

# Homogenization of multi-institutional chest x-ray images in various data transformation schemes

Hyeongseok Kim<sup>Ⓞ, a, †</sup>, Seoyoung Lee<sup>Ⓞ, b, †</sup>, Woo Jung Shim,<sup>c</sup>  
Min-Seong Choi,<sup>c</sup> and Seungryong Cho<sup>a, b, d, e, \*</sup>

<sup>a</sup>KAIST Institute for Artificial Intelligence, Korea Advanced Institute of Science and Technology, Daejeon, Republic of Korea

<sup>b</sup>Korea Advanced Institute of Science and Technology, Department of Nuclear and Quantum Engineering, Daejeon, Republic of Korea

<sup>c</sup>AI Research Center, Radisen Co., Ltd., Seoul, Republic of Korea

<sup>d</sup>KAIST Institute for Health Science and Technology, Korea Advanced Institute of Science and Technology, Daejeon, Republic of Korea

<sup>e</sup>KAIST Institute for IT Convergence, Korea Advanced Institute of Science and Technology, Daejeon, Republic of Korea

## Abstract

**Purpose:** Although there are several options for improving the generalizability of learned models, a data instance-based approach is desirable when stable data acquisition conditions cannot be guaranteed. Despite the wide use of data transformation methods to reduce data discrepancies between different data domains, detailed analysis for explaining the performance of data transformation methods is lacking.

**Approach:** This study compares several data transformation methods in the tuberculosis detection task with multi-institutional chest x-ray (CXR) data. Five different data transformations, including normalization, standardization with and without lung masking, and multi-frequency-based (MFB) standardization with and without lung masking were implemented. A tuberculosis detection network was trained using a reference dataset, and the data from six other sites were used for the network performance comparison. To analyze data harmonization performance, we extracted radiomic features and calculated the Mahalanobis distance. We visualized the features with a dimensionality reduction technique. Through similar methods, deep features of the trained networks were also analyzed to examine the models' responses to the data from various sites.

**Results:** From various numerical assessments, the MFB standardization with lung masking provided the highest network performance for the non-reference datasets. From the radiomic and deep feature analyses, the features of the multi-site CXRs after MFB with lung masking were found to be well homogenized to the reference data, whereas the others showed limited performance.

**Conclusions:** Conventional normalization and standardization showed suboptimal performance in minimizing feature differences among various sites. Our study emphasizes the strengths of MFB standardization with lung masking in terms of network performance and feature homogenization.

© The Authors. Published by SPIE under a Creative Commons Attribution 4.0 International License. Distribution or reproduction of this work in whole or in part requires full attribution of the original publication, including its DOI. [DOI: [10.1117/1.JMI.10.6.061103](https://doi.org/10.1117/1.JMI.10.6.061103)]

**Keywords:** chest radiograph; deep learning; data transformation; data homogenization; radiomics.

Paper 22168SSRR received Jul. 7, 2022; accepted for publication Apr. 3, 2023; published online Apr. 26, 2023.

---

\*Address all correspondence to Seungryong Cho, [scho@kaist.ac.kr](mailto:scho@kaist.ac.kr)

†Both authors contributed equally to this work.

## 1 Introduction

Learning method estimations may not generalize well when the joint distribution of inputs and task labels changes substantially.<sup>1-3</sup> A network architecture trained on a given domain data set may perform poorly when other domain data are tested. There exist a host of research articles that address this domain discrepancy, and the so-called “domain adaptation” is a subfield of transfer learning<sup>4,5</sup> that addresses this problem. A good summary and review of the domain adaptation techniques in chest x-ray (CXR) imaging can be found in the paper of Çallı et. al.<sup>6</sup> Domain adaptation methods basically seek domain-invariant features or representations between the source and target domains.<sup>7-11</sup> These approaches, however, assume that the source domain, from which the samples for testing come, can be appropriately formed and specified. This usually means that enough data from the source domain, specified by a consistent set of scanning hardware, scanning environment, scanning protocol, etc., can be recruited for domain adaptation.<sup>12-14</sup>

The situation in which we have a particular interest is from the global healthcare disparity perspectives, and it is not supportive for forming such a domain. It is often the case that CXR is performed without satisfying the scanning protocols in medically underserved areas and populations. A lack of enough electric power supply, inadequate data acquisition setting, and absence of licensed technologists are possible causes of inconsistent image quality of CXRs. Considering rather unpredictable scanning conditions, the CXR images at hand may not constitute a well-defined domain in such cases. Artificial intelligence (AI)-enabled techniques are fast evolving in medical fields including automated detection and diagnosis of diseases. They will surely help reducing global healthcare service disparity. It is considered an important area of research and development for such tools to be deployed in the field with their optimal performance uncompromised. The purpose of this study is to implement and compare instance-based data transformation methods in that line of research, which is therefore highly relevant to the special issue of global health, equity, bias, and diversity in AI in medical imaging.

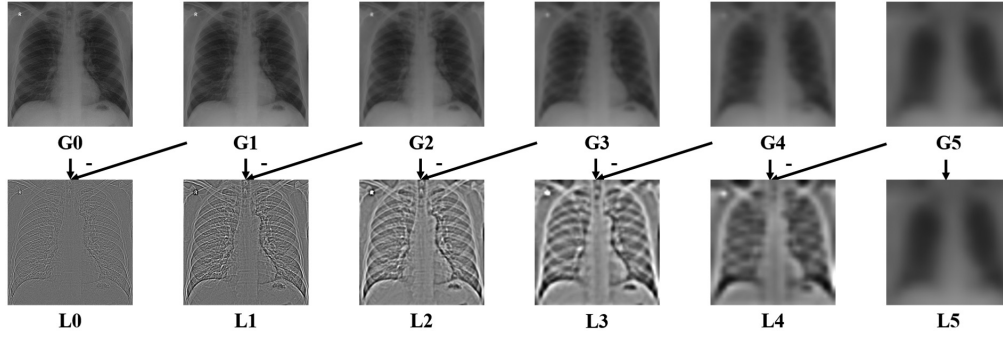
An out-of-distribution detection task that verifies whether a test sample belongs to the predefined source domain can help check the availability of the trained network.<sup>6,15</sup> However, to increase the utility of the learned models, a single data instance-based approach, such as input data transformation, is desirable. In this work, we focus on the input transformation approaches and provide a missing link that can explain which method would be more powerful in deep-learning-based detection tasks.

Preprocessing methods are commonly used for deep network performance enhancement<sup>16</sup> and efficient deep network training.<sup>12</sup> To reduce data discrepancies, data transformations in the scope of histogram modification techniques, including histogram equalization, matching, clipping, and normalization,<sup>12,17</sup> have been investigated. Nevertheless, a claim that such global histogram modification methods cannot harmonize texture differences has been made.<sup>17</sup> Indeed, data preprocessing steps embracing the multi-frequency characteristics or local features of x-ray images have been reported to improve computer-aided detection<sup>18-20</sup> and deep-learning-based detection tasks.<sup>17,21,22</sup> However, it is still questionable whether the results of the network in clinical settings that have different distributions than those used in training can be trusted.<sup>23-25</sup> This study compares several data transformation methods in the CXR-based tuberculosis (TB) detection task and provides strong evidence for why one method is superior to the others through radiomics analysis and deep feature analysis.

## 2 Material and Methods

### 2.1 Data Transformation

We implemented three types of transformation algorithms: data normalization, data standardization, and multi-frequency-based (MFB) data standardization. For data standardization and MFB data standardization, we further split each method into two different schemes: with and without lung masks. Therefore, the total number of methods implemented in this work is five. Data normalization and standardization use the transformation Eqs. (1) and (2), respectively.



**Fig. 1** Example Laplacian pyramid of a CXR image.

$$X_{N,i} = \frac{X_{\max}^{\text{ref}} - X_{\min}^{\text{ref}}}{X_{\max} - X_{\min}} (X_i - X_{\min}) + X_{\min}^{\text{ref}}, \quad (1)$$

$$X_{S,i} = \frac{\sigma^{\text{ref}}}{\sigma} (X_i - \mu) + \mu^{\text{ref}}, \quad (2)$$

where  $X_i \in R$  represents a CXR image input pixel value. For an effective normalization of CXR images coming from different systems, a dynamic normalization with histogram analysis is desirable.<sup>12</sup> Specifically, a DICOM image may contain a constant bias or a letter mark having outlying pixel values. Although some DICOM images provide information that can be utilized to confine pixel value ranges, not all DICOM images provide this information, and windowing parameters are therefore user-specific. In this study, as an attempt to remove such outlying pixels, we first calculated the cumulative histogram  $f: R \rightarrow [0,1]$  of a given CXR image  $X \in R^N$ . Then,  $X_{\min}$  and  $X_{\max}$  are such that  $f(X_{\min}) = 0.02$  and  $f(X_{\max}) = 0.98$ . In the following homogenization processes, we first applied the above normalization and then calculated image statistics to make images have similar pixel value ranges regardless of the transformation methods. We localized pixels with values that are below  $X_{\min}$  or above  $X_{\max}$ , and those outlier pixels are excluded from the statistics calculation.  $\mu$  and  $\sigma$  in Eq. (2) indicate the mean and the standard deviation values of the input image after this preprocessing, which are then adjusted to the reference mean  $\mu^{\text{ref}}$  and standard deviation  $\sigma^{\text{ref}}$ . When standardizing images with lung masks, image statistics were calculated only within the masked lung region, whereas the transformation was applied to all of the image pixels.

The MFB data standardization starts with the Laplacian pyramid decomposition,<sup>26</sup> which is widely used in CXR image enhancement.<sup>27</sup> The Laplacian pyramid decomposition results in the multiscale representation of a CXR image, and its reverse reconstruction process reproduces the original image. Conventional MFB data enhancement techniques strengthen specific frequency bands of a CXR image, whereas the MFB data standardization aims to make a CXR image similar to the target domain at each multi-frequency band.<sup>18</sup> Figure 1 shows an example Laplacian pyramid of a CXR image. We specified the Gaussian and the Laplacian pyramid images at the  $k$ 'th level by  $G_k$  and  $L_k$ , respectively. Lower frequency information tends to be stored at the higher level of the Laplacian pyramid image by design. For MFB standardization of an input image, we first applied the Laplacian pyramid decomposition to a set of training CXR images. For each training image, the mean  $\mu_k$  and standard deviation  $\sigma_k$  values of  $L_k$  were then calculated. Finally, the reference mean  $\mu_k^{\text{ref}}$  and standard deviation  $\sigma_k^{\text{ref}}$  values were calculated by averaging those  $\mu_k$  and  $\sigma_k$  values. The MFB standardization process iteratively adjust  $\mu_k$  and  $\sigma_k$  values of an input image to  $\mu_k^{\text{ref}}$  and  $\sigma_k^{\text{ref}}$  values. As mentioned above, pixels inside the masked lung regions were used for the statistics calculation when the lung masking scheme was adopted. The iterative procedures are summarized in Algorithm 1. We set the total number of iterations  $N$  to 50 in this study. We empirically set the maximum level of Laplacian pyramids by 5, which results in a  $16 \times 16$  array size at the fifth level. The ResNet-18<sup>28</sup> can downsample the input image with an array size of  $512 \times 512$  to  $16 \times 16$ , which is equivalent to the minimum array size that is manageable when level 5 is used in the Laplacian pyramid decomposition.

**Algorithm 1** MFB standardization algorithm

---



---

```

1: Inputs  $X, \mu_k^{\text{ref}}, \sigma_k^{\text{ref}}, N$ 
2:  $X_M^1 \leftarrow X$ 
3: for  $i \leftarrow 1$  to  $N$  do
4:   Decompose  $L_k^i$  from  $X_M^i$ 
5:   Calculate  $\mu_k^i$  and  $\sigma_k^i$  from  $L_k^i$ 
6:    $L'_k \leftarrow (L_k^i - \mu_k^i) \times \frac{\sigma_k^{\text{ref}}}{\sigma_k^i} + \mu_k^{\text{ref}}$ 
7:   Reconstruct  $X_M^{i+1}$  from  $L'_k$ 
8: end for
9: Output  $X_M^{N+1}$ 

```

---



---

**Table 1** Datasets information for lung segmentation network training.

Site	Number of samples
JSRT (training)	197
JSRT (validation)	50
Montgomery (validation)	138
China (validation)	566

**2.2 Lung Segmentation**

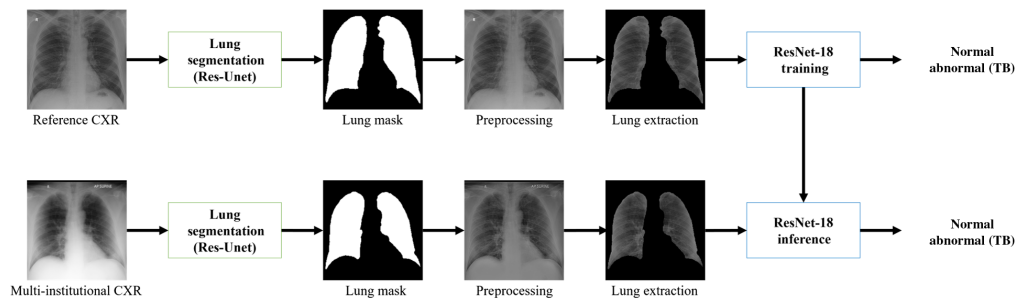
To compare the effects on the detection performance of image transformation methods in conjunction with masked lung statistics, we need a lung segmentation tool. Lung segmentation was conducted using a separate deep neural network from the network that is used for TB detection. A network with a Res-UNet structure<sup>29</sup> was trained with the JSRT CXR dataset<sup>30</sup> and its mask dataset.<sup>31</sup> Additionally, we used the China and Montgomery datasets<sup>32</sup> for additional validation. The training code was implemented using PyTorch libraries<sup>33</sup> on a system with a GeForce RTX 3090. We summarize the details of the data set used for lung segmentation in Table 1.

**2.3 Data Preparation for TB Detection**

We used multi-institutional CXR data from seven clinical sites to train and test the deep neural network for TB detection. Anonymized datasets were collected from the clinically cooperative institutions of the Radisen AI Research Center. This study was approved by the institutional review boards, and informed consent was waived. We summarize dataset information in Table 2. It should be noted that the reference site data provide the target domain examples and other sites provide the source domain examples and that the source domains are diverse in terms of CXR modality, bit-depths, and scanning protocols, of which details are unavailable. The reference dataset is composed of half normal CXR and half abnormal CXR diagnosed with TB. For non-reference data, there are class imbalances between TB sample sizes and normal sample sizes. Because data skewness affects performance metrics,<sup>34</sup> we calculated the performance metrics after under-sampling normal data, so the sample sizes of the two classes become equal. The under-sampling was randomly performed and repeated, so statistical analyses are feasible. Meanwhile, we used all of the data for the feature analyses because the class imbalance itself does not have a significant influence on the feature extraction.

**Table 2** Multi-institutional datasets information.

Site	Modality	Data bits	Number of normal samples	Number of TB samples
Reference (training)	DR	16	956	956
Reference (test)	DR	16	250	250
A (test)	CR	16	250	77
B (test)	CR	14	248	41
C (test)	DR	12	244	136
D (test)	CR	12	285	40
E (test)	CR	12	247	37
F (test)	CR	10	297	26

**Fig. 2** Diagram of overall procedures for comparison study.

## 2.4 TB Detection Network Training

The overall workflow of the CXR-based TB detection in this work is shown in Fig. 2. In the training phase, lung regions were first identified using the trained lung segmentation network. After appropriate image cropping, each image was downsized to an array of  $512 \times 512$  using bilinear interpolation. Various data transformation methods were then applied to each image and then the lung regions of the processed images were used for the network training. In Fig. 2, we omitted the image cropping process for the simplicity of presentation. ResNet-18 architecture was used for the TB detection network. Training details are presented in Appendix A.

## 2.5 TB Detection Performance Evaluation

In the inference phase, CXR images from seven different sites in Table 2 were tested. For testing each TB detection network trained by the data that went through a specific data transformation, the same data transformation method was applied to the input CXR image. For example, we applied the MFB data standardization with lung masking to the test dataset when we evaluated the performance of a TB detection network that was trained by the data transformed by the MFB standardization with lung masking. For the evaluation, receiver operating characteristic (ROC) and precision-recall (PR) curves were used. We also calculated an F1 score of the reference test results with 20 different threshold values within  $[0, 1]$ . The threshold value that provides the highest F1 score in the reference test results was applied to other datasets to calculate F1 scores and recalls. For the reference dataset, we bootstrapped the performance metrics for statistical analysis. Random bootstrapping was performed up to a thousand times, which was determined so that the standard deviation from the bootstrapping results stay within 2% of difference from the Delong's estimated standard deviation in the area under the curve (AUC) value of ROC

curves.<sup>35,36</sup> For non-reference data (from sites A to F), we randomly repeated the under-sampling a thousand times, which is the same number of times that bootstrapping was performed.

Localization of suspicious regions in the CXR image was also performed using Grad-Cam++,<sup>37</sup> which can provide the activation map for TB detection. The bounding boxes were drawn by radiologists for suspicious areas in the entire data set, and we evaluated how well the obtained saliency maps match the radiologists' insights. We achieved Grad-Cam++ images at the deepest feature layer and calculated the weighted intersection over attention (WIOA) values with respect to the radiologists' bounding box information. Grad-CAM++ images and WIOA values were produced from the M3d-CAM PyTorch library.<sup>38</sup> Details on the procedure for calculating the WIOA can be found in [Appendix B](#).

## 2.6 Radiomic Features

To observe the distributions of various datasets, we first extracted the radiomic features<sup>39–42</sup> of CXR images transformed by the aforementioned methods. For each processed CXR image, we masked out the region outside of the lung. PyRadiomics<sup>43</sup> was used to extract features from the preprocessed lung images. We calculated 93 radiomic features, which can be grouped into 6 different series including: first-order statistics, gray-level co-occurrence matrix (GLCM), gray-level dependence matrix (GLDM), gray-level run-length matrix, gray-level size zone matrix, and neighboring gray tone difference matrix. The last five features can be grouped into second-order statistics. We did not include shape features because the shape of the lung in the CXR is not the target of data homogenization.

We calculated the Mahalanobis distance<sup>44,45</sup> as a numerical assessment of the homogenization ability of the radiomic features. The Mahalanobis distance  $D_M$  is a distance between a point  $x$  and a distribution  $D$  with a mean of  $\vec{\mu}$  and a covariance matrix  $S$  and is defined as

$$D_M(\vec{x}) = \sqrt{(\vec{x} - \vec{\mu})^T S^{-1} (\vec{x} - \vec{\mu})}. \quad (3)$$

We calculated the distances between each sample of test sites (A to F) and the distribution of reference data. The distances of first- and second-order radiomics were calculated separately.

For a better visualization of the features, we performed a principal component analysis (PCA) of the calculated radiomics. We chose PCA for visualization to compare the effects of data harmonization on radiomic features in the common coordinate system. Because the radiomic features of images were extracted in the same way regardless of the dataset and harmonization methods, such a comparison is legitimate.

## 2.7 Deep Features

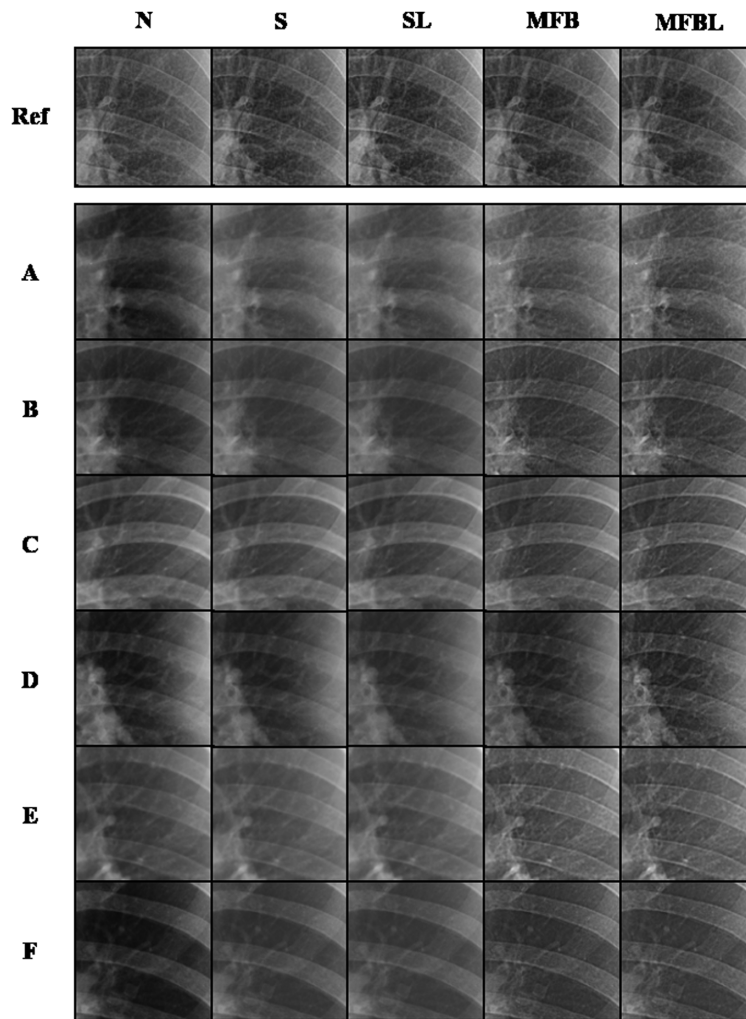
The architecture of a typical convolutional neural network (CNN)<sup>46,47</sup> consists of a series of convolutional layers and pooling layers. We focused on the final layer of the CNN because it stores all of the essential information extracted from the input image in the form of a feature vector. In this work, we used the latent feature vector of the final layer of the TB detection network (Sec. 2.4), which is a vector with a size of 512. We used t-distributed stochastic neighbor embedding (t-SNE)<sup>48</sup> to visualize the latent feature vector. The t-SNE helps with understanding how the network responds to the test data under various data transformation through visualizing the feature clusters.

## 3 Results

### 3.1 Data Transformation

Figure 3 shows the example network inputs of lung-segmented images processed by different transformation methods from various sites. N, S, SL, MFB, and MFBL indicate data normalization, data standardization without and with lung masking, MFB data standardization without and with lung masking, respectively. In Fig. 3, the data N method resulted in non-uniform lung





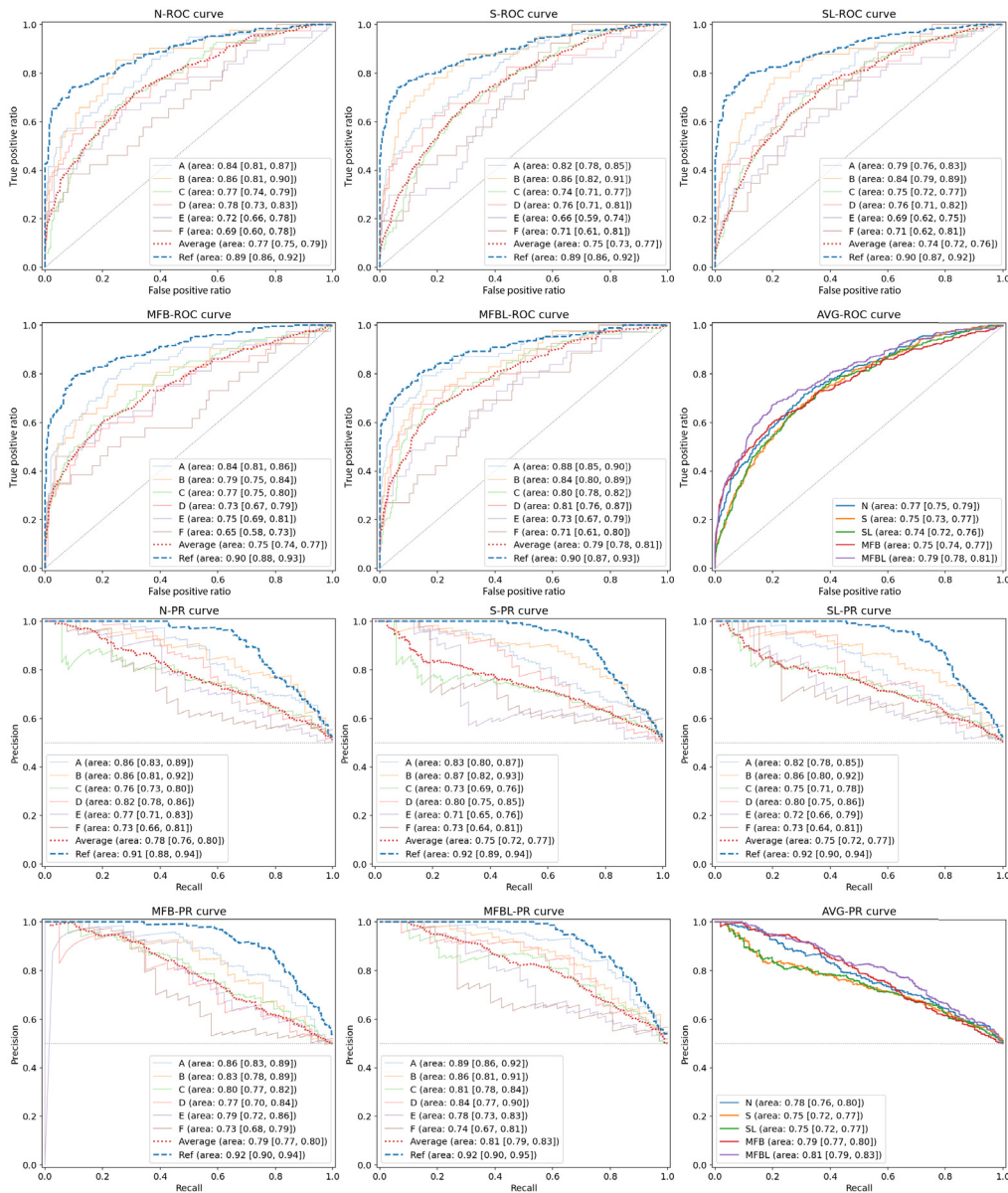
**Fig. 3** Patches extracted from inputs to TB detection networks with different data transformations. Images in the same column are displayed with the same display window.

region brightness. It is observed that MFB methods generally provide a more similar appearance overall with the reference compared with the N, S and SL in the same display window setting. N, S and SL resulted in rather coarse image textures in other site datasets compared with those in the reference, whereas MFB methods produced finer textures similar to the reference. However, the MFB without a lung masking scheme provides suboptimal visual texture similarity in some data, for example, site D in Fig. 3. We included a full breakdown of the time spent on various preprocessing techniques in [Appendix C](#).

### 3.2 TB Detection Performance

In Fig. 4, we show the ROC and the PR curves of the network outputs from different data transformation methods. The AUC values of the ROC and the average precision (AP) values of the PR curves for the reference data showed marginal network performance differences among various data transformations. Network performances, however, largely varied for the six other site datasets. The MFBL method resulted in the minimum gap between the reference curve and the site-average curve in both ROC and PR, whereas S and SL showed poorer results in both AUC and AP.

Table 3 summarizes the network performance evaluation results over the sites in terms of F1 score, recall, and WIOA. Here, please note that the average values in the table are average scores



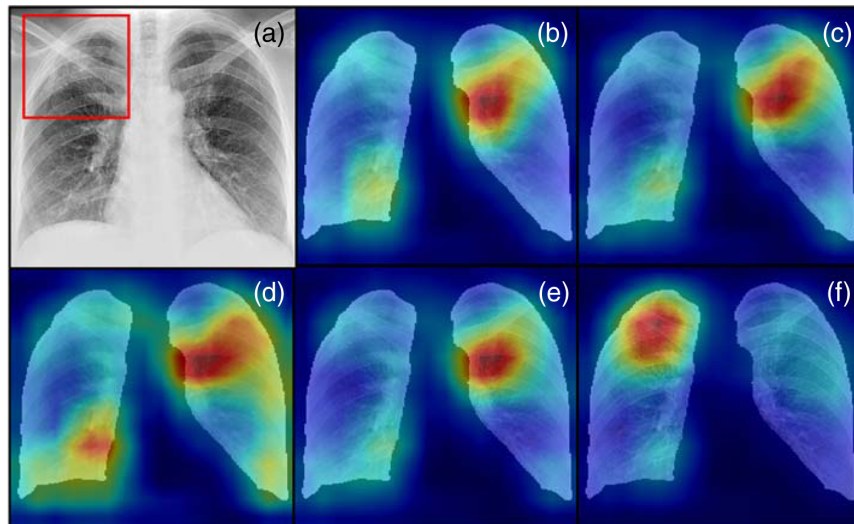
**Fig. 4** ROC and PR curves with different data transformations. Each line shows the ROC or PR curve for each site. The area value in the legend means the AUC and AP values for ROC and PR curves, respectively. Values in square brackets represent the 95% confidence interval. Average lines were produced by collecting every result from site A to site F.

from sites A to F, and we present the standard deviation of those six scores. In the F1 score result, the MFBL method showed the highest values for every non-reference site except F. It is noted that site variation of the scores is minimum in the MFB and MFBL results, which implies more robustness of the network performance. In the recall score, average recall scores of the N, S, and SL methods did not exceed 0.5, which means they failed to detect true TB cases at a >50% chance. Considering that we balanced the classes by undersampling, the chance would be lower than random calls. The WIOA value of a non-trained network specified as random in Table 3 was around 0.31. The WIOA values of the reference dataset with different networks were around 0.60, and the MFBL methods provided slightly lower values in the non-reference datasets as well. Figure 5 shows examples of bounding boxes and Grad-CAM++ images. As shown in Fig. 5(f), a hot spot in the saliency map from the MFBL network goes well with the radiologist’s bounding box.



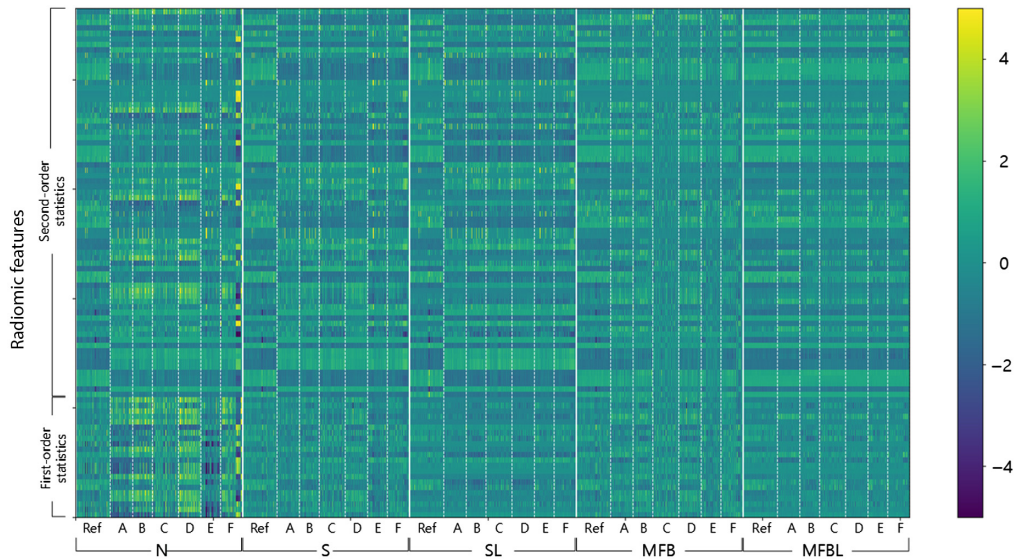
**Table 3** Network quantitative evaluation summary.

		Ref	A	B	C	D	E	F	Average
F1	N	0.80	0.62	0.55	0.17	0.46	0.59	0.21	0.42 ± 0.20
	S	0.82	0.62	0.57	0.22	0.14	0.63	0.32	0.45 ± 0.17
	SL	0.83	0.45	0.50	0.22	0.10	0.61	0.27	0.36 ± 0.18
	MFB	0.83	0.77	0.71	0.51	0.56	0.65	0.56	0.63 ± 0.09
	MFBL	0.81	0.78	0.75	0.70	0.75	0.68	0.54	0.72 ± 0.09
Recall	N	0.75	0.47	0.39	0.10	0.30	0.49	0.12	0.27 ± 0.17
	S	0.77	0.49	0.41	0.13	0.25	0.62	0.19	0.31 ± 0.19
	SL	0.78	0.30	0.34	0.13	0.07	0.59	0.15	0.23 ± 0.19
	MFB	0.80	0.75	0.66	0.36	0.50	0.62	0.46	0.53 ± 0.14
	MFBL	0.77	0.87	0.73	0.62	0.75	0.76	0.46	0.70 ± 0.14
WIOA	N	0.61	0.54	0.51	0.35	0.37	0.53	0.41	0.45 ± 0.08
	S	0.59	0.49	0.42	0.36	0.35	0.50	0.38	0.42 ± 0.07
	SL	0.60	0.45	0.40	0.33	0.33	0.49	0.35	0.39 ± 0.07
	MFB	0.59	0.57	0.51	0.39	0.48	0.58	0.48	0.50 ± 0.07
	MFBL	0.62	0.61	0.57	0.48	0.52	0.62	0.54	0.56 ± 0.06
	*Random	0.39	0.32	0.33	0.24	0.30	0.41	0.28	0.31 ± 0.06

**Fig. 5** (a) Bounding box generated by radiologist and example GCam++ images of set D with different data transformations, (b) N, (c) S, (d) SL, (e) MFB, and (f) MFBL.

### 3.3 Radiomic Features

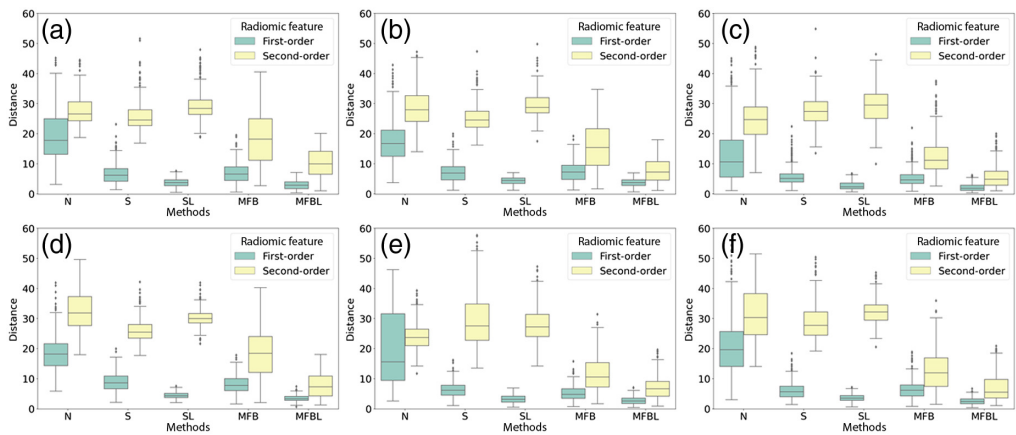
For a visual comparison, we calculated the Z-scores of the features and presented a heatmap of the Z-scores. Figure 6 shows the corresponding heatmap, with each row representing one feature and each column representing one CXR image. The heatmap is divided into five sections on the horizontal axis; each section denotes each data transformation method. If Z-scores of a certain



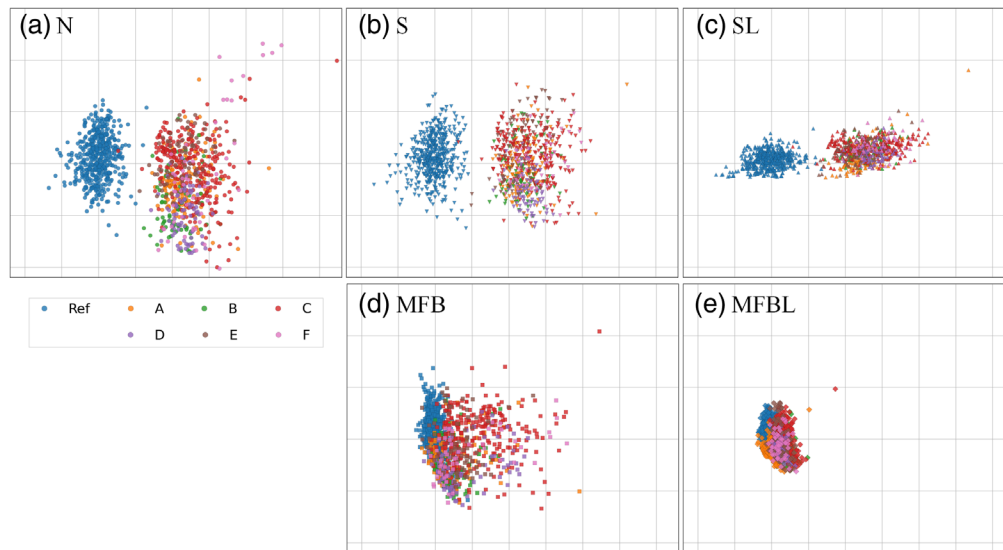
**Fig. 6** Heatmap depicting z-scores of 93 radiomic features for various datasets with a conventional N and four different homogenization methods. The vertical axis denotes radiomic features, divided into first-order statistics and second-order statistics including GLCM, and GLDM. The horizontal axis denotes samples. Note that the MFBL method showed the best homogenization results throughout various features.

radiomic feature are similar across the samples from different sites, it can be said that the homogenization performance of the method is proper in terms of that specific radiomic feature. The N cannot harmonize data from the multi-site data. S and SL, worked as intended for the first-order statistics. However, there was a clear discrepancy between reference test data and other multi-site data in terms of higher-order radiomic features. MFB succeeds in reducing the discrepancies for various sites both in first- and second-order radiomic features. The use of lung masks seems to be more effective in terms of harmonization performance. The heatmap suggests that the MFBL can harmonize not only the histogram characteristic but also textural features.

Figure 7 is a box-and-whisker plot of Mahalanobis distances between the features of test and reference sites. A larger distance implies that a corresponding feature is not well harmonized, albeit after a certain data transformation method, and vice versa. The implications of the box plots of distances are in the same vein as the results shown qualitatively in the heatmap (Fig. 6).



**Fig. 7** Box-and-whisker plots of the Mahalanobis distances between test radiomic features and the reference radiomic features. (a)–(f) The site A, B, C, D, E, and F. N, S, and SL cannot harmonize the second-order statistics well, whereas the performance of the MFBL was the best for all sites.



**Fig. 8** Visualizations of radiomic features after data transformation. (a) N, (b) S, (c) SL, (d) MFB, and (e) MFBL. Visualizations are done by PCA of high-dimensional radiomic feature vectors.

The N has a limited performance on feature harmonization overall. S and SL can harmonize only first-order statistics and not second-order features. MFBL resulted in the shortest feature distances; implying a superior homogenization performance.

Note that the smaller discrepancy in radiomics feature space does not always guarantee better performance of the classification network. For example, although the S improved the first-order feature in all cases compared with N, the performance of the network is generally better for the normalized images (see Fig. 4). Nevertheless, we confirmed that conventional N and S failed to harmonize the textural features of the images as compared to the MFB and thus cannot substantially reduce the effect of reference data bias.

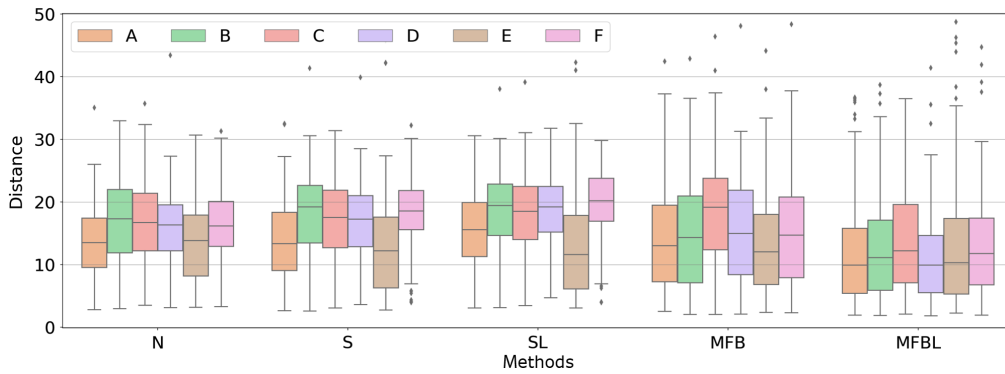
Figure 8 contains visualizations of radiomic features after five different data transformation processes. There are distinct differences between the radiomic features of the reference site and other sites that have undergone conventional N and S. However, after MFB, especially with lung masks [Fig. 8(e)], it is clear that the radiomic features of the samples from different sites are well harmonized with the reference group.

### 3.4 Deep Features

Table 4 summarizes a list of the calculated Mahalanobis distances of the multi-site features to the reference distribution used for the network training. Figure 9 shows the box-and-whisker plots corresponding to Table 4. The MFBL method shows the closest distance regardless of the test

**Table 4** Mahalanobis distances ( $D_M$ ) for multi-site data homogenization. The target distribution was the reference train dataset. A larger distance indicates an outlier to the target distribution.

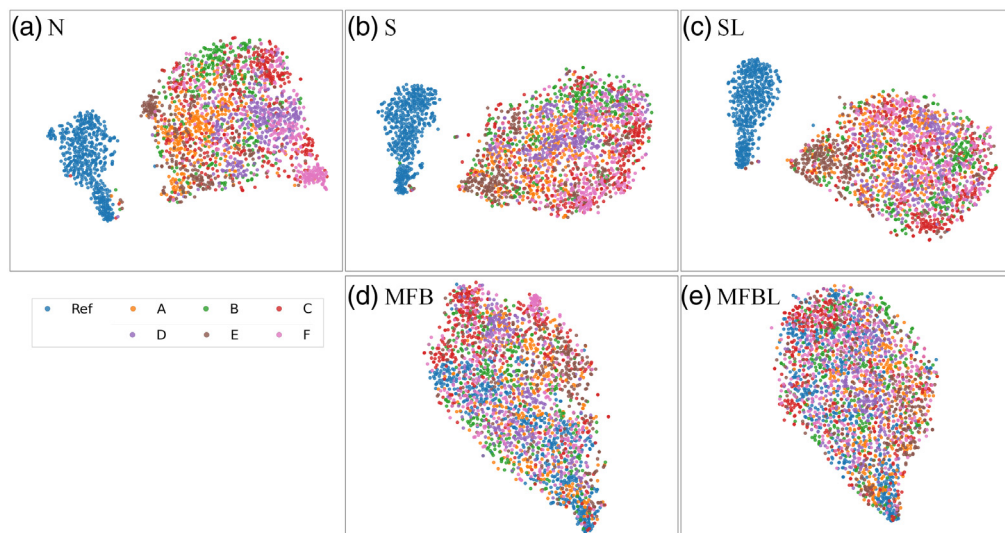
	A	B	C	D	E	F
N	13.55 ± 5.458	16.82 ± 6.530	16.86 ± 6.473	15.62 ± 5.640	13.73 ± 6.248	16.77 ± 5.109
S	13.62 ± 6.123	17.86 ± 6.655	17.33 ± 6.526	16.85 ± 5.810	12.80 ± 7.181	18.44 ± 5.157
SL	15.38 ± 5.939	18.37 ± 6.315	18.17 ± 6.666	18.63 ± 5.039	12.83 ± 7.769	19.98 ± 4.810
MFB	13.93 ± 7.837	14.83 ± 8.461	18.20 ± 8.367	15.46 ± 8.028	13.46 ± 7.749	14.67 ± 7.791
MFBL	11.66 ± 7.681	12.54 ± 7.789	13.76 ± 8.294	10.72 ± 6.124	12.54 ± 8.831	12.55 ± 7.010



**Fig. 9** Box-and-whisker plot of the Mahalanobis distance of the deep features. Distances were calculated between each point in a dataset to the reference distribution.

data. This implies that the deep feature harmonization ability of the MFBL method is the highest and other harmonization methods may have failed to match the data distributions. Similar to the result of radiomic feature analysis, the distance in the deep feature space is not necessarily inversely correlated to the network classification performance. However, the distance can be used as a metric to determine the network’s trustworthiness for data having different distributions from the training data.

The t-SNE visualization results are presented in Fig. 10. After N, S, and SL [Fig. 10(a)–10(c)], the intrasite differences were reduced, but the homogenization between reference data and other sites was not successful; multi-site data forms a distinct cluster from the reference. From radiomic feature analysis (Sec. 3.3), we found that N, S and SL could not reduce data distribution between reference and other sites. The t-SNE result implies that the network was not able to handle the reference data bias and finally resulted in a biased model. From the point of view of the TB detection network learned from the reference dataset, it can be seen that data from other sites are still treated as out-of-distribution. Although the network may have achieved a somewhat satisfying performance after N, S or SL (Sec. 3.2), the latent vector analysis does not provide strong support for such an improvement. Meanwhile, both versions of MFB methods can harmonize multi-site data in terms of the network’s deep feature, not forming any distinct cluster [Figs. 10(d) and 10(e)].



**Fig. 10** t-SNE visualizations of deep-embedded features after conventional N and four different data transformations. (a) N, (b) S, (c) SL, (d) MFB, and (e) MFBL. Note that, for N and S, the deep features are not well harmonized.

## 4 Discussion

In this study, we used a single dataset from a single clinical site for training and tried to harmonize other datasets with the reference data exploiting data transformation methods. Using multi-institutional datasets for training in conjunction with data augmentation is also a viable option for a network's generalizability enhancement. This approach can be interpreted as one that increases the diversity of feature space of the training dataset, hopefully covering the feature space of a new dataset. Pixel-level transformations, including contrast or brightness adjustment and spatial-level transformation, such as shift and rotation, were used in an example work.<sup>13</sup> Our study shows the importance of image features at multiscale representations in the detection network training. Developing data augmentation strategies with image frequency modulation is planned as our future study.

It is observed that radiomic and deep features from the multi-site dataset after the N, S, or SL method did not agree well with the features of the reference test data. However, it is shown repeatedly that the features within the multi-sites become similar to each other after N, S and SL [see Figs. 8(a)–10(c), Figs. 10(a)–10(c)]. It is perhaps due to the fact that the textures of test data from A to F are similar in the original images. As shown in Fig. 4, the reference image is rather sharp and accordingly high-frequency-emphasized, whereas the images of other sites after N, S and SL are blurrier in a similar fashion.

Deep learning has achieved great success in various image processing tasks, and domain adaptation is one of the benefited applications. Unsupervised deep-neural-network-based domain adaption methods aim to transform the data distribution of the source domain to the target domain.<sup>7,10,49</sup> By such a domain transformation, data harmonization between chest radiographs acquired from various conditions can be achieved. However, there are two shortcomings with such translation methods, which can be major obstacles, especially in medically underserved regions.

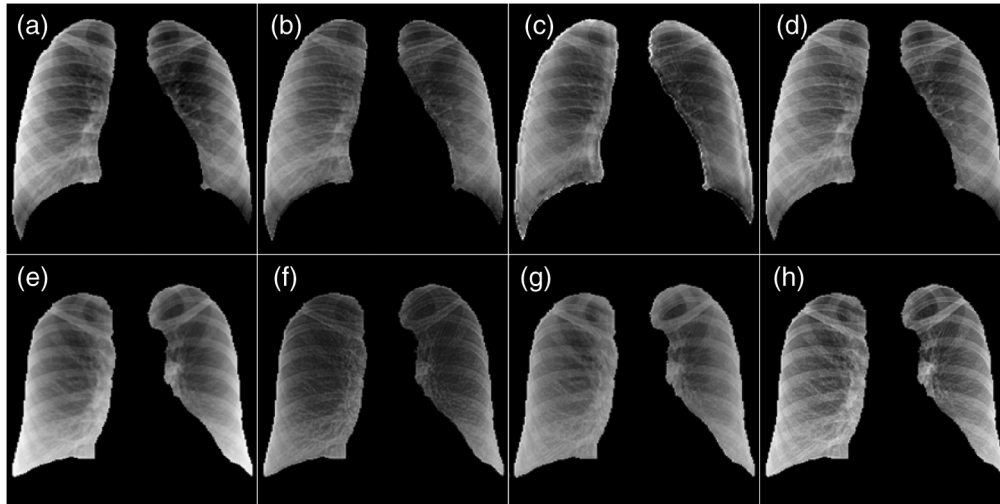
One reason is that deep learning techniques require a considerable amount of data. Recent studies, motivated by generative models, utilize the generative adversarial networks (GAN)<sup>50</sup>-based model to generate domain-matched data from the target to the source.<sup>51–54</sup> The performance of generative models heavily depends on the number of images for training, and the data-efficient models still require hundreds of images.<sup>55</sup>

Another point is that deep neural networks are computationally heavy. The problem remains for the few-shot learning techniques because the CNNs have a large number of parameters to be trained. The computational burden of training the network in a new environment inevitably becomes an obstacle to rapid diagnosis. If the chest x-ray scanner is placed on a mobile system, it is impractical to mount such a high-performance computing device to the system.

There are alternative versions of GAN such as one- or few-shot domain transfer models that require smaller datasets to train. Because we are interested in instance-based data harmonization, we implemented a one-shot GAN learning model.<sup>56</sup> A network was trained to align the domain features of dataset A to the reference data. Lung-masked images were used for training. The process was accelerated by a single NVIDIA GeForce GTX 1080 Ti graphics processing unit, and it took two days to train a single one-shot domain translation network. Two distinct networks were trained on two different samples of dataset A, and generative-model-based domain translation resulted in significantly varying outcomes, depending on which instance was used for one-shot training (Please see Fig. 11). However, the transformations in this paper's scope are fully instance-based and thus free from such complications.

In this study, we analyzed the data features after various harmonizations. For the quantitative analysis, we calculated the Mahalanobis distance between the deep features of reference data and the test data points. The conclusion was that the MFBL, which resulted in the shortest distance, is the most efficient harmonization. Unfortunately, it was difficult to establish the direct correlation between the distance in feature space and the degree of harmonization. For example, the average deep feature distance of N (13.73) is farther than that of S (12.80) in the case of dataset E (see Table 4), whereas the network performance in terms of AUC and AP was superior for the N case. Still, there is an overall tendency that the shorter the distance is, the better the network performance metric becomes. Although further research is needed to determine the criteria for a reliable and credible model that is less affected by the reference data distribution, our suggested feature analysis may provide a basis for such discussion.





**Fig. 11** Comparison of the deep-learning-based one-shot domain translation and the MFBL transformation. Each image corresponds to (a, e) normalized data from dataset A, (b, f) outputs of a network trained on (a), (c, g) outputs of a network trained on (e), and (d, h) MFBL.

We would like to emphasize that this study contributes to understanding the data homogenization processes through feature analyses and recommends the MFBL method as the instance-based data transformation method potentially for the CXR images acquired at medically underserved areas. As the special issue of global health, equity, bias, and diversity in AI in medical imaging is pursued, this contribution will help to reduce global healthcare disparity and diversity.

## 5 Conclusions

In this work, we implemented various instance-based data transformations to reduce data discrepancy for the multi-institutional use of a trained deep-learning prediction model. A CNN-based TB detection network was trained using the reference site data, and the TB detection performance was tested for the remaining six sites after applying N, S, SL, MFB, and MFBL. MFBL outperformed other methods in terms of numerical criteria including AUC, AP, F1 score, and WIOA. For the radiomic feature analysis, we calculated the Mahalanobis distance and performed dimensionality reduction. We found that S and SL match the histogram-based features well but fail to match the texture-related second-order statistics. On the other hand, the textural features of the multi-site CXRs after MFBL were well homogenized to the reference data. The deep features of the trained network were analyzed through the same method, and the MFBL showed the best harmonization performance. Conventional N and S, on the other hand, did not lessen the distribution gap between multi-site datasets and were accordingly unsuccessful in deep feature harmonization. Our study emphasizes the strengths of the MFBL, especially its comparative advantage on network performance and ability to lessen the disparity between various data distributions.

## 6 Appendix A: TB Detection Network Training Details

We initialized every model corresponding to each data transformation method with the pre-trained ResNet-18 provided by PyTorch, and the same random seed was used for training. A dropout layer having a dropout ratio of 0.6 was added before the last linear layer. We used the same computing resources that were used for the lung segmentation network training, and training details are summarized in Table 5. The early stopping condition was determined by monitoring the validation losses with the reference test dataset.

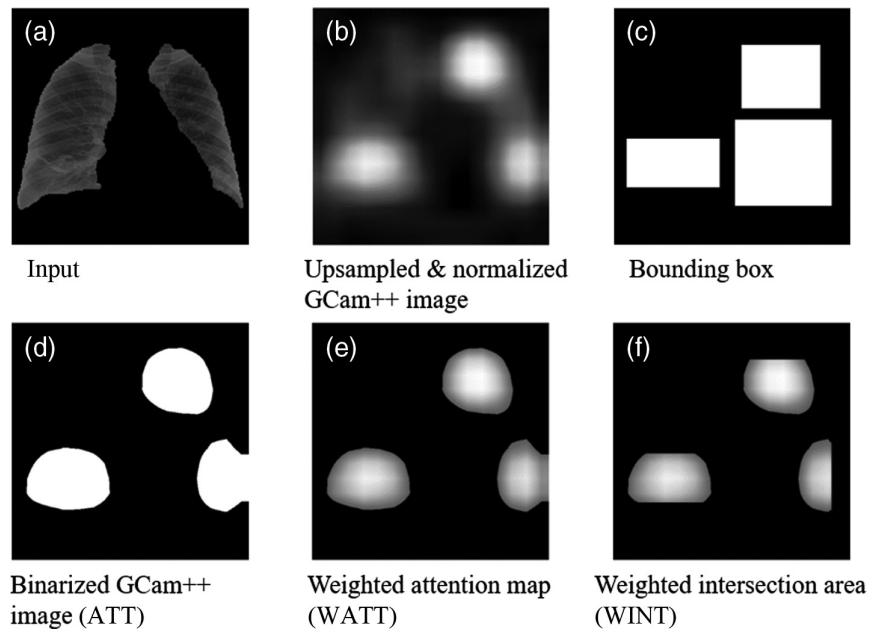
**Table 5** TB detection network training details.

Task	TB detection
Loss function	BCE with logits loss
Optimizer	SGD with momentum
Learning scheduler	Cosine annealing warm restart <sup>57</sup>
Maximum learning rate	1e <sup>-4</sup>
Batch size	128
Number of total epochs	1200

## 7 Appendix B: Calculation of WIOA

In Fig. 12, we show the procedure for calculating the WIOA value in an exemplary case. The Grad-Cam++ image is first segmented by the Otsu method,<sup>58</sup> which results in a binarized attention map ATT. Then, a weighted attention map WATT is generated by multiplying the Grad-Cam++ image and ATT. The multiplication of the WATT and the bounding boxes provides the weighted intersection area WINT, and the WIOA value is finally calculated by

$$\text{WIOA} = \frac{\sum_x \text{WINT}(x)}{\sum_x \text{WATT}(x)}. \quad (4)$$



**Fig. 12** (a) Example input image to explain WIOA value calculation, (b) the GCam++ image of the corresponding input image, achieved at the deepest feature layer, (c) the bounding box of TB regions drawn by a radiologist, (d) the binarized GCam++ image with thresholding, (e) the weighted attention map, (b) multiplied with (d), (f) the weighted intersection area, and (e) multiplied with (c). The WIOA value was defined as the summation of the pixel values of (f) divided by that of (e).

## 8 Appendix C: Data Preprocessing Time

Table 6 provides a detailed analysis of the time required for each data-transforming procedure. It is important to mention that we preprocessed the entire dataset prior to the actual training and saved the results separately. Through this strategy, we could avoid increasing the training time. Specifically, the preprocessing of  $\sim 2000$  reference data required an additional 4 minutes to complete.

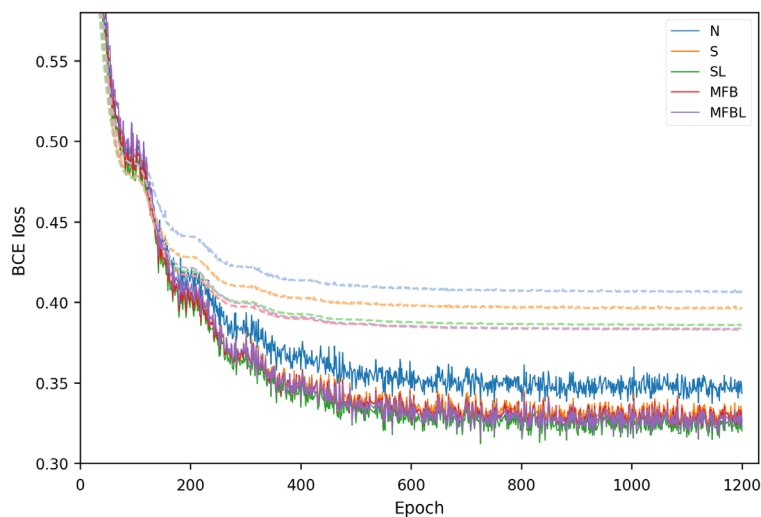
As is summarized above, in addition to the reference computation time (0.28 s), only a few 100 milliseconds are additionally required for the lung segmentation and MFB. We believe this increase in computation time would not hamper its use in clinical practices.

**Table 6** Data preprocessing time. All measurements were done in intel Xeon® CPU E3-1270v5/GTX 1080 Ti platform.

Preprocessing	Approximate time (s)
Dicom read	0.004
N	0.28
Lung segmentation	0.04
Downsampling processes	0.25
S	0.004
MFB	0.13

## 9 Appendix D: Learning Curve

In terms of convergence in the training phase, all of the implemented harmonization methods showed a similar speed of convergence, as shown in Fig. 13. The training phase used 1200 epochs in all cases. Solid lines and dashed lines represent the training phase and the validation phase, respectively.



**Fig. 13** Training and validation loss plots of the networks.

## Disclosures

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

## Acknowledgments

This work was supported in part by the National Research Foundation of Korea (Grant Nos. NRF-2019M2A2A4A05031487 and NRF-2020R1A2C2011959), the Korea Medical Device Development Fund grant funded by the Korea government (the Ministry of Science and ICT, the Ministry of Trade, Industry and Energy, the Ministry of Health and Welfare, and the Ministry of Food and Drug Safety) (Grant Nos. 1711137888 and KMDF\_PR\_20200901\_0012), and the Ministry of Trade, Industry and Energy (MOTIE, Korea) (Grant No. 20014921).

## References

1. R. K. Samala et al., “Generalization error analysis for deep convolutional neural network with transfer learning in breast cancer diagnosis,” *Phys. Med. Biol.* **65**(10), 105002 (2020).
2. I. M. El Naqa et al., “Lessons learned in transitioning to AI in the medical imaging of COVID-19,” *J. Med. Imaging* **8**(S1), 010902 (2021).
3. J. P. Cohen et al., “On the limits of cross-domain generalization in automated X-ray prediction,” in *Proc. Third Conf. Med. Imaging with Deep Learn.* 121, pp. 136–155 (2020).
4. A. Chaddad, L. Hassan, and C. Desrosiers, “Deep CNN models for predicting COVID-19 in CT and x-ray images,” *J. Med. Imaging* **8**(S1), 014502 (2021).
5. J. Manokaran et al., “Detection of COVID-19 from chest x-ray images using transfer learning,” *J. Med. Imaging* **8**(S1), 017503 (2021).
6. E. Çallı et al., “Deep learning for chest X-ray analysis: a survey,” *Med. Image Anal.* **72**, 102125 (2021).
7. H. Zhao et al., “On learning invariant representations for domain adaptation,” in *36th Int. Conf. Mach. Learn. ICML 2019*, pp. 12985–12999 (2019).
8. R. Wang, P. Chaudhari, and C. Davatzikos, “Embracing the disharmony in medical imaging: a simple and effective framework for domain adaptation,” *Med. Image Anal.* **76**, 102309 (2022).
9. S. J. Pan et al., “Domain adaptation via transfer component analysis,” *IEEE Trans. Neural Networks* **22**(2), 199–210 (2011).
10. Y. Ganin and V. Lempitsky, “Unsupervised domain adaptation by backpropagation,” in *32nd Int. Conf. Mach. Learn. ICML 2015*, Vol. 2 (2015).
11. J. Jiang et al., “Self-derived organ attention for unpaired CT-MRI deep domain adaptation based MRI segmentation,” *Phys. Med. Biol.* **65**(20), 205001 (2020).
12. S. Gündel et al., “Robust classification from noisy labels: integrating additional knowledge for chest radiography abnormality assessment,” *Med. Image Anal.* **72**, 102087 (2021).
13. Y. X. Tang et al., “Automated abnormality classification of chest radiographs using deep convolutional neural networks,” *NPJ Digital Med.* **3**(1), 70 (2020).
14. J. Zheng et al., “Pairwise domain adaptation module for CNN-based 2-D/3-D registration,” *J. Med. Imaging* **5**(2), 021204 (2018).
15. P. Márquez-Neila and R. Sznitman, “Image data validation for medical systems,” *Lect. Notes Comput. Sci.* **11767**, 329–337 (2019).
16. D. Singh and B. Singh, “Investigating the impact of data normalization on classification performance,” *Appl. Soft Comput.* **97**, 105524 (2020).
17. M. S. Vidya et al., “Local and global transformations to improve learning of medical images applied to chest radiographs,” *Proc. SPIE* **10949**, 1094936 (2019).
18. R. H. H. M. Philipsen et al., “Localized energy-based normalization of medical images: application to chest radiography,” *IEEE Trans. Med. Imaging* **34**(9), 1965–1975 (2015).
19. R. H. H. M. Philipsen et al., “Automated chest X-ray reading for tuberculosis in the Philippines to improve case detection: a cohort study,” *Int. J. Tuberc. Lung Dis.* **23**(7), 805–810 (2019).

20. K. Murphy et al., “Computer aided detection of tuberculosis on chest radiographs: an evaluation of the CAD4TB v6 system,” *Sci. Rep.* **10**, 5492 (2020).
21. T. De Moor et al., “Automated lesion detection and segmentation in digital mammography using a u-net deep learning network,” *Proc. SPIE* **10718**, 1071805 (2018).
22. K. Murphy et al., “COVID-19 on chest radiographs: a multireader evaluation of an artificial intelligence system,” *Radiology* **296**(3), E166–E172 (2020).
23. J. T. Y. Wu et al., “Developing and validating multi-modal models for mortality prediction in COVID-19 patients: a multi-center retrospective study,” *J. Digital Imaging* **35**, 1514–1529 (2022).
24. M. Roberts et al., “Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans,” *Nat. Mach. Intell.* **3**(3), 199–217 (2021).
25. A. J. DeGrave, J. D. Janizek, and S. I. Lee, “AI for radiographic COVID-19 detection selects shortcuts over signal,” *Nat. Mach. Intell.* **3**(7), 610–619 (2021).
26. P. J. Burt and E. H. Adelson, “The Laplacian pyramid as a compact image code,” *Readings Comput. Vision* 671–679 (1987).
27. P. Vuytsteke and E. P. Schoeters, “Multiscale image contrast amplification (MUSICA),” *Proc. SPIE* **2167**, 551–560 (1994).
28. K. He et al., “Deep residual learning for image recognition,” in *Proc. IEEE Comput. Soc. Conf. Comput. Vision Pattern Recognit.*, pp. 770–778 (2016).
29. X. Huang et al., “Metal artifact reduction on cervical CT images by deep residual learning,” *Biomed. Eng. Online* **17**(1), 175 (2018).
30. J. Shiraishi 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.* **174**(1), 71–74 (2000).
31. 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.* **10**(1), 19–40 (2006).
32. S. Jaeger et al., “Two public chest X-ray datasets for computer-aided screening of pulmonary diseases,” *Quantum Imaging Med. Surg.* **4**(6), 475–477 (2014).
33. A. Paszke et al., “Automatic differentiation in pytorch,” in *NIPS 2017 Work. Autodiff.* (2017).
34. K. K. I. Williams, “The effect of class imbalance on precision-recall curves,” *Neural Comput.* **33**(4) (2021).
35. E. R. DeLong, D. M. DeLong, and D. L. Clarke-Pearson, “Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach,” *Biometrics* **44**(3), 837–845 (1988).
36. X. Sun and W. Xu, “Fast implementation of DeLong’s algorithm for comparing the areas under correlated receiver operating characteristic curves,” *IEEE Signal Process. Lett.* **21**(11), 1389–1393 (2014).
37. A. Chattopadhyay et al., “Grad-CAM++: generalized gradient-based visual explanations for deep convolutional networks,” in *Proc. - 2018 IEEE Winter Conf. Appl. Comput. Vision*, pp. 839–847 (2018).
38. K. Gotkowski et al., “M3d-CAM: a PyTorch library to generate 3D data attention maps for medical deep learning,” in *Bildverarbeitung für die Medizin* pp. 217–222 (2021).
39. R. J. Gillies, P. E. Kinahan, and H. Hricak, “Radiomics: images are more than pictures, they are data,” *Radiology* **278**(2), 563–577 (2016).
40. S. Rizzo et al., “Radiomics: the facts and the challenges of image analysis,” *Eur. Radiol. Exp.* **2**(1), 36 (2018).
41. S. S. F. Yip and H. J. W. L. Aerts, “Applications and limitations of radiomics,” *Phys. Med. Biol.* **61**(13), R150–R166 (2016).
42. D. Kontos, R. M. Summers, and M. L. Giger, “Special section guest editorial: Radiomics and deep learning,” *J. Med. Imaging* **4**(4), 041301 (2018).
43. J. J. M. van Griethuysen et al., “Computational radiomics system to decode the radiographic phenotype,” *Cancer Res.* **77**(21), e104–e107 (2017).



44. R. De Maesschalck, D. Jouan-Rimbaud, and D. L. Massart, "The Mahalanobis distance," *Chemom. Intell. Lab. Syst.* **50**(1), 1–18 (2000).
45. K. Lee et al., "A simple unified framework for detecting out-of-distribution samples and adversarial attacks," in *Adv. Neural Inf. Process. Syst.*, pp. 7167–7177 (2018).
46. J. Schmidhuber, "Deep learning in neural networks: an overview," *Neural Networks* **61**, 85–117 (2015).
47. Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature* **521**(7553), 436–444 (2015).
48. L. van Der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.* **9**, 2579–2605 (2008).
49. S. Guo et al., "Multi-level semantic adaptation for few-shot segmentation on cardiac image sequences," *Med. Image Anal.* **73**, 102170 (2021).
50. I. J. Goodfellow et al., "Generative adversarial nets," in *Adv. Neural Inf. Process. Syst.*, Vol. 3(Jan.) (2014).
51. M. Y. Liu and O. Tuzel, "Coupled generative adversarial networks," in *Adv. Neural Inf. Process. Syst.* (2016).
52. J. Hoffman et al., "CyCADA: cycle-consistent adversarial domain adaptation," in *35th Int. Conf. Mach. Learn.*, Vol. **5** (2018).
53. E. Tzeng et al., "Adversarial discriminative domain adaptation," in *Proc. - 30th IEEE Conf. Comput. Vis. Pattern Recognit.* (2017).
54. S. W. Huang et al., "AugGAN: cross domain adaptation with GAN-based data augmentation," *Lect. Notes Comput. Sci.* **11213** 731–744 (2018).
55. S. Zhao et al., "Differentiable augmentation for data-efficient GAN training," in *Adv. Neural Inf. Process. Syst.* (2020).
56. S. Benaim and L. Wolf, "One-shot unsupervised cross domain translation," in *Adv. Neural Inf. Process. Syst.*, pp. 2104–2114 (2018).
57. I. Loshchilov and F. Hutter, "SGDR: stochastic gradient descent with warm restarts," in *5th Int. Conf. Learn. Represent. ICLR 2017 - Conf. Track Proc.* (2017).
58. N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Trans Syst Man Cybern* **9**(1), 62–66 (1979).

Biographies of the authors are not available.