as follows: (a) Autoencoder (AE) model (AE-BS) where BS denotes bone suppression; (b) Sequential ConvNet model (ConvNet-BS); (c) Residual learning model (RL-BS); and (d) Residual network model (ResNet-BS). The architectures of these models follow:

AE-BS Model

The AE-BS model is a convolutional denoising AE with symmetrical encoder and decoder layers. The encoder consists of three convolutional layers with 16, 32, and 64 filters, respectively. The size of the input is decreased twice at the encoder layers and increased correspondingly in the decoder layers. As opposed to the conventional denoising AEs, the noise in the proposed AE-BS model represents the bony structures. The model trains on the original CXRs and their bone-suppressed counterparts to predict a bone-suppressed soft-tissue image. Supplementary Figure 2 illustrates the architecture of the proposed AE-BS model.

ConvNet-BS model

The ConvNet-BS model is a sequential model consisting of seven convolutional layers having 16, 32, 64, 128, 256, 512, and 1 filter, respectively. Zero paddings are used to preserve

the dimensions of the input image at all convolutional layers. Lasso regularization (L1) penalties are used at each convolutional layer to induce penalty on weights that seldom contribute to learning meaningful feature representations. This helps in improving model sparsity and generalizing to unseen data. The deepest convolutional layer with the sigmoidal activation produces the bone-suppressed soft tissue image. Supplementary Figure 3 illustrates the architecture of the proposed ConvNet-BS model.

RL-BS model

The architecture of the RL-BS model consists of eight convolutional layers having 8, 16, 32, 64, 128, 256, 512, and 1 filter, respectively. Zero paddings are used at all convolutional layers to preserve the dimensions of the input image. The RL-BS model learns the residual error between the predicted bone-suppressed image and its corresponding ground truth. The deepest convolutional layer produces bone-suppressed images. Supplementary Figure 4 shows the architecture of the proposed RL-BS model. The RL-BS model learns the residual error between the predictions and ground truth to produce bone-suppressed images.

ResNet-BS model

The architecture of the proposed ResNet-BS model is illustrated in Figure 1. The residual design utilizes shortcuts to skip over layers thereby eliminating learning convergence issues due to vanishing gradients. This facilitates reusing previous layer activations until the weights are updated in the adjacent layer. These shortcuts lead to improved convergence and optimization and help to construct deeper models.

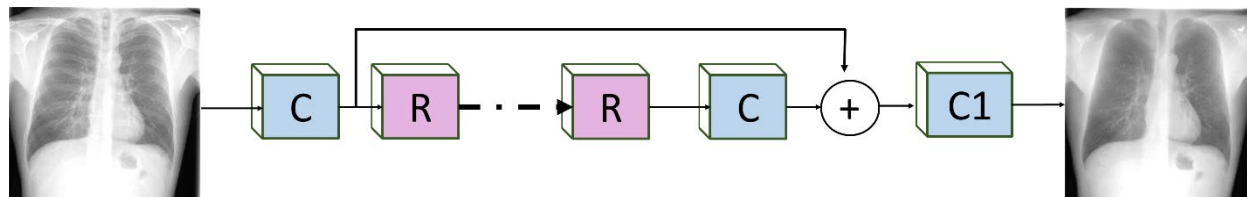


Figure 1. The architecture of the proposed ResNet-BS model. The convolutional block is denoted by C, having 64 filters of size 3×3 and zero paddings to preserve input dimensions. R denotes the modified ResNet block where the final ReLU activation is removed together with the batch normalization layer. The proposed model has 16 ResNet blocks. The deepest convolutional layer C1 with a single filter, zero paddings, and sigmoidal activation, predicts the bone-suppressed image.

Inspired by³¹, ReLU activation layers are not used outside the residual blocks. This literature³¹ also demonstrates that batch normalization leads to loss of information and reduces the range tractability of activations. Hence, the batch normalization layer and the final ReLU activation are removed from each ResNet block. A sequence of 16 ResNet blocks are used, each having 64 filters of size 3×3 and zero paddings to preserve original image dimensions. Scaling layers with a scaling factor of 0.1 are added after the deepest convolutional layer in each ResNet block to scale down the residuals before adding them back to the convolutional path³². The deepest convolutional layer with the sigmoidal activation predicts the bone-suppressed image.

Evaluating bone suppression models

The bone suppression models are trained to suppress the bony structures in the CXRs and produce soft-tissue images. This can be treated as an image denoising problem where the bones are considered noise. To obtain superior bone suppression results, we aim to reduce the error between the predicted bone-suppressed image and its ground truth and maximize the structural similarity. The selection of the loss function plays a prominent role in the bone suppression task.

In this study, the performance of the proposed bone suppression models is evaluated through constructing a loss function that benefits from the combination of mean absolute error (MAE) and multi-scale structural similarity index measure (MS-SSIM) losses, herein referred to as *combined loss*. Other pixel-based evaluation metrics used in this study include peak signal-to-noise ratio (PSNR) and structural similarity index measure (SSIM). The mean-squared error (MSE), also known as L2 loss, is a pixel loss measure that computes the sum of the squared distance between the predicted image and its ground truth. However, MSE does not interpret the quality of the predicted image. The MAE, otherwise called L1 loss computes the sum of absolute differences between the ground truth and the predicted image. Studies in the literature reported that, unlike MSE, MAE provides a more natural measure of average error, and is useful in performing inter-comparisons of average model-performance errors³³. PSNR computes the peak signal-to-noise ratio between the predicted and ground truth images. This ratio is used to provide a quantitative assessment of the predicted image. A higher value for PSNR indicates a higher quality of prediction. SSIM provides a measure of similarity between the ground truth and predicted images. A previous study³⁴ reveals that SSIM provides a superior indication of prediction performance as it exemplifies human visual perception. The MS-SSIM measure is an extension of SSIM that computes structural similarity at various scales and combines them. Another study³⁵ reveals that MS-SSIM is an improved measure to use compared to SSIM while characterizing the performance of the models because (a) it is measured over multiple scales, and (b) it is demonstrated to preserve contrast at higher frequencies compared to SSIM. On the other hand, MAE preserves luminance and contrast in the predicted image. The mathematical formulations of these metrics can be found in the literature³³⁻³⁵.

Combined loss function

We propose to train the bone suppression models using a combined loss function that benefits from both MAE and MS-SSIM as follows:

$$Loss_{Combined} = \Omega . MS - SSIM + (1 - \Omega) . MAE \quad (1)$$

We set the value of $\Omega = 0.84$ after empirical evaluations. Greater weight is given to MS-SSIM since we want the bone suppressed image to be highly similar (i.e., least structural alteration) to the ground truth. The MAE is given lower significance in this measure as it focuses on overall luminance and contrast in the image which are expected to change due to bone (white pixels) suppression.

Histogram similarity assessment

The histograms of the ground truth and the bone-suppressed image predicted by the proposed models are plotted and compared to observe their tonal distributions. Various metrics including correlation, intersection, chi-squared distance, Bhattacharyya distance, and Earthmover distance (EMD) are used to compare these histograms and provide a measure of similarity. The higher the value of correlation and intersection, the closer (or more similar) is the histogram of the image pairs. This implies the histogram of the predicted bone-suppressed image closely matches that of the ground truth. For distance-based metrics including chi-squared, Bhattacharyya, and EMD, a smaller value indicates a superior match between the histogram pairs, signifying that the predicted bone-suppressed image closely matches that of the ground truth. The mathematical formulations of these metrics can be found in the literature³⁶.

Classification models

In this study, an ImageNet-pretrained VGG-16 model²² is retrained on a large collection of CXRs combined using RSNA CXR and pediatric pneumonia CXR data collections producing sufficient diversity in terms of image acquisition and patient demographics to learn the characteristics of abnormal and normal lungs. This VGG-16 model is truncated at its deepest convolutional layer and appended with a global average pooling (GAP) layer, a dropout layer with an empirically determined dropout ratio (0.5), and an output layer with two nodes to predict probabilities of the input CXRs as showing normal lungs or other pulmonary abnormalities. This CXR modality-specific retraining helps in improving the specificity of the network weights conforming to the CXR classification task under study. This approach is followed to learn CXR modality-specific characteristics about the normal lungs and an extensive selection of pulmonary abnormalities. The modality-specific knowledge would be relevant to be transferred to the CXR classification task as compared to using the ImageNet weights from the natural image processing domain. A previous study³⁷ shows the benefits of using CXR modality-specific models retraining toward improving classification and localization performance and model generalization.

During this training step, the data are split at the patient-level into 90% for training and 10% for testing. We allocated 10% of the training data for validation using a fixed seed value. This CXR-VGG-16 model is fine-tuned on the original Shenzhen and Montgomery TB CXR collection (baseline models) and their bone-suppressed counterparts (bone-suppressed models) to classify them as showing normal lungs or pulmonary TB manifestations. The bone-suppressed datasets are constructed by using the best-performing bone suppression model among the proposed models.

Toward this classification task, four-fold cross-validation is performed in which the CXRs in the Shenzhen and Montgomery collections are split at the patient-level into four equal folds. The hyperparameters of the models are tuned while training on the three folds and validating with the fourth fold. The validation process is repeated with each fold, resulting in four different models. During model training, data are augmented with random horizontal and vertical pixel shifts (-5 to 5 pixels), horizontal mirroring, and rotations (-10

to 10 degrees) to introduce data diversity into the training process and reduce overfitting to the training data. Class weights are used to penalize majority classes and reduce class-imbalance errors. The models are trained and evaluated using stochastic gradient descent (SGD) optimization to estimate learning error and classification performance. Callbacks are used to store checkpoints of the models. The model weights delivering superior performance with their respective validation fold are used for further analysis.

The ground truth disease annotations for the Montgomery TB dataset were provided by an expert radiologist with more than 45 years of experience. The ground truth disease annotations for a subset ($n = 68$) of the Shenzhen TB dataset were provided by another expert radiologist with more than 30 years of experience. The web-based, VGG Image Annotator tool³⁸ was used by the radiologists to independently annotate the collections. The radiologists were asked to annotate TB-consistent ROIs using rectangular bounding boxes. These annotations were exported to JSON for subsequent analyses.

The performance of the models is quantitatively compared using the following metrics and analyzed for statistical significance: (a) Accuracy; (b) AUC; (c) Sensitivity; (d) Specificity; (e) Precision; (f) F-measure; and (g) Matthews correlation coefficient (MCC). The predictions of the best-performing models trained on the baseline and bone-suppressed data are interpreted through CRM-based visualization. A Windows® system with Intel Xeon CPU and NVIDIA GeForce GTX 1070 graphics card and Keras DL framework with Tensorflow backend is used to train the models. The trained models and codes will be made available upon request.

Statistical analysis

We performed statistical analyses to identify the existence of a statistically significant difference in performance between the models. For the bone suppression task, we used 95% confidence intervals (CI) as the “Wilson” score interval for the MS-SSIM metric to compare the performance of the proposed bone suppression models and estimate their precision through the error margin. For the classification task, we used one-way analysis of variance (ANOVA)³⁹ to investigate if there exists a statistically significant difference in the MCC values obtained using the baseline and bone-suppressed models. Before performing one-way ANOVA analysis, we conducted Shapiro-Wilk and Levene’s tests⁴⁰ to check if the assumptions of data normality and variance homogeneity are satisfied. We used R statistical software (Version 3.6.1) to perform these evaluations.

Results

Bone Suppression

Recall that the proposed bone suppression models are trained on the augmented JSRT dataset and its bone-suppressed counterpart. The performance of the trained models is evaluated with the cross-institutional NIH-CC-DES test set ($n = 27$). The performance achieved by the various bone suppression models is shown in Table 1.

Table 1. Performance achieved by the proposed bone suppression models using the cross-institutional NIH-CC-DES test set. Data in parenthesis are 95% CI for the MS-SSIM values measured as the “Wilson” score interval. Combined loss = $0.16 * MAE + 0.84 * MS-SSIM_{loss}$. The best performances are denoted by bold numerical values in the corresponding columns. The ResNet-BS model statistically significantly outperformed the AE-BS model in all categories ($p < 0.05$) and the ConvNet-BS and RL-BS models for the PSNR metric ($p < 0.05$). For other metrics, the ResNet-BS model demonstrated superior performance than the CNN-BS and RL-BS models.

Model	Combined loss	MAE	MS-SSIM _{loss}	PSNR	SSIM	MS-SSIM
AE-BS	0.0251	0.0212	0.0258	30.462	0.921	0.9742 (0.8759, 1.0)
ConvNet-BS	0.0217	0.0198	0.0221	30.952	0.935	0.9779 (0.8867, 1.0)
RL-BS	0.0211	0.0219	0.021	31.749	0.937	0.979 (0.8901, 1.0)
ResNet-BS	0.0167	0.014	0.0172	34.068	0.949	0.9828 (0.9022, 1.0)

It is observed that the 95% CI for the MS-SSIM metric achieved by the ResNet-BS model demonstrates a tighter error margin and hence higher precision compared to the other models. The ResNet-BS model demonstrated the least values for the combined loss, MAE, and MS-SSIM_{loss} and superior values for PSNR, SSIM, and MS-SSIM. The ResNet-BS model statistically significantly outperformed the AE-BS model ($p < 0.05$) and the ConvNet-BS and RL-BS models for the PSNR metric ($p < 0.05$). For other metrics, the ResNet-BS model demonstrated superior performance than the CNN-BS and RL-BS models.

Figure 2 shows the final bone suppression images along with the original unsuppressed CXR from a normal CXR in the NIH-CC DES test set. All approaches appear to show substantial suppression of the bony structures in the apical regions. For differentiation among them, quantitative indices are needed. A quantitative comparison of the bone suppressed CXR images in Fig. 2 is provided by histogram similarity comparisons in Fig. 3 and Table 2 that follow. Based on the comparison findings, the ResNet-BS model was used in subsequent analyses.

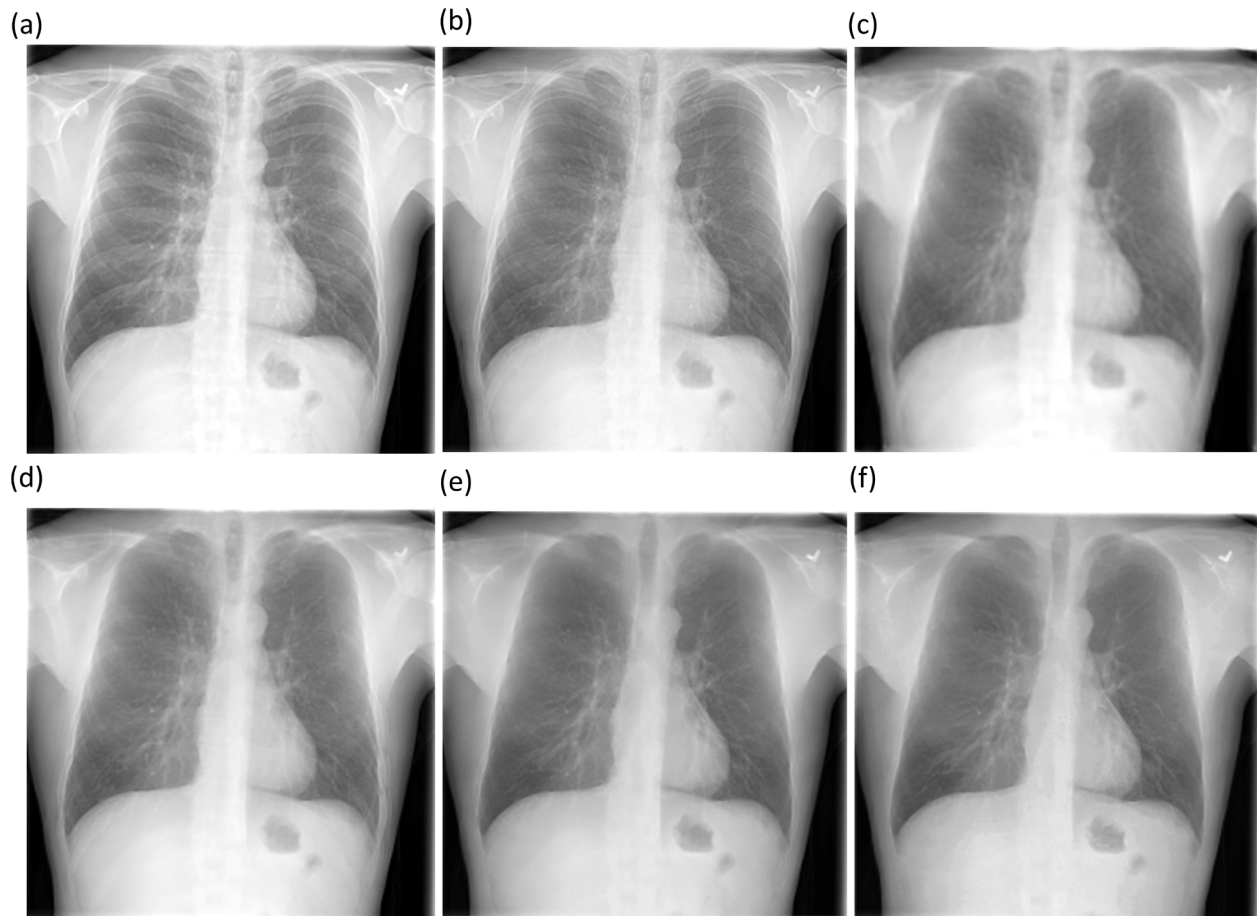


Figure 2. Bone-suppressed CXR images predicted by the proposed models using a CXR sample from the cross-institutional NIH-CC DES test set. (a) Original CXR; (b) AE-BS model; (c) ConvNet-BS model; (d) RL-BS model; (e) ResNet-BS model; and (f) Ground truth.

Figure 3 shows several comparisons of the histogram of the images predicted using the bone suppression models and the histogram of the ground truth using the sample CXR from Figure 2. It is observed from Fig. 3(d) that the histogram of the bone-suppressed image predicted by the ResNet-BS model closely matched the ground truth compared to the histogram obtained with other models. We assessed the similarity of the histograms of the predicted images to the ground truth through several performance metrics including correlation, intersection, chi-squared distance, Bhattacharyya distance, and EMD as shown in Table 2.

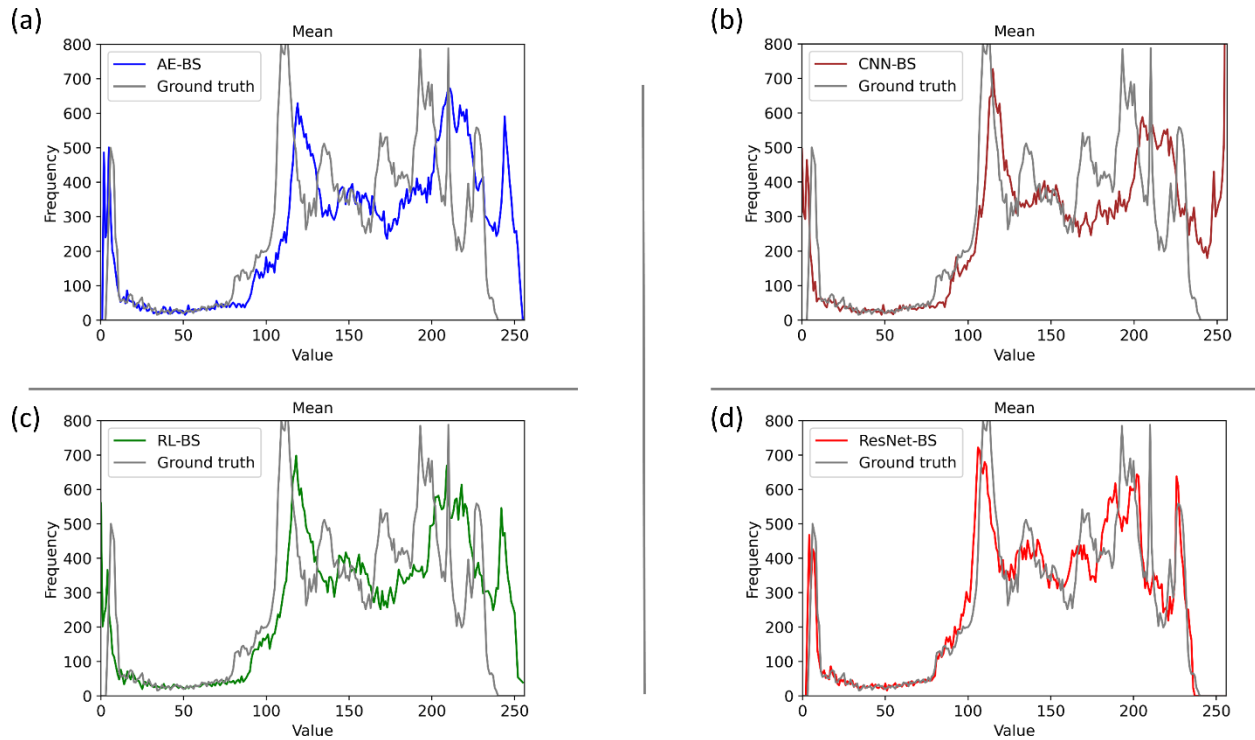


Figure 3. Comparing the histogram of the predicted image using the proposed bone suppression models and the ground truth using the sample CXR from Figure 2. (a) Ground truth and AE-BS model; (b) Ground truth and CNN-BS model; (c) Ground truth and RL-BS model; and (d) Ground truth and ResNet-BS model.

Table 2. Histogram similarity assessment. The similarity of the histograms of the predicted images using the bone suppression models and their corresponding ground truths are measured. Bold numerical values denote superior performance in respective rows.

Method	Histogram pairs				
	GT-GT	GT-AE-BS	GT-ConvNet-BS	GT-RL-BS	GT-ResNet-BS
Correlation	1	0.4368	0.4406	0.4644	0.6723
Intersection	10.627 3	7.1058	7.1681	7.2151	9.2880
Chi-squared distance	0	122.59	80.9075	60.30	1.7931
Bhattacharyya distance	0	0.4288	0.4272	0.4249	0.3595
EMD	0	0.0141	0.0135	0.0114	0.0089

We observed from Table 2 that the similarity of the ground truth to itself resulted in a value of 0 for all the distance measures and a value of 1 for the correlation metric. This demonstrates a perfect match. Higher values for the correlation and intersection metrics computed using the GT-ResNet-BS histogram pair demonstrate that the histogram of the

ResNet-BS-predicted bone-suppressed image closely matches that of the ground truth image. For distance-based metrics including chi-squared, Bhattacharyya, and EMD, a smaller value indicates a superior match between the histogram pairs. This signifies that compared to other models, the bone-suppressed image predicted by the ResNet-BS closely matches that of the ground truth.

The best performing ResNet-BS model is further used to suppress bones in Shenzhen and Montgomery TB CXR collections. Supplementary Figure 5 shows the bone-suppressed instances of a sample CXR from the Shenzhen and Montgomery CXR collections. It is observed that the ResNet-BS model generalized to the Shenzhen and Montgomery TB CXR collections that are not seen by the model during training or validation. The bone shadows are completely suppressed and the resolution of the CXRs is preserved.

Recall that for the classification task, the CXRs in the Shenzhen and Montgomery TB CXR collections are split at the patient-level into four equal folds for performing cross-validation studies. The mean performance of the cross-validated models is given in Table 3. It is observed that the classification performance achieved with the bone-suppressed models using the Shenzhen and Montgomery TB CXR collections is superior compared to the baseline models. The bone-suppressed models demonstrated superior values for all performance metrics.

Table 3. Mean performance achieved by the cross-validated models using the bone-suppressed and non-bone-suppressed (baseline) CXR instances of the Shenzhen and Montgomery TB CXR dataset. Acc = Accuracy; Sens. = Sensitivity; Spec. = Specificity; Prec. = Precision; F = F-measure. One-way ANOVA is performed using the MCC values obtained by the baseline and bone-suppressed cross-validated models to analyze for the existence of a statistically significant difference in performance. Bold values denote superior performances in corresponding columns for the Shenzhen and Montgomery TB CXR collections. The performance of the bone-suppressed models is statistically significantly superior ($p < 0.05$) to the baseline models in all categories.

Dataset	Model	ACC	AUC	Sens.	Spec.	Prec.	F	MCC
Shenzhen (n = 326)	Bone suppressed	0.8879± 0.0247	0.9535± 0.0186	0.8805± 0.0205	0.8954± 0.0423	0.8949± 0.0376	0.8873± 0.0233	0.7765± 0.0492
	Baseline	0.8304± 0.0117	0.8991± 0.0268	0.8068± 0.0203	0.8537± 0.0345	0.8469± 0.0265	0.8259± 0.0089	0.6620± 0.0238
Montgomery (n = 58)	Bone suppressed	0.9230± 0.0312	0.9635± 0.0106	0.8772± 0.0708	0.9687± 0.0625	0.9706± 0.0588	0.9188± 0.0345	0.8539± 0.0581
	Baseline	0.7701± 0.0820	0.8567± 0.0870	0.7991± 0.1931	0.7411± 0.0342	0.7517± 0.0274	0.7682± 0.1039	0.5537± 0.1761

We analyzed the baseline and bone-suppressed models in the Shenzhen and Montgomery TB CXR collection using one-way ANOVA of the MCC values for the existence of a statistically significant difference in their performance. One-way ANOVA assumes normality of data and homogeneity of variances. For the Shenzhen TB CXR collection, we plotted the residual quantiles against the normal distribution as shown in Supplementary

Figure 6 to check if the normality of the data assumption holds good. We observed that the residuals are normally distributed and approximately followed the 45-degree line of reference signifying that the assumption of data normality is satisfied. We observed that no evident relationships existed between the residuals and the mean value of MCC obtained by the baseline and bone-suppressed models. This shows that the assumption of homogeneity of data variances is satisfied. It is observed that the p -value for Shapiro-Wilk and Levene's analysis is greater than 0.05 (Shapiro-Wilk (p) = 0.7780 and Levene (p) = 0.4268). This signifies that the assumptions of the normal distribution of data and variance homogeneity hold valid. Hence, we performed one-way ANOVA to investigate if there exists a statistically significant difference in the MCC values obtained by the cross-validated baseline and bone-suppressed models. We have the baseline and bone-suppressed models, and each model has four observations for the MCC metric. Thus, the F-statistic is mentioned as F (1, 6). It is observed that a statistically significant difference existed in the MCC values obtained by the baseline and bone-suppressed models ($F(1, 6) = 17.58, p = 0.00573$). This demonstrated that the MCC values obtained by the bone-suppressed models are statistically superior compared to the baseline models.

We performed a similar statistical significance analysis with the Montgomery TB CXR collection. We analyzed the MCC values obtained by the cross-validated baseline and bone-suppressed models for statistical significance using one-way ANOVA. To check if the assumption of data normality is satisfied, we plotted the residual quantiles against that of the normal distribution as shown in Supplementary Figure 7. As in Supplementary Figure 6, the results for the analysis of the bone-suppressed images from the Montgomery TB CXR collection are significantly better. We observed $p > 0.05$ when performing Shapiro-Wilk and Levene's analysis (Shapiro-Wilk (p) = 0.6767 and Levene (p) = 0.808). These p -values underscored that the assumptions of data normality and homogeneity of variances are satisfied. Hence, we performed one-way ANOVA to investigate if there existed a statistically significant difference in the MCC values obtained by the baseline and bone-suppressed models using the Montgomery TB CXRs to classify them as showing normal lungs or pulmonary TB manifestations. We observed that a statistically significant difference existed in the MCC values obtained by the cross-validated baseline and bone-suppressed models ($F(1, 6) = 10.48, p = 0.0177$). This demonstrated that the MCC values obtained by the bone-suppressed models using the Montgomery TB CXR collection are statistically superior compared to the baseline models.

Supplementary Figure 8 and Supplementary Figure 9 show the following visualizations obtained using the best-performing cross-validated bone-suppressed model respectively using the Shenzhen and Montgomery TB CXR collection: (a) confusion matrix; (b) AUC-ROC curve; and (c) Normalized Sankey diagram. Recall that the bone-suppressed models demonstrated statistically superior values for all performance metrics compared to their baseline counterparts.

We also used CRMs to interpret the predictions of the best-performing baseline and bone-suppressed models using the Shenzhen and Montgomery TB CXR collections to localize TB-consistent findings. Figure 4(a) and Figure 4(d) show instances of original CXRs

respectively from the Shenzhen and Montgomery TB CXR collections. The expert ground truth annotations are shown with red bounding boxes. Figure 4(b) and Figure 4(e) show how the best-performing baseline models interpret their prediction toward localizing TB-consistent ROI. Figure 4(c) and Figure 4(f) show the TB-consistent ROI localized by the best-performing bone-suppressed models. It is observed that the bone-suppressed models demonstrated superior TB-consistent ROI localization compared to the baseline models. From Figure 4(b) and Figure 4(e), it is observed that the baseline models are learning the surrounding context but not meaningful features. The TB-consistent ROI localization achieved by the bone-suppressed models conformed to the expert knowledge of the problem as observed from Fig. 4(c) and Fig. 4(f) and showed that it learned meaningful, salient feature representations.

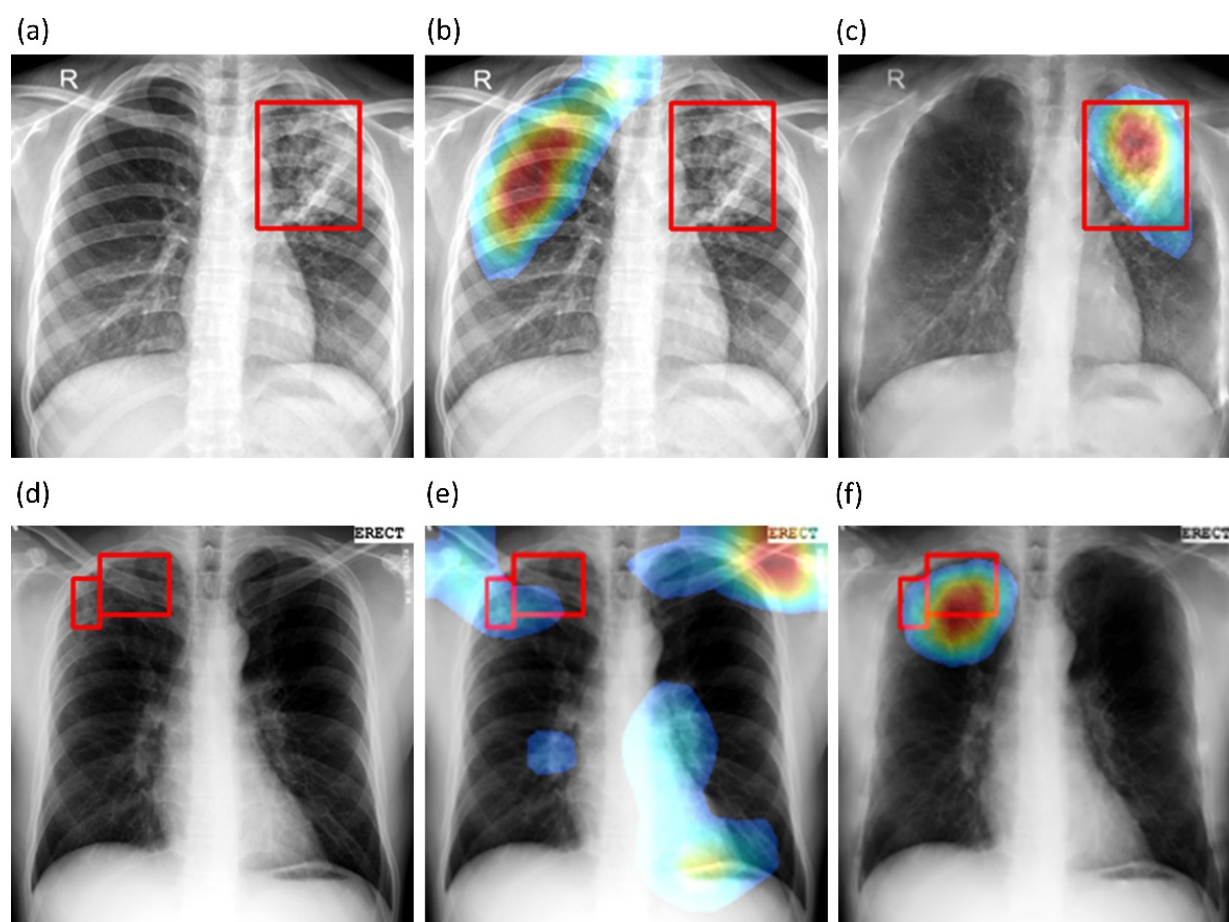


Figure 4. CRM-based TB-consistent ROI localization achieved by the best-performing baseline and bone-suppressed model respectively using a sample CXR from the Shenzhen and Montgomery TB CXR collection. (a) and (d) CXR instances of Shenzhen and Montgomery TB CXRs respectively with expert ground truth annotations (shown with red bounding boxes); (b) and (e) TB-consistent ROI localization achieved using the best-performing baseline model; (c) and (f) TB-consistent ROI localization by the best-performing bone-suppressed model.

We further used the CRM algorithm and the best-performing bone-suppression models to visualize the overall pulmonary location of TB manifestations in Shenzhen and Montgomery

TB CXR collections. The average CRMs for the two datasets are shown in Figure 5. The steps taken to generate the average CRMs independently for the Shenzhen and Montgomery TB collection are (a) The average of CRMs were computed for the TB class in each dataset; (b) The average of the ground truth lung masks for the Montgomery²⁴ and Shenzhen TB CXR⁴¹ collections were computed, and (d) a bitwise-AND operation was performed using the average CRMs and the averaged lung masks to visualize the activations in the lung ROI. The average CRMs appeared quite interesting and showed that the Shenzhen TB-positive group had primarily upper lobe CXR abnormalities. The average CRM obtained using the Montgomery TB CXR collection also showed upper lung predominance as well as other zones.

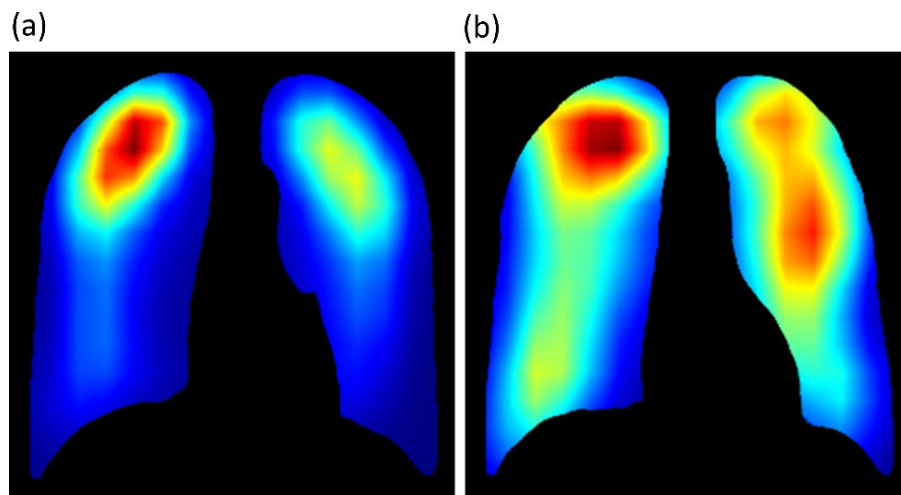


Figure 5. Average CRM computed for the TB class using the (a) Shenzhen and (b) Montgomery TB CXR collection.

We visualized the learned features from the Shenzhen and Montgomery TB CXR collections by the best-performing baseline and bone-suppressed models using t-SNE⁴². The t-SNE is a dimensionality reduction technique that helps to visualize the learned feature space by embedding high-dimensional images into low dimensions while maintaining the pairwise distances of the points. The 512-dimensional vector extracted from the GAP layer of the baseline and bone-suppressed models is plugged into t-SNE to visualize feature embeddings in the two-dimensional space. From Supplementary Figure 10, it is observed that the feature space learned by the bone-suppressed models demonstrated a better and more compact clustering of the normal and TB class features. Such feature clustering facilitates markedly superior separation between the classes as compared to the baseline models. This improved behavior is observed with the bone-suppressed models fine-tuned on both Shenzhen and Montgomery TB CXR collections.

Discussion

Observations made from this study include the need for (i) CXR modality-specific model pretraining, (ii) model customization suiting the problem, (iii) statistical validation,

(iv) localization studies with expert annotations conforming to the problem, and (v) feature embedding visualization.

CXR modality-specific pretraining: Previous studies reveal that compared to using ImageNet weights, CXR modality-specific model pretraining results in learning meaningful modality-specific features that can be transferred to improve performance in a relevant classification task^{37,43}. We performed CXR modality-specific pretraining using a selection of various publicly available CXR data collections to introduce sufficient diversity into the training process in terms of acquisition methods, patient population, and other demographics, to help the models broadly learn significant features from CXRs showing normal lungs and other pulmonary abnormalities. The learned knowledge is transferred to improve convergence and performance in a relevant classification task to classify CXRs as showing normal or TB manifestations. This approach may have helped the DL models to distinguish salient radiological manifestations of normal lungs and TB-consistent findings.

Model customization: Residual networks are one of the most commonly used backbones for computer vision tasks including segmentation, classification, and object detection⁴⁴. The use of residual blocks helps construct and train deeper models since they alleviate the problem of vanishing gradients. In this study, we explored the use of residual networks in the context of an image denoising problem where the bony structures in the CXRs are considered noise. Through empirical evaluations, we observed that the proposed ResNet-BS model outperformed other models by demonstrating superior values for the PSNR, SSIM, and MS-SSIM metrics. The bone-suppressed image predicted by the ResNet-BS model effectively suppressed the bony structures and the image appeared sharp while preserving the soft-tissues, rendering it suitable for lung disease screening/diagnosis.

Statistical validation: Studies in the literature that accomplish bone suppression in CXRs have not performed a quantitative assessment of the bone-suppressed images by comparing them to their respective ground truths⁴⁵. Statistical analysis would help evaluate model performance based on quantitative measures and help distinguish between realistic and uncertain assumptions. In our study, we performed histogram-based similarity assessments using several performance metrics including correlation, intersection, and other distance measures including chi-squared distance, Bhattacharyya distance, and EMD to statistically demonstrate the closeness of the predicted bone-suppressed images with the ground truth. This led to the observation that, unlike other proposed bone suppression models, the histogram of the bone-suppressed CXRs predicted by the ResNet-BS model closely matched their respective ground truth images. We performed statistical analysis using 95% CI as the “Wilson” score interval to investigate the existence of a statistically significant difference in performance between the bone suppression models. We also performed one-way ANOVA analyses to observe the existence of a statistically significant difference in the classification performance using the baseline and bone-suppressed models. To this end, we observed a statistically significant difference ($p < 0.05$) existed in the MCC values obtained using the baseline and bone-suppressed models toward classifying the CXRs in the Shenzhen and Montgomery TB CXR collections. This demonstrated that the bone-suppressed models that

are trained and evaluated individually on the Shenzhen and Montgomery TB CXR collections statistically significantly outperformed their baseline counterparts.

Localization studies: We observed from the CRM-based localization study that the model accuracy is not related to its disease-specific ROI localization ability. The baseline models revealed poor TB-consistent ROI localization compared to the bone-suppressed models. It is observed that the bone-suppressed models learned meaningful feature representations conforming to the expert knowledge of the problem under study. On the other hand, the baseline models also seem to have learned the surrounding context irrelevant to the problem to classify the CXRs to their respective classes. The average CRMs obtained using the Shenzhen and Montgomery TB CXR datasets, collected from TB clinics in two different countries, showed upper lung predominance. These observations conform to the findings in the literature⁴⁶ that discusses that 58% of patients with sputum-positive TB had upper lobe infiltrates. Another study⁴⁷ demonstrated that reactivation TB was especially common in the posterior segment of the upper lobe and superior segment of the lower lobe. On frontal CXRs, those segments can appear to be in the mid-zone. The improved CRM localization achieved using the bone-suppressed models could be attributed to the fact that the suppression of bones helped to detect TB-consistent findings that often manifest in the apical lung regions so that their visibility is not obstructed by the occlusion of ribs and clavicles¹⁶, thereby increasing model sensitivity.

Feature embedding visualization: We visualized the feature space learned by the baseline and bone-suppressed models using the t-SNE dimensionality reduction algorithm that embeds the learned high-dimensional features into the 2-D space. To this end, we observed that the bone-suppressed model demonstrated a compact clustering of the features learned for the TB and normal classes. The decision boundary between the normal and TB categories are well defined, showing that meaningful feature embeddings are learned by the bone-suppressed models.

Limitations: This study, however, suffers from the following limitations: (a) To train and validate the proposed bone-suppression models, we used limited data that may not encompass a wide range of bone structure variability. With the increased availability of bone-suppressed CXRs, it would be possible to train deeper architectures with sufficient data diversity to build confidence in the models and improve their generalization to real-world data, and (b) The impact of this approach on patient triage and treatment planning can only be theorized. Deriving guidance for them is beyond the scope of this study.

In sum, models trained on bone-suppressed CXRs improved detection of TB-consistent findings resulted in compact clustering of the data points in the feature space signifying that bone suppression improved the model sensitivity toward TB classification. The models proposed in this study are not limited to improving TB detection. The results suggest that the proposed ResNet-BS bone suppression model could be extended to other CXR applications such as improved performance in detecting and differentiating lung nodules, pneumonia, COVID-19, and other pulmonary abnormalities. This could further enhance the utility of digital CXRs for the evaluation of pulmonary disorders for underserved

patients in low-resource or remote locations. We believe our results will improve human visual interpretation of TB findings, as well as automated detection in AI-driven workflows.

Acknowledgments

This research was supported by the Intramural Research Program of the National Library of Medicine, National Institutes of Health.

References

1. Britto, C. J., Brady, V., Lee, S. & Dela Cruz, C. S. Respiratory Viral Infections in Chronic Lung Diseases. *Clinics in Chest Medicine* (2017). doi:10.1016/j.ccm.2016.11.014
2. Jaeger, S. *et al.* Automatic screening for tuberculosis in chest radiographs: a survey. *Quant. Imaging Med. Surg.* **3**, 89–99 (2013).
3. Hogeweg, L., Sanchez, C. I. & Van Ginneken, B. Suppression of translucent elongated structures: Applications in chest radiography. *IEEE Trans. Med. Imaging* (2013). doi:10.1109/TMI.2013.2274212
4. Kuhlman, J. E., Collins, J., Brooks, G. N., Yandow, D. R. & Broderick, L. S. Dual-energy subtraction chest radiography: What to look for Beyond calcified nodules. *Radiographics* (2006). doi:10.1148/rg.261055034
5. Karargyris, A., Antani, S. & Thoma, G. Segmenting anatomy in chest x-rays for tuberculosis screening. in *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS* (2011). doi:10.1109/IEMBS.2011.6091917
6. Loog, M. & Van Ginneken, B. Bony structure suppression in chest radiographs. in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (2006). doi:10.1007/11889762_15
7. Suzuki, K., Abe, H., MacMahon, H. & Doi, K. Image-processing technique for suppressing ribs in chest radiographs by means of massive training artificial neural network (MTANN). *IEEE Trans. Med. Imaging* (2006). doi:10.1109/TMI.2006.871549
8. Nguyen, H. X. & Dang, T. T. Ribs suppression in chest X-Ray images by using ICA method. in *IFMBE Proceedings* (2015). doi:10.1007/978-3-319-11776-8_47
9. Freedman, M. T., Lo, S. C. B., Seibel, J. C. & Bromley, C. M. Lung nodules: Improved detection with software that suppresses the rib and clavicle on chest radiographs. *Radiology* (2011). doi:10.1148/radiol.11100153
10. Oda, S. *et al.* Performance of radiologists in detection of small pulmonary nodules on chest radiographs: Effect of rib suppression with a massive-training artificial neural network. *Am. J. Roentgenol.* (2009). doi:10.2214/AJR.09.2431
11. Li, F. *et al.* Small lung cancers: Improved detection by use of bone suppression imaging - Comparison with dual-energy subtraction chest radiography. *Radiology* (2011). doi:10.1148/radiol.11110192
12. Li, F., Engelmann, R., Pesce, L., Armato, S. G. & MacMahon, H. Improved detection of focal pneumonia by chest radiography with bone suppression imaging. *Eur. Radiol.* (2012).

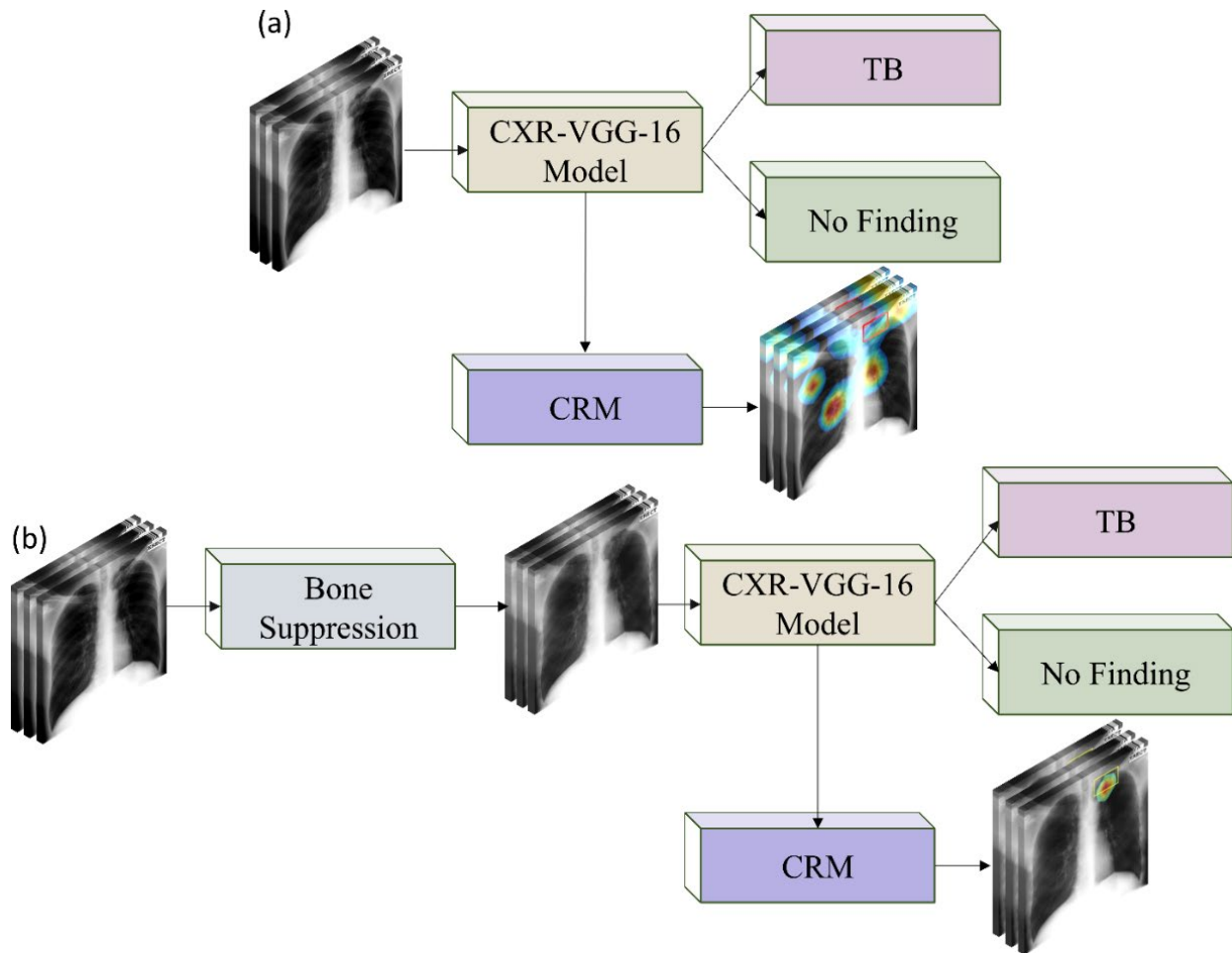
doi:10.1007/s00330-012-2550-y

13. Yang, W. *et al.* Cascade of multi-scale convolutional neural networks for bone suppression of chest radiographs in gradient domain. *Med. Image Anal.* (2017). doi:10.1016/j.media.2016.08.004
14. Gordienko, Y. *et al.* Deep learning with lung segmentation and bone shadow exclusion techniques for chest X-ray analysis of lung cancer. in *Advances in Intelligent Systems and Computing* (2019). doi:10.1007/978-3-319-91008-6_63
15. Shiraishi, J. *et al.* 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. *Am. J. Roentgenol.* (2000). doi:10.2214/ajr.174.1.1740071
16. Nachiappan, A. C. *et al.* Pulmonary tuberculosis: Role of radiology in diagnosis and management. *Radiographics* (2017). doi:10.1148/rg.2017160032
17. Hogeweg, L. *et al.* Fusion of local and global detection systems to detect tuberculosis in chest radiographs. in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* **6363 LNCS**, 650–657 (2010).
18. Maduskar, P., Hogeweg, L., Philipsen, R., Schalekamp, S. & van Ginneken, B. Improved texture analysis for automatic detection of tuberculosis (TB) on chest radiographs with bone suppression images. in *Medical Imaging 2013: Computer-Aided Diagnosis* (2013). doi:10.1117/12.2008083
19. Li, H. *et al.* High-Resolution Chest X-Ray Bone Suppression Using Unpaired CT Structural Priors. *IEEE Trans. Med. Imaging* (2020). doi:10.1109/TMI.2020.2986242
20. Matsubara, N., Teramoto, A., Saito, K. & Fujita, H. Bone suppression for chest X-ray image using a convolutional neural filter. *Australas. Phys. Eng. Sci. Med.* (2019). doi:10.1007/s13246-019-00822-w
21. Chen, Y. *et al.* Bone suppression of chest radiographs with cascaded convolutional networks in wavelet domain. *IEEE Access* (2019). doi:10.1109/ACCESS.2018.2890300
22. Simonyan, K. & Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. in *3rd International Conference on Learning Representations (ICLR2015)* 1–14 (2014). doi:10.1016/j.infsof.2008.09.005
23. Nishio, M., Noguchi, S., Matsuo, H. & Murakami, T. Automatic classification between COVID-19 pneumonia, non-COVID-19 pneumonia, and the healthy on chest X-ray image: combination of data augmentation methods. *Sci. Rep.* (2020). doi:10.1038/s41598-020-74539-2
24. Jaeger, S. *et al.* Two public chest X-ray datasets for computer-aided screening of pulmonary diseases. *Quant. Imaging Med. Surg.* **4**, 475–477 (2014).
25. Budapest University of Technology and Economics (BME). Bone Shadow Eliminated Images of the JSRT Database. (2013). Available at: <https://www.mit.bme.hu/eng/events/2013/04/18/boneshadow-eliminated-images-jsrt-database>. (Accessed: 6th March 2020)
26. Candemir, S. *et al.* Atlas-based rib-bone detection in chest X-rays. *Comput. Med. Imaging Graph.* (2016). doi:10.1016/j.compmedimag.2016.04.002

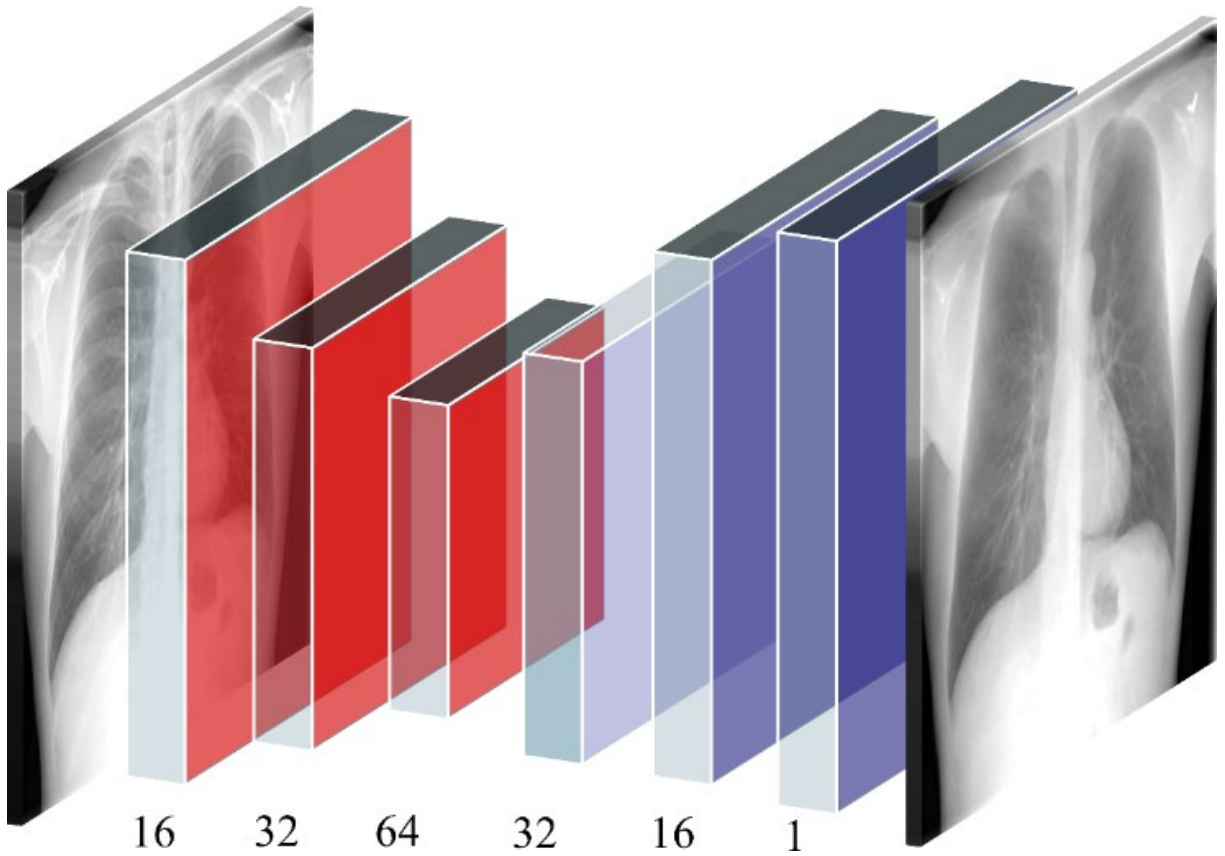
27. Kim, I., Rajaraman, S. & Antani, S. Visual interpretation of convolutional neural network predictions in classifying medical image modalities. *Diagnostics* (2019). doi:10.3390/diagnostics9020038
28. Wang, X. *et al.* ChestX-ray8: Hospital-scale Chest X-ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases. in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 1–19 (2017). doi:10.1109/CVPR.2017.369
29. Shih, G. *et al.* Augmenting the National Institutes of Health Chest Radiograph Dataset with Expert Annotations of Possible Pneumonia. *Radiol. Artif. Intell.* (2019). doi:10.1148/ryai.2019180041
30. Kermany, D. S. *et al.* Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning. *Cell* **172**, 1122–1124.e9 (2018).
31. Lim, B., Son, S., Kim, H., Nah, S. & Lee, K. M. Enhanced Deep Residual Networks for Single Image Super-Resolution. in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops* (2017). doi:10.1109/CVPRW.2017.151
32. Szegedy, C., Ioffe, S. & Vanhoucke, V. Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. *arXiv cs.CV*, (2016).
33. Willmott, C. J. & Matsuura, K. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Clim. Res.* (2005). doi:10.3354/cr030079
34. Brunet, D., Vrscay, E. R. & Wang, Z. On the mathematical properties of the structural similarity index. *IEEE Trans. Image Process.* (2012). doi:10.1109/TIP.2011.2173206
35. Wang, Z., Simoncelli, E. P. & Bovik, A. C. Multi-scale structural similarity for image quality assessment. in *Conference Record of the Asilomar Conference on Signals, Systems and Computers* (2003). doi:10.1109/acssc.2003.1292216
36. Open Source Computer Vision. Histogram Comparison. (2020). Available at: https://docs.opencv.org/3.4/d8/dc8/tutorial_histogram_comparison.html. (Accessed: 3rd March 2020)
37. Rajaraman, S., Kim, I. & Antani, S. K. Detection and visualization of abnormality in chest radiographs using modality-specific convolutional neural network ensembles. *PeerJ* (2020). doi:10.7717/peerj.8693
38. Dutta, A. & Zisserman, A. The VIA annotation software for images, audio and video. in *MM 2019 - Proceedings of the 27th ACM International Conference on Multimedia* (2019). doi:10.1145/3343031.3350535
39. Gastwirth, J. L., Gel, Y. R. & Miao, W. The Impact of Levene ' s Test of Equality of Practice. *Stat. Sci.* **24**, 343–360 (2009).
40. Levene, H. Levene test for equality of variances. *Contrib. to Probab. Stat.* (1960).
41. Stirenko, S. *et al.* Chest X-Ray Analysis of Tuberculosis by Deep Learning with Segmentation and Augmentation. in *2018 IEEE 38th International Conference on Electronics and Nanotechnology, ELNANO 2018 - Proceedings* (2018). doi:10.1109/ELNANO.2018.8477564

42. Van Der Maaten, L. & Hinton, G. Visualizing Data using t-SNE. *J. Mach. Learn. Res.* (2008). doi:10.1007/s10479-011-0841-3
43. Islam, M. T., Aowal, M. A., Minhaz, A. T. & Ashraf, K. Abnormality Detection and Localization in Chest X-Rays using Deep Convolutional Neural Networks. *arXiv* (2017).
44. He, K., Zhang, X., Ren, S. & Sun, J. Deep Residual Learning for Image Recognition. in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 7, 770–778 (2016).
45. Schalekamp, S. *et al.* Bone suppression increases the visibility of invasive pulmonary aspergillosis in chest radiographs. *PLoS One* (2014). doi:10.1371/journal.pone.0108551
46. Geng, E., Kreiswirth, B., Burzynski, J. & Schluger, N. W. Clinical and radiographic correlates of primary and reactivation tuberculosis: A molecular epidemiology study. *J. Am. Med. Assoc.* (2005). doi:10.1001/jama.293.22.2740
47. Bhalla, A., Goyal, A., Guleria, R. & Gupta, A. Chest tuberculosis: Radiological review and imaging recommendations. *Indian J. Radiol. Imaging* 25, 213 (2015).

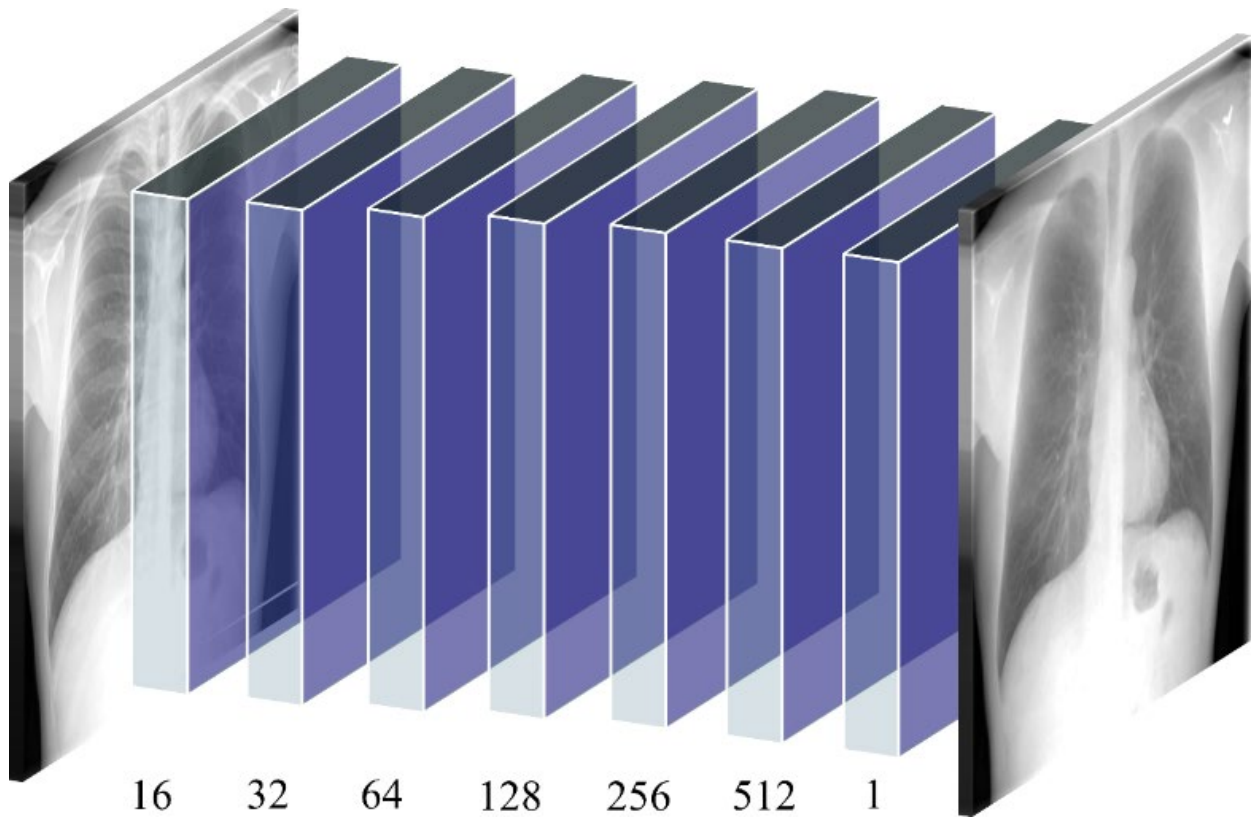
SUPPLEMENTARY MATERIAL



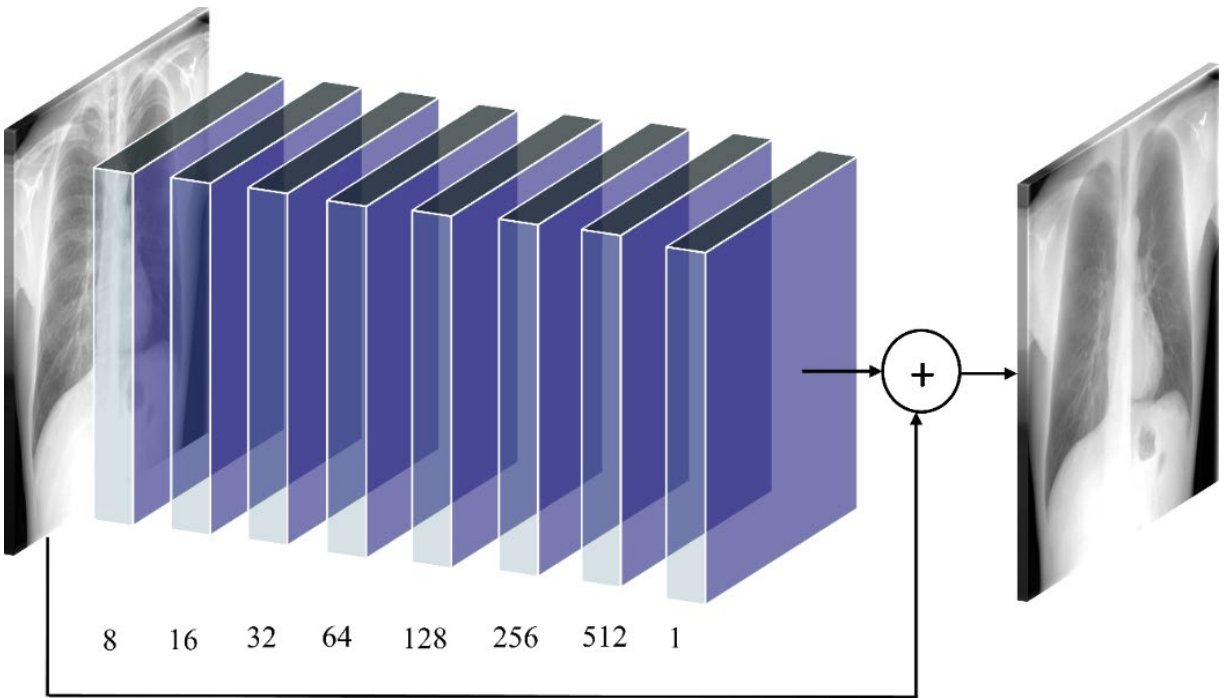
Supplementary Figure 1. The workflow of the proposed approach. (a) The CXR-VGG-16 model is finetuned on the original, non-bone-suppressed CXRs in Shenzhen and Montgomery TB CXR collections to classify them as showing normal lungs or pulmonary TB manifestations. The model predictions are recorded and interpreted through CRMs; (b) The original CXRs are fed into the best-performing bone suppression model that suppresses the bones in the Shenzhen and Montgomery TB CXRs to produce bone-suppressed CXRs. The CXR-VGG-16 model is then fine-tuned on these bone-suppressed CXRs to classify them as showing normal lungs or pulmonary TB manifestations. The classification and localization performance of the baseline, non-bone-suppressed (a), and bone-suppressed models (b) are then compared.



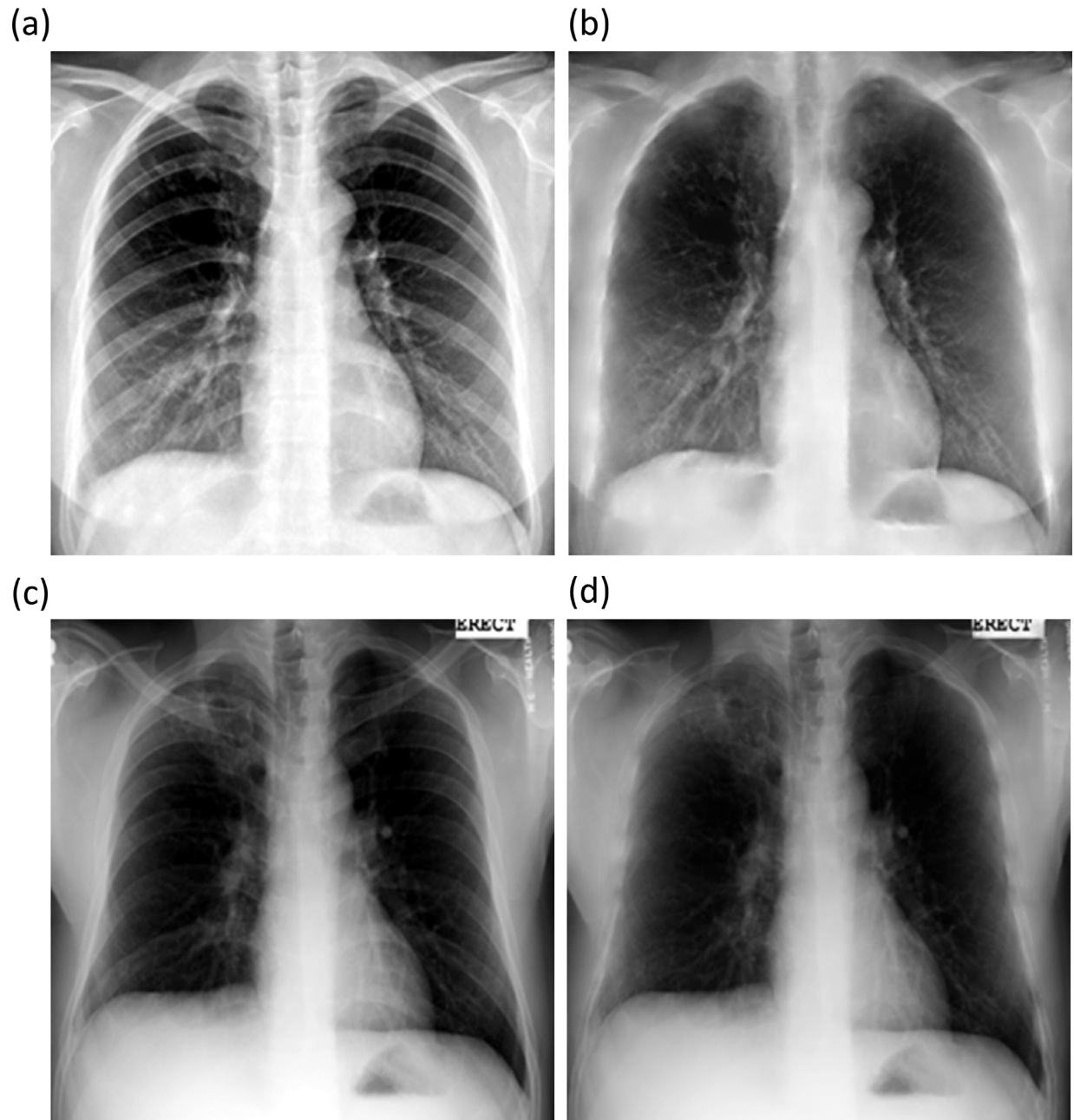
Supplementary Figure 2. The architecture of the proposed AE-BS model. The AE-BS model has a symmetrical convolutional encoder (shown with red-colored boxes) and decoder (shown with blue-colored boxes) architecture.



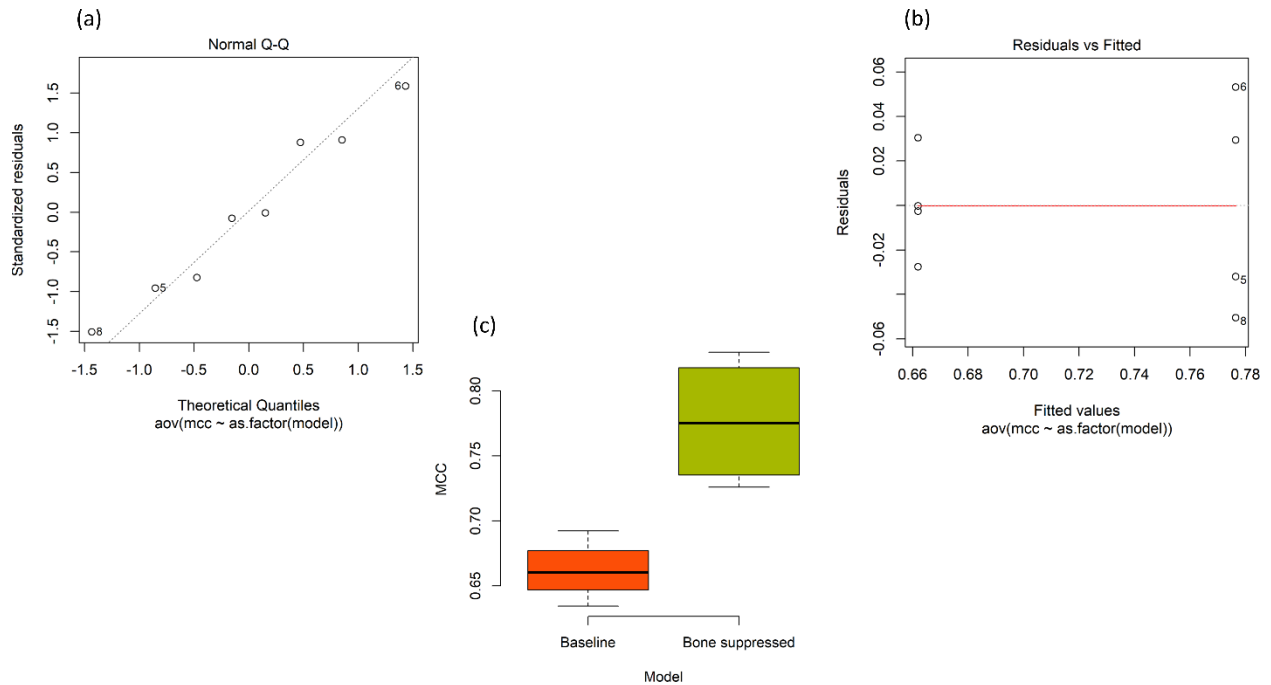
Supplementary Figure 3. The architecture of the proposed ConvNet-BS model. The ConvNet-BS model has seven convolutional layers (shown with blue-colored boxes) with zero paddings to preserve original input dimensions. The deepest convolutional layer with the sigmoidal activation produces bone-suppressed soft-tissue images.



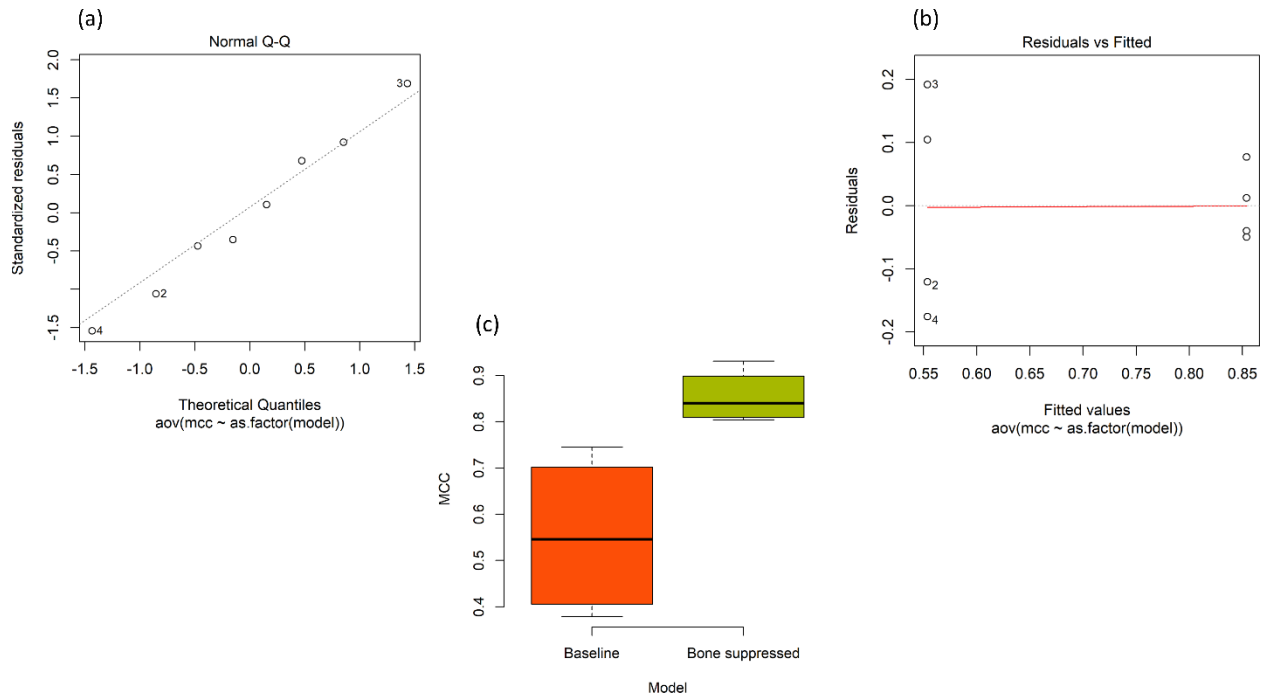
Supplementary Figure 4. The architecture of the proposed RL-BS model.



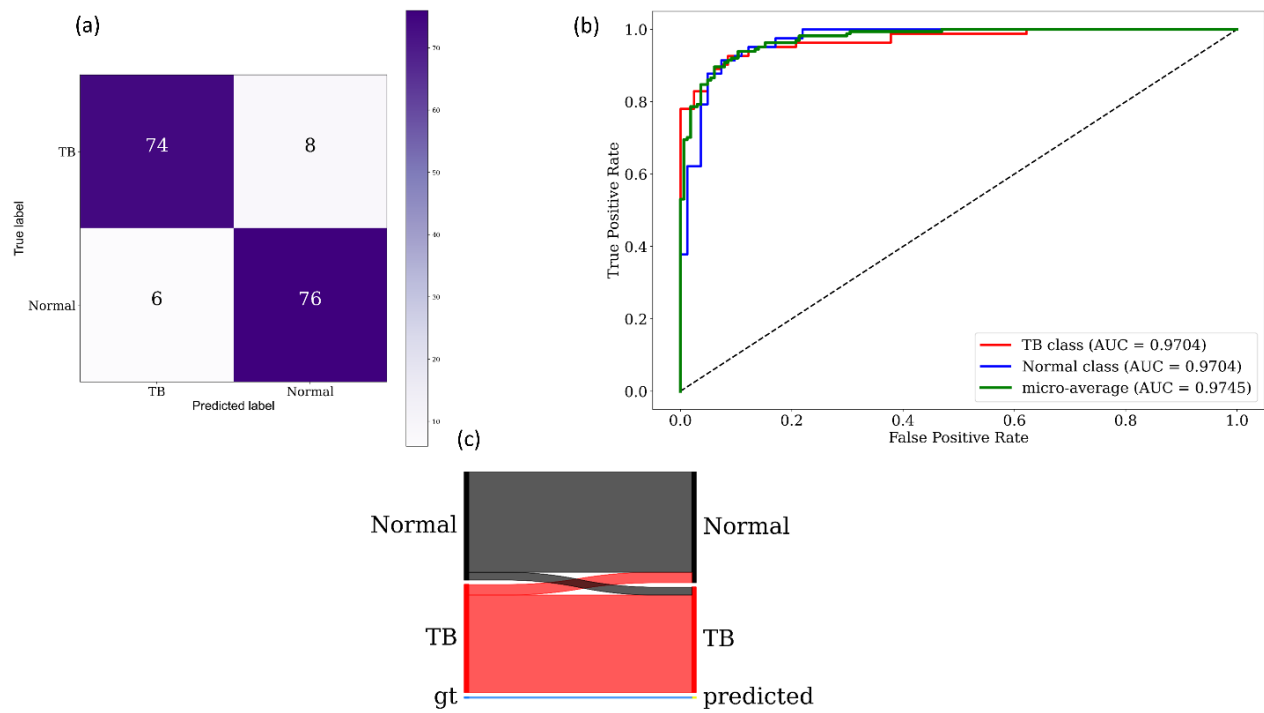
Supplementary Figure 5. Bone-suppressed CXRs predicted by the ResNet-BS model using a sample CXR from the Shenzhen and Montgomery TB collection. (a) Shenzhen abnormal CXR; (b) Predicted bone-suppressed image; (c) Montgomery abnormal CXR; and (d) Predicted bone-suppressed image.



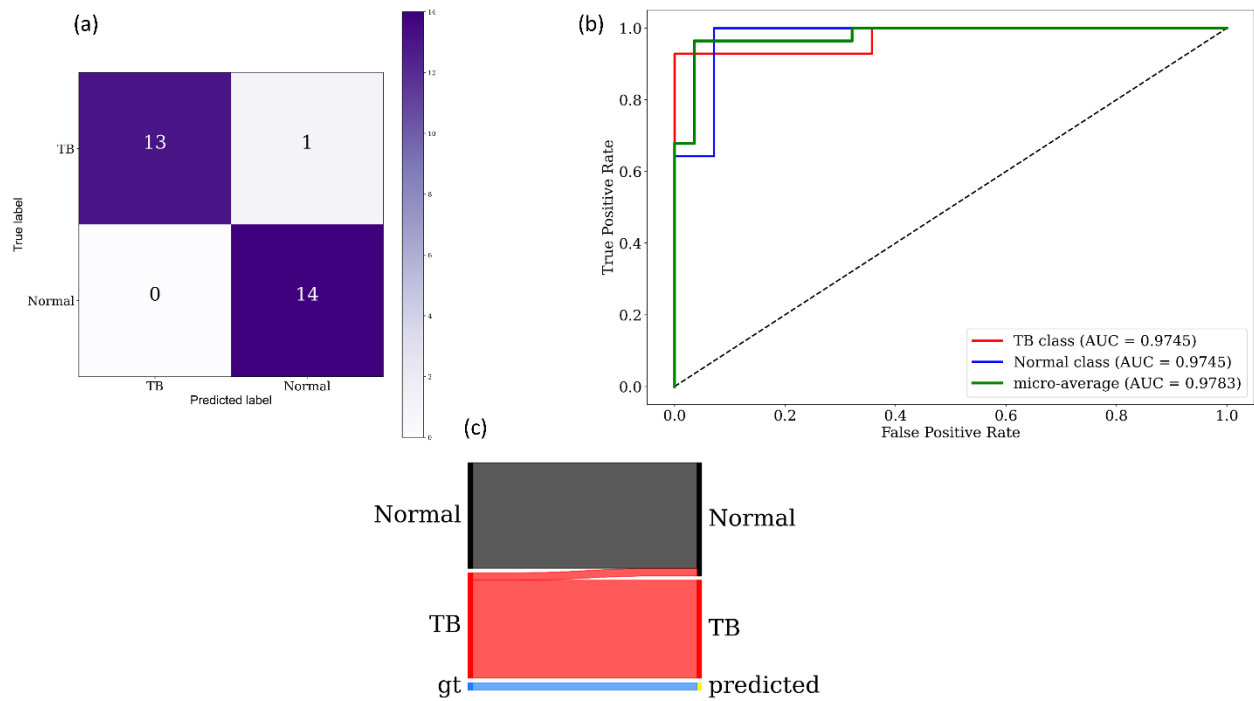
Supplementary Figure 6. Statistical analyses using cross-validated models trained and evaluated of the Shenzhen TB CXR collection. (a) Residual plot showing that the data follow the normal distribution; (b) Relationship plot between the residuals and the mean value of Mathews Correlation Coefficient (MCC) obtained by the baseline and bone-suppressed models to investigate for homogeneity of variances; and (c) Mean plot for the MCC scores obtained by the baseline and bone-suppressed models. Error bars represent standard errors. The differences are statistically significant ($p < 0.05$).



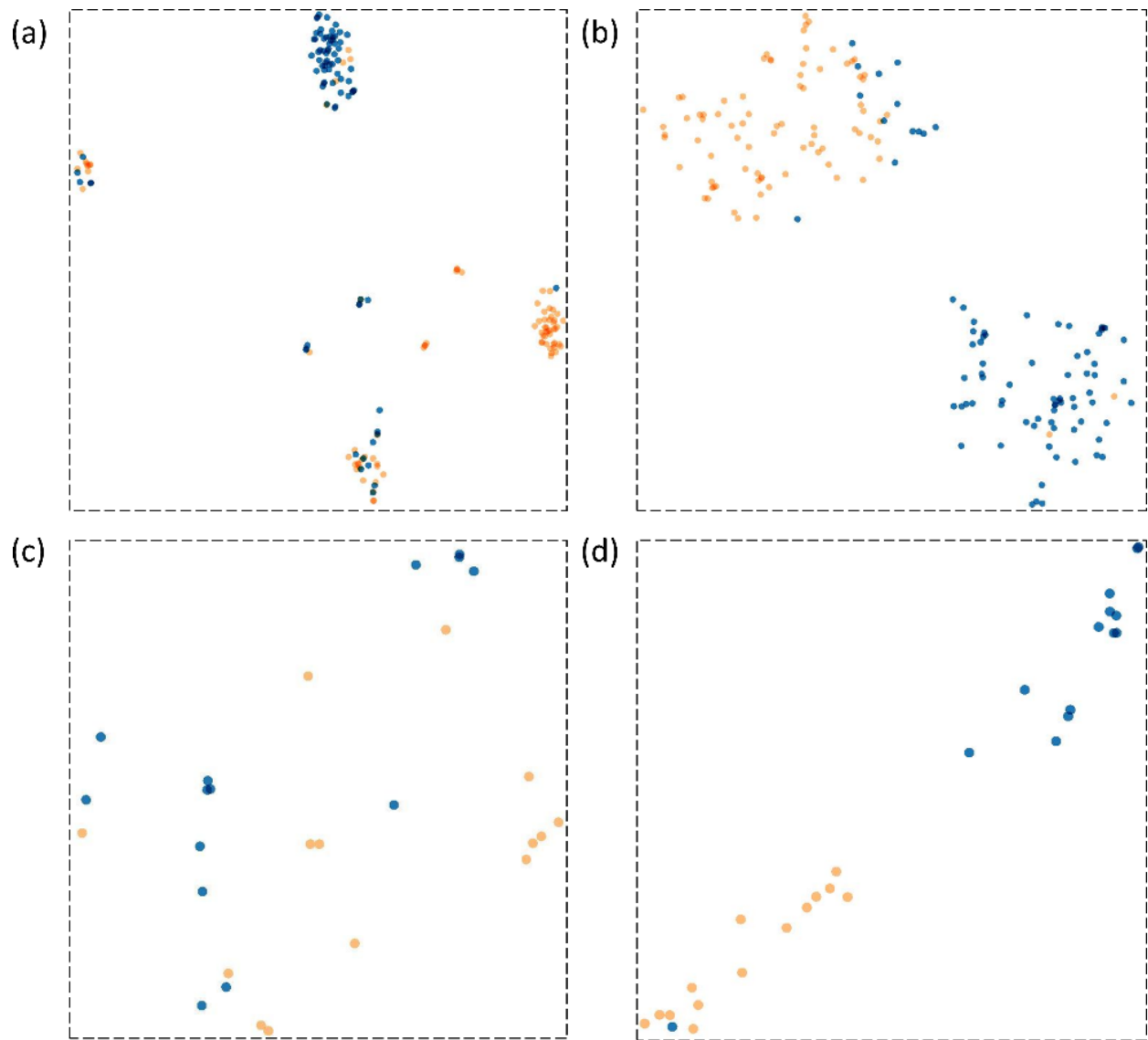
Supplementary Figure 7. Statistical analyses using cross-validated models trained and evaluated on the Montgomery TB CXR collection. (a) Residual plot showing that the data follow the normal distribution; (b) Relationship plot shows no evident relationship exists between the residuals and the mean value of MCC obtained by the baseline and bone-suppressed models, and (c) Mean plot for the MCC scores obtained by the baseline and bone-suppressed models. Error bars represent standard errors. The differences are statistically significant ($p < 0.05$).



Supplementary Figure 8. Performance visualization using the best-performing cross-validated bone-suppressed model that is trained and evaluated on the Shenzhen TB CXR collection. (a) Confusion matrix; (b) AUC-ROC curves; and (c) Normalized Sankey flow diagram.



Supplementary Figure 9. Performance visualization using the best-performing cross-validated bone-suppressed model that is trained and evaluated on the Montgomery TB CXR collection. (a) Confusion matrix; (b) AUC-ROC curves; and (c) Normalized Sankey flow diagram.



Supplementary Figure 10. Feature embedding visualization with the t-SNE algorithm. The blue-colored points denote the feature embeddings for the TB class and the orange-colored points denote that of the normal class. (a) and (c) show the t-SNE visualization obtained with the best-performing baseline models respectively using the Shenzhen and Montgomery TB CXR collections; (b) and (d) show the t-SNE visualization obtained with the best-performing bone-suppressed models respectively using the Shenzhen and Montgomery TB CXR collections.

Table 1. Demographic study. Details including patient count, sex, and the count of abnormal and normal images available for various datasets used in this study are shown. NA denotes Not Available. A total of 33,497 CXRs are included. Of these, 22,654 are abnormal with 394 being positive for TB (1.74% of abnormal, 1.18% of the entire sample).

Dataset	Total		Images	
	Male	Female	Normal	Abnormal
JSRT CXR	119	128	93	154
Pediatric pneumonia CXR	NA	NA	1493	4273
RSNA CXR	17006	12888	8851	17833
Shenzhen TB CXR	449	213	326	336
Montgomery TB CXR	64	74	80	58