IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 20 November 2023, accepted 18 December 2023, date of publication 22 December 2023,
date of current version 9 January 2024.

Digital Object Identifier 10.1109/ACCESS.2023.3346315

RESEARCH ARTICLE

ThoraX-PriorNet: A Novel Attention-Based Architecture Using Anatomical Prior Probability Maps for Thoracic Disease Classification

MD. IQBAL HOSSAIN[1], **MOHAMMAD ZUNAED**[1], **(Student Member, IEEE),**
MD. KAWSAR AHMED[1], **S. M. JAWWAD HOSSAIN**[1], **ANWARUL HASAN**[2,3], **(Member, IEEE),**
AND TAUFIQ HASAN[1,4], **(Senior Member, IEEE)**
[1]mHealth Laboratory, Department of Biomedical Engineering, Bangladesh University of Engineering and Technology, Dhaka 1205, Bangladesh
[2]Department of Mechanical and Industrial Engineering, Qatar University, Doha, Qatar
[3]Biomedical Research Center, Qatar University, Doha, Qatar
[4]Center for Bioengineering Innovation and Design (CBID), Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD 21218, USA

Corresponding authors: Taufiq Hasan (taufiq@bme.buet.ac.bd) and Anwarul Hasan (ahasan@qu.edu.qa)

This work was supported by the Open Access funding provided by the Qatar National Library.

ABSTRACT Computer-aided disease diagnosis and prognosis based on medical images is a rapidly emerging field. Many Convolutional Neural Network (CNN) architectures have been developed by researchers for disease classification and localization from chest X-ray images. It is known that different thoracic disease lesions are more likely to occur in specific anatomical regions compared to others. This article aims to incorporate this disease and region-dependent prior probability distribution within a deep learning framework. We present the ThoraX-PriorNet, a novel attention-based CNN model for thoracic disease classification. We first estimate a disease-dependent spatial probability, i.e., an *anatomical prior*, that indicates the probability of occurrence of a disease in a specific region in a chest X-ray image. Next, we develop a novel attention-based classification model that combines information from the estimated *anatomical prior* and automatically extracted chest region of interest (ROI) masks to provide attention to the feature maps generated from a deep convolution network. Unlike previous works that utilize various self-attention mechanisms, the proposed method leverages the extracted chest ROI masks along with the probabilistic *anatomical prior* information, which selects the region of interest for different diseases to provide attention. The proposed method shows superior performance in disease classification on the NIH ChestX-ray14 dataset compared to existing state-of-the-art methods while reaching an area under the ROC curve (%AUC) of 84.67. Regarding disease localization, the anatomy prior attention method shows competitive performance compared to state-of-the-art methods, achieving an accuracy of 0.80, 0.63, 0.49, 0.33, 0.28, 0.21, and 0.04 with an Intersection over Union (IoU) threshold of 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, and 0.7, respectively. The proposed ThoraX-PriorNet can be generalized to different medical image classification and localization tasks where the probability of occurrence of the lesion is dependent on specific anatomical sites.

INDEX TERMS Anatomical prior, anatomy-aware attention, chest X-ray, thoracic disease classification.

I. INTRODUCTION

Thoracic disorders are one of the major health concerns worldwide as the heart and lungs, two vital human organs, are located within the thorax. In 2017, around 544·9 million people were affected by chronic respiratory illness [1], a thoracic disease, leading to 3.9 million deaths [2]. Various medical imaging modalities, e.g., X-ray, Magnetic Resonance Imaging (MRI), and Computed Tomography (CT) can diagnose different thoracic disorders. The chest

The associate editor coordinating the review of this manuscript and approving it for publication was Tao Huang.

X-ray (CXR) remains the most commonly performed and widely available radiological diagnostic method to assess and diagnose thoracic diseases. The chest radiograph is an X-ray projection image of the thoracic cavity used to diagnose conditions affecting the chest, its contents, and nearby structures. It is one of the most effective and low-cost methods for diagnosing thoracic diseases. Since CXR is a projection imaging method providing a 2D image of the 3D thoracic structure, anatomical structures are overlapped in the resulting image. Therefore, diagnosis of diseases with CXR image highly depends on the skill and experience of the radiologist [3]. However, in many underserved regions of the world, the number of skilled radiologists is insufficient. In such scenarios, automated CXR image interpretation using artificial intelligence (AI) can significantly benefit health systems. This is true even if the algorithms are not making full autonomous decisions and are only used to assist physicians.

However, it is of paramount importance for the machine learning models to be explainable for the radiologists to trust them. Thus, providing an accurate location for the predicted pathologies is a prerequisite for computer-aided diagnosis. However, due to the lack of pixel-level ground truth annotation data, the deep learning models suffer from sub-optimal optimizations. A number of weakly supervised disease localization methods over the recent years have been proposed to solve this problem. In the literature, different attention-based approaches [4], [5], [6] have been used for medical disease diagnosis, where the model traditionally learns to identify and focus on the regions of interest containing the lesions using activated feature maps from the classifiers. However, these methods are data-driven and are generally agnostic to the human anatomy and its dependence on identifying the diseased regions. They do not take into account the typical occurrence areas for a specific pathology, and thus, they often fail to predict the lesion region as recognized by radiologists. Intuitively, radiologists do not search all the parts when diagnosing chest X-ray images of a patient for thoracic diseases. Instead, they concentrate on the areas related to the symptoms of the disease of a patient.

Different thoracic disease lesions have unique characteristics and are identified in specific regions of a chest radiographic image. For example, when identifying pneumonia, a radiologist looks for white spots in the lungs that show the characteristics of infection. In contrast, the opacity features of pleural effusion manifest in the pleural space, not inside the lung region. Similarly, the cardiomegaly pathology is associated with the heart. Thus, we may consider that the diagnostic features of different thoracic diseases have a higher probability of occurrence in certain anatomical regions of the chest X-ray. Consequently, specific disease features may have a zero probability of occurrence in certain anatomical regions (e.g., observing consolidation features outside the lungs). Therefore, to reliably detect and localize thoracic diseases, we not only require deep learning-based models to learn the disease-specific features but also to focus on the specific anatomical regions where the likelihood of the disease is highest. However, the existing literature studies predict only the most discriminative areas for the pathology localization and classification of a patient without considering the prior distribution knowledge of the regions where a pathology most repeatedly appears. Although Chen et al. [7] and Kamal et al. [8] utilized lung segmentation-based attention mechanisms, disease-specific anatomical prior knowledge was not considered within the attention mechanism and abnormality localization.

Considering the limitations of previous works in this area, we propose a novel model architecture using two types of attentions: chest region of interest mask-based attention and disease-specific anomaly-based attention for disease classification. The main contributions of this paper are as follows:

- We propose the concept of a novel probabilistic anatomical prior map that provides a spatial probability distribution of a disease occurrence within X-ray images. To the best of our knowledge, the idea of a disease-specific anatomical prior probability maps generated using an aggregation of disease ROI masks has not been explored in previous research works.
- We developed an end-to-end model ThoraX-PriorNet, a novel attention-based architecture that focuses on specific regions of an X-ray image informed by both disease-specific anatomical prior probability maps and lung region-of-interest (ROI) masks.
- We conducted a thorough experimental evaluation to compare the performance of the proposed ThoraX-PriorNet model with the existing methods. Detailed ablation studies conducted using the anatomy prior attention module (APAM) demonstrate the effectiveness of the proposed method in accurately detecting thoracic diseases.

The rest of our document is organized as follows. Section II reviews the related works in the thoracic disease classification and weakly supervised localization tasks. Section III presents our proposed approach in detail. Section IV discusses our experimental settings, such as datasets, data preparation, training scheme, and so on. We conduct comprehensive experiments in Section V, including ablation studies, performance comparison with state-of-the-art methods, statistical analysis, and so on, both for classification and localization tasks. In section VI, we conclude this paper.

II. RELATED WORKS

A. ATTENTION

Attention mechanisms that selectively attend to zones of an image with a high probability of exhibiting particular diseases can yield a substantial performance improvement for machine learning models [9]. Chest X-rays are frequently employed for diagnosing respiratory and cardiovascular conditions, precise interpretation of these images is imperative for effective treatment [10], [11], [12], [13], [14], [15]. Attention modules available in the computer vision literature can

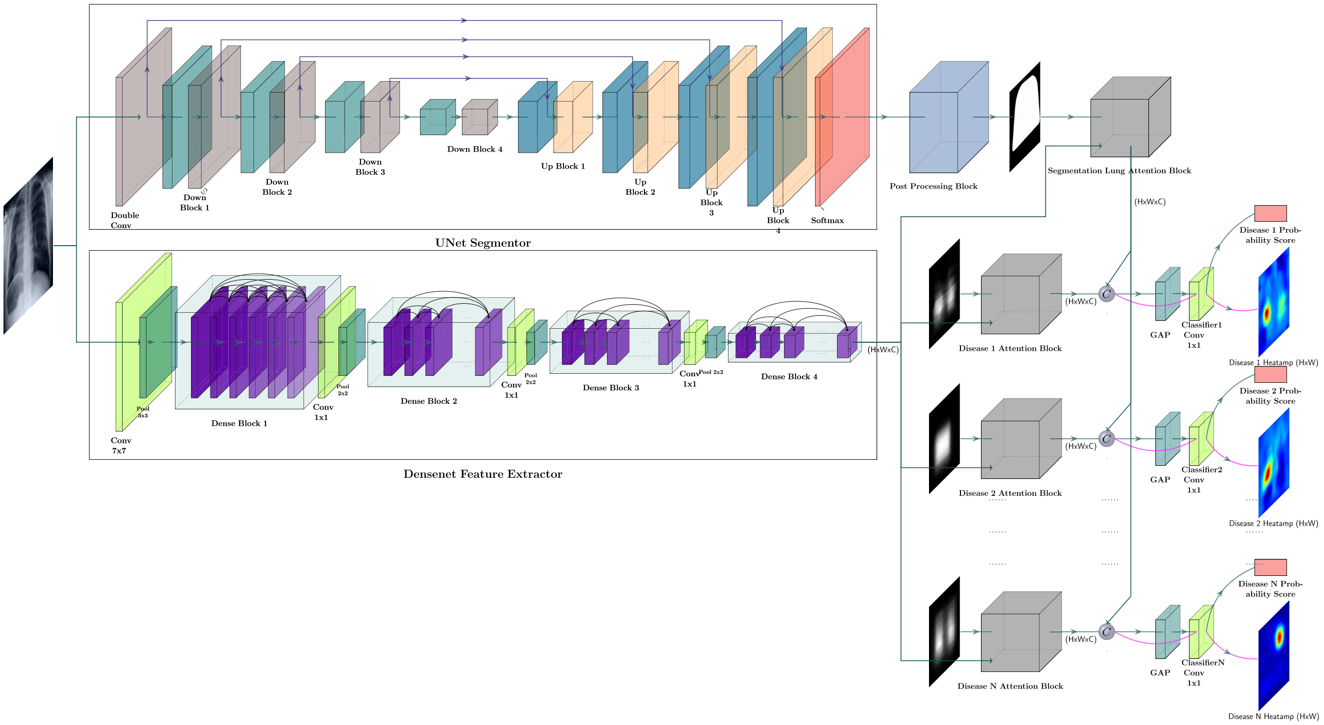


FIGURE 1. A schematic of the proposed ThoraX-PriorNet architecture for disease classification from CXR utilizing both lung segmentation attention and disease-specific attention. The model consists of three components: the lung segmentation attention module, the disease-specific attention module, and then concatenation for classification. The lung segmentation U-Net model generates a lung ROI mask, which is then used to provide lung mask guided attention. The disease-specific probability mask is used along feature map to provide disease specific attention. Finally, the concatenated feature maps are used to make the final disease classification.

be divided into two main categories. One includes the Squeeze-and-Excitation (SE) approach that adaptively recalibrates channel-wise feature responses by explicitly modeling inter-dependencies between channels [16]. The other is Gather-Excite (GE) method which efficiently aggregates feature responses from a large spatial extent and excites, redistributing the pooled information to local features [17]. Chen et al. [18] presented a non-local (NL) attention module to utilize the local relationship for capturing long-range dependencies. Wang et al. [19] have introduced a triplet attention model that can learn channel-wise, element-wise, and scale-wise attention simultaneously. This approach helps to capture distinctive information relevant to the task of classifying thorax diseases. Ullah et al. [10] incorporated channel-wise attention as layer in multiple positions in the their feed forward network for Covid-19 classification. Zhang et al. [20] presented attention guided with different parts of lung. Kamal et al. [8] used lung segmentation mask to provide attention in the lung region in a chest X-ray image. To overcome the domain mismatch of lung segmentation dataset they used GAN model to segment lung that was later used for providing attention.

Though providing attention modules in network enhances model performance, most existing approaches mainly focus on learning the attention map using global CXR images, without considering disease specific lung regions. Aiming to address this constraint, the proposed method generated disease specific probabilistic map from the provided bounding box annotation. Then, we provided probabilistic map guided and lung mask guided attention to focus at specific regions in chest X-ray image for thoracic disease evaluation.

B. WEAKLY SUPERVISED LEARNING

Achieving success in supervised learning demands sophisticated network engineering and an enormous quantity of precisely labeled training data [21]. Weakly supervised learning is becoming increasingly important in medical chest X-ray analysis as it can alleviate the need of extensive and precise annotations required for supervised learning. Wang et al. [22] introduce the ChestX-ray14 dataset, together with a baseline for evaluating weakly supervised lesion localization. Furthermore, numerous studies have previously investigated disease localization on CXR images [22], [23], [24], without directly utilizing ROI labels. Notably, prior research on localization such as Ye et al.'s [14] use of probabilistic-CAM Pooling and Ouyang et al.'s [6] use of hierarchical attention for weakly supervised abnormality localization have incorporated attention mechanisms in their architectures. In their study, Ullah et al. [10] utilized grad-CAM to produce a COVID-19 heatmap, with the aim of presenting classification outcomes that are supported by clinical evidence, and thus applicable to clinical practice. Employing saliency techniques, such as Class Activation Mapping (CAM), Grad-CAM [25], Grad-CAM++ [26], Eigen-CAM [27], and similar methods, to produce heatmaps can prove to be highly beneficial in furnishing clinical evidence. Rozenberg et al. [28] achieved high localization performance in regimes by learning to localize the areas with a limited annotation derived from a small fraction masked. Zhu et al. [29] proposed a convolutional attention-based network named PCAN that is pathology-aware and capable of capturing the variations in lesion size and location by generating pixel-wise diagnoses and pixel-wise weights. Han et al. [30] leverage two views, i.e., radiomic and global image features, for training the framework for classifying and localizing thoracic diseases. To extract the radiomic features, they have exploited Grad-CAM generated by the image classifier backbone through a feedback loop mechanism. Xiao et al. [31] improved the performance of ViTs by pre-training with 266,340 chest X-rays using Masked Autoencoders, reconstructing missing pixels from a small part of each image. Li et al. [32] utilized an adaptive ViT with a DenseNet architecture with a feature pyramid structure to design the inter-patch and patch-wise long-range dependencies and obtain fine-grained feature maps.

However, the previous methods from the literature depend on the discriminative power of deep-learning convolutional networks and predict the area of a chest X-ray that is most responsible for classification as lesion area without considering the prior knowledge of the distribution of disease occurrence area in a chest X-ray image. Instead, we developed an end-to-end novel attention-based architecture named ThoraX-PriorNet, which focuses on specific regions of a chest X-ray image guided by typical disease-specific spatial anatomical prior probability maps.

III. PROPOSED METHOD

This section describes our proposed approach, where we have used a deep learning-based novel classification architecture, named ThoraX-PriorNet, that utilizes both the chest ROI mask and a disease-specific anatomical prior probability map for pathology classification and localization. We also describe in detail the extraction of the chest ROI mask and the generation of a disease-specific anatomical prior probability map.

A. GENERATING DISEASE-SPECIFIC ANATOMICAL PRIOR PROBABILITY MAP

We compute the disease-specific anatomical prior probability maps by identifying the spatial regions of the CXR images where the lesions are most likely to occur. To construct this map, we use the NIH Chest X-ray dataset, which includes 880 bounding-box annotated images identifying the regions of the abnormality [22]. First, we create a binary image keeping the bounding-box interior spatial values equal to 1 and the rest equal to 0 for a particular disease. Out of the eight pathologies, seven pathologies (atelectasis, effusion, infiltrate, mass, nodule, pneumonia, and pneumothorax) can occur symmetrically in the lungs. Leveraging this behavior, we apply horizontal flipping to bounding boxes of these seven types of diseases to generate new annotations. We then take the sum of all binary images of a particular disease to generate unnormalized probability map. Finally, we normalize pixel values of the unnormalized probability map by dividing them by the maximum pixel value within that probability map. The normalized mask is used in the network as the anatomical prior probability map for providing disease-specific attention.

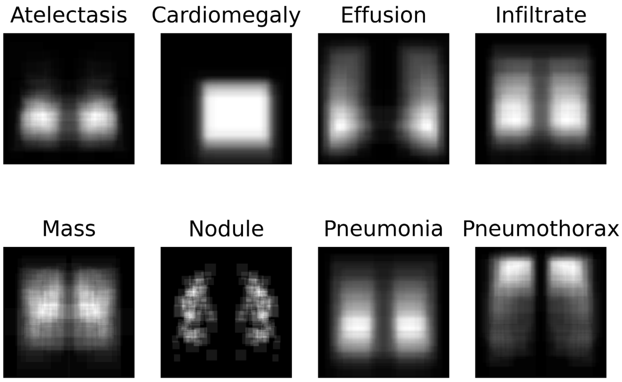


FIGURE 2. Disease-specific anatomical prior probability maps generated for the 8 diseases for which the bounding box annotations are available in the NIH dataset.

First, we obtain the unnormalized raw probability map. Let $I_c^k(i,j)$ indicate the pixel position (i,j) of the k^{th} constructed binary mask image from the bounding box annotated ground truth image for the disease class c. The disease-specific anatomical prior probability map M_c^p is generated as follows.

$$\hat{M}_c(i,j) = \sum_{k=1}^{N_c} I_c^k(i,j) \tag{1}$$

where N_c indicates the number of CXR images available for the disease class c. Next, we normalize the raw map $\hat{M}_c$ to obtain the final anatomical prior probability map by,

$$M_c^p(i,j) = \frac{\hat{M}_c(i,j)}{\max\left(\hat{M}_c\right)} \tag{2}$$

Here, the max operation identifies the maximum pixel value of the raw probability map $\hat{M}_c$. Finally, these anatomical prior probability maps were generated for all eight diseases for which the bounding box annotations are available. Fig. 2 shows the generated disease-specific anatomical prior probability maps for the eight abnormalities. In the strictest sense, the obtained maps $M_c^p(i,j)$ do not represent an actual probability distribution. Firstly, this is because the regions are obtained from the bounding box information that is larger than the actual disease regions. Secondly, obtaining a probability distribution requires that the integration over the entire map should equal unity. In actual implementation, the map's relative intensity values are more important than the absolute values. For, disease classes whose bounding-box annotations are not available, we used $M_c^p(i,j) = 1$.

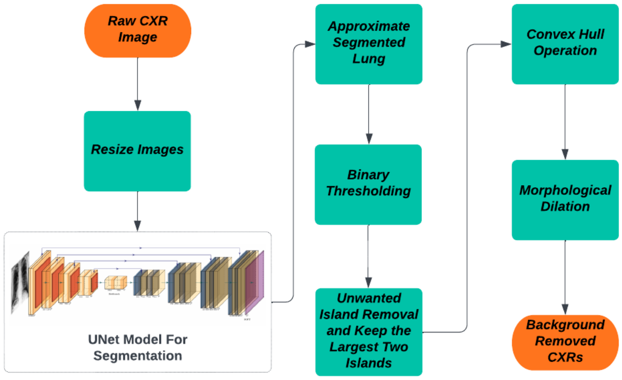


FIGURE 3. A flow-diagram of the chest ROI mask generation module.

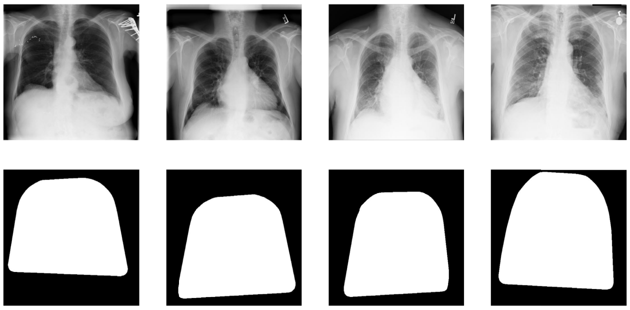


FIGURE 4. Examples of some generated chest ROI masks. Top panel: Example input CXR images, Bottom panel: Corresponding chest ROI masks of the example CXR images.

B. CHEST ROI MASK GENERATION

We employ the well-established U-net [33] segmentation model to extract the lung regions from the input CXR images. We train the model using the 247 images from the JSRT dataset [34]. The segmentation model produces undesirable small islands in the case of some images. To address these issues, we binarize and apply post-processing to the segmentation results to remove the unwanted islands based on the anatomical characteristics of the lungs. Since all other islands are small compared to the lung islands, we filter out the largest two islands representing the right and left lung. The sternum region is also important for some thoracic diseases and contains crucial information for classification. To retain this region, we use the convex hull operation [35]. Finally, we use morphological expansion to retain further information from the pleural regions. The overall chest ROI mask generation flow chart is provided in Fig. 3. Some of the CXR images and their corresponding generated masks are shown in Fig. 4. These postprocessing operations are represented by the postprocessing block in the ThoraX-PriorNet full architecture in Fig. 1.

C. ANATOMICAL PRIOR ATTENTION MODULE (APAM)

In this section, we describe the anatomical prior attention module (APAM), which takes a feature map and a mask (chest ROI mask or anomaly probability map) as inputs to generate an attention map by providing spatial attention to the feature map. An illustration of the APAM framework is demonstrated in Fig. 5. First, we multiply the feature map with the input mask to generate a masked feature map. Later, we take the weighted sum of the feature map and masked feature map to retain information from the region outside the mask since some disease predictions may depend on the feature of the unmasked region. The weights are generated from the feature map and the masked feature map through a CNN. To learn the weights, we use a network similar to the channel-wise attention module described in [36]. However, unlike [36], we aggregate spatial information from both the feature map and the masked feature map.

Let $F \in \mathbb{R}^{C\times H\times W}$ be the feature map generated by the backbone CNN network and $M_{inp} \in \mathbb{R}^{1\times H\times W}$ be the input mask (chest ROI mask or anomaly probability map) resized to the spatial dimension of feature map F. We pass the feature map F into two pooling layers: global average pooling (AvgPool) and global max pooling (MaxPool). The two corresponding outputs from these pooling layers are denoted as F_{avg} and F_{max} respectively, where $F_{avg}, F_{max} \in \mathbb{R}^{C\times 1\times 1}$. Again, let $F_m \in \mathbb{R}^{C\times H\times W}$ be the masked feature map which is produced after we multiply the feature map F with the input mask M_{inp}. We obtain $M_{avg}, M_{max} \in R^{C\times 1\times 1}$ after passing M through the global average pooling and global max pooling layers in a similar way.

$$F_m = F \odot M_{inp} \tag{3}$$
$$F_{avg} = \text{AvgPool}\,(F) \tag{4}$$
$$F_{max} = \text{MaxPool}\,(F) \tag{5}$$
$$M_{avg} = \text{AvgPool}\,(M) \tag{6}$$
$$M_{max} = \text{MaxPool}\,(M) \tag{7}$$

Here, $\odot$ denotes element wise multiplication. Furthermore, instead of shared multi-layered perceptron (MLP), we use separate MLPs for all four spatial context descriptors $\left(F_{avg}, F_{max}, M_{avg}, M_{max}\right)$. After passing the spatial context descriptors through the CNN, the network produces the required channel weighting values, $W \in \mathbb{R}^{C\times 1\times 1}$. The mathematical equation for generating the channel weighting values, W is provided below:

$$\begin{aligned} W =& \text{CS}\Big(\text{CLR}_1\left(F_{avg}\right) + \text{CLR}_2\left(F_{max}\right) \\ &+ \text{CLR}_3\left(M_{avg}\right) + \text{CLR}_4\left(M_{max}\right)\Big) \end{aligned} \tag{8}$$

Here, CLR$_1$,CLR$_2$, ...,CLR$_4$ indicate the blocks of sequential convolutional layer, and leaky ReLU activation layer and then CS indicates block of sequential convolutional layer followed by sigmoid activation layer. In CS block, we use the sigmoid activation function so that the components of weight W are within the range [0, 1]. For the CLR blocks, we use the leaky ReLU with a negative slope of 0.2 to mitigate the vanishing gradient problem [37]. Finally, we generate the attention map $A \in \mathbb{R}^{C\times H\times W}$ from the weighted sum of F

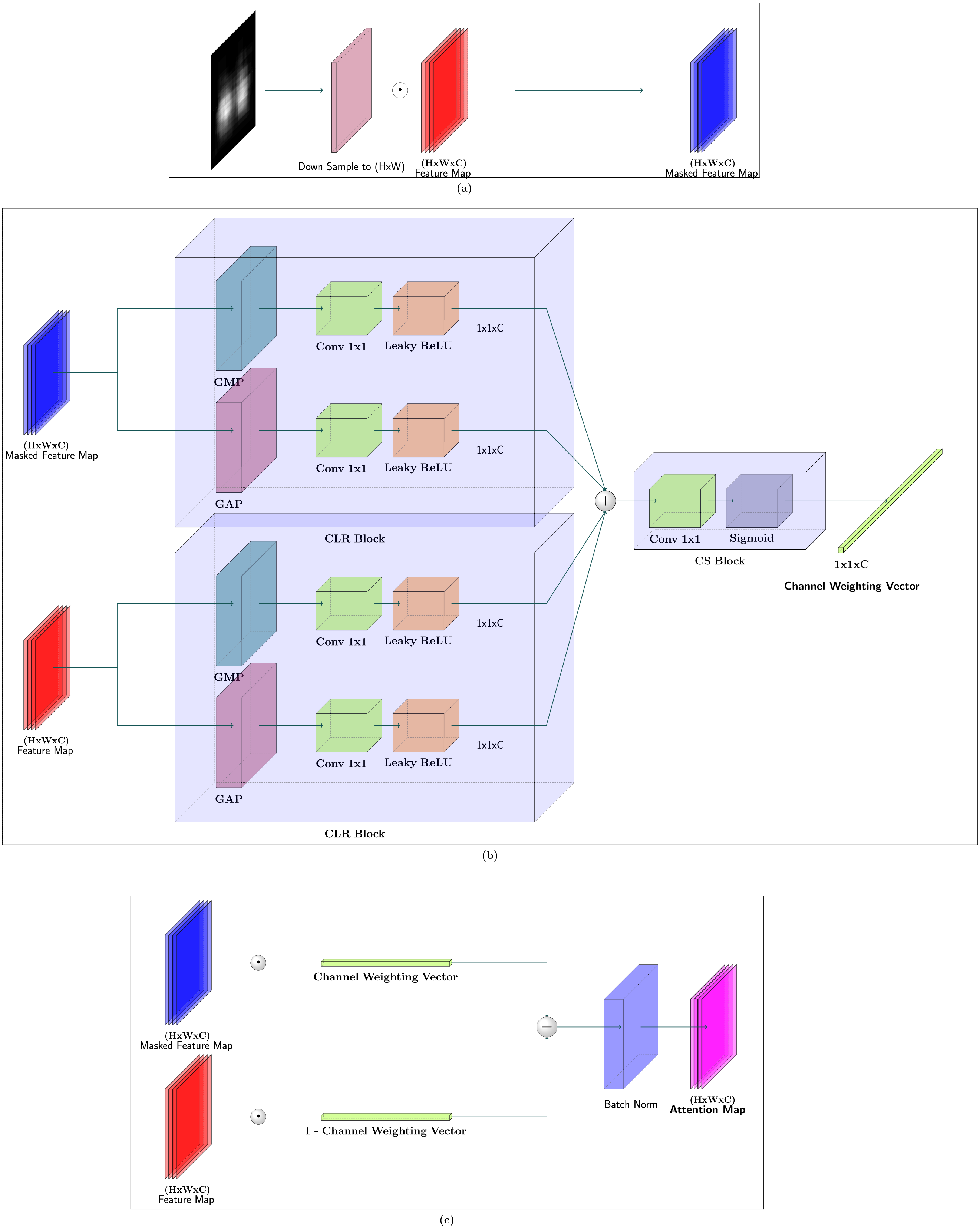


FIGURE 5. A schematic diagram of the Anatomy Prior Attention Module (APAM): A) Mask is multiplied with the feature map to generate a masked feature map; B) Featuremap and Masked Featuremap is being used to produce channel weighting vector. Here, GMP = Global Max Pooling and GAP = Global Average Pooling; C) Channel weighting vector is being used to produce final weighted featuremap i.e, Attention Map.

and F_m using the formula below:

$$A = W \odot F + (1 - W) \odot F_m \tag{9}$$

D. CLASSIFICATION AND LOCALIZATION

At first, we extract a feature map from the input image with a CNN backbone. We have used DenseNet-121 [38] as backbone for feature extraction. Then we use APAM to generate an attention map from the extracted feature map. For generating attention maps from the feature map, we have used the image-specific chest ROI mask described previously with APAM to generate ROI attention map A_{ROI}.

$$A_{ROI} = W_{ROI} \odot F + (1 - W_{ROI}) \odot F_{ROI} \tag{10}$$

Here, W_{ROI} is the weight generated by APAM from feature map F and masked feature map W_{ROI}. Then, we have used K (K = number of abnormalities) numbers of disease-specific anatomy prior probability maps with APAM to generate K disease-specific attention maps A_p^c.

$$A_p^c = W_p^c \odot F + \left(1 - W_p^c\right) \odot F_p^c, \; c \in \{1, \ldots, K\} \tag{11}$$

Here, W_p^c is the weight generated by APAM from feature map F and masked feature map F_p^c of abnormality c. Then for predicting the probability of each disease, the image-specific ROI attention map and the disease-specific attention map of that particular disease are channel-wise concatenated to produce a disease-specific concatenated map.

$$A_{cat}^c = \text{concat}(A_{ROI}, A_p^c), \; c \in \{1, \ldots, K\} \tag{12}$$

Here, $A_{cat}^c \in \mathbb{R}^{2C \times H \times W}$. These concatenated maps are passed through individual global pooling and then 1 × 1 convolutional layers sequentially to generate the probability of that disease. And we have used the same convolutional layers on the concatenated feature maps to generate individual heatmap using CAM method. The schematic of proposed architecture of ThoraX-PriorNet is shown in Fig. 1.

E. LOSS FUNCTION

We concatenate the predicted raw values from each of the pathology-specific classifiers and pass them through a sigmoid layer to generate the probabilities, $p^s = [p_1^s, \ldots, p_i^s, \ldots, p_c^s]$. Here, c represents the number of pathologies presented in a dataset. The ground truth vectors of each chest X-ray are expressed as an c-dimensional label vector, $L = [l_1, \ldots, l_i, \ldots, l_c]$, where $l_i \in \{0, 1\}$. l_i denotes whether there is any pathology, i.e., 1 for presence and 0 for absence. We optimize the weight parameters of our model by minimizing the binary cross-entropy loss, defined as,

$$\mathcal{L} = -\frac{1}{c}\sum_{i=1}^{c}\left[l_i \log\left(p_i^s\right) + (1 - l_i)\log\left(1 - p_i^s\right)\right] \tag{13}$$

IV. IMPLEMENTATIONAL DETAILS

A. DATA RESOURCES

We evaluate the proposed ThoraX-PriorNet architecture on the NIH ChestX-Ray14 and CheXpert datasets. These data resources are briefly described below.

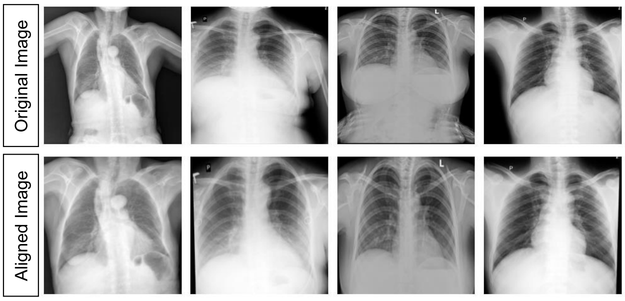


FIGURE 6. Examples of original chest X-ray images and aligned chest X-ray images.

1) NIH CHESTX-RAY14

The NIH ChestX-Ray14 contains 112, 120 frontal chest X-ray images from 30,805 unique patients [22]. All these images are annotated for 15 classes (14 diseases along with "No Findings"). Within this dataset, 880 images are specially annotated by a bounding box for the localization of 8 diseases. In our classification experiments, we use 70%, 10%, and 20% data for training, cross-validation, and testing, respectively. We train and test our model on the classification data for all 15 classes. On the other hand, we use the bounding-box annotated data of the 8 classes to assess the disease localization performance of our model. Note that there is no patient overlap between all the training, validation, and testing sets. The 880 images with bounding box information are not utilized in training or validation splits.

2) CHEXPERT

The CheXpert dataset [39] is a chest X-ray dataset containing class label annotation of 14 classes (13 diseases along with "No Findings"). Other than positive and negative labels for each class, the dataset also contains an uncertainty label for some images. The dataset consists of 224,316 chest X-ray images for training and 230 chest X-ray images for validation. We use only frontal view chest X-ray images from this dataset. If we consider only images with a frontal view, there are about 200,000 chest X-ray images for training and 200 images for validation in the dataset. We use this dataset for the classification of five thoracic diseases, namely, atelectasis, cardiomegaly, consolidation, edema, and pleural effusion.

B. DATA PREPARATION

The chest X-ray images from a dataset generally have diverse variations, such as rotations, shifts, and different scales, making it challenging for the deep-learning models to localize the lesion areas. To address this problem, we utilize the alignment module [40] to perform spatial alignment on all the images as well as on the bounding box images for generating abnormality masks. Given the input image I, the alignment module ϕ transforms I to $\phi(I)$. The canonical chest X-ray image, known as the target image T, is generated by randomly selecting two thousand normal chest X-ray images and averaging them to a single image. To provide $\phi(I)$ with an aligned structure, we minimize the feature reconstruction loss [41] between $\phi(I)$ and T. The backbone of the alignment module consists of ResNet-18 architecture. The output of the alignment network is the affine transformation parameters. Finally, the affine transformation is applied to the original chest X-ray images to generate aligned chest X-ray images. Fig. 6 shows some examples of original and aligned X-ray images.

We first normalize the pixel values of chest X-ray images with the mean and standard deviation of pixels from the ImageNet dataset [42]. Next, we resize the image to 586 × 586 pixels. Afterward, the training images are randomly cropped to 512 × 512 pixels [29], [43]. The validation and test images are center-cropped to 512 × 512 pixels. We use the same resizing and cropping method for the corresponding anatomy prior maps and chest ROI masks. Following [44] and [43], we use test time augmentation by utilizing average probabilities of ten cropped sub-images (four corner crops and one central crop and the horizontally flipped version of them) as the final prediction. In the case of CheXpert dataset preparation (image augmentation, dealing with class imbalance, uncertain labels, etc.), we use the same procedure described in [14]. We use the same disease-specific anatomy prior maps computed from the NIH dataset for the CheXpert dataset.

C. TRAINING PARAMETERS

The Table 1 shows the hyperparameters used for training and evaluation of the deep learning model. These include the number of epochs, batch size, loss function, optimizer, learning rate, learning rate scheduler, and weight decay rate. We have utilized the exponential moving average scheme with an alpha rate of 0.997 for updating the model weight. In addition, we have performed gradient accumulation with a step of eight iterations.

D. ACTIVATION MAP AND BOUNDING BOX GENERATION

We use class activation maps (CAM) for heatmap generation. For the generation of bounding boxes from the heatmap map, to evaluate localization performance, we first convert the activation map or heatmap to a binary mask using binary thresholding with a threshold value of 127. Next, we use the algorithm introduced by [45] to find the contours of the regions inside the binary mask and prepare bounding boxes around the contours by taking extreme boundary values of the contours as the edge of our bounding boxes.

E. EVALUATION METRICS

We use ROC-AUC (Receiver Operating Characteristic-Area Under Curve), also abbreviated as AUC, to measure the classification performance of our model on the NIH test data. Furthermore, we use the ratio of the number of cases with correct localization against the total number of cases in each class to report the localization performance of our models on 880 bounding-box annotated data of the NIH dataset. Here, we use IoU (Intersection over Union) between the predicted bounding box and ground-truth to detect correct localization following prior work [6], [22], [46]. In this case, the localization result is regarded as correct if $IoU > T(IoU)$ where $T(IoU)$ is the threshold for localization.

We have chosen the model with the highest AUC score on the validation split for inference on the test dataset. The loss and AUC curves during training and validation on the NIH ChestX-Ray14 dataset are given in Fig. 7.

TABLE 1. Hyperparameters of the deep learning model used for training and evaluation.

Name of Variable	Values
Epochs	50
Batch size	16
Loss Function	Binary cross entropy loss
Optimizer	Adam
Learning rate	0.0001
Learning rate scheduler	Exponential, 0.75 per 4 epochs

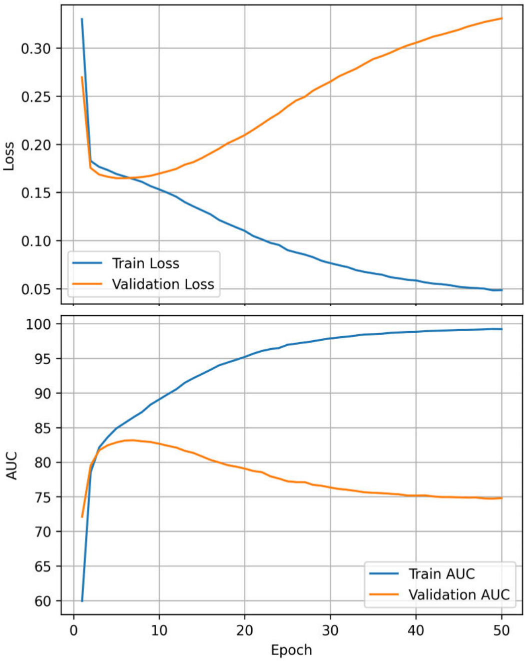


FIGURE 7. Illustration of the training and validation loss and AUC curves on the NIH ChestX-Ray14 dataset.

V. EXPERIMENTAL RESULTS

A. DISEASE CLASSIFICATION

1) ABLATION STUDY

We have conducted several ablation studies on the NIH ChestX-ray14 dataset of our trained model for different thoracic abnormalities. First, we evaluate the impact of attention masks, i.e., probabilistic abnormality mask and chest ROI mask, on the classification performance. Table 2 shows the reported results. The baseline model showed a mean AUC (%) score of 84.30, which performed better for

TABLE 2. Ablation Study: Impact of different types of attention masks on the AUC (%) scores of our trained models on the NIH dataset. The best results are shown in red font.

Method	AM (AbM)	AM (LM)	Atel	Card	Effu	Infil	Mass	Nodu	Pne1	Pne2	Cons	Edem	Emph	Fib	PT	Her	Mean
Baseline			**82.98**	90.20	88.25	72.32	86.50	**80.86**	75.99	88.99	**81.50**	**90.80**	92.56	81.88	80.58	86.91	84.30
ThoraX- PriorNet		✓	82.61	89.78	88.25	72.30	86.83	80.76	75.48	**89.35**	81.04	90.50	**92.85**	81.99	81.00	88.16	84.35
ThoraX- PriorNet	✓		82.47	**90.58**	88.20	72.24	**86.89**	80.64	**76.41**	88.74	81.30	90.67	92.82	82.33	80.64	**91.72**	**84.69**
ThoraX- PriorNet	✓	✓	82.68	90.16	**88.35**	**72.34**	86.73	80.70	76.38	88.98	81.16	90.78	92.70	**82.56**	**81.29**	90.53	84.67

Here, AM (AbM) = APAM Utilizing Probabilistic Abnormality Mask, AM (LM) = APAM Utilizing Chest ROI Mask, Atel = Atelectasis, Card = Cardiomegaly, Effu = Effusion, Infi = Infiltration, Nodu = Nodule, Pne1 = Pneumonia, Pne2 = Pneumothorax, Cons = Consolidation, Edem = Edema, Emph = Emphysema, Fibr = Fibrosis, PT = Pleural Thickening, Her = Hernia

TABLE 3. Ablation Study: Impact of input image spatial resolution on the AUC (%) scores of our trained models on the NIH dataset. The best results are shown in red font.

Method	Atel	Card	Effu	Infil	Mass	Nodu	Pne1	Pne2	Cons	Edem	Emph	Fib	PT	Her	Mean
224x224	82.54	90.57	**88.35**	72.29	86.39	78.01	**77.00**	87.96	**81.89**	**90.98**	92.38	81.75	80.04	**91.90**	84.43
368x368	**82.92**	**90.75**	88.30	72.23	**86.96**	80.03	76.82	88.18	81.30	90.66	**92.96**	82.37	80.59	90.93	84.64
512x512	82.68	90.16	**88.35**	**72.34**	86.73	**80.70**	76.38	**88.98**	81.16	90.78	92.70	**82.56**	**81.29**	90.53	**84.67**

Here, Atel = Atelectasis, Card = Cardiomegaly, Effu = Effusion, Infi = Infiltration, Nodu = Nodule, Pne1 = Pneumonia, Pne2 = Pneumothorax, Cons = Consolidation, Edem = Edema, Emph = Emphysema, Fibr = Fibrosis, PT = Pleural Thickening, Her = Hernia

TABLE 4. Ablation Study: Effect of resizing feature and anatomy prior maps on the AUC (%) scores of our trained models on the NIH dataset. The best results are shown in red font.

Method	Atel	Card	Effu	Infil	Mass	Nodu	Pne1	Pne2	Cons	Edem	Emph	Fib	PT	Her	Mean
16x16	82.56	90.28	**88.15**	71.81	**87.12**	80.50	76.24	**89.06**	80.89	90.49	**93.07**	**83.35**	81.02	**90.85**	**84.66**
32x32	82.42	90.47	88.01	71.78	86.71	80.38	76.40	88.89	80.77	90.69	92.75	83.24	80.69	89.21	84.46
48x48	**82.96**	**90.52**	88.14	**72.06**	86.76	**80.82**	**76.46**	89.00	**81.01**	**91.09**	93.03	81.78	**81.05**	90.29	84.64

Here, Atel = Atelectasis, Card = Cardiomegaly, Effu = Effusion, Infi = Infiltration, Nodu = Nodule, Pne1 = Pneumonia, Pne2 = Pneumothorax, Cons = Consolidation, Edem = Edema, Emph = Emphysema, Fibr = Fibrosis, PT = Pleural Thickening, Her = Hernia

TABLE 5. Comparison of AUC (%) Scores of our best performing model with state-of-the-art methods on the NIH dataset. The best results are shown in red font.

Model	Atel	Card	Effu	Infi	Mass	Nodu	Pne1	Pne2	Cons	Edem	Emph	Fibr	PT	Hern	Mean
LSTM-Net [47]	73.30	85.80	80.60	67.50	72.70	77.80	69.00	80.50	71.70	80.60	84.20	75.70	72.40	82.40	76.73
TieNet [48]	73.20	84.40	79.30	66.60	72.50	68.50	72.00	84.70	70.10	82.90	86.50	79.60	73.50	87.60	77.24
AGCL [23]	75.57	88.65	81.91	68.92	81.36	75.45	72.92	84.99	72.83	84.75	90.75	81.79	76.47	87.47	80.27
Ho *et al.* [49]	79.50	88.70	87.50	70.30	83.50	71.60	74.20	86.30	78.60	89.20	87.50	75.60	77.40	83.60	80.96
CARL [50]	78.10	88.00	82.90	70.20	83.40	77.30	72.90	85.70	75.40	85.00	90.80	83.00	77.80	91.70	81.59
Liu *et al.* [51]	79.80	89.03	83.56	71.40	82.49	77.73	73.86	86.95	75.50	84.95	93.36	81.86	77.60	85.89	81.77
CheXNet [52]	77.95	88.16	82.68	68.94	83.07	78.14	73.54	85.13	75.42	84.96	92.49	82.19	79.25	93.23	81.80
DualCheXNet [53]	78.40	88.80	83.10	70.50	83.80	79.60	72.70	87.60	74.60	85.20	94.20	83.70	79.60	91.20	82.36
LLAGNet [5]	78.30	88.50	83.40	70.30	84.10	79.00	72.90	87.70	75.40	85.10	93.90	83.20	79.80	91.60	82.37
Wang *et al.* [19]	77.90	89.50	83.60	71.00	83.40	77.70	73.70	87.80	75.90	85.50	93.30	83.80	79.10	**93.80**	82.57
Yan *et al.* [43]	79.24	88.14	84.15	70.95	84.70	**81.05**	73.97	87.59	75.98	84.70	94.22	83.26	80.83	93.41	83.01
Luo *et al.* [44]	78.91	90.69	84.18	71.84	83.76	79.85	74.19	**90.63**	76.81	86.10	93.96	**83.81**	80.36	93.71	83.49
Arias-Garzon *et al.* [54]	80.43	88.93	86.89	70.10	83.63	78.92	75.07	85.59	80.17	87.71	85.72	81.68	77.67	82.48	81.79
Ouyang *et al.* [6]	77.00	87.00	83.00	71.00	83.00	79.00	72.00	88.00	74.00	84.00	94.00	83.00	79.00	91.00	81.79
SDFN [55]	78.10	88.50	83.20	70.00	81.50	76.50	71.90	86.60	74.30	84.20	92.10	83.50	79.10	91.10	81.47
Keidar *et al.* [56]	80.64	**90.88**	86.94	70.60	83.93	77.07	**76.53**	85.54	80.43	89.20	90.87	81.47	78.02	91.80	83.14
MANet [57]	81.43	89.35	86.30	70.04	83.36	77.76	75.29	85.46	80.23	88.56	85.23	82.82	76.82	92.10	82.84
PCAN [29]	79.10	88.70	84.10	71.10	83.90	80.90	74.60	88.10	75.90	85.40	**94.40**	81.90	80.60	92.80	83.00
Proposed model	**82.68**	90.16	**88.35**	**72.34**	**86.73**	80.70	76.38	88.98	**81.16**	**90.78**	92.70	82.56	**81.29**	90.53	**84.67**

Here, Atel = Atelectasis, Card = Cardiomegaly, Effu = Effusion, Infi = Infiltration, Nodu = Nodule, Pne1 = Pneumonia, Pne2 = Pneumothorax, Cons = Consolidation, Edem = Edema, Emph = Emphysema, Fibr = Fibrosis, PT = Pleural Thickening, Hern = Hernia

TABLE 6. Comparison of disease classification AUC Scores (%) of the proposed model and SOTA models on the CheXpert dataset. The best results are shown in red font.

Model	Atelectasis	Cardiomegaly	Edema	Consolidation	Effusion	Mean
MANet [57]	81.35	86.61	92.22	91.59	89.86	88.33
Arias-Garz´on *et al.* [54]	81.74	84.24	94.06	90.74	**94.31**	89.02
Keidar *et al.* [56]	**86.42**	87.39	91.97	88.23	91.73	89.15
Irvin *et al.* [39]	85.80	83.20	94.10	89.90	93.40	89.30
Pham *et al.* [58]	82.50	85.50	93.00	93.70	92.30	89.40
ViT-S/16 [31]	83.50	81.80	92.50	**94.50**	93.20	89.20
Zhu *et al.* [29]	84.80	86.50	90.80	91.20	94.00	89.50
Proposed model	86.21	**88.11**	**94.15**	92.26	92.36	**90.62**

TABLE 7. Ablation Study: Impact of different types of attention masks with respect to disease localization performance using different T(IoU) thresholds on the NIH dataset. The best results are shown in red font.

T(IoU)	Method	AM (AbM)	AM (LM)	Atel	Card	Effu	Infil	Mass	Nodu	Pne1	Pne2	Mean
0.1	Baseline			0.6556	**1.0000**	0.8105	0.8293	0.7765	0.2911	0.7833	0.7732	0.7399
	ThoraX-PriorNet		✓	0.6889	**1.0000**	0.7974	0.8455	0.7176	0.4177	0.7917	0.7423	0.7504
	ThoraX-PriorNet	✓		**0.7333**	**1.0000**	**0.8366**	0.8293	**0.7882**	**0.5696**	0.8083	**0.8454**	**0.8013**
	ThoraX-PriorNet	✓	✓	**0.7333**	**1.0000**	0.8235	**0.8780**	0.7294	0.4810	**0.8917**	0.7835	0.7901
0.2	Baseline			0.4333	**0.9726**	0.6209	0.6179	**0.6000**	0.1139	0.5750	0.5979	0.5664
	ThoraX-PriorNet		✓	0.4889	0.9041	0.6209	0.6829	0.5765	0.1772	0.6167	0.5052	0.5715
	ThoraX-PriorNet	✓		**0.5722**	0.8493	**0.7255**	0.5854	0.5882	**0.3038**	**0.7000**	**0.6701**	0.6243
	ThoraX-PriorNet	✓	✓	0.5667	0.8973	0.6928	**0.7236**	0.5765	0.2532	0.6917	0.6082	**0.6262**
0.3	Baseline			0.2889	**0.7329**	0.4183	0.4553	**0.4706**	0.0380	0.4417	0.4330	0.4098
	ThoraX-PriorNet		✓	0.3444	0.6849	0.4248	**0.5447**	0.4235	0.0759	0.5000	0.3918	0.4238
	ThoraX-PriorNet	✓		0.4056	0.5342	0.5033	0.4715	0.4471	**0.1646**	**0.6083**	**0.5052**	0.4550
	ThoraX-PriorNet	✓	✓	**0.4222**	0.5685	**0.5163**	0.5285	0.4588	0.1013	0.5667	0.4845	**0.4559**
0.4	Baseline			0.1667	**0.3425**	0.2418	0.3089	0.3529	0.0253	0.3083	0.3093	0.2570
	ThoraX-PriorNet		✓	0.2167	0.3288	0.2549	0.3333	**0.3882**	0.0127	0.3667	0.2784	0.2724
	ThoraX-PriorNet	✓		**0.3222**	0.2192	0.2876	0.3415	0.3529	**0.0886**	**0.4833**	**0.3814**	0.3096
	ThoraX-PriorNet	✓	✓	0.2833	0.2603	**0.3464**	**0.4065**	0.3176	0.0380	0.4667	**0.3814**	**0.3125**
0.5	Baseline			0.0611	0.1233	0.0980	0.1870	0.2353	0.0127	0.1833	0.2062	0.1384
	ThoraX-PriorNet		✓	0.1111	**0.1438**	0.1307	0.2439	0.2235	0.0127	0.2917	0.1649	0.1653
	ThoraX-PriorNet	✓		0.1556	0.0959	0.1634	0.2358	**0.2706**	**0.0253**	0.2667	0.2268	0.1800
	ThoraX-PriorNet	✓	✓	**0.1833**	0.1233	**0.2026**	**0.2602**	0.2353	**0.0253**	**0.3583**	**0.2784**	**0.2083**
0.6	Baseline			0.0278	0.0685	0.0260	0.1301	0.0588	**0.0127**	0.1333	**0.1340**	0.0739
	ThoraX-PriorNet		✓	0.0389	**0.0890**	0.0392	0.1301	0.0824	0.0000	0.2083	0.0515	0.0799
	ThoraX-PriorNet	✓		**0.0833**	0.0411	0.0588	0.0894	**0.1176**	0.0000	0.1500	0.1134	0.0817
	ThoraX-PriorNet	✓	✓	0.0722	0.0411	**0.1111**	**0.1707**	0.0941	0.0000	**0.2667**	0.0973	**0.1061**
0.7	Baseline			0.0056	**0.0274**	0.0196	0.0325	0.0235	0.0000	0.0667	0.0412	0.0271
	ThoraX-PriorNet		✓	0.0111	0.0000	0.0065	0.0244	**0.0353**	0.0000	**0.1083**	0.0309	0.0271
	ThoraX-PriorNet	✓		**0.0167**	0.0068	0.0261	0.0407	**0.0353**	0.0000	0.0750	**0.0619**	0.0328
	ThoraX-PriorNet	✓	✓	0.0222	**0.0274**	**0.0458**	**0.0813**	0.0235	0.0000	0.0125	0.0309	**0.0445**

Here, AM (AbM) = APAM Utilizing Probabilistic Abnormality Mask, AM (LM) = APAM Utilizing Chest ROI Mask, Atel = Atelectasis, Card = Cardiomegaly, Effu = Effusion, Infi = Infiltration, Nodu = Nodule, Pne1 = Pneumonia, Pne2 = Pneumothorax

TABLE 8. Ablation Study: Impact of input image spatial resolution with respect to disease localization performance using different T(IoU) thresholds on the NIH dataset. The best results are shown in red font.

T(IoU)	Method	Atel	Card	Effu	Infil	Mass	Nodu	Pne1	Pne2	Mean
0.1	224x224	0.6278	**1.0000**	0.7908	0.8618	0.5176	0.1266	0.8250	0.5773	0.6659
	368x368	0.7111	**1.0000**	0.8170	0.8374	0.7059	0.2405	0.8000	0.7143	0.7283
	512x512	**0.7611**	**1.0000**	**0.8366**	**0.8699**	**0.7412**	**0.5822**	**0.8667**	**0.7629**	**0.8026**
0.2	224x224	0.4611	**1.0000**	0.6209	0.6260	0.3176	0.0127	**0.6917**	0.3402	0.5088
	368x368	0.5278	0.9932	0.6667	0.6748	0.5294	0.0380	0.6000	0.5816	0.5764
	512x512	**0.5667**	0.8973	**0.6928**	**0.7236**	**0.5765**	**0.2532**	**0.6917**	**0.6082**	**0.6262**
0.3	224x224	0.3000	**0.9863**	0.4575	0.4390	0.2118	0.0000	0.5417	0.2474	0.3980
	368x368	0.3500	0.8767	0.4706	0.5285	0.3765	0.0000	0.4583	0.3980	0.4323
	512x512	**0.4667**	0.7945	**0.4902**	**0.4634**	**0.5059**	**0.1646**	**0.5583**	**0.4639**	**0.4884**
0.4	224x224	0.1722	**0.9452**	0.2614	**0.3089**	0.1647	0.0000	0.3917	0.1237	0.2960
	368x368	0.2444	0.6304	0.2614	0.3821	0.2118	0.0000	0.2833	0.3163	0.2912
	512x512	**0.3222**	0.6164	**0.2941**	0.2683	**0.3765**	**0.0506**	**0.4000**	**0.3402**	**0.3335**
0.5	224x224	0.1056	**0.7260**	0.1242	**0.2520**	0.1176	0.0000	0.2000	0.0825	0.2010
	368x368	0.1111	0.2397	0.0980	0.2602	0.0706	0.0000	0.2000	**0.2347**	0.1518
	512x512	**0.1722**	0.3904	**0.1438**	0.1789	**0.2941**	**0.0127**	**0.3167**	0.2268	**0.2169**
0.6	224x224	0.0556	**0.4658**	0.0719	0.1626	0.0235	0.0000	0.1083	0.0412	**0.1161**
	368x368	0.0444	0.0890	0.0458	0.1382	0.0588	0.0000	0.1167	**0.1224**	0.0769
	512x512	**0.0722**	0.0411	**0.1111**	**0.1707**	**0.0941**	0.0000	**0.2667**	0.0973	0.1061
0.7	224x224	**0.0222**	**0.1644**	0.0131	0.0650	0.0000	0.0000	**0.0583**	0.0103	0.0417
	368x368	0.0111	0.0137	0.0196	0.0732	**0.0235**	0.0000	0.0250	**0.0510**	0.0271
	512x512	**0.0222**	0.0274	**0.0458**	**0.0813**	**0.0235**	0.0000	0.0125	0.0309	**0.0445**

Here, Atel = Atelectasis, Card = Cardiomegaly, Effu = Effusion, Infi = Infiltration, Nodu = Nodule, Pne1 = Pneumonia, Pne2 = Pneumothorax

classifying diseases like Atelectasis, Nodules, Consolidation, and Edema. The baseline denotes the vanilla DenseNet121 model without incorporating the APAM block. Afterward, we added the APAM block and gradually used the different types of attention masks. Table 2 demonstrates that all three ThoraX-PriorNet variants achieve better classification scores than the baseline. We obtained the most significant jump in classification results when we used APAM with the probabilistic disease-specific masks, i.e., an AUC (%) score of 84.69. Incorporating both types of attention masks yields a slightly lower score, i.e., a percentage AUC score of 84.67. But it improves the performance for pathologies like Effusion, Infiltration, Fibrosis, and Pleural Thickening.

TABLE 9. **Ablation Study: Effect of resizing feature and anatomy prior maps with respect to disease localization performance using different T(IoU) thresholds on the NIH dataset. The best results are shown in red font.**

T(IoU)	Method	Atel	Card	Effu	Infil	Mass	Nodu	Pne1	Pne2	Mean
0.1	16x16	0.7111	0.9932	0.8301	**0.8699**	0.7294	0.5190	0.8417	0.7523	0.7809
	32x32	**0.7611**	**1.0000**	**0.8366**	0.8211	0.7176	0.4810	0.8583	**0.8247**	0.7876
	48x48	**0.7611**	**1.0000**	**0.8366**	**0.8699**	**0.7412**	**0.5822**	**0.8667**	0.7629	**0.8026**
0.2	16x16	0.5667	0.8767	**0.6928**	**0.6992**	**0.6235**	0.1899	0.6750	0.5979	0.6152
	32x32	**0.6111**	0.9110	0.6405	0.6260	0.5529	0.2532	**0.7083**	**0.6495**	0.6191
	48x48	0.5889	**0.9795**	0.6340	0.6341	**0.6235**	**0.3038**	0.6833	0.5567	**0.6255**
0.3	16x16	0.4056	0.5890	0.4706	**0.5366**	0.4941	0.1013	0.5167	0.4536	0.4459
	32x32	0.4389	0.5342	0.4837	0.4797	0.4353	0.1139	0.5500	**0.4845**	0.4400
	48x48	**0.4667**	**0.7945**	**0.4902**	0.4634	**0.5059**	**0.1646**	**0.5583**	0.4639	**0.4884**
0.4	16x16	0.2667	0.2877	0.2680	**0.3984**	**0.3765**	0.0380	0.3667	**0.4027**	0.3005
	32x32	0.3000	0.1986	**0.3072**	0.2602	0.2588	**0.0633**	0.3750	0.3711	0.2668
	48x48	**0.3222**	**0.6164**	0.2941	0.2683	**0.3765**	0.0506	**0.4000**	0.3402	**0.3335**
0.5	16x16	0.1500	0.1644	0.1569	**0.2845**	0.2825	**0.0127**	0.2750	0.2165	0.1928
	32x32	**0.1944**	0.0753	**0.1765**	0.1463	0.2000	**0.0127**	0.2083	**0.2680**	0.1602
	48x48	0.1722	**0.3904**	0.1438	0.1789	**0.2941**	**0.0127**	**0.3167**	0.2268	**0.2169**
0.6	16x16	0.0667	0.0616	0.0392	**0.1301**	**0.1529**	0.0000	**0.1917**	**0.1546**	**0.1000**
	32x32	0.0611	0.0342	**0.0784**	0.0976	0.1294	0.0000	0.1417	0.1237	0.0833
	48x48	**0.0778**	**0.1712**	0.0523	0.0894	0.1176	0.0000	0.1583	0.1031	0.0962
0.7	16x16	0.0000	0.0205	0.0131	**0.0569**	0.0235	0.0000	0.0750	0.0616	0.0314
	32x32	**0.0056**	0.0068	**0.0458**	0.0325	0.0235	0.0000	0.0417	**0.0619**	0.0272
	48x48	0.0000	**0.0753**	0.0131	0.0244	**0.0352**	0.0000	**0.0833**	0.0412	**0.0341**

Here, Atel = Atelectasis, Card = Cardiomegaly, Effu = Effusion, Infi = Infiltration, Nodu = Nodule, Pne1 = Pneumonia, Pne2 = Pneumothorax

Next, we conducted an ablation study to explore the impact of the input image sizes on the classification performance. We resize the input image into three different sizes: 256×256, 420×420, and 586×586 and crop 224×224 patches for 256×256, 368×368 patches for 420×420, and 512×512 patches for 586×586 as inputs, respectively (random crop during training, center crop during inference). Results on the NIH Chest X-ray dataset are shown in Table 3. We can see that increasing input image resolution improves the classification performance. However, the improvement range from 368×368 to 512×512 is lower compared to 224×224 to 368 × 368. More specifically, we observe that the increase in AUC score for small lesions, such as nodules, is significant in the higher resolution.

Finally, we conducted an ablation study on the spatial dimension of the feature map and the attention masks. We downscale and upsample the attention masks and feature maps, respectively, to an intermediate size before using them in the APAM block. For the input image dimension of 512×512, the feature map size is 16×16. We also performed experiments by resizing the feature map to 32×32 and 48×48. The results are reported in Table 4. We observe that increasing the spatial dimension of the final feature map does not yield improvements in the classification performance. The 48×48 model has the same classification performance level as the 16×16 model. However, the 48×48 model improves the localization performance, which will be demonstrated in a later section.

2) PERFORMANCE COMPARISON WITH SOTA METHODS

Table 5 compares the AUC score of ThoraX-PriorNet with other state-of-the-art (SOTA) models on NIH ChestX-ray14 dataset. Here, we observe that the proposed model's performance is superior to existing SOTA methods in terms of the mean AUC score. More specifically, it has shown performance improvement in diseases like Atelectasis, Effusion, Infiltration, Mass, Consolidation, Edema, and Pleural Thickening.

Table 6 shows the comparison of our proposed model with existing state of the art models on CheXpert dataset. Here, we have used the same probabilistic masks which were generated for training on NIH Chest X-ray dataset and for providing disease guided attention. The results show that the proposed method provides superior results for diseases like- cardiomegaly and edema, whereas performance on atelectasis, consolidation, and effusion are slightly less than the compared approaches. However, the overall mean AUC score is better compared to the other models. Our method shows an AUC score of 90.62%.

B. ABNORMALITY LOCALIZATION

1) ABLATION STUDY

We have also conducted several ablation studies on the NIH ChestX-ray14 dataset to explore the impact of different aspects of our trained model on localization performance. First, we evaluate the impact of different types of attention masks. The results are reported in Table 7. We can observe a notable performance improvement after including the APAM module. Our proposed ThoraX-PriorNet outperforms the baseline model by large margins in all T(IoU) thresholds. APAM block utilizing both disease probabilistic maps and chest ROI maps achieves overall better results, especially in the higher thresholds compared to the APAM block using only one type of attention mask.

The impact of different input image resolutions on the localization performance is demonstrated in Table 8. We can

TABLE 10. Comparison of disease localization accuracy of the best performing proposed model with state-of-the-art methods. The best results are shown in red font.

T(IoU)	Method	Atel	Card	Effu	Infil	Mass	Nodu	Pne1	Pne2	Mean
0.1	Cai *et al.* [4]	0.68	0.97	0.65	0.52	0.56	0.46	0.65	0.43	0.61
	Li *et al.* [46]	0.59	0.81	0.72	0.84	0.68	0.28	0.22	0.37	0.57
	Liu *et al.* [40]	0.39	0.90	0.63	0.85	0.69	0.38	0.30	0.39	0.60
	Ouyang *et al.* [6]	0.78	0.97	0.82	0.85	0.78	0.56	0.76	0.48	0.75
	Han *et al.* [30] †	0.72	0.96	0.88	0.93	0.74	0.45	0.65	0.64	0.75
	Han *et al.* [59] ‡	0.61	0.95	0.65	0.82	0.50	0.13	0.79	0.28	0.59
	Li *et al.* [32]	0.64	**1.00**	0.75	0.79	0.69	0.07	0.79	0.39	0.64
	Zhu *et al.* [29]	**0.84**	**1.00**	**0.86**	**0.94**	**0.82**	**0.49**	**0.90**	0.38	0.78
	Rozenberg *et al.* [28] †	0.77	**1.00**	0.84	0.94	0.70	0.44	0.91	0.73	0.79
	Proposed model	0.73	**1.00**	0.82	0.88	0.73	0.48	0.89	**0.78**	**0.80**
0.2	Cai *et al.* [4]	0.51	0.90	0.52	0.44	0.47	0.27	0.54	0.24	0.49
	Han *et al.* [30] †	0.55	0.89	0.78	0.85	0.62	0.31	0.52	0.54	0.63
	Han *et al.* [59] ‡	0.41	0.91	0.41	0.59	0.26	0.05	0.57	0.19	0.42
	Li *et al.* [32]	0.40	**1.00**	0.66	**0.74**	0.43	0.01	**0.69**	0.28	0.53
	Zhu *et al.* [29]	0.47	0.68	0.45	0.48	0.26	0.05	0.35	0.23	0.37
	Proposed model	**0.57**	0.90	**0.69**	0.72	**0.58**	**0.25**	**0.69**	**0.61**	**0.63**
0.3	Cai *et al.* [4]	0.33	0.85	0.34	0.28	0.33	0.11	0.39	0.16	0.35
	Li *et al.* [46]	0.34	0.26	**0.52**	**0.72**	0.40	0.09	0.00	0.23	0.32
	Liu *et al.* [40]	0.34	0.71	0.39	0.65	0.48	0.09	0.16	0.20	0.38
	Ouyang *et al.* [6]	0.34	0.40	0.27	0.55	**0.51**	**0.14**	0.42	0.22	0.36
	Han *et al.* [30] †	0.39	0.85	0.60	0.67	0.43	0.21	0.40	0.45	0.50
	Han *et al.* [59] ‡	0.28	0.79	0.22	0.38	0.12	0.01	0.41	0.05	0.28
	Li *et al.* [32]	0.21	**1.00**	0.44	0.53	0.27	0.00	0.55	0.19	0.40
	Zhu *et al.* [29]	**0.43**	0.34	0.33	0.57	0.48	0.04	**0.60**	0.13	0.36
	Proposed model	0.42	0.57	**0.52**	0.53	0.46	0.10	0.57	**0.48**	**0.49**
0.4	Cai *et al.* [4]	0.23	0.73	0.18	0.20	0.18	0.03	0.23	0.11	0.24
	Han *et al.* [30] †	0.24	0.81	0.42	0.54	0.34	0.13	0.28	0.32	0.39
	Han *et al.* [59] ‡	0.17	0.54	0.13	0.18	0.07	0.01	0.26	0.02	0.17
	Li *et al.* [32]	0.10	**0.98**	0.27	**0.47**	0.18	0.00	0.38	0.13	0.31
	Zhu *et al.* [29]	0.23	0.10	0.16	0.37	0.37	0.01	0.33	0.10	0.21
	Proposed model	**0.32**	0.62	**0.29**	0.27	**0.38**	**0.05**	**0.40**	**0.34**	**0.33**
0.5	Cai *et al.* [4]	0.11	0.60	0.10	0.12	0.07	**0.03**	0.17	0.08	0.17
	Li *et al.* [46]	0.18	0.10	**0.27**	0.46	0.18	**0.03**	0.00	0.11	0.17
	Liu *et al.* [40]	**0.19**	0.53	0.19	**0.47**	**0.33**	**0.03**	0.08	0.11	**0.24**
	Han *et al.* [30] †	0.16	0.77	0.29	0.35	0.24	0.09	0.15	0.22	0.28
	Han *et al.* [59] ‡	0.08	0.32	0.05	0.09	0.05	0.00	0.12	0.01	0.09
	Li *et al.* [32]	0.05	**0.87**	0.13	0.34	0.12	0.00	0.33	0.10	0.24
	Zhu *et al.* [29]	0.09	0.01	0.11	0.19	0.21	0.00	0.17	0.05	0.10
	Proposed model	0.18	0.12	0.20	0.26	0.24	**0.03**	**0.36**	**0.28**	0.22
0.6	Cai *et al.* [4]	0.03	0.44	0.05	0.06	0.05	**0.01**	0.05	0.07	0.10
	Han *et al.* [30] †	0.09	0.74	0.19	0.16	0.18	0.04	0.11	0.14	0.21
	Han *et al.* [59] ‡	0.02	0.15	0.03	0.04	0.03	0.00	0.06	0.00	0.04
	Li *et al.* [32]	0.01	**0.60**	0.06	**0.23**	0.06	0.00	0.17	0.04	**0.15**
	Proposed model	**0.07**	0.04	**0.11**	0.17	**0.09**	0.00	**0.27**	**0.10**	0.11
0.7	Cai *et al.* [4]	0.01	0.17	0.01	0.02	0.01	0.00	0.02	0.02	0.03
	Li *et al.* [46]	**0.09**	0.01	0.07	**0.28**	0.08	**0.01**	0.00	0.05	0.07
	Liu *et al.* [40]	0.08	**0.30**	**0.09**	0.25	**0.19**	**0.01**	0.04	**0.07**	**0.13**
	Han *et al.* [30] †	0.05	0.54	0.09	0.11	0.12	0.02	0.07	0.06	0.13
	Han *et al.* [59] ‡	0.01	0.04	0.01	0.02	0.01	0.00	0.03	0.00	0.02
	Li *et al.* [32]	0.01	0.26	0.02	0.10	0.02	0.00	**0.10**	0.02	0.07
	Proposed model	0.02	0.03	0.05	0.08	0.02	0.00	0.01	0.03	0.04

† utilized bounding box information ‡ scores for input image resolution of 224x224
Here, Atel = Atelectasis, Card = Cardiomegaly, Effu = Effusion, Infi = Infiltration, Nodu = Nodule, Pne1 = Pneumonia, Pne2 = Pneumothorax

observe that increasing the spatial dimension of the input image enhances the localization performance greatly. More specifically, we observe that increasing spatial dimension shows greater performance improvement in the localization tasks for diseases with small spatial features (e.g., mass, nodule, pneumothorax). However, large lesions, such as cardiomegaly, are not benefited. The impact on localization performance due to different dimensions of the intermediate size of feature maps and attention maps is reported in Table 9. Similar to the input spatial dimension, we can observe that the 48 × 48 model achieved overall better localization performance compared to other models.

2) PERFORMANCE COMPARISON WITH SOTA METHODS

Table 10 shows the quantitative comparison of the localization score of ThoraX-PriorNet with previous SOTA models. Note that Han et al.and Rozenberg et al.utilize bounding box information in their pipeline. As a result, their model is not directly comparable to ours and other SOAT models. In spite of that, our proposed method shows comparable performance

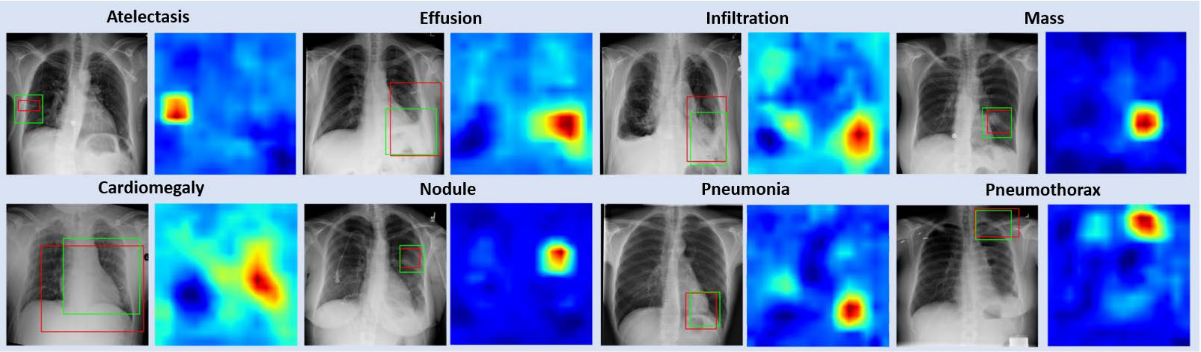


FIGURE 8. Examples of some disease localization by our proposed method. The first column of each sample: Input CXR image with the ground truth bounding box (red color) and the predicted bounding box (green color). The second column of each sample: Corresponding activation map from the proposed model.

TABLE 11. Statistical analysis between the baseline and proposed model for a 10-fold cross-validation using Nadeau and Bengio's corrected t-test Method [60].

Method	Test-1	Test-2	Test-3	Test-4	Test-5	Test-6	Test-7	Test-8	Test-9	Test-10	Mean ± Std	p-value
Baseline	84.56	84.38	84.13	84.81	84.21	**84.09**	84.79	84.12	84.65	84.33	84.41 ± 0.26	0.017
Proposed	**84.66**	**84.60**	**84.51**	**84.97**	**84.65**	84.03	**84.99**	**84.34**	**84.82**	**84.58**	**84.61 ± 0.27**	

at lower T(IoU) thresholds despite not using the bounding box supervision. Our proposed ThoraX-PriorNet achieved improvements of 2.56%, 18.87%, 22.50%, and 6.45% at IoU of 0.1, 0.2, 0.3, and 0.4, respectively, compared to the localization performances of existing methods. In other IoU thresholds, our model achieves slightly lower but competitive scores.

We have extracted the activation maps for eight different diseases from the NIH ChestX-ray8 dataset and plotted them in Fig. 8 to visualize the localization of the proposed model. The red boxes denote the ground truth boxes, while the green boxes denote the predicted boxes. We can observe that our model can identify and localize the abnormal findings.

C. STATISTICAL ANALYSIS

To perform statistical analysis, we have conducted a 10-fold cross-validation and used Nadeau and Bengio' corrected t-test method [60] for calculating the p-values. The results for the baseline and the proposed method are reported in Table 11. The baseline model achieves an average AUC (%) score of 84.41 with a standard deviation of 0.26, while our proposed method achieves 84.61±0.27. The statistical result yields a p-value of 0.017, denoting the improvement of the proposed method compared to the baseline.

D. COMPUTATIONAL COMPLEXITY ANALYSIS

The average time to process a single chest X-ray image during the testing phase, along with the floating point operation computation, for the input image dimension of 512×512 is reported in Table 12. Our proposed ThoraX-PriorNet takes an average of 8.74 ms to process a test image and requires 28.1 GFLOPS compute power to perform this task.

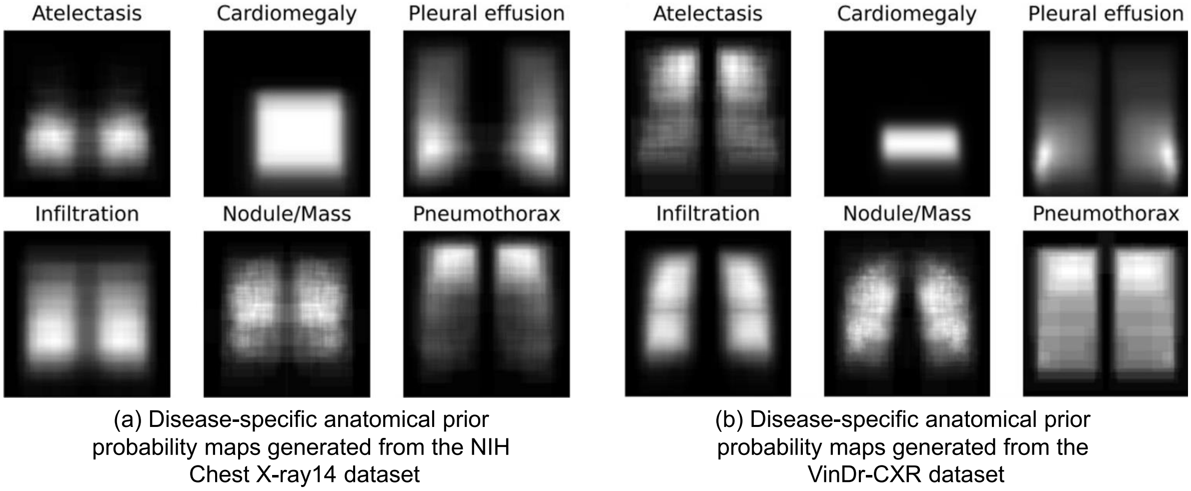
TABLE 12. Computational cost parameters by ThoraX-PriorNet for a single image on 512 × 512 dimension during the test phase on the NIH ChestX-Ray14 dataset.

Method	Time (ms)	FLOPs (G)	AUC
ThoraX-PriorNet	8.74	28.1	84.67

E. ANALYSIS OF GENERAZIABILITY OF THE PROBABILISTIC ABNORMALITY MASKS

Different chest X-ray-based thoracic disease datasets may have diverse affine variations, such as rotations, shifts, and different scales. To address the affine variations, we have utilized the alignment module [40]. In addition, the chest X-ray datasets may have intrinsic variations among them due to patient demographics, geographical diversity, class imbalances, different exposure settings and imaging protocols, scanner intrinsic variations, and so on, inherent to medical datasets. However, in our experiments, we are not utilizing or training different datasets together, a task that is reserved for domain adaptation and generalization methods [61], [62], [63]. Here, we are generating the disease-prior masks by taking and aggregating the referenced spatial positions from the bounding boxes to get a probabilistic map. The domain variations due to exposure shift, different imaging protocols, or machine intrinsic variations are not propagated through the generated abnormality masks. However, we do acknowledge that the number and quality of the ground truth bounding boxes, class imbalance, patient demographics, or geographical diversity may have an effect on the generated probabilistic map, which may influence the performance of the proposed model.

We have conducted experiments to evaluate the generalizability of the disease-prior probabilistic abnormality masks generated from a particular thoracic disease dataset. For this



(a) Disease-specific anatomical prior probability maps generated from the NIH Chest X-ray14 dataset

(b) Disease-specific anatomical prior probability maps generated from the VinDr-CXR dataset

FIGURE 9. Disease-specific anatomical prior probability maps. Abnormality masks generated from NIH (left) and VinDr-CXR (right) datasets.

experiment, we have chosen the NIH chest X-ray14 [22] and the VinDr-CXR dataset [64], as they have provided bounding box annotations. We have performed the experiment for the six common pathologies between them, i.e., Atelectasis, Cardiomegaly, Pleural Effusion, Infiltration, Nodule/Mass, and Pneumothorax. For the NIH chest X-ray14 dataset, we have merged the Nodule and Mass classes into a single Nodule/Mass class, similar to the VinDr-CXR dataset. The VinDr-CXR dataset has a much higher number of available ground truth bounding boxes compared to the NIH chest X-ray14. The generated disease prior masks from NIH chest X-ray14 and VinDr-CXR are given in Fig. 9. Note that we have utilized only the bounding boxes from the official training split of the VinDr-CXR to generate the disease masks.

First, we train the vanilla DenseNet-121 model on both datasets without using the aligned images. Afterward, we train the vanilla DenseNet-121 with the aligned images. Finally, we train our proposed ThoraX-PriorNet, utilizing the dataset-specific abnormality masks from the NIH chest X-ray14 and VinDr-CXR datasets, one at a time. The results are reported in Table 13. The average improvement is calculated as follows:

$$(\%\mathrm{RI})_{\mathrm{avg}} = \frac{1}{n}\sum_{i=1}^{n}\frac{S_i - S_i^{\mathrm{ref}}}{S_i^{\mathrm{ref}}} * 100 \tag{14}$$

Here, n is the number of thresholds, S_i is the performance at a particular threshold i, and S_i^{ref} is the performance of the vanilla DenseNet-121 at threshold i. We can observe that adding the alignment module improves the performance of the vanilla DenseNet-121 on both datasets. Our proposed ThoraX-PriorNet achieves significantly improved scores compared to the vanilla DenseNet-121 using either of abnormality masks. However, we can notice that the disease-prior masks from the VinDr-CXR dataset yield the highest performance in both cases. Especially on the VinDr-CXR test dataset, the improvement for ThoraX-PriorNet is 39.77% with the VinDr-CXR disease-prior mask, compared to 17.52% with the NIH chest X-ray14 disease-prior masks. We hypothesize that this is due to two reasons. First, it is due to the quality of the probabilistic maps, as VinDr-CXR has a much higher number of available bounding box annotations. Second, the demographic and class ratio difference between NIH chest X-ray14 and VinDr-CXR may have an effect on the performance. Nevertheless, considering the average improvement in performance compared to vanilla DenseNet-121 with and without aligned images, our proposed model can achieve a significant improvement with either of the disease-prior probabilistic abnormality masks, proving the efficacy of utilizing the APAM block.

F. ROC CURVES

The performance of the clinical diagnostic systems is primarily measured by their specificity and sensitivity. The ROC curves are generally used to assess the diagnostic performance of a clinical system by converting the continuous test results into the decision of the presence or absence of pathology and to demonstrate the trade-off between clinical sensitivity and specificity for every possible cut-off for the clinical test. The ROC curves for each pathology on the NIH chest X-ray dataset are shown in Fig. 10 to visually represent the diagnostic performance of the proposed method.

G. ANALYSIS OF RANDOM CROPPING AUGMENTATION

We have utilized the random cropping augmentation following previous studies [29], [65], as the random cropping augmentation has shown improved performance in thoracic disease detection in literature. In addition, we have also performed the alignment of images (where the images are transformed to align their spatial structure with the anchor image [40]) to ensure that the random cropping technique reliably encompasses all regions of interest within the images.

TABLE 13. Evaluation of the generalizability of the disease-specific anatomical prior probability maps across different thoracic disease datasets.

Dataset	Method	Abnormality masks used		T(IoU)							Average of
		NIH CXR14	VinDr-CXR	0.1	0.2	0.3	0.4	0.5	0.6	0.7	per threshold %RI
NIH CXR14	DenseNet-121			0.4387	0.3077	0.2036	0.1170	0.0671	0.0359	0.0082	REF
	DenseNet-121+Align			0.4665	0.3552	0.2457	0.1540	0.0676	0.0178	0.0064	+0.35%
	ThoraX-PriorNet	✓		**0.6897**	**0.5446**	**0.3921**	**0.2226**	**0.1212**	**0.0487**	0.0074	+60.51%
	ThoraX-PriorNet		✓	0.6798	0.5336	0.3666	0.2028	0.0943	0.0476	**0.0139**	**+60.63%**
VinDr-CXR	DenseNet-121			0.4008	0.2752	0.1702	0.0646	0.0275	0.0209	0.0060	REF
	DenseNet-121+Align			0.4016	0.2933	0.2019	0.1043	0.0432	0.0136	0.0000	+1.29%
	ThoraX-PriorNet	✓		0.4536	0.3452	0.2325	0.1262	0.0413	0.0192	0.0006	+17.52%
	ThoraX-PriorNet		✓	**0.4565**	**0.3536**	**0.2468**	**0.1446**	**0.0598**	**0.0198**	**0.0033**	**+39.77%**

Here, %RI = %Relative improvement

TABLE 14. Effect of Random Cropping augmentation on our proposed method with or without test time augmentation.

TTA	RCrop	Atel	Card	Effu	Infil	Mass	Nodu	Pne1	Pne2	Cons	Edem	Emph	Fib	PT	Her	Mean
		81.59	89.97	**88.27**	**72.17**	85.02	76.74	76.16	**87.98**	**81.63**	90.38	**92.55**	80.74	79.52	88.98	83.69
	✓	**82.23**	**90.24**	88.16	71.81	**86.05**	**77.75**	**76.83**	87.69	81.58	**90.83**	91.95	**81.35**	**79.96**	**91.89**	**84.16**
TTA	RCrop	Atel	Card	Effu	Infil	Mass	Nodu	Pne1	Pne2	Cons	Edem	Emph	Fib	PT	Her	Mean
✓		81.74	89.68	88.00	**72.36**	84.70	77.09	76.14	87.46	81.45	90.53	92.30	79.42	79.59	90.03	83.61
✓	✓	**82.54**	**90.57**	**88.35**	72.29	**86.39**	**78.01**	**77.00**	**87.96**	**81.89**	**90.98**	**92.38**	**81.75**	**80.04**	**91.90**	**84.43**

Here, TTA = Test time augmentation, RCrop = Random Cropping, Atel = Atelectasis, Card = Cardiomegaly, Effu = Effusion, Infi = Infiltration, Nodu = Nodule, Pne1 = Pneumonia, Pne2 = Pneumothorax, Cons = Consolidation, Edem = Edema, Emph = Emphysema, Fib = Fibrosis, PT = Pleural Thickening, Her = Hernia

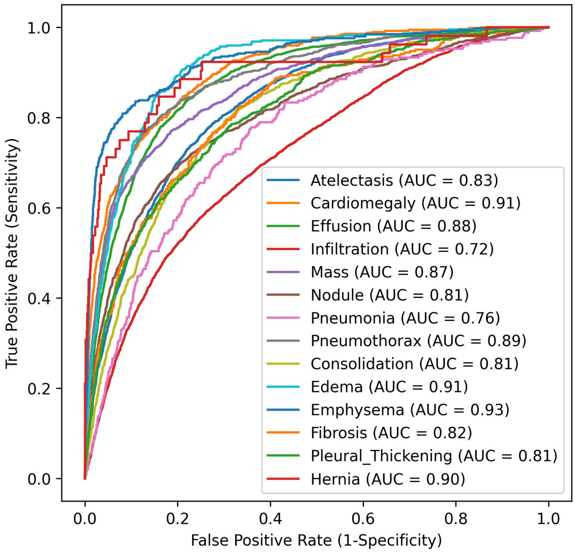


FIGURE 10. ROC curves of thoracic diseases on the NIH ChestX-Ray14 dataset.

The anchor image is constructed by taking an average of 2000 normal images. In Fig 11, we plot five different random cropping windows of size 512×512 on the anchor image of 586×586 dimensions (four outmost corners and one centered). We can observe that the random cropping windows can encompass the region of interest.

We have also conducted experiments to assess the impact of random cropping augmentation on the performance. The results are reported in Table 14. We can observe that the model performs better when random cropping is utilized. It is intuitive because the random cropping technique significantly augments the training data. Another benefit of utilizing random cropping during training is that we can use test time augmentations (TTA) that consist of different random cropping windows. We have followed the procedure mentioned in [43] and [44] and applied TTA based on random cropping, i.e., utilizing average probabilities of ten cropped sub-images (four corner crops and one central crop and the horizontally flipped version of them) as the final prediction. The results are reported in Table 14. We can observe that TTA with random cropping can enhance the performance further.

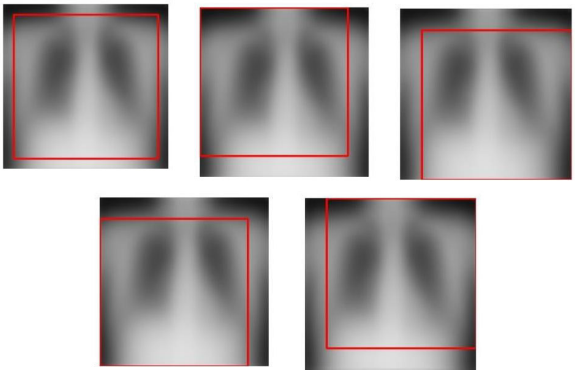


FIGURE 11. Five different random cropping windows on the anchor image. The red window represents the random cropping window.

H. DISCUSSIONS

We make several observations by analyzing the extensive experimental evaluation results described in the previous sections. Our studies show that incorporating attention mechanisms like the proposed ThoraX-PriorNet can enhance the performance of thoracic disease classification and localization. The classification accuracy has improved from 84.30% to 84.67% for the inclusion of both chest ROI mask and disease-specific mask-based attention in the ThoraX-PriorNet architecture. The improvement in the case of localization is by a more noticeable margin from 0.74 to 0.80 with an IoU threshold of 0.1, 0.56 to 0.63 with an

IoU threshold of 0.2, 0.41 to 0.49 with an IoU threshold of 0.3, 0.26 to 0.33 with an IoU threshold of 0.4, 0.14 to 0.22 with an IoU threshold of 0.5, 0.07 to 0.11 with an IoU threshold of 0.6, and 0.03 to 0.04 with an IoU threshold of 0.7. We can also observe that utilizing increased input image spatial resolution or increased feature map dimension shows more notable performance improvement in the localization tasks for diseases with small spatial features (e.g., mass, nodule, pneumothorax).

In addition, we have performed the statistical analysis and found the results statistically significant. We have also conducted experiments on the generalizability of the disease-specific prior probabilistic abnormality masks generated from a specific dataset. We observe that though the quality and quantity of the ground truth boxes can affect the generated probabilistic map, our proposed attention mechanism based on the disease-specific probabilistic abnormality masks can achieve superior performance compared to vanilla deep learning architecture.

VI. CONCLUSION

In this work, we present a novel architecture, ThoraX-PriorNet, providing attentions with disease-specific anatomy prior probability maps and chest ROI masks to simultaneously address the CXR image classification and abnormality localization problem. We evaluated our method on two publicly available datasets, NIH ChestX-ray14 and Stanford CheXpert and compared the results with recent state-of-the-art methods.Extensive experiments show that the model, ThoraX-PriorNet performs better by a good margin when considering both classification and localization tasks in a single model and also in the constraint of multiple datasets.

ACKNOWLEDGMENT

(Md. Iqbal Hossain and Mohammad Zunaed contributed equally to this work.)

REFERENCES

[1] J. B. Soriano, P. J. Kendrick, K. R. Paulson, V. Gupta, E. M. Abrams, R. A. Adedoyin, and T. B. Adhikari, "Prevalence and attributable health burden of chronic respiratory diseases, 1990–2017: A systematic analysis for the global burden of disease study 2017," *Lancet Respiratory Med.*, vol. 8, no. 6, pp. 585–596, 2020.

[2] *Rise in Global Deaths and Disability Due to Lung Diseases Over Past Three Decades*. Accessed: Dec. 18, 2023. [Online]. Available: https://www.bmj.com/company/newsroom/rise-in-global-deaths-and-disability-due-to-lung-diseases-over-past-three-decades/

[3] B. S. Kelly, L. A. Rainford, S. P. Darcy, E. C. Kavanagh, and R. J. Toomey, "The development of expertise in radiology: In chest radiograph interpretation 'expert' search pattern may predate 'expert' levels of diagnostic accuracy for pneumothorax identification," *Radiology*, vol. 280, no. 1, pp. 252–260, 2016.

[4] J. Cai, L. Lu, A. P. Harrison, X. Shi, P. Chen, and L. Yang, "Iterative attention mining for weakly supervised thoracic disease pattern localization in chest X-rays," in *Proc. MICCAI*, 2018, pp. 589–598.

[5] B. Chen, J. Li, G. Lu, and D. Zhang, "Lesion location attention guided network for multi-label thoracic disease classification in chest X-rays," *IEEE J. Biomed. Health Informat.*, vol. 24, no. 7, pp. 2016–2027, Jul. 2020.

[6] X. Ouyang, S. Karanam, Z. Wu, T. Chen, J. Huo, X. S. Zhou, Q. Wang, and J.-Z. Cheng, "Learning hierarchical attention for weakly-supervised chest X-ray abnormality localization and diagnosis," *IEEE Trans. Med. Imag.*, vol. 40, no. 10, pp. 2698–2710, Oct. 2021.

[7] B. Chen, Z. Zhang, J. Lin, Y. Chen, and G. Lu, "Two-stream collaborative network for multi-label chest X-ray image classification with lung segmentation," *Pattern Recognit. Lett.*, vol. 135, pp. 221–227, Jul. 2020.

[8] U. Kamal, M. Zunaed, N. B. Nizam, and T. Hasan, "Anatomy-XNet: An anatomy aware convolutional neural network for thoracic disease classification in chest X-rays," *IEEE J. Biomed. Health Informat.*, vol. 26, no. 11, pp. 5518–5528, Nov. 2022.

[9] A. M. Obeso, J. Benois-Pineau, M. S. García Vázquez, and A. Á. R. Acosta, "Visual vs internal attention mechanisms in deep neural networks for image classification and object detection," *Pattern Recognit.*, vol. 123, Mar. 2022, Art. no. 108411.

[10] Z. Ullah, M. Usman, S. Latif, and J. Gwak, "Densely attention mechanism based network for COVID-19 detection in chest X-rays," *Sci. Rep.*, vol. 13, no. 1, p. 261, Jan. 2023.

[11] M. Innat, M. F. Hossain, K. Mader, and A. Z. Kouzani, "A convolutional attention mapping deep neural network for classification and localization of cardiomegaly on chest X-rays," *Sci. Rep.*, vol. 13, no. 1, p. 6247, Apr. 2023.

[12] G. Hong, X. Chen, J. Chen, M. Zhang, Y. Ren, and X. Zhang, "A multi-scale gated multi-head attention depthwise separable CNN model for recognizing COVID-19," *Sci. Rep.*, vol. 11, no. 1, pp. 1–13, Sep. 2021.

[13] S. Guendel, S. Grbic, B. Georgescu, S. Liu, A. Maier, and D. Comaniciu, "Learning to recognize abnormalities in chest X-rays with location-aware dense networks," in *Proc. Iberoamer. Congr. Pattern Recognit.*, 2018, pp. 757–765.

[14] W. Ye, J. Yao, H. Xue, and Y. Li, "Weakly supervised lesion localization with probabilistic-CAM pooling," 2020, *arXiv:2005.14480*.

[15] I. M. Baltruschat, H. Nickisch, M. Grass, T. Knopp, and A. Saalbach, "Comparison of deep learning approaches for multi-label chest X-ray classification," *Sci. Rep.*, vol. 9, no. 1, pp. 1–10, Apr. 2019.

[16] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-excitation networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 8, pp. 2011–2023, Aug. 2020.

[17] J. Hu, L. Shen, S. Albanie, G. Sun, and A. Vedaldi, "Gather-excite: Exploiting feature context in convolutional neural networks," in *Proc. Adv. Neural Inf. Process Syst.*, vol. 31, 2018, pp. 9423–9433.

[18] B. Chen, Y. Huang, Q. Xia, and Q. Zhang, "Nonlocal spatial attention module for image classification," *Int. J. Adv. Robotic Syst.*, vol. 17, no. 5, Sep. 2020, Art. no. 172988142093892.

[19] H. Wang, S. Wang, Z. Qin, Y. Zhang, R. Li, and Y. Xia, "Triple attention learning for classification of 14 thoracic diseases using chest radiography," *Med. Image Anal.*, vol. 67, Jan. 2021, Art. no. 101846.

[20] R. Zhang, F. Yang, Y. Luo, J. Liu, J. Li, and C. Wang, "Part-aware mask-guided attention for thorax disease classification," *Entropy*, vol. 23, no. 6, p. 653, May 2021.

[21] S. Roy, T. Meena, and S.-J. Lim, "Demystifying supervised learning in healthcare 4.0: A new reality of transforming diagnostic medicine," *Diagnostics*, vol. 12, no. 10, p. 2549, Oct. 2022.

[22] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, "ChestX-ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3462–3471.

[23] Y. Tang, X. Wang, A. P. Harrison, L. Lu, J. Xiao, and R. M. Summers, "Attention-guided curriculum learning for weakly supervised classification and localization of thoracic diseases on chest radiographs," in *Proc. Int. Workshop Mach. Learn. Med. Imag.*, 2018, pp. 249–258.

[24] L. Yao, J. Prosky, E. Poblenz, B. Covington, and K. Lyman, "Weakly supervised medical diagnosis and localization from multiple resolutions," 2018, *arXiv:1803.07703*.

[25] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 618–626.

[26] A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, "Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2018, pp. 839–847.

[27] M. Bany Muhammad and M. Yeasin, "Eigen-CAM: Visual explanations for deep convolutional neural networks," *Social Netw. Comput. Sci.*, vol. 2, no. 1, pp. 1–14, Feb. 2021.

[28] E. Rozenberg, D. Freedman, and A. A. Bronstein, "Learning to localize objects using limited annotation, with applications to thoracic diseases," *IEEE Access*, vol. 9, pp. 67620–67633, 2021.

[29] X. Zhu, S. Pang, X. Zhang, J. Huang, L. Zhao, K. Tang, and Q. Feng, "PCAN: Pixel-wise classification and attention network for thoracic disease classification and weakly supervised localization," *Computerized Med. Imag. Graph.*, vol. 102, Dec. 2022, Art. no. 102137.

[30] Y. Han, C. Chen, A. Tewfik, B. Glicksberg, Y. Ding, Y. Peng, and Z. Wang, "Knowledge-augmented contrastive learning for abnormality classification and localization in chest X-rays with radiomics using a feedback loop," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2022, pp. 1789–1798.

[31] J. Xiao, Y. Bai, A. Yuille, and Z. Zhou, "Delving into masked autoencoders for multi-label thorax disease classification," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2023, pp. 3577–3589.

[32] F. Li, L. Zhou, Y. Wang, C. Chen, S. Yang, F. Shan, and L. Liu, "Modeling long-range dependencies for weakly supervised disease classification and localization on chest X-ray," *Quant. Imag. Med. Surgery*, vol. 12, no. 6, pp. 3364–3378, Jun. 2022.

[33] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. MICCAI*, 2015, pp. 234–241.

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

[35] F. P. Preparata and S. J. Hong, "Convex hulls of finite sets of points in two and three dimensions," *Commun. ACM*, vol. 20, no. 2, pp. 87–93, Feb. 1977.

[36] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. ECCV*, 2018, pp. 3–19.

[37] A. L. Maas, A. Y. Hannun, and A. Y. Ng, "Rectifier nonlinearities improve neural network acoustic models," in *Proc. ICML*, vol. 30, no. 1, 2013, p. 3.

[38] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2261–2269.

[39] J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghgoo, R. Ball, K. Shpanskaya, J. Seekins, D. A. Mong, S. S. Halabi, J. K. Sandberg, R. Jones, D. B. Larson, C. P. Langlotz, B. N. Patel, M. P. Lungren, and A. Y. Ng, "CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison," in *Proc. AAAI Conf. Artif. Intell.*, Jul. 2019, vol. 33, no. 1, pp. 590–597.

[40] J. Liu, G. Zhao, Y. Fei, M. Zhang, Y. Wang, and Y. Yu, "Align, attend and locate: Chest X-ray diagnosis via contrast induced attention network with limited supervision," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 10631–10640.

[41] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Proc. ECCV*, 2016, pp. 694–711.

[42] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[43] C. Yan, J. Yao, R. Li, Z. Xu, and J. Huang, "Weakly supervised deep learning for thoracic disease classification and localization on chest X-rays," in *Proc. ACM Int. Conf. Bioinf., Comput. Biol., Health Informat.*, Aug. 2018, pp. 103–110.

[44] L. Luo, L. Yu, H. Chen, Q. Liu, X. Wang, J. Xu, and P.-A. Heng, "Deep mining external imperfect data for chest X-ray disease screening," *IEEE Trans. Med. Imag.*, vol. 39, no. 11, pp. 3583–3594, Nov. 2020.

[45] S. Suzuki and K. Abe, "Topological structural analysis of digitized binary images by border following," *Comput. Vis., Graph., Image Process.*, vol. 29, no. 3, p. 396, Mar. 1985.

[46] Z. Li, C. Wang, M. Han, Y. Xue, W. Wei, L.-J. Li, and L. Fei-Fei, "Thoracic disease identification and localization with limited supervision," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8290–8299.

[47] L. Yao, E. Poblenz, D. Dagunts, B. Covington, D. Bernard, and K. Lyman, "Learning to diagnose from scratch by exploiting dependencies among labels," 2017, *arXiv:1710.10501*.

[48] X. Wang, Y. Peng, L. Lu, Z. Lu, and R. M. Summers, "TieNet: Text-image embedding network for common thorax disease classification and reporting in chest X-rays," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 9049–9058.

[49] T. K. Ho and J. Gwak, "Multiple feature integration for classification of thoracic disease in chest radiography," *Appl. Sci.*, vol. 9, no. 19, p. 4130, Oct. 2019.

[50] Q. Guan and Y. Huang, "Multi-label chest X-ray image classification via category-wise residual attention learning," *Pattern Recognit. Lett.*, vol. 130, pp. 259–266, Feb. 2020.

[51] F. Liu, Y. Tian, Y. Chen, Y. Liu, V. Belagiannis, and G. Carneiro, "ACPL: Anti-curriculum pseudo-labelling for semi-supervised medical image classification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 20665–20674.

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

[53] B. Chen, J. Li, X. Guo, and G. Lu, "DualCheXNet: Dual asymmetric feature learning for thoracic disease classification in chest X-rays," *Biomed. Signal Process. Control*, vol. 53, Aug. 2019, Art. no. 101554.

[54] D. Arias-Garzón, J. A. Alzate-Grisales, S. Orozco-Arias, H. B. Arteaga-Arteaga, M. A. Bravo-Ortiz, A. Mora-Rubio, and J. M. Saborit-Torres, "COVID-19 detection in X-ray images using convolutional neural networks," *Mach. Learn. Appl.*, vol. 6, Dec. 2021, Art. no. 100138.

[55] H. Liu, L. Wang, Y. Nan, F. Jin, Q. Wang, and J. Pu, "SDFN: Segmentation-based deep fusion network for thoracic disease classification in chest X-ray images," *Computerized Med. Imag. Graph.*, vol. 75, pp. 66–73, Jul. 2019.

[56] D. Keidar et al., "COVID-19 classification of X-ray images using deep neural networks," *Eur. Radiol.*, vol. 31, no. 12, pp. 9654–9663, 2021.

[57] Y. Xu, H.-K. Lam, and G. Jia, "MANet: A two-stage deep learning method for classification of COVID-19 from chest X-ray images," *Neurocomputing*, vol. 443, pp. 96–105, Jul. 2021.

[58] H. H. Pham, T. T. Le, D. Q. Tran, D. T. Ngo, and H. Q. Nguyen, "Interpreting chest X-rays via CNNs that exploit hierarchical disease dependencies and uncertainty labels," *Neurocomputing*, vol. 437, pp. 186–194, May 2021.

[59] Y. Han, G. Holste, Y. Ding, A. Tewfik, Y. Peng, and Z. Wang, "Radiomics-guided global-local transformer for weakly supervised pathology localization in chest X-rays," *IEEE Trans. Med. Imag.*, vol. 42, no. 3, pp. 750–761, Mar. 2023.

[60] C. Nadeau and Y. Bengio, "Inference for the generalization error," *Mach. Learn.*, vol. 52, no. 3, pp. 239–281, Sep. 2003.

[61] H. Wang and Y. Xia, "Domain-ensemble learning with cross-domain mixup for thoracic disease classification in unseen domains," *Biomed. Signal Process. Control*, vol. 81, Mar. 2023, Art. no. 104488.

[62] H. Guan and M. Liu, "Domain adaptation for medical image analysis: A survey," *IEEE Trans. Biomed. Eng.*, vol. 69, no. 3, pp. 1173–1185, Mar. 2022.

[63] M. Zunaed, M. Aynal Haque, and T. Hasan, "Learning to generalize towards unseen domains via a content-aware style invariant model for disease detection from chest X-rays," 2023, *arXiv:2302.13991*.

[64] H. Q. Nguyen et al., "VinDr-CXR: An open dataset of chest X-rays with radiologist's annotations," *Sci. Data*, vol. 9, p. 429, Jan. 2022.

[65] Q. Guan, Y. Huang, Y. Luo, P. Liu, M. Xu, and Y. Yang, "Discriminative feature learning for thorax disease classification in chest X-ray images," *IEEE Trans. Image Process.*, vol. 30, pp. 2476–2487, 2021.

MD. IQBAL HOSSAIN received the B.Sc. degree in biomedical engineering from the Bangladesh University of Engineering and Technology (BUET), Bangladesh, in 2022. He is currently pursuing the Ph.D. degree in imaging science with Washington University in St. Louis. Since 2022, he has been a Research Assistant with the mHealth Laboratory, Biomedical Engineering Department, BUET. His research interests include explainable artificial intelligence and medical computer vision.

MOHAMMAD ZUNAED (Student Member, IEEE) received the B.Sc. and M.Sc. degrees in electrical and electronic engineering from the Bangladesh University of Engineering and Technology. He was a Lecturer with the Electrical and Electronic Engineering Department, Daffodil International University. He is currently a Research Assistant with the mHealth Laboratory, Bangladesh University of Engineering and Technology, under the supervision of Dr. Taufiq Hasan.

MD. KAWSAR AHMED received the B.Sc. degree in biomedical engineering from the Bangladesh University of Engineering and Technology (BUET), Bangladesh, in 2021. He is currently a Lecturer with the Department of Biomedical Engineering, BUET. His research interests include machine learning/AI for biomedical engineering, medical imaging, medical instrumentation, and device design.

S. M. JAWWAD HOSSAIN received the B.Sc. degree in biomedical engineering from the Bangladesh University of Engineering and Technology (BUET), Bangladesh, in 2022. His research interests include computer vision and machine learning.

ANWARUL HASAN (Member, IEEE) received the Ph.D. degree in mechanical engineering from the University of Alberta, Canada, in 2010. He is currently an Associate Professor with the Department of Mechanical and Industrial Engineering and the Biomedical Research Center, Qatar University. Previously, he was an Assistant Professor of biomedical and mechanical engineering with the American University of Beirut, Lebanon, and a Visiting Assistant Professor and an NSERC Postdoctoral Fellow with Harvard University and the Massachusetts Institute of Technology, USA. His current research interests include biomaterials, tissue engineering, 3D bioprinting, diabetic wound healing, cancer biochips, machine learning, and artificial intelligence in health care applications.

TAUFIQ HASAN (Senior Member, IEEE) received the B.Sc. and M.Sc. degrees in electrical and electronic engineering (EEE) from the Bangladesh University of Engineering and Technology (BUET) and the Ph.D. degree in electrical engineering from The University of Texas at Dallas. He was a member of the Center of Robust Speech Systems (CRSS), The University of Texas at Dallas. He was a Research Scientist with the Robert Bosch Research and Technology Center, Palo Alto, CA, USA. He is currently with the Department of Biomedical Engineering, BUET, as an Associate Professor, where he leads the mHealth Research Group. He is also affiliated with the Center for Bioengineering Innovation and Design (CBID), Department of Biomedical Engineering, Johns Hopkins University. His research interests include biomedical signal/image analysis and medical device design.

Open Access funding provided by ‘Qatar National Library’ within the CRUI CARE Agreement